United Nations
Educational, Scientific and
Cultural Organization

Intergovernmental
Oceanographic
Commission

Manuals and Guides 73

Guidelines for a Data Management Plan

UNESCO 2016

Manuals and Guides 73

Guidelines for a Data Management Plan

UNESCO 2016

IOC Manuals and Guides, 73
Version 1
March 2016

EXECUTIVE SUMMARY

A data management plan is a formal document outlining how research data will be managed, stored, documented and secured throughout a research project as well as planning for what will happen to the data after completion of the project.

The data management plan is intended to provide descriptive details of the data, the processes, the decisions, as well as identifying roles and responsibilities. This also includes a long-term data sharing and preservation plan to ensure data are publicly accessible beyond the life of the project. A data management plan is often a requirement of funding agencies.

The IODE encourages all researchers to prepare a data management plan for research projects that will collect marine data and to ensure the data generated by research projects be permanently archived in the IODE network of National Oceanographic Data Centres (NODCs).

For bibliographic purposes this document should be cited as follows:
Guidelines for a Research Data Management Plan. Paris. Intergovernmental Oceanographic Commission of UNESCO, 16pp. 2016. (IOC Manuals and Guides, 73) (English.) (IOC/2016/MG/73)

Printed in France

TABLE OF CONTENTS

Page

1. Introduction 5

2. What is a Data Management Plan? 6
2.1 Initial Planning 6
2.2 Data System Coordination and Planning 7

3. Components of a Data Management Plan 8
3.1 Data to be collected and managed 8
3.2 Documenting the data that will be collected: Metadata 8
3.3 Data formats 9
3.4 Data Assembly and Processing 10
3.5 Data Archival and Preservation 11
3.6 Data Dissemination 12

ANNEX

Annex I : References

1. INTRODUCTION

Data management activities require coordination and planning not only during the data acquisition phase, but the subsequent assembly and curation of the data. This document provides guidance on steps to prepare a data management plan, the activities to consider and suggested actions. Of course, a plan must also be executed to be of any value, and this will require that the data management component be adequately resourced, including staffing.

This document is intended to be a **brief description for preparing a data management plan**. More comprehensive documents that fill out details of the points mentioned here are provided in Annex 1. Readers are encouraged to consult these before completing a data management plan required for their purposes.

While this document was created primarily for use by research or governmental organizations, many of the principles will also apply to data collected by commercial enterprises. Wherever possible, these data should be moved into the public domain as quickly as possible so that they can contribute to the global archives and subsequent analyses that these archives support.

When the data acquisition, assembly, processing and dissemination activities are spread over different groups, coordination of objectives and execution is very important. In every case clarity of the roles and responsibilities of the members of the different groups should be agreed on as early as possible in the project lifespan, and ideally before data acquisition has started. Statements of all responsibilities should be included in the data management plan.

Definitions of some data management terms as part of a data system, used in this document are as follows:

- **Acquisition (collection).** This encompasses the activities by which the data are collected in-situ or remotely. This often includes processes to filter noise, convert from engineering to physical units and even some data smoothing processing. Typically these are actions that do the initial data processing in preparation for post-acquisition processing.

- **Assembly.** This includes the activities of combining into a single data set, all of the data collected during the acquisition phase. The data may be from a single instrument, or multiple, depending on the nature of the acquisition activities. Appropriate metadata are combined with the data at this stage. Sensor calibration, whether done prior to acquisition or as part of data assembly, should be clearly described in the plan.

- **Processing.** This includes all of the data and metadata processing needed to prepare the data for the archives. Typically, it includes quality checks, duplicate resolution, perhaps unit conversions, format changes, etc. Some of these may have occurred in earlier stages and appropriate metadata describing this should be brought together at this stage.

- **Archiving.** This includes insertion of the data into archives that will provide for later dissemination. Actions can include the typical create, modify, replace, delete actions in an archive. Ensuring long-term storage and preservation is also

part of this activity.

- **Dissemination.** This activity includes requests for data and the responses to those requests. It includes identification of appropriate repositories and data catalogue systems. Capabilities such as on-line and off-line data access, reformatting and download are included in this aspect of data management planning.

2. WHAT IS A DATA MANAGEMENT PLAN?

A data management plan is a formal document that outlines how to handle data collected or generated in the course of a research project and what happens to these data during its life-cycle. The goal of the data management plan is to ensure that data are properly collected, documented, made accessible, and preserved for future use. A data management plan is often a requirement of funding agencies.

> The data management plan describes the data that will be collected, the data management practices that will be used, the policies that will apply to the data, who will have custody of the data, who will have access to the data, and who will be responsible for the preservation of the data.

