D3.1 – Definition of data formats and metadata structure

Authors: Torill Hamre, Øystein Godøy, Sjur Lid Ringheim, Stein Tronstad, Helge Sagen, Kjetil Fjellheim, Svein Østerhus, Morten Krogstad, Aave Lepland, Nina Nordlund, Svein Kristiansen, Roald Vang, Jan Karud, Nils Valland, Benjamin Pfeil

Revision: 0.11
Date: 2014-05-15

Approved by Steering Group 2014-05-28

Revision history

Version	Date	Comment	Responsible
0.1	2013-05-08	Initial topic layout.	Torill Hamre
0.2	2013-05-14	First version of terminology, reference list and major metadata standards (DIF, ISO 19115)	Torill Hamre
0.3	2013-05-22	Revised terminology, added references and background information.	Torill Hamre
0.4	2013-05-29	Added description of the pre-project metadata structure.	Torill Hamre
0.5	2013-10-16	Revised definition of data set and included list of relevant metadata standards and vocabularies after annual meeting.	Øystein Godøy, Torill Hamre
0.6	2013-10-31	Referred to concepts document for terminology, and added and first version of anticipated products list from partners.	Torill Hamre
0.7	2013-11-22	Inserted description of related initiatives and projects, with links to home pages, documentation and web portals.	Torill Hamre
0.8	2013-12-17	Updated after comments from Kjetil Fjellheim, Helge Sagen and Øystein Godøy. Inserted more references and links to related initiatives (e.g. DOKIPY), and made clarifications (e.g. question for partner's data formats, terminology).	Kjetil Fjellheim, Helge Sagen, Øystein Godøy, Torill Hamre
0.9	2014-02-11	Updated after input from UiN, NGU, NPI, FFI, UiT, NINA, IMR.	Morten Krogstad, Aave Lepland, Stein Tronstad, Nina Nordlund, Svein Kristiansen, Roald Vang, Kjetil Fjellheim
0.10	2014-03-31	Updated after input from NIVA, UiO, NBIC (Norwegian Biodiversity Information Center) Artsdatabanken, UIB, UniResearch. Prepared for internal review.	Jan Karud, Joel Durant, Nils Valland, Benjamin Pfeil, Torill Hamre
0.11	2014-05-15	Incorporated comments from quality control.	Torill Hamre

Contents

1 Introduction

1.1 Background

The Norwegian Ministry of Fisheries and Coastal affairs have identified the importance of an operational national service and competence centre for marine data. The national infrastructure, the Norwegian Marine Data Centre (NMDC), will be the focal access point for marine data in Norway. The conceptual model for NMDC shown in Figure 1 will evolve from existing institutional data centres in Norway by integrating, merging, coordinating and expanding activities and services. One of these local centres is the Norwegian Marine Data centre (NMD), which has the national responsibility of archiving oceanographic data in addition to acoustic, biological and other data at the Institute of Marine Research (IMR). Several other Norwegian institutions also have data management activities and online data distribution systems, among others, the Norwegian Meteorological Institute (Met.no), University of Tromsø (UiT), Norwegian Polar Institute (NPI), Norwegian Institute for Water research (NIVA) and Norwegian Institute for Nature Research (NINA).

NMDC aims to serve the marine science community with seamless access to documented marine data sets covering waters of Norwegian interest [1]. The national infrastructure long-term data storage will be organised in three main activities, user contribution of data, data products/storage and data applications for metadata and data exchange. The OECD's principles for free and unrestricted access to research data will act as an underlying principle for data access.

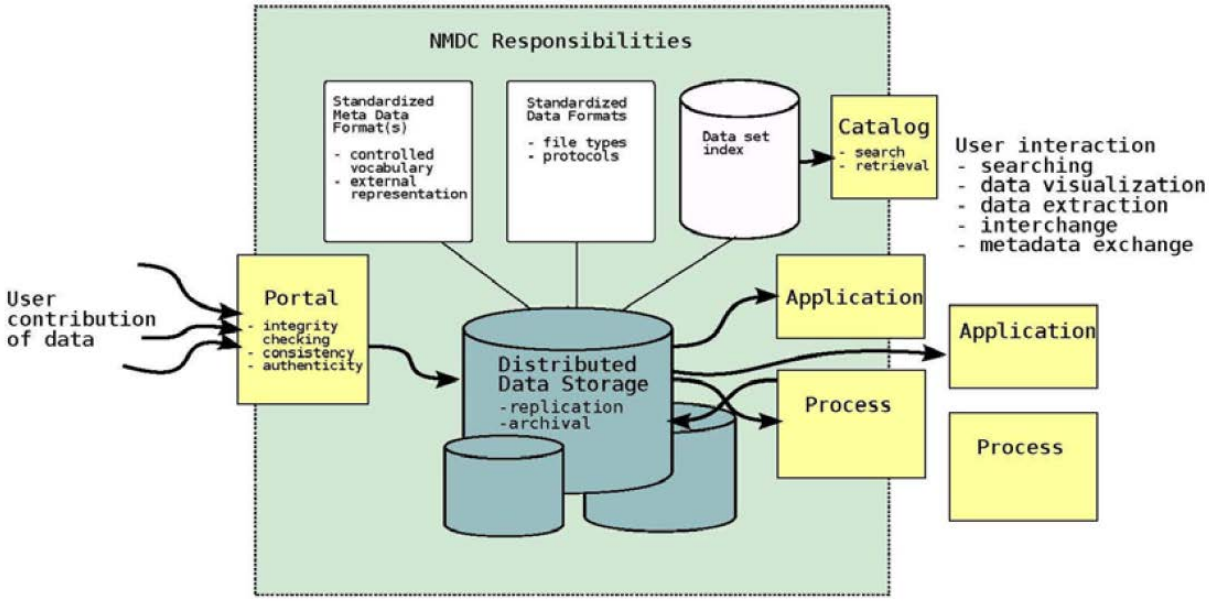

Figure 1 The conceptual data model for the NMDC.

Data sets for the NMDC are collected during field experiments or regular monitoring programmes, by fixed or moving instruments mounted on land or deployed in the ocean. Each (data providing) partner has his own established data acquisition, pre-processing and data quality chain that make the data available for internal use within the partner organisation. Some organisations also offer access to their data repositories through

interfaces, but not necessarily in a standard format or through a standard data exchange protocol. To be compatible with the NMDC, interface software will have to be installed locally on top of the databases at the specific data repository. These distributed repositories will have to comply with the guidelines of best data management practices for their data management service to be accepted as a local repository. Data products activities in the NMDC project will focus on harmonising quality control routines made upon the data sets, develop databases for data storage and conversion software making exchange of data easier. This includes making the currently non digital data sets available in digital form through the same standard interfaces and in the same standard data formats.

The NMDC project will also offer access to partners' data sets through a common web portal, where users can search for data sets based on their temporal and geographic coverage. Data that naturally are presented in an aggregated way (e.g. in a map) will be made available by implementing services like Web Map Service (WMS), Web Feature Service (WFS) and Web Coverage Service (WCS) which comply with OGC standards. The commonly used protocol OPeNDAP (Open Source Project for a Network Data Access Protocol) [18] is one of the supported protocols the infrastructure will use to facilitate direct data access.

A national data catalogue containing metadata information on data sets and time series will be established at NMDC based on existing work carried out by the Norwegian Research Council and the pre-project on NMD [21]. The lists of data sets and time series are organised in three reports, important marine time series [17], important climatology time series [18], and important terrestrial and inland waters time series [19]. This will serve as a basis for an updated national catalogue in the current NMDC project.

The research infrastructure will manage both historical and near real-time data. The latter requires data to be exchanged with low latency and will need a message oriented infrastructure for the distribution of the data. This infrastructure will collect data from instruments and distribute it both to end users and to special purpose services for quality control and automatic archiving. These real-time data could then again be distributed to vessels along the Norwegian coast over the AIS network as special purpose AIS messages (MetHyd) allowing important information to be available on board.

1.2 Objectives of report

This report will describe standard metadata and data formats that are candidate formats for use in the NMDC and make an overall recommendation for which format(s) to use both within the system and to external parties. The recommendation will be based on a review of established standard metadata and data formats used by the NMDC partners and in the marine scientific community. The report will also propose a set of standard data discovery and delivery protocols that the NMDC should support to distribute data sets to users.

