Norwegian Meteorological Institute

Data Management Handbook template for MET and partners in S-ENDA written in asciidoc

Nina E. Larsgård, Elodie Fernandez, Morten W. Hansen, ...

Abstract

Abstract will come here..

Table of Contents

Revision history

Version	Date	Comment	Responsible
2.0	2021-??-??	New version based on original MET DMH	Nina E. Larsgård, Elodie Fernandez, Morten W. Hansen, Matteo De Stefano …

1. Introduction

The purpose of the Data Management Handbook (DMH) is threefold:

1. to provide an overview of the principles for data management to be employed.
2. to help personnel identify their roles and responsibilities for good data management.
3. to provide personnel with practical guidelines for carrying out good data management.

Data management is the term used to describe the handling of data in a systematic and cost-effective manner. The data management regime should be continuously evolving, to reflect the evolving nature of data collection. Therefore this DMH is a living document that will be revised and updated from time to time in order to maintain its relevance.

The DMH is a strategic governing document and should be used as part of the quality framework the organisation is using.

The first version of this DMH was created to focus on the management of dynamic geodata. Dynamic geodata is weather, environment and climate-related data that changes in space and time and is thus descriptive of processes in nature. Examples are weather observations, weather forecasts, pollution (environmental toxins) in water, air and sea, information on the drift of cod eggs and salmon lice, water flow in rivers, driving conditions on the roads and the distribution of sea ice. Dynamic geodata provides important constraints for many decision-making processes and activities in society.

This introduction (Chapter 1) lays forth the background and principles for the data management regime. Chapters 2-5 describe the implementation of the main building blocks: structuring and documenting data (Chapter 2), data services (Chapter 3), portals and documentation aimed at users (Chapter 4) and governance issues (Chapter 5). Each chapter starts with a brief statement of its purpose, followed by a description of what is implemented at the organisation at present, as well as the planned developments for the short-term (<2 years) and expected developments for the longer term (2-5 years).

Practical guidelines for carrying out good data management are addressed in Chapter 6 and especially in the Quick Manual for Data Providers. The Quick Manual is a concise, informal HOW-TO for data management practitioners.

The intended audience for this DMH is any personell involved in the process of making data available for the end user. This process can be viewed as a value chain that moves from the producer of the data to the end user, this is further described in the section Section 1.1.2 (Figure 1).

The handbook can be used in three ways:

1. Read the Introduction (Chapter 1) to find out the background and principles of data management;
2. Read Chapters 2-5 to learn about how data management is currently implemented and how it is expected to evolve in the next few years;
3. Read the Quick Manual for Data Providers for guidelines on what to do in real-life cases; alternatively read Chapter 6 for typical workflow examples.

1.1. The principles of data management for dynamic geodata

Principles of standardised data documentation, publication, sharing and preservation have been formalised in the FAIR Guiding Principles for scientific data management and stewardship [RD3 [https://www.nature.com/articles/sdata201618]] through a process facilitated by FORCE11.

FAIR - findability, accessibility, interoperability and reusability

By following these FAIR principles it is easier to reach a common approach to data management, or a unified data management model. One of the main motivations for implementing a unified data management is to better serve the users of the data. Primarily, this can be approached by making user needs and requirements the guide for determining what data we provide and how. For example, it will be described below how the specification of datasets should be determined by user needs. By implementing the data management practices described here, it is expected that users will benefit from:

- the ease of discovering, viewing and accessing all the datasets that are offered by the institute;
- standardised ways of accessing data, including downloading or streaming data, which reduces the need for special solutions on the user side;
- reducing their own data storage needs, by accessing just what they need;
- easy and standardised access to remote datasets and catalogues, when using their own visualisation/analysis tools;
- the ability to compare and combine data from internal and external sources (through metadata catalogues);
- the ability to apply common data transformations, like spatial, temporal and variables subsetting and reprojection, before downloading anything;
- the ease of building specific metadata catalogues and data portals that include data from the institute and can target a specific user community;
- the access to datasets which can be integrated in their internal and external workflows through standardised web services.

1.1.1. External data management requirements and forcing mechanisms

Any organisation that strives to implement FAIR data management model has to relate to external forcing mechanisms concerning data management at several levels. At the national level, the organisation must comply with national regulations as decided by the government. Some of these are indications of expected behaviour (e.g. OECD regulations) and some are implemented through a legal framework. The Norwegian government has over time promoted free and open sharing of public data. Mechanisms for how to do this are governed by the Geodataloven (implemented as Geonorge), which is a national implementation of the European INSPIRE directive (to be amended in 2019). INSPIRE defines a federated multinational Spatial Data Infrastructure (SDI) for the European Union, similar to NSDI in the USA or UNSDI under the United Nations. The goal is to provide a standardised access to data and provide the necessary tools to be able to work with the data in a unified manner. In short, these legal frameworks require standardised documentation (at discovery and use level; these concepts are described later) and access (through specified protocols) to the data identified.

Other external requirements and forcing mechanisms that are organisation-specific are listed under section X.X.X

1.1.2. The data value chain

The process of getting the data from the producer of the data to the end consumer can be viewed as a value chain. An example of a data value chain is presented in Figure 1. Typically, data from a wide variety of providers are used in the value chain. Traditionally, the data used have been transmitted on request from one data centre to another, and used in the specific processing chains that requested the data. The focus on reuse of data in various contexts has been missing.

Figure 1. Value chain for data

Data sets and metadata are what travels through the value chain, and at the end of the data management value chain are the users of the data (aka. data consumers, see Section 1.2.2.1), who may be either external or internal to the institute.

1.1.3. Dataset

A dataset is a collection of data. In the context of the data management model, the storage mode of the dataset is irrelevant, since access mechanisms can be decoupled from the storage layer as experienced by a data consumer. Typically, a dataset represents a number of variables in time and space. A more detailed definition is provided in the Glossary of Terms. In order to best serve the data through the web services developed, the following guidance is given for defining datasets:

1. A dataset can be a collection of variables stored in, for example, a relational database or as flat files.
2. A dataset is defined as a number of spatial and/or temporal variables.
3. A dataset should be defined by the information content and not the production method. This implies that the output of, for example, a numerical model may be divided into several datasets that are related. This is also important in order to efficiently serve the data through web services. For instance, model variables defined on different vertical coordinates should be separated as linked datasets, since some OGC services (e.g. WMS) are unable to handle mixed coordinates in the same dataset.
4. A good dataset does not mix feature types, e.g. do not combine trajectories and gridded data in one dataset.