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

2.1 Initial Planning

When planning an activity, people who will be tasked with managing the data should be brought in at the very earliest stage. At this point it is not necessary to identify which group is finally responsible for the data management. That can be decided later once the following basic information is known.

The following questions will need to be addressed:

1. What kind of data will be collected, using what instruments and platforms, and approximately how much and how often the data will ‘come ashore’ or be available for subsequent processing steps?
2. Will the data come through a telecommunications system directly to a data management system or come through a Principal Investigator (PI)? Will all the

data be in digital form?
3. What processing, data availability, and restriction timeframes are to be met by the project? Is there a data policy that describes this? Is data publication (e.g. in a persistent metadata catalogue, with assignment of a Digital Object Identifier (DOI) required?
4. Are there relevant standards for metadata, processing, and delivery that should be used by the data management system of the activity?
5. Are there reasons other than technical ones why data and information should not be made available to the general public as soon as possible?
6. What metadata must accompany the data entering the data management system and subsequently be provided to data requesters (e.g. measurement units, instrument characteristics, sampling and analytical protocols, quality control procedures, data identification, publications, etc.)?

2.2 Data System Coordination and Planning

Having the above information, the data management personnel should document a data management system and discuss this with the planners of the research activity. This discussion should bring out more details and will allow refinement of the data plan. At this stage it should be possible to start identifying which organizations will be participants in the data management workflow and their roles. Prospective organizations will need to make commitments to support the plan including identifying which of their staff will be involved and what financial resources will be needed.

For activities that include many participants, it is common to establish a data management team. Committed organizations should nominate an individual to participate in a data management team for the activity. For smaller activities, the size of the team may be smaller, however, it is vital that there is strong and frequent communication between the data management team and the activity organizers.

The activity planners should select a chair(s) for the data management team. The team then needs to decide on how they will coordinate their activities. This may be done through meetings in a location, virtual meetings, or other electronic ways to share and coordinate actions. An initial meeting of members is advised to discuss the work involved and details of the plan. Thereafter, the team can decide the appropriate means of coordination.

At this point, details of how the data system will function and the respective roles and responsibilities of participants need to be clearly established. The best way to do this is for one or a few individuals to write a data management plan with as much detail as possible that describes the data system activities, including roles and responsibilities of all partners. The plan should describe the details of how the data are assembled from the collectors, what processing is needed, and how data will be made available. The objectives for timeliness of data delivery from collectors, for processing and delivery need to be documented. Any restrictions on data or metadata assembly or delivery must be clearly stated. This draft needs to be completed and circulated quickly, and preferably before data collection begins. Depending on timing, it may be necessary to draft the plan in stages concentrating on the data assembly component first especially if data are starting to arrive.

The plan needs review by both the data management team and the scientists or committee managing the activity. Everyone needs to be in agreement with the plan. Once this agreement is confirmed, work to execute the plan can begin. Arrangements should be made to communicate the plan with research scientists during project

meetings (virtual or face-to-face), to ensure that all project participants are aware of the plan, and to provide opportunities to modify the plan as appropriate for changing research.

The data management plan, once agreed upon, should be readily available to the general public. The usual practice is to place this document on a web site built for the activity.

3. COMPONENTS OF A DATA MANAGEMENT PLAN

The data management plan should be a concise document that addresses the following activities:

3.1 Data to be collected and managed

This includes a description of the expected types of data, samples, collections and other materials that will be produced during the project.

Examples could include:

- Conductivity and temperature from moorings and shipboard CTD surveys
- Currents from ADCP and moorings.
- Weather observations (wind speed and direction, temperature, humidity, precipitation, and cloud cover).
- Species occurrences, and additional environmental or biometric measurements
- Model output and code
- Data from laboratory experiments

3.2 Documenting the data that will be collected: Metadata

Careful description of data is essential to enable data discovery but also to discover errors that were due to, for example, systemic instrument faults. Metadata forms should therefore be developed that capture the basic documentation required to interpret the resultant data.

For a cruise, this includes generation of a cruise report (e.g. Cruise Summary Report form) and a sampling event log recording all instrument activities and deployments.

Mooring deployments should be documented by detailed configuration reports including sensor and instrument components.

Analogous documentation should be provided for other research modes (models and experimental research).

Established protocols should be followed when possible, and the appropriate references cited.

Documentation describing sampling and analytical protocols is essential to enable accurate interpretation by colleagues wishing to collaborate with the original providers of the data, or subsequent future users of the data.