1.3 Terminology

The following terms are used within this report (extracted from [2]):

- *Data set* – A data set is a collection of measurements that are acquired by one of more instruments within a well defined and known geographical area and time period, that are described by an informative and (preferably) unique title, a version number (alphanumeric), an abstract summarising the content, marked up with keywords from designated vocabularies, and that have a unique identifier within the NMDC system. Examples of data sets include all measurements collected during a field experiment (e.g. scientific cruise) or seismic campaign, all measurements collected by a meteorological station or weather ship, drifting of fixed buoy, glider or cluster of

acoustical cabled moorings, as well as biological or physical samples.

- *Metadata* – A metadata record is a file of information, usually presented as an XML document, which captures the basic characteristics of a data or information resource. It represents the who, what, when, where, why and how of the resource. Metadata are often divided in Discovery metadata and use metadata where the first describes overall characteristics of the data set while the latter is required to actually use the data set without previously knowing the contents or speaking to the data provider. See http://www.fgdc.gov/metadata for details.
- *Spatial data* – Any data with a direct or indirect reference to a specific location or geographic area. (INSPIRE Directive).
- *Spatial data set* – An identifiable collection of spatial data. (INSPIRE Directive).
- *Discovery service* – Distinct part of the functionality that is provided by an entity through interfaces for the inquiry of the nature and content of a spatial resource.
- *Download service* - Distinct part of the functionality that provides access to the full extent of geographic and thematic information in a data set. By access to the full extent is meant access to features described with feature attributes and real world coordinates. A download service may include elements of a feature type service, a coordinate conversion/transformation service and schema transformation service.
- *Portrayal service* – Distinct part of the functionality that provides a graphical presentation of a data set (in part or whole) or a combination (overlay) of multiple data sets transformed to a common map projection and spatial resolution.

1.4 References

[1] Norwegian Marine Data Centre, Project plan, 1 August 2012 – 31 July 2017, NFR Contract 108849.

[2] Norwegian Marine Data Centre, Acronyms and Concepts, version 0.1.

[3] Norwegian Marine Data Centre, Deliverable D2.2 Use cases, version 0.3.

[4] Norwegian Marine Data Centre, Deliverable D2.3 System Requirements, version 0.3.

[5] ISO/TC-211, ISO 19115:2003 Geographic information - Metadata. 2003.

[6] ISO/TC-211, ISO 19119 :2005 Geographic information - Services. 2005.

[7] ISO 19139:2007, Geographic information - Metadata – XML Schema Implementation.

[8] OpenGIS Catalogue Service Specification, Version 2.0.2, OGC 07-006r1. Online at http://portal.opengeospatial.org/files/?artifact_id=20555 (accessed on 22 April 2010).

[9] WMO, 2008. WMO Information System Functional Architecture "Managing and Moving Weather, Water and Climate Information in the 21st Century".

[10] Directory Interchange Format (DIF) Writer's Guide, 2013. Global Change Master Directory. National Aeronautics and Space Administration. Online at http://gcmd.nasa.gov/add/difguide/ (accessed 26 April 2013).

[11] INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119, V. 1.3, 2013-10-29. Online at http://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf (accessed 7 May 2014).

[12] WMO Core Metadata Profile version 1.2 Guidelines on the use of Metadata for WIS Document version: 0.1 (DRAFT) Author: Jeremy Tandy (IPET-MDI) Date: 12 November 2010. Online at

http://wis.wmo.int/2010/metadata/version_1-2/WMO%20Core%20Metadata%20Profile%20v1-2%20Guidance%20Documentation%20v0.1%20%28DRAFT%29.pdf (accessed 26 April 2013).

[13] NORMAP System Design Document. Øystein Godøy. Version 3. 8 September 2001.

[14] NetCDF Climate and Forecast (CF) Metadata Conventions Version 1.6, 5 December, 2011. Online at http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.6/cf-conventions.pdf (accessed 22 May 2013).

[15] OPeNDAP, www.opendap.org, 2013 (accessed 27 May 2013).

[16] NMDC pre-project, http://talos.nodc.no:8080/infrastruktur, 2013 (accessed 16 December 2013).

[17] Norges forskingsråd, 2004, Lange tidsserier for miljøovervåkning og forskning Rapport nr, 3 Viktige marine dataserier.

[18] Norges forskingsråd, 2003, Lange tidsserier for miljøovervåkning og forskning. Rapport nr, 1 Viktige klimadataserier.

[19] Norges forskingsråd, 2003, Lange tidsserier for miljøovervåkning og forskning.Rapport nr.2, Viktige terrestriske og limniske dataserier.

[20] Olsen, L.M., G. Major, K. Shein, J. Scialdone, S. Ritz, T. Stevens, M. Morahan, A. Aleman, R. Vogel, S. Leicester, H. Weir, M. Meaux, S. Grebas, C.Solomon, M. Holland, T. Northcutt, R. A. Restrepo, R. Bilodeau,2013. NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 8.0.0.0.0. Online at http://gcmd.nasa.gov/learn/keyword_list.html (accessed 27 May 2013).

[21] GeoTIFF Format Specification GeoTIFF Revision 1.0. Online at http://www.remotesensing.org/geotiff/spec/geotiffhome.html (accessed 28 May 2013).

[22] Hierarchical Data Format - Earth Observing System (HDF-EOS). Online at http://nsidc.org/data/hdfeos/ (accessed 28 May 2013).

[23] ESRI, 1998. ESRI Shapefile Technical Description An ESRI White Paper - July 1998.

[24] CSV format. http://en.wikipedia.org/wiki/Comma-separated_values (accessed 29 May 2013).

[25] Clemens Portele (editor), 2007. OpenGIS® Geography Markup Language (GML) Encoding Standard, Version: 3.2.1. OGC document OGC 07-036.

[26] Darwin Core, http://rs.tdwg.org/dwc/ (accessed 16 October 2013).

[27] The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14. Online at http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm (accessed 31 October 2013).

[28] OGC Web Map Service (WMS) Implementation Specification, version 1.3.0, OGC 06-042, Online at http://portal.opengeospatial.org/files/?artifact_id=14416 (accessed 21 April 2010).

[29] OGC Web Feature Service Implementation Specification, version 1.1.0, OGC 04-094. Online at http://www.opengeospatial.org/standards/wfs (accessed 5 January 2012).

[30] OGC Web Coverage Service (WCS) Implementation Specification, version 1.1.2, OGC 07-067r5, Online at http://portal.opengeospatial.org/files/?artifact_id=27297 (accessed 22 April 2010).

[31] SOSI http://www.kartverket.no/Standarder/SOSI/

[32] SEG Y rev 1 Data Exchange format, SEG Technical Standards Committee, 2002. http://www.seg.org/documents/10161/77915/seg_y_rev1.pdf

[33] Saarenmaa, H. et al., 2014, EU BON Deliverable 2.1 (D2.1) Architectural design, review and guidelines for using standards.

[34] Pearlman J., Albert Williams III, and Pauline Simpson (editors), 2013, Facilitating Open Exchange of Data and Information. Report of the Research Coordination Network RCN:OceanObsNetwork. Online at

https://darchive.mblwhoilibrary.org/bitstream/handle/1912/5937/RCN_Open_data_report_final.pdf?sequence=1 (accessed 6 May 2014).

[35] Olsen, L.M., G. Major, K. Shein, J. Scialdone, S. Ritz, T. Stevens, M. Morahan, A. Aleman, R. Vogel, S. Leicester, H. Weir, M. Meaux, S. Grebas, C.Solomon, M. Holland, T. Northcutt, R. A. Restrepo, R. Bilodeau, 2013. NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 8.0.0.0.0.

[36] EDMED, HOW TO PREPARE ENTRIES FOR EDMED, 1999. Online at https://www.bodc.ac.uk/data/information_and_inventories/edmed/documents/edmed1.pdf (accessed 8 May 2014)

[37] World Wide Web Consortium, 1999, Hypertext Transfer Protocol -- HTTP/1.1. Online at http://www.w3.org/Protocols/rfc2616/rfc2616.html (accessed 9 May 2014)

[38] NMDC home page, http://www.nmdc.no.