Most importantly, a dataset should be defined to meet a consumer need. This means that the specification of a dataset should follow not only the content guidelines just listed, but also address the user needs for delivery, security and preservation.

1.1.4. Metadata

Metadata is a broad concept. In our data management model the term “metadata” is used in several contexts, specifically the five categories that are briefly described in Table 1.

Table 1. Brief introduction to different types of metadata.

Type	Purpose	Description	Examples
Discovery metadata	Used to find relevant data	Discovery metadata are also called index metadata and are a digital version of the library index card. They describe who did what, where and when, how to access data and potential constraints on the data. They shall also link to further information on the data like site metadata. Discovery metadata are thus WIS metadata.	ISO 19115 GCMD DIF
Use metadata	Used to understand data found	Use metadata describe the actual content of a dataset and how it is encoded. The purpose is to enable the user to understand the data without any further communication. They describe the content of variables using standardised vocabularies, units of variable, encoding of missing values, map projections, etc.	Climate and Forecast (CF) Convention BUFR GRIB

Type	Purpose	Description	Examples
Site metadata	Used to understand data found	Site metadata are used to describe the context of observational data. They describe the location of an observation, the instrumentation, procedures, etc. To a certain extent they overlap with discovery metadata, but also extend discovery metadata. Site metadata can be used for observation network design. Site metadata can be considered a type of use metadata.	WIGOS OGC O&M StInfoSys
Configuration metadata	Used to tune portal services for datasets for users	Configuration metadata are used to improve the services offered through a portal to the user community. This can be e.g. how to best visualise a product.	
System metadata	Used to understand the technical structure of the data management system and track changes in it	System metadata covers e.g. technical details of the storage system, web services, their purpose and how they interact with other components of the data management system, available and consumed storage, number of users and other KPI elements etc.	SysDok

The tools and facilities used to manage the information contained in the metadata are further described in Chapter 2.

1.1.5. A data management model based on the FAIR principles

This model is based on the model of the Arctic Data Centre, which adheres to the FAIR principles.

For its implementation, the data management model is built upon the following principles:

- **Standardisation** – compliance with established international standards;

- **Interoperability** – enabling machine-to-machine interfaces and standardised documentation and encoding of data;
- **Integrity** – ensuring that data and access to them can be maintained over time, ensuring the user receives the same data each time;
- **Traceability** – documentation of the provenance of a dataset, i.e., all actions taken to produce and maintain the dataset and the usage of the data in downstream systems;
- **Modularisation** – enabling replacement of one component of the system without necessitating other changes.

The model's basic functions fall into three main categories:

1. **Documentation of data** using discovery and use metadata. The documentation identifies who, what, when, where, and how, and shall make it easy for consumers to find and understand data. This requires application of information containers and utilisation of controlled vocabularies and ontologies where textual representation is required. It also covers the topic of data provenance which is used to describe the origin and all actions done on a dataset. Data provenance is closely linked with workflow management. Furthermore, it covers the relationship between datasets. Application of ontologies in data documentation is closely linked to the concept of linked data.
2. **Publication and sharing of data** focuses on making data accessible to consumers internally and externally. Application of standardised approaches is vital, along with cost efficient solutions that are sustainable. Direct integration of data in applications for analysis through data streaming minimises the complexity and overhead in dissemination solutions. This category also covers persistent identifiers for data.
3. **Preservation of data** includes short and long term management of data, which secures access and availability throughout the lifespan of the data. Good solutions in this area depend on expected and actual usage of the data. Preservation of data includes the concept of data life cycle, i.e., the documented flow of data from initial storage through to obsolescence and permanent archiving (or deletion) and preserving the metadata for the same data (even after deleting).

1.2. Introduction to the data management at [insert organisation here]

1.2.1. Background at [insert organsiation here]

1.2.2. External data management requirements and forcing mechanisms specific to [insert organisation here]

1.2.3. Human roles in data management

1.2.3.1. Data consumer

The Data Consumer may be a scientist or student, employee of a governmental agency, consultant or some other person with a professional or personal interest in the data provided. Data consumers may be internal or external to the entities providing and managing the data.

At [add institute or organisation here] the following are examples of data consumers:

- Researcher working on a project *

An overview of the users of the data covered by this DMH, including a categorisation and a non-exhaustive list of known users, is given in Appendix B where the diversity of [insert organisation here] data users and their demands on our data services is shown. Users include consumers of both data and metadata.

1.2.3.2. Data provider

The Data Provider is generating datasets managed by the data management system described in this document. Data providers can be internal or external to the system. They should be able to maintain the datasets they have committed.

At [add institute or organisation here] the following are examples of data providers:

*

1.2.3.3. Data Management Roles

Between the data providers and data consumers are the processes that manage and deliver the datasets (cf. [img-value-chain]). A number of human roles may be defined with responsibilities that, together, ensure that these processes are carried out in accordance with the data management requirements of the organisation. The definition and filling of these roles depend heavily on the particular organisation, and each organisation must devise its own best solution.

Data management roles in use at [insert institute or organisation here]:

Role	Description	Responsibility

1.2.4. Summary of data management requirements

The data management regime described in this DMH follows the Arctic Data centre model and shall ensure that:

1. There are relevant metadata for all datasets, and both data and metadata are available in a form and in such a way that they can be utilised by both humans and machines.

	Link to relevant section	Example
There are sufficient metadata for each dataset for both discovery and use purposes.		

	Link to relevant section	Example
Discovery metadata are indexed and can be retrieved from available services in a standard way and with standard protocols.		
There are interfaces for discovery, visualisation and download, as well as portals for human access, that operate seamlessly across institutions.		
The data are described in a relevant, standardised and managed vocabulary that supports machine-machine interfaces.		
Datasets have attached a unique and permanent identifier that enables traceability.		
Datasets have licensing that ensures free use and reuse wherever possible.		
Datasets are available for download in a standard form according to the FAIR guiding principles and through standard protocols that are accepted and utilised in the user environment.		