A valuable source of information on marine metadata can be found in the "Marine Metadata Interoperability" web site (http://www.marinemetadata.org).

It is strongly recommended to use an internationally agreed Metadata Standard as this will facilitate interoperability and exchange. Specifically the ISO 19115 Geographic information -- Metadata -- Part 1: Fundamentals (http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798) is a widely utilized metadata standard that combines format and content standards (see also: https://marinemetadata.org/guides/mdatastandards).

For the preparation of metadata a wide variety of tools are available. Many of these can be found here: https://marinemetadata.org/tools. A widely used tool is GeoNetwork (http://www.geonetwork-opensource.org) which is a catalogue application to manage spatially referenced resources. It provides powerful metadata editing and search functions as well as an interactive web map viewer.

A highly desirable element in the description of data is the use of Vocabularies and Standards References. Through their use, data users will have a crystal clear understanding of the data. These can be grouped as follows:

- **Vocabularies**: A vocabulary is a set of terms (words, codes,…) that is used by a specific community. They provide a way to make sure that members of the community all agree on the meaning of a term or code. (see also https://marinemetadata.org/conventions). An example is the BODC parameter dictionary (see https://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/)

- **Ontologies**: An ontology is a representation of knowledge, generally of a particular domain, written with a standardized, structured syntax that describes the relationship between concepts, also called resources, that serve to characterize the domain. (see https://marinemetadata.org/guides/vocabs/ont/definition).

- **Thesauri** are similar to ontologies in that they can describe hierarchical and associative relationships between terms. However, they are generally used to facilitate indexing and retrieval of written and recorded items. (see https://marinemetadata.org/guides/vocabs/voctypes/voccat/thesaurus)

3.3 Data formats

Data acquisition systems of instrumentation have a wide variety of formats in which the information is stored. It is inevitable that in the course of processing the data, they will be transformed to other data structures. Data format types should ideally be considered and decided upon *before* the commencement of data collection. A Data Management Plan needs to document the forms that the data take as they are received, processed/transformed and passed to subsequent users. It is particularly important to

have clear descriptions of data transformation algorithms, such as from engineering to physical units, and from non-standard field names to standards.

For example, a temperature measurement coming out of an instrument may be stored in a field called T in the instrument documentation. A transformation of the value in the field T to a field named TEMP needs to be documented. This can take the form of a document that shows the mapping of information in incoming fields to fields in the subsequent data format. The mapping could include conversion algorithms or whatever other information described how values are transformed.

In the same vein, transformations at data centres from incoming formats to data base fields need also to be mapped and documented.

In brief, the Data Management Plan should provide details of **data formats and standards** including:

- Description of the format in which the data or products will be stored.
- Do data formats conform to an open standard and/or are they proprietary?
- What standards (e.g. vocabularies, units, analysis, etc.) will be used?
- Will the IODE/JCOMM recommended standards and best practice be used? (see: http://www.oceandatastandards.org or http://www.oceandatapractices.org)
- Description of any non-standard formats and how to convert to standard formats.
- Mappings of one format to another when data transformations are done.

3.4 Data Assembly and Processing

There are many operations that are under consideration in data assembly and processing. Essentially this describes the complete set of operations that are applied to incoming data to transform it into a form suitable for the existing or planned archives. As part of this, incoming data should be preserved as received to serve as a reference if questions arise later (such as sometimes happens when backtracking errors).

Additionally, it may be necessary to preserve the software used in the data processing. This can be quite complicated because, of course, software is written in time dependent languages that run on time dependent versions of operating systems. How best to preserve the processing operations embodied in software needs consideration.

Within the framework of the Data Management Plan the following questions will need to be addressed:

- Who is responsible for assembling which data and metadata?
- Do all of the data arrive at once, or do they come in different time frames? If the latter, how are the relationships between data collections preserved? For example, if water samples are collected for chemical analyses in conjunction with standard water temperature measurements made by a CTD, it is common for the chemical analyses to take some time. How will these two sets of measurements be identified as being collected coincidently in time?
- Will all of the required metadata accompany the data, or arrive separately?
- Are there real-time data (those arriving directly from the platform or instrument and rapidly after acquisition) and/or delayed mode (those typically passing through some manipulation by a PI and usually with some time delay after

acquisition)?
- What is the routing and format of data coming to an assembly centre(s)?
- How much data and of which types are expected?
- What are the processing steps (e.g. reformatting, unit conversions, information transformation, quality assessment, duplicates identification, software version control, etc.) from reception to archive? Are these processes well documented and is this documentation readily available to a user?
- What feedback is required to be given to data providers as processing and archiving is achieved?
- Is there any need to monitor processing and show such results? Do these results need to go to JCOMM and/or IODE?
- Does the same variable received from different sources and instruments, pass through identical processing?
- How are real-time and delayed mode data reconciled; that is, how will the two versions be recognized as coming from the same initial measurement?
- Are original values (those as received) preserved, and if so, how?
- What file naming conventions will be used (including standard use of vocabulary, date style, punctuation and numbers)
- Is it necessary and if so, how will any code used in processing the data be preserved and stored?