[39] JSON home page, http://json.org/ (accessed 9 May 2014)

[40] GeoJSON home page, http://geojson.org/geojson-spec.html#introduction (accessed 9 May 2014)

[41] Houbie, F. And L. Bigagli (eds), 2010, OGC® Catalogue Services Standard 2.0 Extension Package for ebRIM Application Profile: Earth Observation Products. OGC® Implementation Standard. OGC 06-131r6, version 1.0.0.

[42] ISO 19115-2:2009 Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data.

[43] Open Geospatial Consortium Inc., OpenGIS Geography Markup Language (GML) Application Schema for Earth Observation Products. 0GC 06-080r4, version 1.0.

[44] Open Geospatial Consortium Inc., CSW-ebRIM Registry Service - Part 1: ebRIM profile of CSW. OGC 07-110r4, version 1.0.1 (Corrigendum 1).

[45] OASIS, ebXML Registry Information Model. Version 3.0. OASIS Standard, 2 May, 2005.

[46] Eaton,B., J. Gregory, B. Drach, K. Taylor and S. Hankin, NetCDF Climate and Forecast (CF) Metadata Conventions, Version 1.6, 5 December, 2011.

[47] ECS, Release 7 Implementation Earth Science Data Model for the ECS Project. May 2004.

1.5 Outline of report

The remainder of the report is organised as follows. Section 2 briefly describes some relevant international geo-spatial metadata and interoperability initiatives and projects. Section 3 identifies relevant metadata standards, including the structure used in the NMDC pre-project. This section also describes the metadata structure that will be used in the forthcoming NMDC system, while an example metadata file in XML format is included in Appendix A. Section 4 identifies a list of data formats that are used by partners for NMDC products, and the formats that will be offered to external clients. Section 5 outlines candidate metadata and data exchange protocols. Section 6 concludes with a summary of recommendations for metadata and data standards, as well as protocols for metadata and data exchange for the NMDC system.

2 International geo-spatial metadata initiatives and projects

Many international initiatives and projects address similar issues to the NMDC, namely those of defining a common and standards-based metadata structure for geographic data sets and of developing a joint infrastructure for discovery, access and presentation of such data. This section briefly describes some of these initiatives and projects, outlining their objectives and if available, including links to available standards, best practices, vocabularies and/or data portals. (The listing is in alphabetical order and does not imply any prioritization or ranking.)

2.1 Data Observation Network for Earth (DataONE)

DataONE aims to develop "a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data"[1]. At the time of writing, the consortium is comprised of fourteen partners from academia, environmental protection, libraries, data centres, data and knowledge networks. DataONE is supported by the US National Science Foundation (NSF), and is organised as a set of Member Nodes (contributing data) and Coordinating Nodes (mirroring data), and in thematic Working Groups conducting research, developing infrastructure, and engaging users through dedicated workshops and educational activities. Besides access to data, DataONE offers training activities and resources, best practices, and software tools to help developing Data Management Plans[2] as well as an Investigator Toolkit[3] for searching, accessing and contributing data in the DataONE network. DataONE supports multiple metadata standards, including Dublin Core, Dublin Kernel, and DataCite Kernel[4].

- Home page: http://www.dataone.org/
- Metadata catalogue: https://cn.dataone.org/onemercury/
- Metadata structure: http://mule1.dataone.org/ArchitectureDocs-current/design/WhatIsData.html

2.2 Building the European Biodiversity Observation Network (EU BON)

The main objective of EU BON is to build a substantial part of the Group on Earth Observation's Biodiversity Observation Network (GEO BON). In light of the new Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services (IPBES), such a network and approach are imperative for attaining efficient processes of data collation, analysis and provisioning to stakeholders.

A key feature of EU BON will be the delivery of near-real-time relevant data – both from on-ground observation and remote sensing – to the various stakeholders and end users ranging from local to global levels. Doing so will require (1) the establishment and adoption of new data standards and integration techniques, (2) harmonized data collection, and (3) the development of new approaches and strategies for future biodiversity monitoring and assessment. We will develop and provide practical indicators and interpretation tools for endusers from, e.g., agriculture to nature conservation. EU BON will support national and international authorities, as well as private stakeholders and the general public with integrated and scientifically sound biodiversity data analyses. The project intends to develop a full-scale

[1] http://www.dataone.org/what-dataone

[2] https://dmp.cdlib.org/

[3] http://www.dataone.org/investigator-toolkit

[4] https://wiki.auckland.ac.nz/display/eResearchNZ/DataONE,Dryad+and+EZID

model for a durable mechanism for higher level integration of biodiversity information providers and users through a network of networks approach scalable from local to global biodiversity observation systems.

EU BON proposes two related networking levels, (1) a science-based social network, comprising and linking the communities of practice engaged in collecting, managing, analyzing, and utilizing biodiversity observations and data, and (2) a technological network of interoperating IT infrastructures and systems that store and distribute information of all kinds held by multiple organisations and partners, and to provide a platform for data analysis and interpretation. For resource efficiency, the establishment of the GEO biodiversity information network will build on existing infrastructures and efforts to integrate monitoring schemes and their data across Europe and internationally.

In doing that, EU BON will have the following specific objectives:

- advancing the technological/informatics infrastructures for GEO BON, by moving existing biodiversity networks towards standards-based, service-oriented approaches and cloud computing, enabling full interoperability through the GEOSS Common Infrastructure;
- improving the range and quality of the methods and tools for assessment, analysis, and visualization of biodiversity and ecosystem information, particularly focussing on predictive modelling, identification of drivers of change, and biodiversity indicators, and to support priority setting;
- Home page: http://eubon.eu/.

EU BON has also reviewed several biodiversity e-infrastructure projects [33], including:

- BioFresh (Biodiversity of Freshwater Ecosystems: Status, Trends, Pressures, and Conservation Priorities);
- BIO_SOS (Biodiversity Multi-Source Monitoring System from Space to Species);
- BioVeL (Biodiversity Virtual e-Laboratory);
- EBONE (European Biodiversity Observation Network);
- EUMON (EU-wide monitoring methods and systems of surveillance for species and habitats of Community interest);
- KNEU (KNEU – Developing a Knowledge Network for EUropean expertise on biodiversity and ecosystem services to inform policy making economic sectors);
- MS.MONINA (Multi-Scale Service for Monitoring Natura 2000 Habitats of European Community Interest);
- PESI (Pan-European Species directories Infrastructure);
- pro-iBiosphere (pro-iBiosphere);
- ViBRANT (Virtual Biodiversity Research and Access Network for Taxonomy)

2.3 European Global Ocean Observing System (EuroGOOS)

EuroGOOS is an International Non-Profit Organisation committed to establishing operational oceanography services across Europe, to provide the European component of the Global Ocean Observing System (GOOS). EuroGOOS is currently comprised of 36 members originating from 18 European countries. The members include national government agencies and research organisations, which all provide operational oceanographic services or conduct marine research in national and/or regional ocean waters.

Six Working Groups have been defined:

- Coastal Ocean and Shelf Seas Modelling Working Group (COSMO WG),
- Data Management, Exchange and Quality Working Group (DATA-MEQ),
- EuroGOOS Product Working Group (EPWG)
- ECOMF Strategic Partnership Working Group (ESPWG)
- Science Advisory Working Group (SAWG)
- Technology Plan Working Group (TPWG)

to organise and develop strategies and standards in the respective domains, and to set priorities for the tasks and actions to be carried out to establish the EuroGOOS system through a set of regional ocean observing systems (ROOS). The following ROOS have been established: the Arctic (Arctic ROOS), the Baltic (BOOS), the North West Shelf (NOOS), the Ireland-Biscay-Iberian area (IBI-ROOS) and the Mediterranean (MONGOOS). In addition, an MoU is set up with the Black Sea GOOS, which acts as a ROOS for the Black Sea region.

The developed strategies and standards are actively disseminated and promoted to relevant European and national operational enties and funding, to raise awareness and amplify impact. EuroGOOS also works closely with the international operational oceanographic community, in particular with JCOMM and global programs like GODAE[5], ARGO[6], OceanSITES (section 2.14) and SOOP[7].