2. There is an organisation that provides for the management of each dataset throughout its lifetime (life cycle management).

	Link to relevant section	Example
There is documentation that describes physical storage, lifetime of each dataset, degree of storage redundancy, metadata consistency methods, how dataset versioning is implemented and unique IDs to ensure traceability. The organisation provides seamless access to data from distributed data centres through various portals.		

	Link to relevant section	Example
The above and a business model at dataset level are described in a Data Management Plan (DMP)		

3. There are services or tools that provide the following functionalities on the datasets:

	Link to relevant section	Example
Transformations, including: subsetting; slicing of gridded data sets to points, sections, profiles; reprojection; resampling; reformatting		
Visualisation (time series, mapping services, etc.)		
Aggregation		
Upload of new datasets (including enabling and configuring data access services)		

2. Structuring and Documenting Data for Efficient Discovery and Use

In order to properly find, understand and use dynamic geodata, standardised encoding and documentation are required, i.e., metadata. Both discovery metadata and use metadata can be embedded in the files produced for a dataset through utilisation of self-explaining file formats. If properly done by the data provider, publication and preservation of data through services is simplified and can be automated.

An essential prerequisite for structuring and documenting data is the specification of the dataset(s). The dataset is the basic building block of our data management model; all the documentation and services described in this DMH are built on datasets. It is the first step in structuring one's data for efficient management, and it is mandatory.

2.1. Structuring and documenting data at [insert organisation here]

2.1.1. Current practice in structuring and documenting data

Table 2. Data types available at [insert institution here], with the fileformats supported. The primary fileformat is marked in **bold**

Supported file formats/structures	Datatype	Available metadata	Examples
Comments			

2.1.2. Planned developments in the near-term (< 2 years)

2.1.3. Expected evolution in the longer term (> 2 years)

3. Data Services

The purpose of this chapter is to describe the services existing and to be developed that benefit from the standardisation performed in the previous step. The information structures described in the previous chapter pave the way for efficient data discovery and use through tools and automated services. Implementation of the services must be in line with the institute's delivery architecture.

This includes:

1. data ingestion, storing the data in the proper locations for long term preservation and sharing;
2. data cataloging, extracting the relevant information for proper discovery of the data;
3. configuration of visualisation and data publication services.

When planning and implementing data services, there are a number of external requirements that constrain choices, especially if reuse of solutions nationally and internationally is intended and wanted for the data in question. At the national level, important constraints are imposed by the national implementation of the INSPIRE directive through Norge digitalt.

3.1. Implementation of data services at [institution]

3.1.1. Planned developments in near-term (< 2 years)

3.1.2. Expected evolution in the longer term (> 2 years)

3.1.3. Legacy data organisation/background/historical developement

4. User Portals and Documentation

The purpose of this chapter is to describe the human interfaces data consumers would use to navigate data and the related services. A portal is an entry point for data consumers, enabling them to discover and search for datasets and services, and providing sufficient documentation and guidance to ensure that they are able to serve themselves using the interactive and machine interfaces offered.

Here, we can distinguish between a general portal for all publishable datasets from the institution and targeted portals that offer a focused selection of data, which may include external datasets. Targeted portals cater to specific user groups and may have a limited lifetime, but also can be long-term commitments.

4.1. Implementation of user portals at [institution]

Table 3. User portals in use at [institution]

User portal	Description	General or targeted portal	Data consumer
Example			

4.1.1. Example portal, current implementation

4.1.1.1. Planned developments in near-term (< 2 years)

4.1.1.2. Expected evolution in the longer term (> 2 years)

5. Data Governance

This chapter describes how we organise and steer data management activities in order to ensure that:

1. the guidelines described above are implemented throughout the organisation;
2. our data management practices are in line with and contribute to the institute's strategic aims;
3. our data management regime is subject to review, analysis and revision in a timely manner.

These higher level aspects of data management are often referred to as data governance. A useful definition is: *"Data governance ... is the overall management of the availability, usability, integrity and security of data used in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures and a plan to execute those procedures."*

In this chapter we address many aspects of this definition, but a full description of data governance touches on management structures that are beyond the scope of this handbook.

5.1. Data life cycle management

Data life cycle management is steered by documentation describing how data generated or used in an activity will be handled throughout the lifetime of the activity and after the activity has been completed. This is living documentation that follows the activity and specifies what kind of data will be generated or acquired, how the data will be described, where the data will be stored, whether and how they can be shared, and how they will be retired (archived or deleted). The purpose of life cycle management is to safeguard the data, not just during their "active" period but also for future reuse of the data, and to facilitate cost-effective data handling.

This DMH recommends the following two concepts of life cycle management to be implemented for the institution:

- Extended discovery metadata for data in internal production chains. These are metadata elements that provide the necessary information for life cycle management just described.
- A data management plan document. A DMP is expected for datasets produced in external projects, but may also be useful for internal datasets, as a supplement to the extended discovery metadata.

The goal is that life cycle management information shall be readily available for every dataset managed by the institute. How these concepts are implemented are described in the institution specific subsection below

5.1.1. Data Management Plan

A Data Management Plan (DMP) is a document that describes textually how the data life cycle management will be carried out for datasets used and produced in specific projects. Generally, these are externally financed projects for which such documentation is required by funding agencies. However, larger internal projects covering many datasets may also find it beneficial to create a specific document of this type.

Currently, agencies funding R&D (such as NFR and the EU) do not strictly require a DMP from the start of

any project. However, for projects in the geosciences, data management is an issue that must be addressed, and the agencies strongly recommend a DMP solution. For example, NFR publishes guidelines [https://www.forskningsradet.no/en/Adviser-research-policy/open-science/open-access-to-research-data/] for the contents of a DMP, including links to tools (templates and online services); these guidelines are recommended for any data management project or activity and will in time become a requirement according to NFR.