An important task in the data management chain is verifying the quality of the sampled data. Depending on the data type (and instrument) methodologies exist for this purpose. To some extent the quality control can be carried out automatically (by computer) but in a number of cases this needs to be done manually. It is recommended that standard methodologies should be applied to quality control the data and to indicate the results (the quality flags) of the verification process.

A variety of data quality control manuals have been published. Many of these are available from the OceanDataPractices web site (http://www.oceandatapractices.net) developed and maintained by the IODE programme of the IOC. Another example is The SeaDataNet Quality Control Manual (http://www.seadatanet.org/Standards-Software/Data-Quality-Control).

Data processing descriptions do exist for some international data systems. One example is for the Argo profiling float system. An overview of the components and responsibilities can be found at ftp://www.jcommops.org/argo/Doc/Handbook_1.1.pdf . Other components and greater detail can be found at http://www.argodatamgt.org/Documentation under the subsections of draft and obsolete reports.

3.5 Data Archival and Preservation

The data and information generated as a result of scientific exploration form a valuable legacy of the programme that collected and used the data for scientific research. While today's portable technology allows nearly anyone to have huge data systems on their portable device this also considerably increases the risk of data loss, as compared to the era prior to personal computing when data were stored in mainframe or mini computer systems that were professionally managed. The Data Management Plan must include a discussion of the responsibilities of every partner of the data system (researcher, intermediate data processor, or archive centre) to ensure for the long-term retention, archival, preservation and disposal of the data.

While data access may be restricted for some time to allow analysis and publication of results, all data should ultimately be archived in permanent archives. Examples are the IODE National Oceanographic Data Centres (NODCs) and ICSU World Data System centres. A very useful aggregation of much of the ocean data collected in the world occurs in the World Ocean Database hosted and managed by the NOAA National Centers for Environmental Information (see https://www.nodc.noaa.gov/OC5/WOD/pr_wod.html). Data Management planners are strongly encouraged to include submission of their programme data to this centre.

Within the framework of the Data Management Plan the following questions will need to be addressed:

- Details of archive update procedures including frequency
- Details of backup procedures (and frequencies), disaster recovery and offsite backup facilities.
- Will the data be submitted to National Oceanographic Data Centres, IODE Associated Data Units or OBIS nodes or other repositories such as DRYAD, DataONE etc.?
- If the data are spread over more than one archive, is there information to tell a user how to reassemble data from the different archives?
- Details of the data retention period and any plans for disposal.

The longevity of research data formats and software versions is an important component of data preservation. The Data Management Plan should address the procedures needed to ensure that research data is secure and retrievable for long term use. Conversion of data into standard interchangeable formats may be necessary for preservation purposes. When selecting file formats for archiving, the formats should ideally be:

- Non-proprietary and not tied to a specific type or version of software
- Open, documented standard
- In common use by the community
- Standard representation (ASCII, Unicode)
- Unencrypted
- Uncompressed

In many cases a research project will also collect large quantities of shipboard information ranging from core analysis, genetic samples, video and images. The final repository for these data should also be included in the Data Management Plan.

3.6 Data Dissemination

Once data have been processed and inserted in the appropriate archive, they should be made available, preferably online. It is at this stage that an agreed upon data exchange policy is important: some data may become available to all users immediately, others may require a temporary embargo (to allow further scientific work, while some may be restricted). However even embargoed or restricted data should be discoverable through metadata. It is important that data are disseminated in a timely fashion.

While it is understood that initial periods of exclusive data use are customary, the ultimate goal should be to make all data openly available as soon as possible and ideally within two years of acquisition. The early sharing of data fosters collaboration

that results in improved data quality and additional opportunities for scientific publication and ultimately future cooperative research proposals.

A new way of making data available is through data publishing/data citation: when writing a scientific publication the author will use data sets. Some of these data sets can be stored individually and given a DOI. As such these data sets can be referred to in a unique and persistent manner. This is also important when data sets are used from large data bases that are continuously managed (and thus changed). We recommend consulting the "Ocean Data Publication Cookbook" (IOC Manuals and Guides No. 64) (see http://www.iode.org/mg64).