- Home page: http://eurogoos.eu/

2.4 European Collaborative Data Infrastructure (EUDAT)

EUDAT is an initiative under the EC-FP7 infrastructure programme. The EUDAT consortium is comprised of 25 organizations from 13 countries, and includes national data centres, technology providers and research. EUDAT aims to develop a cost-effective and high-quality Collaborative Data Infrastructure (CDI) with common data delivery services across multiple scientific domains. This will enable community-specific support services to be developed by existing data centres and service providers, and also offer sustainable data and metadata services for new communities, individual scientists and small organizations that do not have the means or capacity to set up research infrastructure of their own.

In the first part of the project, EUDAT has collected and reviewed requirements and approaches from a selected subset of science domains: (1) linguistics[8], (2) solid earth sciences[9], (3) climate sciences[10], (4) environmental sciences[11], and (5) biological and medical sciences[12]. Based on these consultations, EUDAT has identified four generic services[13], of which one service is a joint metadata catalogue. This service will allow metadata access for data held by EUDAT partners, and in addition harvest metadata using OAI-PHM[14] (with domain specific metadata embedded) from other metadata providers to allow multi-disciplinary searches.

[5] Global Ocean Data Assimilation Experiment, http://www.godae.org/

[6] Global ARGO network, http://www.argo.net/

[7] SHIP-OF-OPPORTUNITY PROGRAMME (SOOP), http://www.jcommops.org/soopip/soopip_overview.html

[8] CLARIN (Common Language and Resource Technology Infrastructure): http://www.clarin.eu/

[9] EPOS (European Plate Observing System: http://www.epos-eu.org/

[10] ENES (European Network for Earth System Modelling): https://is.enes.org/

[11] LIFEWATCH: http://www.lifewatch.eu

[12] VPH (Virtual Physiological Human): http://www.vph-noe.eu/home

[13] http://ijdc.net/index.php/ijdc/article/view/8.1.279/312

[14] http://www.openarchives.org/OAI/openarchivesprotocol.html

- Home page: http://www.eudat.eu/

2.5 European Marine Observation and Data Network (EMODNet)

EMODNET is a "contribution to the EU Integrated Maritime Policy" [15] and in its first phase a series of pilots were developed to "migrate fragmented and inaccessible marine data into interoperable, continuous and publicly available data streams for complete maritime basins". Thematic groups were defined: six thematic assembly groups were set up: (1) hydrography, (2) geology, (3) physics, (4) chemistry, (5) biology and (6) physical habitats.

For the five first of these groups, pilot portals have been developed (see URLs below). The chemistry, physical parameters and hydrography pilots were built using system components and tools from SeaDataNet, such as the Common Data Index (CDI) data discovery and access service. The biology pilot is based on the European Ocean Biogeographic Information System[16] (EurOBIS) and the Integrated Marine Information System[17] (IMIS) using taxonomies from the European Register of Marine Species[18] (ERMS), which constitutes the European component in the World Register of Marine Species[19] (WoRMS).

- Home page: https://webgate.ec.europa.eu/maritimeforum/category/162
- Hydrography pilot portal: http://www.emodnet-hydrography.eu/
- Geology pilot portal: http://www.emodnet-geology.eu/
- Physical Parameters pilot portal: http://www.emodnet-physics.eu/
- Chemistry pilot portal: http://www.emodnet-chemistry.eu/portal/portal/
- Biology pilot portal: http://bio.emodnet.eu/
- Metadata structure Hydrography/Physical/Chemistry: Common Data Index[20] (CDI) format
- Metadata structure Geology: OneGeology-Europe's Geological Metadata Profile[21] (GMP), which is based on ISO 19115 (for data sets, series) [5] and ISO 19119 (for services) [6] and compliant with the INSPIRE Metadata Regulation (Nr. 1205/2008). Existing metadata records were stored in EU-SEASED metadata structure[22], which was further developed in the EC-FP7 GeoSeas[23] project that adopted and adapted the CDI format from SeaDataNet.
- Metadata structure Biology: IMIS[24] metadata structure which takes into account ISO 19115 and EDMED metadata structure[25]. The biology portal can generate ISO 19115, GCMD and EDMED compliant format.

[15] https://webgate.ec.europa.eu/maritimeforum/category/162
[16] http://www.eurobis.org/
[17] http://www.vliz.be/en/node/416
[18] http://www.marbef.org/data/erms.php
[19] http://www.marinespecies.org/
[20] http://www.seadatanet.org/Standards-Software/Metadata-formats/CDI
[21] http://www.onegeology.org/wmscookbook/2_7.html
[22] http://www.eu-seased.net/frameset_nonflash.asp
[23] http://www.geo-seas.eu/content/content.asp?menu=0020000_000000
[24] https://webgate.ec.europa.eu/maritimeforum/system/files/FINAL%20REPORT_EMODNET_BIOLOGY.pdf
[25] http://www.eurobis.org/standards

2.6 Global Change Master Directory (GCMD)

GCMD is a metadata catalogue operated by NASA, providing access to information about thousands of environmental data sets, as well as numerous software tools and packages, plus descriptions of ancillary resources such as data centers, projects or campaigns, platforms and instruments. Each resource (data set, software, ancillary description) is described using DIF [10] and marked up using the GCMD keyword vocabulary[26] describing: Science and Services Keywords [35], Data Centers, Projects, Instruments, Platforms, Locations, Horizontal Data Resolution, Vertical Data Resolution, Temporal Data Resolution and URL Content Types. The GCMD metadata catalogue also offers an online editor to enable registered users to input their own metadata and store it in the catalogue.

- Home page: http://gcmd.nasa.gov/

2.7 Global Ocean Observing System (GOOS)

GOOS is "a permanent global system for observations, modelling and analysis of marine and ocean variables to support operational ocean services worldwide"[27]. Observations are collected from a number of platforms, among others, Argo floats, drifting buoys, moorings, and ships (research, commercial, other), which together measure a wide range of physical, chemical and biological parameters. Observations are collected and processed in real and quasi-real time, quality controlled and made available for analysis of the current condition of the ocean and as input for numerical models predicting future states.

GOOS is realised by member states' government agencies, navies and oceanographic research institutions, and is supported by large international organisations like the Intergovernmental Oceanographic Commission (IOC), United Nations Environment Programme (UNEP) , World Meteorological Organization (WMO) and International Council for Science (ICSU).

GOOS is a distributed system, GOOS is a distributed system, and links to the numerous data centres and portals providing data access can be found at the web pages of the Global Observing Systems Information Center (GOSIC).

- Home page: http://www.ioc-goos.org/
- GOSIC links for data access: http://gosic.org/goos

2.8 Infrastructure for Spatial Information in the European Community (INSPIRE)

INSPIRE is an initiative to establish an European Spatial Data Infrastructure to facilitate and ease sharing of environmental spatial information among public sector organisations, increase public access to spatial information and assist in policy-making across boundaries in Europe.

The INSPIRE Directive[28] defines a framework for a Spatial Data Infrastructure (SDI), where an "'infrastructure for spatial information' means metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use; and coordination and monitoring mechanisms, processes and procedures, established, operated or made available in accordance with this Directive". INSPIRE defines how the SDI should be implemented through a series of Implementation Rules for: (1) metadata, (2) data specifications, and (3) service types. The INSPIRE metadata standard is a profile of ISO 19115, and it core elements are described in section 3.1.2. The data

[26] http://gcmd.nasa.gov/learn/keywords.html

[27] http://www.ioc-goos.org/index.php?option=com_content&view=article&id=12&Itemid=26&lang=en

[28] http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002

specifications define a set of spatial themes for different types of data sets, among others,

- Addresses, Administrative Units, Cadastral Parcels, Coordinate Reference Systems, Geographical Grid Systems, Geographical Names, Hydrography, Protected Sites, Transport Networks (Annex I[29])
- Agricultural and Aquaculture Facilities, Area Management/Restriction/Regulation Zones and Reporting Units, Atmospheric Conditions and Meteorological Geographical Features, Bio-geographical Regions, Buildings, Elevation, Energy Resources, Environmental Monitoring Facilities, Geology, Habitats and Biotopes, Human Health and Safety, Land Cover, Land Use, Mineral Resources, Natural Risk Zones, Oceanographic geographical features, Orthoimagery, Population Distribution, Production and Industrial Facilities, Sea Regions, Soil, Species Distribution, Statistical Units, Utility and Government Services (Annexes II and III[30])

The INSPIRE Directive also defined a set of network services, for access and manipulation of spatial data sets: Download Services, View Services, Discovery Services, Schema Transformation Network Service, Co-ordinate Transformation Services. To provide access to data sets from the member states, the European Commission has established the INSPIRE GEOPORTAL as a common access point.