5.2. Data governance data at [insert organisation here]

5.2.1. Current implementation

5.2.1.1. Organisational Roles

5.2.1.2. Status DMH

5.2.1.3. Status Discovery metadata:

5.2.1.4. Status DMP:

5.2.2. Planned developments in the near-term (< 2 years)

Revise DMH annually or when needed.

5.2.3. Expected evolution in the longer term (> 2 years)

Revise DMH annually or when needed.

5.2.4. Creating NetCDF-CF files

By documenting and formatting your data using NetCDF following the CF conventions [https://cfconventions.org/] and the Attribute Convention for Data Discovery (ACDD) [https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3], MMD files can be automatically generated from the NetCDF files. The CF conventions is a controlled vocabulary providing a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. The ACDD vocabulary describes attributes recommended for describing a NetCDF dataset to data discovery systems. See, e.g., netCDF4-python docs [https://unidata.github.io/netcdf4-python/], or xarray docs [http://xarray.pydata.org/en/stable/user-guide/io.html] for documentation about how to create netCDF files.

The ACDD recommendations should be followed in order to properly document your netCDF-CF files. The below tables summarize required and recommended ACDD and some additional attributes that are needed to properly populate a discovery metadata catalog which fulfills the requirements of international standards (e.g., GCMD/DIF, the INSPIRE and WMO profiles of ISO19115, etc.).

5.2.4.1. Notes

Keywords describe the content of your dataset following a given vocabulary. It should be provided as a comma separated list with a short name defining the vocabulary used followed by the actual keyword, i.e., `short_name:keyword`. Example:

```
:keywords = "GCMDSK:Earth Science > Atmosphere > Atmospheric radiation, NORTHEMES:Weather and climate" ;
```

In addition, `keywords_vocabulary` provides information about the vocabulary defining the keywords. Example:

```
:keywords_vocabulary = "GCMDSK:GCMD Science Keywords:https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/sciencekeywords,
GCMDPROV:GCMD Providers:https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/providers,
GCMDLOC:GCMD Locations:https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/locations" ;
```

See https://adc.met.no/node/96 for more information about how to define the ACDD keywords.

A data **license** provides information about any restrictions on the use of the dataset. To support a linked data approach, the `license` element should be supported by a `license_resource` element, providing a link to the license definition. Example:

```
:license = "CC-BY-4.0" ;
:license_resource = "http://spdx.org/licenses/CC-BY-4.0" ;
```

5.2.4.2. List of Attributes

This section provides lists of ACDD elements that are required and recommended, as well as some extra elements that are needed to fully support our data management needs. The right columns of these tables provide the MET Norway Metadata Specification (MMD) [https://htmlpreview.github.io/?https://github.com/metno/mmd/blob/master/doc/mmd-specification.html] fields that map to the ACDD (and our extension to ACDD) elements. Please refer to MMD [https://htmlpreview.github.io/?https://github.com/metno/mmd/blob/master/doc/mmd-specification.html] for definitions of these elements, as well as controlled vocabularies that should be used. Note that the below tables are automatically generated - check https://github.com/metno/py-mmd-tools/blob/master/py_mmd_tools/mmd_elements.yaml if anything is unclear.

The following ACDD elements are required:

ACDD Attribute	Repetition	MMD equivalent
id	No repetition allowed	metadata_identifier
date_created	Use date_metadata_modified with comma separation	last_metadata_update>update>datetime
date_metadata_modified	Use date_metadata_modified with comma separation	last_metadata_update>update>datetime
title	Use ACDD extension for translations	title>title

summary	Use ACDD extension for translations	abstract>abstract
geospatial_lat_max	No repetition allowed	geographic_extent>rectangle>north
geospatial_lat_min	No repetition allowed	geographic_extent>rectangle>south
geospatial_lon_max	No repetition allowed	geographic_extent>rectangle>east
geospatial_lon_min	No repetition allowed	geographic_extent>rectangle>west
keywords	Comma separated list	keywords>keyword
keywords_vocabulary	Comma separated list	keywords>vocabulary

The following ACDD elements are recommended:

ACDD Attribute	**Repetition**	**Default**	**MMD equivalent**
time_coverage_start	Comma separated list		temporal_extent>start_date
time_coverage_end	Comma separated list		temporal_extent>end_date
geospatial_bounds	No repetition allowed		geographic_extent>polygon
processing_level	No repetition allowed		operational_status
license	No repetition allowed		use_constraint>identifier
['creator_role', 'contributor_role']	Comma separated list	Investigator	personnel>role
['creator_name', 'contributor_name']	Comma separated list	Not available	personnel>name
creator_email	Comma separated list	Not available	personnel>email
creator_institution	Comma separated list	Not available	personnel>organisation
institution	Comma separated list		data_center>data_center_name>short_name
institution	Comma separated list		data_center>data_center_name>long_name
publisher_url	Comma separated list		data_center>data_center_url
project	Semicolon separated list		project>short_name
project	Semicolon separated list		project>long_name
platform	Comma separated list		platform>short_name

platform	Comma separated list		platform>long_name
platform_vocabulary	Comma separated list		platform>resource
instrument	Comma separated list		platform>instrument>short_name
instrument	Comma separated list		platform>instrument>long_name
instrument_vocabulary	Comma separated list		platform>instrument>resource
source	Semicolon separated list		activity_type
creator_name	Comma separated list		dataset_citation>author
date_created	Comma separated list		dataset_citation>publication_date
title			dataset_citation>title
publisher_name	Comma separated list		dataset_citation>publisher
metadata_link	Comma separated list		dataset_citation>url
references	Comma separated list		dataset_citation>other

The following elements are ACDD extensions that are needed to improve (meta)data interoperability. Please refer to the documentation of MMD [https://htmlpreview.github.io/?https://github.com/metno/mmd/blob/master/doc/mmd-specification.html] for more details:

Extra Attribute	**Repetition**	**Default**	**MMD equivalent**
date_created_type	Comma separated list	Created	last_metadata_update>update>type
collection	Comma separated list		collection
['title_no']	Use ACDD extension for translations		title>title
title_lang	No repetition allowed	en	title>lang
['summary_no']	Use ACDD extension for translations		abstract>abstract
summary_lang	No repetition allowed	en	abstract>lang
dataset_production_status	No repetition allowed		dataset_production_status
access_constraint	No repetition allowed		access_constraint
license_resource	No repetition allowed		use_constraint>resource
contributor_email	Comma separated list	Not available	personnel>email

['contributor_institution', 'contributor_organisation']	Comma separated list	Not available	personnel>organisation
related_dataset_id	Comma separated list		related_dataset>id
related_dataset_relation_type	Comma separated list		related_dataset>relation_type
iso_topic_category	Comma separated list		iso_topic_category
spatial_representation	No repetition allowed		spatial_representation

Acknowledgements

At various stages during the writing of the first version of this handbook, we have solicited comments on the manuscript from coworkers at MET Norway (in alphabetical order): Åsmund Bakketun, Arild Burud, Lara Ferrighi, Håvard Futsæter and Nina Larsgård. Their comments and advice are gratefully acknowledged. In addition, we thank members of the top management at MET Norway, including Lars-Anders Breivik, Bård Fjukstad, Jørn Kristensen, Anne-Cecilie Riiser, Roar Skålin and Cecilie Stenersen, who have provided valuable criticism and advice.

While working on the second version of this handbook, valuable input has come from Matteo De Stefano from NINA. This input has made it possible to transform this handbook into a tool that can be adopted by institutions outside of METNorway.

Glossary of Terms and Names

Term	Description
Application service	TBC
CDM dataset	A dataset that “may be a NetCDF, HDF5, GRIB, etc. file, an OPeNDAP dataset, a collection of files, or anything else which can be accessed through the NetCDF API.” Unidata Common Data Model [https://www.unidata.ucar.edu/software/netcdf-java/v4.6/CDM/index.html]
Configuration metadata	See Configuration metadata definition in Table 2
Controlled vocabulary	A carefully selected list of terms (words and phrases) controlled by some authority. They are used to tag information elements (such as datasets) so that they are easier to search for. (see Wikipedia article [https://en.wikipedia.org/wiki/Controlled_vocabulary]) A basic element in the implementation of the Semantic web.
Data Governance	Tech Target (https://searchdatamanagement.techtarget.com/definition/data-governance). An alternative definition by George Firican: “Data Governance is the discipline which provides all data management practices with the necessary foundation, strategy, and structure needed to ensure that data is managed as an asset and transformed into meaningful information.” (http://www.lightsondata.com/what-is-data-governance/ which also contains several more definitions.)
Data life cycle management	“Data life cycle management (DLM) is a policy-based approach to managing the flow of an information system’s data throughout its life cycle: from creation and initial storage to the time when it becomes obsolete and is deleted.” Excerpt from TechTarget [https://searchstorage.techtarget.com/definition/data-life-cycle-management] article. Alias: life cycle management

Term	Description
Data Management Plan	“A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyse, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.” Stanford Libraries [https://library.stanford.edu/research/data-management-services/data-management-plans]
Data centre	A combination of a (distributed) data repository and the data availability services and information about them (e.g., a metadata catalog). A data centre may include contributions from several other data centres.
Data management	How data sets are handled by the organisation through the entire value chain - include receiving, storing, metadata management and data retrieval.
Data provenance	“The term ‘data provenance’ refers to a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.” (Gupta, 2009). See also Boohers (2015) [https://www.theboohers.org/2015/03/03/provenance/]
Data repository	A set of distributed components that will hold the data and ensure they can be queried and accessed according to agreed protocols. This component is also known as a Data Node.

Term	Description
Dataset	A dataset is a pre-defined grouping or collection of related data for an intended use. Datasets may be categorised by: *Source*, such as observations (in situ, remotely sensed) and numerical model projections and analyses; *Processing level*, such as “raw data” (values measured by an instrument), calibrated data, quality-controlled data, derived parameters (preferably with error estimates), temporally and/or spatially aggregated variables; *Data type*, including point data, sections and profiles, lines and polylines, polygons, gridded data, volume data, and time series (of points, grids, etc.). Data having all of the same characteristics in each category, but different independent variable ranges and/or responding to a specific need, are normally considered part of a single dataset. In the context of data preservation a dataset consists of the data records and their associated knowledge (information, tools). In practice, our datasets should conform to the Unidata CDM dataset definition, as much as possible.
Discovery metadata	See Discovery metadata definition in Table 2
Dynamic geodata	Data describing geophysical processes which are continuously evolving over time. Typically these data are used for monitoring and prediction of the weather, sea, climate and environment. Dynamic geodata is weather, environment and climate-related data that changes in space and time and is thus descriptive of processes in nature. Examples are weather observations, weather forecasts, pollution (environmental toxins) in water, air and sea, information on the drift of cod eggs and salmon lice, water flow in rivers, driving conditions on the roads and the distribution of sea ice. Dynamic geodata provides important constraints for many decision-making processes and activities in society.

Term	Description
FAIR principles	The four foundational principles of good data management and stewardship: *F*indability, *A*ccessibility, *I*nteroperability and *R*eusability. Nature article [RD3 [https://www.nature.com/articles/sdata201618]], FAIR Data Principles [https://www.go-fair.org/fair-principles/], FAIR metrics proposal [https://github.com/FAIRMetrics/Metrics], EU H2020 Guidelines [https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf]
Feature type	A categorisation of data according to how they are stored, for example, grid, time series, profile, etc. It has been formalised in the NetCDF/CF feature type table [https://www.nodc.noaa.gov/data/formats/netcdf/v2.0/#templatesexamples], which currently defines eight feature types.
Geodataloven	“Norwegian regulation toward good and efficient access to public geographic information for public and private purposes.” See Deling av geodata – Geodataloven [https://www.regjeringen.no/no/tema/plan-bygg-og-eiendom/plan—og-bygningsloven/kart/geodataloven/id749728/].
Geonorge	“Geonorge is the national website for map data and other location information in Norway. Users of map data can search for any such information available and access it here.” See Geonorge [https://www.geonorge.no/en/].
Geographic Information System	A geographic information system (GIS) is a system designed to capture, store, manipulate, analyze, manage and present spatial or geographic data. (Clarke, K. C., 1986) GIS systems have lately evolved in distributed Spatial Data Infrastructures (SDI)
Glossary	Terms and their definitions, possibly with synonyms.
Interoperability	The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort.