When designing a data discovery and retrieval service (whether off- or on-line) it is essential to design such a system with the end user in mind, rather than designing an all encompassing interface that allows queries of the highest complexity but confusing to the "average" user. Similarly the data should be provided in easy to use formats.

Within the framework of the Data Management Plan the following questions will need to be addressed:

- Who is responsible for data distribution?
- Describe the data flows (from data provider to regional to global data aggregators)
- Are there any restrictions on the distribution or use of the data and are these clearly described for potential users?
- Are there clearly defined conditions for general release of the data, such as licensing?
- Is there a documented data sharing policy in place for public access?
- Is there a web site for the Project and is this well advertised?
- What data and product services are offered?
- Are the data available using web services (e.g. WMS, WFS, Rest API etc.)?
- What data access protocols will be used to enable data sharing (e.g. THREDDS, OpeNDAP, FTP, etc.)?
- Can users provide feedback?
- Are the data to be published using such forms as DOIs or some other method?
- Is there readily available documentation that describes how the products are made?

ANNEX I
REFERENCES

There are a number of detailed documents that provide additional information on data management planning and data management best practices. These include the following:

***OceanObs'09 Data Management Review*.** This comprises four documents that describe an overview of marine data management, details of what needs consideration in assembling data into an archive, details about considerations for making archive data available, and a look to the future. (OceanObs'09 Proceedings are available online at http://www.oceanobs09.net/)

The IMBER Data Management Cookbook – A Project Guide to good Data practices. This is particularly strong in describing detailed data acquisition planning and the collaboration between data managers and researchers in the data collection process. (available at http://www.imber.info/index.php/Science/Working-Groups/Data-Management/Cookbook)

JCOMM Data Management Plan (JCOMM Technical Report No. 40). This Data Management Plan explains how data management can be conducted to promote the long-term objectives of JCOMM. The plan presents a review of the various components of data management that must be considered as part of JCOMM. (available at http://www.jcomm.info/index.php?option=com_content&view=article&id=29&Itemid=100024)

***BCO-DMO Data Management resources*.** A collection of best practice recommendations for sharing biogeochemical and ecological oceanographic data and metadata from NSF funded projects to the BCO-DMO data system. Influenced by the U.S. NSF requirements, provides a good example of a data policy and the various considerations that support it. (available at http://www.bco-dmo.org/resources)

***Data Management Planning Tool*.** Online tool to develop a data management plan. Mainly intended for US users but can be used by others as well. Requires creation of user account (available at https://dmptool.org/)

***NSF Data Management Plan* (example).**This provides a simple document describing a project, its objectives and how the data will be managed. (available at http://www.dataone.org/sites/all/documents/DMP_Copepod_Formatted.pdf)

Data Management Plan Examples. This provides a list of US data management plans. (available at http://data.library.arizona.edu/data-management-plans/data-management-plan-examples)

Guidelines on Data Management in Horizon 2020 (Version 2,1, 15 February 2016. This document is intended to help applicants and beneficiaries of projects under Horizon 2020 meet their responsibilities as regards research data quality, sharing and security. (available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)

Data Management Plans from the UK Digital Curation Centre (DCC) includes various data management resources and online tools (available at http://www.dcc.ac.uk/resources/data-management-plans)

Data Management Planning (by NERC, UK). Provides outline and guidance documents. All NERC proposals require an Outline Data Management Plan to identify data sets of long term value that should be made available to NERC data centres for archiving and reuse at the end of the fellowship or grant (available at http://www.nerc.ac.uk/research/sites/data/dmp/)

Data Management Plan Tool v.2. The IEDA Data Management Plan (DMP) Tool provides an easy way to generate a DMP for inclusion in NSF proposals. While IEDA's focus is on solid earth data from the marine, terrestrial, and polar environments, the DMP has been designed as a generic form that can be used for proposals being submitted to other NSF Divisions. (available at http://www.iedadata.org/compliance/plan)

USGS Data Management Plans. Includes checklist and examples of USGS Data Management Plans. (available at http://www.usgs.gov/datamanagement/plan/dmplans.php)

Intergovernmental Oceanographic Commission (IOC)
United Nations Educational, Scientific and Cultural Organization
1, rue Miollis, 75732 Paris Cedex 15, France
Tel: + 33 1 45 68 39 83
Fax: +33 1 45 68 58 12
http://ioc.unesco.org

IOC Project Office for IODE
Wandelaarkaai 7/61
8400 Oostende, Belgium
Tel: +32 59 34 21 34
Fax: +32 59 34 01 52
http://www.iode.org