- Home page: http://inspire.ec.europa.eu/
- GEOPORTAL: http://inspire-geoportal.ec.europa.eu/

2.9 International Council for the Exploration of the Sea (ICES)

ICES is a "global organization for enhanced ocean sustainability" [31], and is comprised of "a network of more than 4000 scientists from almost 300 institutes, with 1600 scientists participating in activities annually". The main objective of ICES is "to increase the scientific knowledge of the marine environment and its living resources and to use this knowledge to provide advice to competent authorities" [32]. A key activity is ensuring persistent storage of marine data sets and offer seamless access to these to scientists and decision-makers in public authorities and organisations.

ICES offers a joint data portal with access to most data sets. There are also thematic portals for: (1) trawl surveys (DATRAS), (2) marine chemical and biological data (DOME), (3) eggs and larvae (4) fish stomach, (5) historical plankton, (6) oceanography. These portals are all open. In addition, there are two portals for that data requires registration and login: (a) fish catchments (InterCatch), and (b) fisheries assessment (FishFrame). The Data and Information (DIG) working group has defined recommended metadata elements for the different types data sets stored by ICES, in the "ICES Data Type Guidelines"[33].

- Home page: http://www.ices.dk/
- General data portal: http://ecosystemdata.ices.dk/
- DATRAS http://www.ices.dk/marine-data/data-portals/Pages/DATRAS.aspx
- DOME: http://www.ices.dk/marine-data/data-portals/Pages/DOME.aspx
- Eggs and larvae: http://www.ices.dk/marine-data/data-portals/Pages/Eggs-and-larvae.aspx
- Fish stomach: http://www.ices.dk/marine-data/data-portals/Pages/Fish-stomach.aspx

[29] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_AD_v3.1.pdf
[30] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/annex2and3_guidance_documents.zip
[31] http://www.ices.dk/explore-us/who-we-are/Pages/Who-we-are.aspx
[32] http://www.ices.dk/explore-us/what-we-do/Pages/default.aspx
[33] http://www.ices.dk/marine-data/guidelines-and-policy/Pages/ICES-data-type-guidelines.aspx

- Historical plankton: http://www.ices.dk/marine-data/dataset-collections/Pages/Plankton.aspx
- Oceanography: http://www.ices.dk/marine-data/data-portals/Pages/ocean.aspx ocean.ices.dk

2.10 International Oceanographic Data and Information Exchange (IODE)

IODE is a programme of the IOC (Intergovernmental Oceanographic Commission) of UNESCO that was initiated in 1961 with the aim to "enhance marine research, exploitation and development, by facilitating the exchange of oceanographic data and information between participating Member States, and by meeting the needs of users for data and information products" [34]. IODE runs the Ocean Data Portal (ODP) that provides access to marine data collections and repositories from the member states' national oceanographic data centres. The ODP offers discovery, presentation of metadata and data, as well as data download facilities. The metadata structure is based on ISO 19115 and the WMO Core Profile v.0.1[35]. Data sets can be downloaded in NetCDF format (E2E convention[36]) or viewed in the portal as text (ASCII).

- Home page: http://www.iode.org/
- ODP: http://odp.oceandataportal.net/odp

2.11 Ocean Biogeographic System (OBIS)

OBIS is a global alliance of individuals and organisations that aims to make marine bio-geographic data from all over the world openly and freely available over the World Wide Web. OBIS has evolved from a concept and initial prototype in the late 1990s to a full-fledged web GIS system with a flexible search and retrieval, presentation and downloading of world-wide bio-geographic data. The OBIS metadata catalogue is based on a metadata profile of the Ecological Markup Language[37] (EML), and the data themselves are available in Darwin Core Archive format[38], which is a zipped archive containing one or more data files and the associated metadata. As of 8 October 2013[39] (ref), the OBIS metadata catalogue gives access to 37 million data records, describing 147,552 different species, 164,766 valid marine taxa, and published as 1,374 data sets.

- Home page: http://www.iobis.org/
- Metadata catalogue: http://iobis.org/mapper/
- Metadata structure: Two options[40]: GBIF metadata profile[41] or Dublin Core[42]

2.12 Ocean Data Standards Pilot Project (ODS)

ODS is a cooperation between IODE/JCOMM ETDMP (Joint Commission for Oceanography and Marine Meteorology / Expert Team on Data Management Practises). ODS was established with the objective to define, document and promote a *process* for marine data standard development, and to closely collaborate

[34] http://www.iode.org/
[35] http://www.oceandataportal.org/index.php?option=com_content&task=view&id=41&Itemid=67&catid=9
[36] http://www.oceanobs09.net/blog/?p=419
[37] http://knb.ecoinformatics.org/software/eml/
[38] http://code.google.com/p/gbif-ecat/wiki/DwCArchive
[39] http://www.iobis.org/about/statistics
[40] http://code.google.com/p/gbif-ecat/wiki/DwCArchive#Dataset_metadata
[41] http://rs.gbif.org/schema/eml/eml-gbif-profile.xsd
[42] http://rs.tdwg.org/dwc/

with other organisations such as GEOSS and Ocean.US, and World Data Centers. ODS will also apply and refine this process to define standards proposed by the partners. The process has been defined in a Best Practice document[43]. As part of the standardisation process, the minimum required topics to cover and a template has also been created to ensure standards are developed and documented in a consistent manner[44]. Three new standards have been recommended by Ocean Data Standards so far (the documents are avialble from the link below):

1. Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data
2. Recommendation to Adopt ISO 8601 as the standard for the representation of dates and times in Oceanographic Data Exchange
3. Recommendation to Adopt ISO 3166-1 and 3166-3 Country Codes as the Standard for Identifying Countries in Oceanographic Data Exchange.

- Home page: http://oceandatastandards.org/
- Standards: http://oceandatastandards.org/index.php?option=com_content&view=article&id=36&Itemid=44

2.13 Ocean observation Research Coordination Network (RCN:OceanObsNetwork)

The aim of RCN:OceanObsNetwork is to foster "a broad, multi-disciplinary dialogue for more effective use of ocean observing systems and data/modeling infrastructure to inform societal decisions" [45]. A recent report from a working group on "Facilitating Open Exchange of Data and Information" [34] discussed many important aspects of Open Data and Data Sharing. Among others, the report identified three core issues for Open Data:

"1. Ability to be discovered, accessed and used across domains with different cultural backgrounds;
2. Transparency and information supporting use such as quality and fitness for purpose; and
3. Sustainability for future access."

These issues are overlapping and also relate to the need for providing interoperability between different domains and communities to enable Open Data. Standards for data formats and inter-system communication (i.e. protocols) are crucial in achieving machine-to-machine and data-level interoperability. The report foresees a combination of both standardization organizations standards and de facto standards is needed, and one example suggests a combination of standards for networking (TCP/IP), data formats (HDF5) and metadata (ISO 19115) can provide a moderate level of technical interoperability. This could be a first step on the way to achieving higher levels of interoperability, but it is also recognized that a number of other issues has to be addressed. Some of these include development of middleware for e.g. format translation and information brokering. Capacity building among users on data formats and cross-domains tools is also recommended.

- Home page: http://rcn.iode.org/

[43] The Ocean Data Standards and Best Practices Review Process (2012 revision), http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=9385

[44] http://oceandatastandards.org/index.php?option=com_content&view=article&id=38&Itemid=46

[45] http://rcn.iode.org/images/documents/Pearlman%20RCN%20Poster%20Blue%20Planet.ppt

2.14 OceanSITES Network (OceanSITES)

OceanSITES is a global network of long-term, open-ocean stations or observatories that measure a number of parameters in the full water column, from the surface and down to 5000 meters. The observations cover, among others, meteorology, physical oceanography, transport of water, biogeochemistry, the ecosystem, and geophysics. Many observing sites make their data available in real-time thus giving scientists and the general public instant access to new data.