Term	Description
[linked-data]]Linked data	A method of publishing structured data so that they can be interlinked and become more useful through semantic queries [https://en.wikipedia.org/wiki/Semantic_query], i.e., through machine-machine interactions. (see Wikipedia article [https://en.wikipedia.org/wiki/Linked_data])
Ontology	A set of concepts with attributes and relationships that define a domain of knowledge.
OpenSearch	A collection of simple formats for the sharing of search results (OpenSearch [https://github.com/dewitt/opensearch])
Product	“Product” is not a uniquely defined term among the various providers of dynamical geodata, either nationally or internationally. It is often used synonymously with “dataset.” For the sake of clarity, “product” is not used in this handbook. The term “dataset” is adequate for our purpose.
Semantic web	“The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". W3C [https://www.w3.org/2001/sw/] (see Wikipedia article [https://en.wikipedia.org/wiki/Semantic_Web])
Site metadata	See Site metadata definition in Table 2
Spatial Data Infrastructure	“Spatial Data Infrastructure (SDI) is defined as a framework of policies, institutional arrangements. technologies, data, and people that enables the sharing and effective usage of geographic information by standardising formats and protocols for access and interoperability.” (Tonchovska et al, 2012) SDI has evolved from GIS. Among the largest implementations are: NSDI in the USA, INSPIRE in Europe and UNSDI as an effort by the United Nations. For areas in the Arctic, there is arctic-sdi.org [https://arctic-sdi.org/].
Unified data management	A common approach to data management in a grouping of separate data management enterprises.
Use metadata	See Use metadata definition in Table 2

Term	Description
Web portal	A central website where all users can search, browse, access, transform, display and download datasets irrespective of the data repository in which the data are held.
Web service	Web services are used to communicate metadata, data and to offer processing services. Much effort has been put on standardisation of web services to ensure they are reusable in different contexts. In contrast to web applications, web services communicate with other programs, instead of interactively with users. (See TechTerms article [https://techterms.com/definition/web_service])
Workflow management	Workflow management is the process of tracking data, software and other actions on data into a new form of the data. It is related to data provenance, but is usually used in the context of workflow management systems.
(Scientific) Workflow management systems	A scientific workflow system is a specialised form of a workflow management system designed specifically to compose and execute a series of computational or data manipulation steps, or workflow, in a scientific application. (Wikipedia [https://en.wikipedia.org/wiki/Scientific_workflow_system]) As of today, many different frameworks exist with their own proprietary languages, these might eventually get connected by using a common workflow definition language [https://www.commonwl.org/].

List of Acronyms

This list contains acronyms used throughout the DMH. The column "General/Specific" indicates if the acronyms are used in the general part of the DMH (from the template) or if they are organisation specific.

Acronym	Meaning	General/specific
ACDD	Attribute Convention for Dataset Discovery RD5 [https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3]	
ADC	Arctic Data Centre (ADC [https://pm.met.no/home])	
AeN	Arven etter Nansen (English: Nansen Legacy)	
BUFR	Binary Universal Form for the Representation of meteorological data. WMO standard format for binary data, particularly for non-gridded data (BUFR [https://www.wmo.int/pages/prog/www/WDM/Guides/Guide-binary-1A.html])	
CDM	Unidata Common Data Model (CDM [https://docs.unidata.ucar.edu/netcdf-java/current/userguide/common_data_model_overview.html])	
CF	Climate and Forecast Metadata Conventions (CF [http://cfconventions.org/])	
CMS	Content Management System	
CSW	Catalog Service for the Web (CSW [https://www.ogc.org/standards/cat])	
DAP	Data access protocol (DAP [https://earthdata.nasa.gov/esdis/eso/standards-and-references/data-access-protocol-2])	
DBMS	DataBase Management System (DBMS [https://en.wikipedia.org/wiki/Database#Database_management_system])	

Acronym	Meaning	General/specific
DIANA	Digital Analysis tool for visualisation of geodata, open source from MET Norway (DIANA [https://github.com/metno/diana])	
diana-WMS	WMS implementation in DIANA	
DIAS	Copernicus Data and Information Access Services (DIAS [https://www.copernicus.eu/en/access-data/dias])	
DIF	Directory Interchange Format of GCMD (DIF [https://earthdata.nasa.gov/esdis/eso/standards-and-references/directory-interchange-format-dif-standard])	
DLM	Data life cycle management (DLM [https://searchstorage.techtarget.com/definition/data-life-cycle-management])	
DM	Data Manager	
DMH	Data Management Handbook (this document)	
DMCG	Data Management Coordination Group	
DMP	Data Management Plan (DMP definition [https://en.wikipedia.org/wiki/Data_management_plan], easyDMP tool [https://www.sigma2.no/data-planning])	
DOI	Digital Object Identifier (DOI [https://www.doi.org/index.html])	
eduGAIN	The Global Academic Interfederation Service (eduGAIN [https://edugain.org/])	
ENVRI	European Environmental Research Infrastructures (ENVRI [https://envri.eu/])	
ENVRI FAIR	"Making the ENV RIs data services FAIR." A proposal to the EU's Horizon 2020 call INFRAEOSC-04	

Acronym	Meaning	General/specific
EOSC	European Open Science Cloud (EOSC [https://eosc-portal.eu/])	
ERDDAP	NOAA Environmental Research Division Data Access Protocol (ERDDAP [https://coastwatch.pfeg.noaa.gov/erddap/index.html])	
ESA	European Space Agency (ESA [http://www.esa.int/])	
ESGF	Earth System Grid Federation (ESGF [https://esgf.llnl.gov/])	
EWC	European Weather Cloud ()	
FAIR	Findability, Accessibility, Interoperability and Reusability RD3 [https://www.nature.com/articles/sdata201618]	
FEIDE	Identity Federation of the Norwegian National Research and Education Network (UNINETT) (FEIDE [https://www.feide.no/])	
FFI	Norwegian Defence Research Establishment (FFI [https://www.ffi.no/en])	
FORCE11	Future of Research Communication and e-Scholarship (FORCE11 [https://www.force11.org/])	
GCMD	Global Change Master Directory (GCMD [https://idn.ceos.org/])	
GCW	Global Cryosphere Watch (GCW [https://globalcryospherewatch.org/])	
GeoAccessNO	An NFR-funded infrastructure project, 2015- (GeoAccessNO [https://www.geoaccessno.no/])	
GIS	Geographic Information System	

Acronym	Meaning	General/specific
GRIB	GRIdded Binary or General Regularly-distributed Information in Binary form. WMO standard file format for gridded data (GRIB [https://www.wmo.int/pages/prog/www/WDM/Guides/Guide-binary-2.html])	
HDF, HDF5	Hierarchical Data Format (HDF [https://en.wikipedia.org/wiki/Hierarchical_Data_Format])	
Hyrax	OPeNDAP 4 Data Server (Hyrax [https://www.opendap.org/software/hyrax-data-server])	
IMR	Institute of Marine Research (IMR [https://www.hi.no/en])	
INSPIRE	Infrastructure for Spatial Information in Europe (INSPIRE [https://inspire.ec.europa.eu/])	
ISO 19115	ISO standard for geospatial metadata (ISO 19115-1:2014 [https://www.iso.org/standard/53798.html]).	
IPY	International Polar Year (IPY [https://public.wmo.int/en/bulletin/international-polar-year-2007-2008])	
JRCC	Joint Rescue Coordination centre (Hovedredningssentralen [https://www.hovedredningssentralen.no/english/])	
KDVH	KlimaDataVareHus	Specific to METNorway
KPI	Key Performance Indicator (KPI [https://en.wikipedia.org/wiki/Performance_indicator])	
METCIM	MET Norway Crisis and Incident Management (METCIM [https://metcim.no/])	Specific to METNorway
METSIS	MET Norway Scientific Information System	Specific to METNorway
MMD	Met.no Metadata Format MMD [https://github.com/metno/mmd]	

Acronym	Meaning	General/specific
MOAI	Meta Open Archives Initiative server (MOAI [https://github.com/infrae/moai])	
ncWMS	WMS implementation for NetCDF files (ncWMS [https://reading-escience-centre.github.io/ncwms/])	
NERSC	Nansen Environmental and Remote Sensing Center (NERSC [https://www.nersc.no/])	
NetCDF	Network Common Data Format (NetCDF [https://www.unidata.ucar.edu/software/netcdf/])	
NetCDF/CF	A common combination of NetCDF file format with CF -compliant attributes.	
NFR	The Research Council of Norway (NFR [https://www.forskningsradet.no/en/])	
NILU	Norwegian Institute for Air Research (NILU [https://www.nilu.no/Forsiden/tabid/41/language/en-GB/Default.aspx])	
NIVA	Norwegian Institute for Water Research (NIVA [https://www.niva.no/en])	
NMDC	Norwegian Marine Data Centre, NFR-supported infrastructure project 2013-2017 (NMDC [https://nmdc.no/])	
NorDataNet	Norwegian Scientific Data Network, an NFR-funded project 2015-2020 (NorDataNet [https://www.nordatanet.no/])	
Norway Digital	Norwegian national spatial data infrastructure organisation (Norway Digital [https://www.geonorge.no/en/infrastructure/norway-digital/]). Norwegian: Norge digitalt [https://www.geonorge.no/Geodataarbeid/Norge-digitalt/]	

Acronym	Meaning	General/specific
NORMAP	Norwegian Satellite Earth Observation Database for Marine and Polar Research, an NFR-funded project 2010-2016 (NORMAP [https://normap.nersc.no/])	
NRPA	Norwegian Radiation Protection Authority (NRPA [https://dsa.no/en/])	
NSDI	National Spatial Data Infrastructure, USA (NSDI [https://www.fgdc.gov/nsdi/nsdi.html])	
NVE	Norwegian Water Resources and Energy Directorate (NVE [https://www.nve.no/english/])	
NWP	Numerical Weather Prediction	
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH [https://www.openarchives.org/pmh/])	
OAIS	Open Archival Information System (OAIS [https://en.wikipedia.org/wiki/Open_Archival_Information_System])	
OCEANOTRON	Web server dedicated to the dissemination of ocean in situ observation data collections (OCEANOTRON [https://forge.ifremer.fr/plugins/mediawiki/wiki/oceanotron/index.php/Oceanotron_description])	
OECD	The organisation for Economic Co-operation and Developement. OECD [https://www.oecd.org/]	
OGC	Open Geospatial Consortium (OGC [https://www.ogc.org/])	
OGC O&M	OGC Observations and Measurements standard (OGC O&M [https://www.ogc.org/standards/om])	
OLA	Operational-level Agreement (OLA [https://en.wikipedia.org/wiki/Operational-level_agreement])	

Acronym	Meaning	General/specific
OPeNDAP	Open-source Project for a Network Data Access Protocol (OPeNDAP [https://www.opendap.org/]) - reference server implementation	
PID	Persistent Identifier (PID [https://en.wikipedia.org/wiki/Persistent_identifier])	
RM-ODP	Reference Model of Open Distributed Processing (RM-ODP [https://en.wikipedia.org/wiki/RM-ODP])	
PROV	A W3C Working Group on provenance and a Family of Documents (PROV [https://www.w3.org/TR/prov-overview/])	
SAON	Sustaining Arctic Observing Networks (SAON/IASC [https://iasc.info/data-observations/saon])	
SDI	Spatial Data Infrastructure	
SDN	SeaDataNet [https://www.seadatanet.org/], Pan-European infrastructure for ocean & marine data management	
SIOS	Svalbard Integrated Arctic Earth Observing System	
SIOS-KC	SIOS Knowledge Centre, an NFR-supported project 2015-2018 (SIOS-KC [https://www.sios-svalbard.org/KnowledgeCentre])	
SKOS	Simple Knowledge Organization System (SKOS [https://www.w3.org/2004/02/skos/])	
SLA	Service-level Agreement (SLA [https://en.wikipedia.org/wiki/Service-level_agreement])	
SolR	Apache Enterprise search server with a REST-like API (SolR [https://lucene.apache.org/solr/])	