The OceanSITES data flow is organized through three types of entities: (1) Principal Investigators (PI), (2) Data Assembly Centers (DAC), and (3) Global Data Assembly Centers (GDAC). The typical flow of data is as follows. A PI provides both data and metadata to a DAC, which in turn transforms these into the OceanSITES file format and forwards the files to the GDAC. There are currently two DACs: (1) IFREMER Coriolis[46], and (2) US NDBC[47]. OceanSITES uses the NetCDF as file format with global attributes specification according to recommendations of Unidata NetCDF Attribute Convention for Dataset Discovery, as described in the OceanSITES User's Manual[48].

- Home page: http://www.oceansites.org/

2.15 Ocean Data Interoperability Project (ODIP)

ODIP is a Coordination and Support Action under the EC FP7 Research Infrastructures programme. ODIP has a global consortium of 28 partners from Europe, the US, Australia and the IOC/IODE (Intergovernmental Oceanographic Commission/ International Oceanographic Data and Information Exchange).The ODIP project "aims to establish an EU / USA / Australia/ IOC-IODE coordination platform" and to demonstrate this platform through a series of joint prototypes that will cut across scientific domains, organisations and national boundaries. Recently, the three first prototypes were defined:

"1. ODIP 1: Establishing interoperability between SeaDataNet CDI, US NODC, and IMOS MCP Data Discovery and Access services, making use of a brokerage service, towards interacting with the IODE-ODP en GEOSS portals - lead by Europe,
2. ODIP 2: Establishing deployment and interoperability between Cruise Summary reporting systems in Europe, US and Australia, making use of GeoNetWork, towards interacting with the POGO portal - lead by USA
3. ODIP 3: Establishing a prototype for a Sensor Observation Service (SOS) for selected sensors (SWE), installed at vessels and in real-time monitoring systems - lead by Australia"

These prototypes will build on existing infrastructures, vocabularies, discovery metadata, data formats, data and exchange mechanisms and data formats, OGC and other standards (e.g. O&M, SensorML), existing implementations of standards as well as existing data management and presentation tools[49].

- Home page: http://www.odip.eu/

2.16 Pan-European Infrastructure for Ocean & Marine Data Management

[46] ftp://ftp.ifremer.fr/ifremer/oceansites/

[47] ftp://data.ndbc.noaa.gov/data/oceansites/

[48] http://www.oceansites.org/docs/oceansites_user_manual_version1.2.pdf

[49] http://www.odip.eu/media/odip/org/documents/odip_d3.1-prototypes1-v1.doc

(SeaDataNet)

The Pan-European Infrastructure for Ocean & Marine Data Management (SeaDataNet-1 and SeaDataNet-2) has developed "an efficient distributed Marine Data Management Infrastructure for the management of large and diverse sets of data deriving from in situ and remote observation of the seas and oceans" [50]. All SeaDataNet metadata services use a common metadata structure, the Common Data Index (CDI) format, which is based on ISO 19115 [5]. Furthermore, all SeaDataNet metadata formats use common vocabularies, notably the SeaDataNet Common Vocabularies and the EDMO directory (European Directory of Marine Organisations) to enable the relationships and semantic coherency between the different metadata services.

- Home page: http://www.seadatanet.org/
- Metadata catalogues:
  EDMED(data sets): http://www.bodc.ac.uk/data/information_and_inventories/edmed/search/
  EDMERP(projects): http://seadatanet.maris2.nl/v_edmerp/search.asp
  EDMO(organisations): http://seadatanet.maris2.nl/v_edmo/welcome.asp
  CSR(Cruise summary reports): http://seadata.bsh.de/csr/retrieve/sdn2_index.html
  EDIOS(Ocean-observing systems): http://seadatanet.maris2.nl/v_edios_v2/search.asp
- Metadata structure: Common Data Index[51] (CDI) format

2.17 Svalbard Integrated Observing System – Preparatory Phase (SIOS-PP)

SIOS-PP is project under EC FP7 Infrastructures programme. SIOS aims to establish "a cooperating and transparent research infrastructure which will give better estimates of the future environmental and climate changes in the Arctic"[52]. The SIOS consortium is comprised of 28 partners from Europe and Asia. SIOS is currently designing the system architecture for the infrastructure that will give access to a wide range of environmental data sets to support research and education in the Svalbard region and surrounding Arctic areas. The SIOS metadata catalogue will use a metadata profile based on ISO 19115 [5] and GCMD DIF[53][10] (Global Change Master Directory/Directory Interchange Format).

- Home page: http://www.sios-svalbard.org/

2.18 WMO Information System (WIS)

The WIS aims to be an "extension of WMO's policy on the free and unrestricted international exchange of meteorological and related data and products to the entire domain of such data and products used globally" (WMO Bulletin Vol. 55(4) - Oct 2006). The WIS is comprised of a set of collaborating data centres, which are classified as one of (1) Global Information System Centres (GISCs), (2) Data Collection or Production Centres (DCPs) or (3) National Centres (NCs). The different data centres will communicate using the WIS core network, and WIS will offer a joint interface to all collected, derived and modelled data through its metadata catalogue. The metadata structure is a profile of ISO 19115, and is further described in section 3.

[50] http://www.seadatanet.org/Overview
[51] http://www.seadatanet.org/Standards-Software/Metadata-formats/CDI
[52] http://www.sios-svalbard.org/prognett-sios/About_SIOS/1234130481137
[53] http://gcmd.gsfc.nasa.gov/add/difguide/index.html

- Home page: http://www.wmo.int/pages/prog/www/WIS/

2.19 World Register of Marine Species (WoRMS)

WoRMS aims to "provide an authoritative and comprehensive list of names of marine organisms" worldwide. In addition to the standard names defined by taxonomic experts, WoRMS also provides synonyms to help interpret different taxonomies used in literature. WoRMS is maintained by taxonomic expects using an editorial management system where new terms are quality controlled by a group of experts, with one designated taxonomic expect is in charge of each taxonomic group.

- Home page: http://www.marinespecies.org/index.php

2.20 Recommendations for NMDC

The initiatives and projects described in the preceding subsections are all relevant to NMDC development, and NMDC partners will keep abreast with their work and achievements, and seek to adopt upcoming standards and/or tools to fulfil NMDC objectives. For consistent naming of parameters, units, etc., SeaDataNet, GCMD, WoRMS and INSPIRE are particularly important. For maintaining a standards compliant metadata structure INSPIRE, SeaDataNet and EMODNET will be followed closely, as well as general recommendations stemming from IODE, ICES and ODS.

Best practices for data management, including metadata and data formatting and delivery, will draw upon different projects for different domains. For instance, representation of ocean physical, chemical and biological data sets will be aligned with e.g. SeaDataNet and EMODNET, while for species information key initiatives/projects include OBIS and EU-BON/PESI. For system architecture and infrastructure development, valuable lessons can be learnt from initiatives/projects such as DataONE, EU BON/PESI, EUDAT, EMODNet, GOOS, OBIS, RCN:OceanObsNetwork, SeaDataNet, SIOS and WIS.

3 Metadata structure

3.1 Standard metadata structures

A number of metadata standards exist for geographical data sets. Some standards are generic and aim to be applicable to a wide range of application domains, while others have grown out of one particular domain and is intended to meet the specific needs of this community.

This subsection describes briefly the metadata standards that have been identified in the use cases and system requirements documents [3] [4] and that were used in the NMDC pre-project, i.e. GCMD DIF [10] (see also section 2.6), ISO 19115 [5] (see also section 2.8), and EDMED [36] (see also section 3.1.3). There are also other relevant metadata standards for specific application domains addressed by the NMDC system; see section 3.1.4 for a description of these standards.

The GCMD programme also defines a standard set of keywords that can be used to mark up data sets [20], e.g. to denote which parameter(s) is being contained in the data set. Other standard keyword lists include the vast keyword lists and the relationships between them maintained by the NERC Vocabulary Server[54] (NVS) at the British Oceanographic Data Centre (BODC). The NVS holds among others the keyword lists of SeaDataNet-2 (see also section 2.16). The NVS also offers access to the INSPIRE defined spatial themes[55], which can be used to mark up NMDC data sets in general terms as e.g. "Oceanographic geographical features"[56].

By adhering to established standards for metadata and keyword lists, NMDC aims to deliver geographic metadata and data in formats and through protocols that are compatible with both national and international data sharing initiatives. Norge Digitalt/GeoNorge[57] (Norway Digitalt) works to increase the sharing of geographic data and associated services, not only on national level, but also with authorities in other European countries and with EEA institutions, in line with the INSPIRE Directive.

3.1.1 Global Change Master Directory / Directory Interchange Format (GCMD/DIF)

The Global Change Master Directory (GCMD) has defined the Directory Interchange Format (DIF) metadata standard to describe Earth science data sets [10]. DIF is composed of several *required* fields:

- Entry ID – a unique document identifier of the metadata record
- Entry Title – the title of the data set described by the metadata
- Parameters (Science Keywords) – a list of Earth science keywords describing the data set
- ISO Topic Category - keywords from the ISO 19115 - Geographic Information Metadata Topic Category Code List [5]
- Data Center - the data centre, organisation or institution that distribute the data set
- Summary – describes the data set and its purpose

[54] http://vocab.nerc.ac.uk/

[55] http://vocab.nerc.ac.uk/scheme/INSPIRE_THEMES/current/

[56] http://vocab.nerc.ac.uk/collection/P22/current/28/

[57] http://www.statkart.no/Geonorge/Norge-digitalt/

- Metadata Name – identifies current DIF standard name
- Metadata Version- identifies the current DIF metadata standard version

DIF also defines several *highly recommended* fields:

- Data Set Citation – provides a citation for the data set producer
- Personnel – defines a point of contact for getting more information about the data set or the metadata
- Related URL – allows to give a link to web site with information related to the data sets
- Instrument (Sensor Name) – names the instrument that were used to acquire the data
- Platform (Source Name) – names platform used to acquire the data
- Temporal Coverage – delimits the period the data set was collected within (start-stop dates)
- Paleo-Temporal Coverage –specifies the length of time represented by the data collected (only relevant for paleoclimate or geologic data)
- Spatial Coverage – defines the geographic and vertical (altitude, depth) coverage of the data
- Location – names a place on Earth which the data are collected
- Data Resolution – specifies the resolution of the data
- Project – names the scientific program, field campaign or project that collected the data set
- Quality - information about data quality or quality assurance procedures applied when preparing the data set
- Access Constraints – indicates any constraints for accessing the data set
- Use Constraints – described how the data set may be used
- Distribution - describes available media options, size, data format, and any fees charged for distributing the data set
- Data Set Language - the language used in preparing, storing and describing the data set
- Data Set Progress - describes the production status of the data set w.r.t. completeness

Finally, DIF defines a set of *recommended* fields:

- DIF Revision History – where the metadata author can insert a list of changes made to the DIF record
- (Ancillary) Keyword – additional words or phrases to further describe the data set
- Originating Center – the data centre/producer who originally generated the data set
- Multimedia Sample – can contain e.g. a URL providing a quick-look image of the data set
- References/Publications – lists important bibliographic citations for the data set
- Parent DIF – links to aggregated DIF metadata records (parents) related to the current record
- IDN Node – a field used internally in the GCMD metadata catalogue
- DIF Creation Date – date of metadata record creation
- Last DIF Revision Date – dates of last modification of metadata record

- Future DIF Revision Date – allows to indicate a future date for metadata revision
- Privacy Status – enables the metadata author to restrict free publication of the metadata record

Extensive guidelines for preparing DIF metadata are found in [10]. The list of GCMD Science Keywords contains more than 2500 entries, of which a small subset is included below, to illustrate the different kinds of ocean wave parameters that are available in this keyword:

```
EARTH SCIENCE → OCEANS → OCEAN WAVES
EARTH SCIENCE → OCEANS → OCEAN WAVES → GRAVITY WAVES
EARTH SCIENCE → OCEANS → OCEAN WAVES → ROSSBY/PLANETARY WAVES
EARTH SCIENCE → OCEANS → OCEAN WAVES → SEA STATE
EARTH SCIENCE → OCEANS → OCEAN WAVES → SEICHES
EARTH SCIENCE → OCEANS → OCEAN WAVES → SIGNIFICANT WAVE HEIGHT
EARTH SCIENCE → OCEANS → OCEAN WAVES → STORM SURGE
EARTH SCIENCE → OCEANS → OCEAN WAVES → SURF BEAT
EARTH SCIENCE → OCEANS → OCEAN WAVES → SWELLS
EARTH SCIENCE → OCEANS → OCEAN WAVES → TOPOGRAPHIC WAVES
EARTH SCIENCE → OCEANS → OCEAN WAVES → TSUNAMIS
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE FETCH
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE FREQUENCY
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE HEIGHT
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE LENGTH
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE PERIOD
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE SPECTRA
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE SPEED/DIRECTION
EARTH SCIENCE → OCEANS → OCEAN WAVES → WAVE TYPES
EARTH SCIENCE → OCEANS → OCEAN WAVES → WIND WAVES
```

GCMD has an equally rich classification of parameters within other fields of earth science, among others, for agriculture, atmosphere, biological classification, biosphere, climate indicators. This allows for data providers to mark up their parameters with detailed information on what phenomena their values represent.

A data set may consist of multiple data files. To enable registration of multiple files for the same data set, the NORMAP project[58] proposed to tweak the standard allowing the Related_URL element to be repeated so that each tag can identify one of the needed files [13].

3.1.2 ISO 19115:2003 (Geographic information – Metadata)

ISO 19115:2003 (Geographic information – Metadata) [5] is an extensive specification of metadata for geographic information (data sets) as well as services. An ISO 19115 metadata record provides information about, among others, data set identification, spatio-temporal extent, quality, and distribution. ISO 19115 defines metadata elements as mandatory (M), conditional (C) and optional (O). ISO 19115 defines a Core Metadata element set, which has been used as basis for many community profiles, such as the INSPIRE metadata profile (table below, extracted from [11], pp.7-8).

ISO 19115 Core metadata elements	INSPIRE metadata elements

[58] http://normap.nersc.no/

ISO 19115 Core metadata elements	INSPIRE metadata elements
Data set title (M)	Resource Title
Data set reference date (M) (+)	Temporal Reference
Data set responsible party (O)	Responsible organization (+)
Geographic location of the data set (C)	Geographic Bounding Box (M) (+)
Data set language (M) (+)	Resource Language
Data set character set (C)	- (not used)
Data set topic category (M)	Topic Category
Spatial resolution of the data set (O)	Spatial Resolution
Abstract describing the data set (M)	Resource abstract
Distribution format (O)	-
Additional extent information for the data set (vertical and temporal) (O) (+)	Temporal extent
Spatial representation type (O)	-
Reference system (O)	-
Lineage (O)	Lineage (M) (+)
On-line resource (O)	Resource Locator
Metadata file identifier (O)	-
Metadata standard name (O)	-
Metadata standard version (O)	-
Metadata language (C)	Metadata Language (M) (+)
Metadata character set (C) (+)	-
Metadata point of contact (M)	Metadata point of contact (M) (+)
Metadata date stamp (M)	Metadata Date
	Resource Type (+)
	Unique Resource Identifier (+)
-	Keyword (+)
-	Conformity (+)
-	Conditions for access and use (+)
-	Limitations on public access
(+) More demanding (typically requiring more detailed information)	

ISO 19115 allows multiple files to be registered for the same data set.

ISO 19139 is the standard encoding of ISO 19115 in XML [7]; and defines the XML schemas to use when encoding an ISO 19115 metadata record in XML.

World Meteorological Organization (WMO) has also defined a profile of ISO 19115 [10]. The WMO Core Metadata Profile is a so-called “type-1” profile, i.e. it only restricts the original standard and adds or amends the ISO 19115 Code Lists. The full definition of the WMO Core Metadata Profile is found in [12]. Figure 2 shows its mandatory and optional elements.'