Acronym	Meaning	General/specific
StInfoSys	MET Norway's Station Information System	Specific to METnorway
TDS	THREDDS Data Server (TDS [https://www.unidata.ucar.edu/software/tds/current/])	
THREDDS	Thematic Real-time Environmental Distributed Data Services	
UNSDI	United Nations Spatial Data Infrastructure (UNSDI [https://en.wikipedia.org/wiki/United_Nations_Spatial_Data_Infrastructure])	
UUID	Universally Unique Identifier (UUID [https://en.wikipedia.org/wiki/Universally_unique_identifier])	
W3C	World Wide Web Consortium (W3C [https://www.w3.org/])	
WCS	OGC Web Coverage Service (WCS [https://www.ogc.org/standards/wcs])	
WFS	OGC Web Feature Service (WFS [https://www.ogc.org/standards/wfs])	
WIGOS	WMO Integrated Global Observing System (WIGOS [https://community.wmo.int/activity-areas/wigos])	
WIS	WMO Information System (WIS [https://community.wmo.int/activity-areas/wis])	
WMO	World Meteorological Organisation (WMO [https://public.wmo.int/en])	
WMS	OGC Web Map Service (WMS [https://www.ogc.org/standards/wms])	
WPS	OGC Web Processing Service (WPS [https://www.ogc.org/standards/wps])	
YOPP	Year of Polar Prediction (YOPP Data Portal [https://yopp.met.no/metadata_search])	

Appendix A: List of Referenced Software or Services

Name	Description	Reference
Fimex package including fimex	File Interpolation, Manipulation and EXtraction library for gridded geospatial data	wiki.met.no documentation [https://wiki.met.no/fimex/start] github repository [https://github.com/metno/fimex]
frost2nc	Dump observational time series from KDVH to NetCDF files	github repository [https://github.com/metno/frost2nc]
met_moai	OAI-PMH implementation based on MOAI	github repository [https://github.com/metno/met_moai]
mdharvest	Perl and Python code to harvest discovery metadata using OAI-PMH, OpenSearch and OGC CSW	github repository [https://github.com/steingod/mdharvest]
METSIS-data -ingestion	A generic utility to index MMD dataset, thumbnails to SolR.	github repository: metsis-metadata [https://github.com/metno/metsis-metadata]
METSIS-data -drupal	A module linking the METSIS back-end services to the Drupal CMS	github repository: metsis-metadata [https://github.com/metno/metsis-metadata]
METSIS-station-handling	TBC	TBC
METSIS-ts	WPS or HTTP interface to graphical diagrams.	Not yet openly available, but beta-version is in use in ADC, SIOS, GCW, NorDataNet, YOPP and APPLICATE portals.
MMD XSD	XML Schema document for MMD	github repository: mmd [https://github.com/metno/mmd]
nc_to_mmd.py	Builds MMD metadata from ACDD-compliant NetCDF file attributes.	github repository: py-mmd-tools [https://github.com/metno/py-mmd-tools/tree/master/py_mmd_tools]
NorDataNet validator	Validates NetCDF files for CF and ACDD compliance.	Access URL [https://nordatanet.metsis.met.no/user/login?destination=dataset_validation/form]
threddslso	Extracting discovery metadata from NetCDF/CF files with ACDD to ISO 19115	github repository [https://github.com/Unidata/threddsIso]

Appendix B: Users of MET Norway's Geodata

Users are divided into categories by type of collaboration with MET, not by type of service they consume.

User category	Description	METNorway Examples
Public service collaborators	Use the data as part of collaborative public services. Includes public sector agencies and companies, government-supported enterprises. Specific requirements on the data and delivery, possibly by SLA or OLA.	Examples: NVE (flood warning service), NRK (yr.no, radio, TV), Norwegian Defence Research Establishment (FFI), Avinor (aviation forecasting), research infrastructures (NORMAP, SIOS, NorDataNet), Norwegian Public Roads Administration (SVV), The Norwegian Coastal Administration (BarentsWatch), JRCC (search-and-rescue support), Bane NOR (Railway Infrastructure Company), Norwegian Environment Agency, NERSC (Copernicus Marine Service), NIVA (Norwegian Institute for Water Research), NILU (air quality monitoring), NRPA (radiation emergency response support), IMR (fish eggs/larvae dispersion)
Value-added service providers	Use open data in proprietary services, often commercial. Value-added and repackaging services. Weak requirements on the data.	Little or no collaboration with MET Norway. Examples: StormGeo (forecasting, consultancy), Navtor (ship routing services), offshore operators, news media, Skiforeningen (iMarka), web and mobile app developers, …
Commercial services	Data services provided under commercial contract, SLA, OLA, or similar.	Examples: Statkraft, Copernicus (CMEMS, C3S, CAMS), offshore operators (EXWW/VXWW), …
Direct consumers	Use data in their own activities (e.g., decision-making, research). Both public and private enterprises. Requirements on data and delivery vary; open data use is preferred.	Halo has tailored datasets for public users. Examples: academic researchers, students, commercial researchers, Defence (FFI), NGOs, DnV

User category	Description	METNorway Examples
Other catalogues and portals	Harvest metadata for inclusion in e.g. thematic catalogues and portals, with access to the data. Both public and private enterprises. No explicit collaboration. Require structured metadata and data access.	Examples: earthengine.google.com, research infrastructures, …
Internal users	The organisation's own activities, covering the full value chain shown in Figure 1. Requires structured data and metadata.	Involves data used for both state-mandated and externally financed activities. Includes a majority of the institute's co-workers in VDiv, SUV, FoU, ObsKlim and IT. Covers: numerical forecast production, emergency model production, climate modelling, collection and QC of operational observations (in situ, remote sensing), analysis of climate records (archived observations, reanalyses, model projections), model development, product development, data dissemination, communication, consultancy, advisories … and data management.