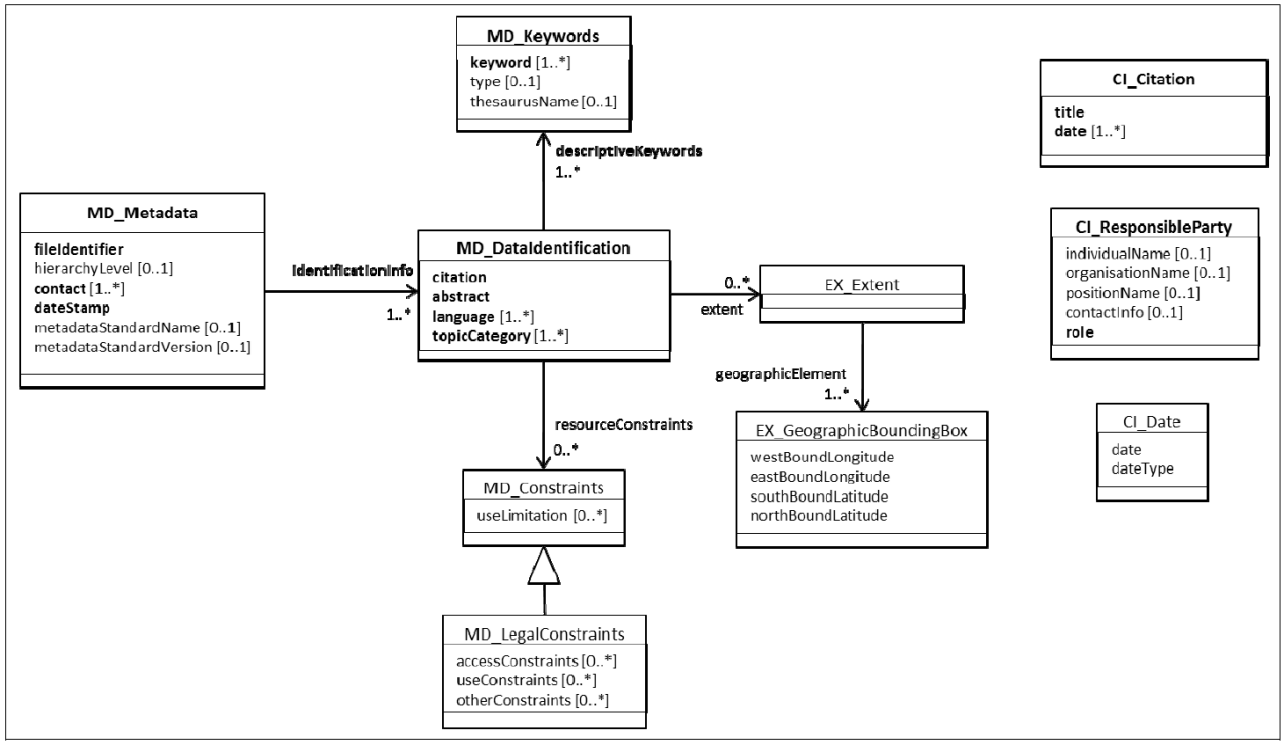

Figure 2 The WMO Core Profile[59] including both optional and mandatory (bold) elements.

ISO 19115-2:2009 [42] defines a profile of ISO 19115 for Earth Observation (EO) and gridded datasets. This profile includes metadata elements for the sensor used for data acquisition, the geometry of the data and the production process used to derived geographic datasets from the raw data. OGC has defined an application profile of Geography Markup Language (GML) for EO products [43]. This profile is an OGC Best Practice documents with metadata elements for different types of EO products, organised as GML modular schemas.

3.1.3 European Directory of Marine Environmental Data (EDMED)

EDMED was first developed within the European Marine Science and Technology (MAST) framework in 1991. EDMED has later been further developed within the SEA-SEARCH (2003-2005) and SeaDataNet (2006-2011) projects, and is at present being maintained as part of the SeaDataNet-2 project. EDMED defines a metadata structure [36] which is implemented in the EDMED metadata catalogue[60].The profile prescribes the following metadata fields for a data set:

- Dataset-Name - the name of the data set,
- Time-Period - the first and last date of the values in the data set (i.e. start and end date)
- Geographic coverage - general description of the data set's location and coverage, using text (names)

[59] http://wis.wmo.int/2012/metadata/WMO_Core_Metadata_Profile_v1.3_Specification_Part_1_v1.0FINALcorrected.pdf

[60] http://www.bodc.ac.uk/data/information_and_inventories/edmed/search/

or lat,lon bounding box

- Project - the associated project, if any,
- Parameter keywords - list of free text keywords describing the parameters measured/derived.
- Instruments - list of instruments types or measurement techniques used to collect data set,
- Summary - informative overview of the data set and suitable usage,
- Reference - refers to any information sheets or literature that describes the data further,
- Data-Website - URL to a web site with more information about the data set or where the data set itself can be downloaded,
- Originator - name of the organisation(s) or person(s) that have the main responsibility for the data set,
- Centre - title and acronym of the centre holding the data
- Storage-Medium - the number of and type of storage mediums where the data set is stored,
- Availability - describes rules applying for data access, e.g. of the data set is freely available or what licenses apply,
- Supply-Details - in what form/format/media the data set can be provided in, and any delivery costs,
- Contact - the name/position with contact detail, or contact information for a help desk,
- Completed by - name and contact information (phone/fax/e-mail) for the person filling in these metadata,
- Currency-Date - the data the metadata (for the data set) was last verified, or initially entered into the system (i.e. the EDMED catalogue), and
- Revision-Date - the date these metadata were last revised.

In addition, EDMED defines a structure for data holding centres and data contacts. For a data holding centre, the following fields are needed;

- Center-Name - the English name for the data holding centre,
- Center-Host - the national name for the centre (optional),
- Visit-address - the visiting address of the centre,
- Country - the country the centre is located in,
- Centre-Website - an URL for more information about the centre,
- Description - a brief description of the centre,
- Currency-Date - the data the metadata (for the centre) was last verified, or initially entered into the system (i.e. the EDMED catalogue), and
- Revision-Date - the date these metadata were last revised.

while for data contact, the following fields are mandatory;

- Contact-Name - name of the person that can be contacted for more information about the data set,
- Contact-Title - this person's role or title within the organisation, e.g. "Metadata manager",
- Post-address - the full postal address of the contact person,
- Phone - phone number of the contact person,
- Fax - fax number of the contact person,
- Email - the e-mail address of the contact person,

- Currency-Date - the data the metadata (for the data contact) was last verified, or initially entered into the system (i.e. the EDMED catalogue), and
- Revision-Date - the date these metadata were last revised.

EDMED allows data providers to define their own keywords to mark up e.g. parameters in their data sets. However, to simplify search across multiple data collating centres, four common types of keywords have been defined: (1) Host Country, (2) Data Theme, (3) Coastal Zone, and (4) Ocean/Sea Area. For Data Theme, the following keywords are predefined for physical oceanographic variables:

```
PH    PHYSICAL OCEANOGRAPHY
P1    Hydrography (e.g. T,S) – near surface only
P2    Hydrography (e.g. T,S)
P3    Currents – drift – dispersion - tracers
P4    Waves – swell
P5    Sea level (& bottom pressure, IES)
P6    Underwater acoustics
P7    Optical measurements
```

By combining the general pre-defined EDMED keywords with more specific keyword defined by the data collating centres, a common search across several centres can be conducted by the pre-defined keywords and more specific information about parameter content can be retrieved by the centre specific keywords.

3.1.4 Other relevant metadata standards

OGC and ISO Observations and Measurements[61] (O&M): This international standard defines a conceptual schema for observations, and for features involved in sampling when making observations. These provide models for the exchange of information describing observation acts and their results, both within and between different scientific and technical communities.

CSDGM - Metadata Profile for Shoreline Data[62]: Metadata Profile of CSDGM for Shoreline Data “addresses variability in the definition and mapping of shorelines by providing a standardised set of terms and data elements required to support metadata for shoreline and coastal data sets. The profile also includes a glossary and bibliography.”

Digital Object Identifier (DOI)[63]: A Digital Object Identifier (DOI), i.e., a "digital identifier of an object" is a globally unique persistent identifier that can be resolved within the DOI system to obtain information about the object including descriptive metadata. The DOI system enables the construction of automated services and transactions.

Taxon Concept transfer Schema[64] (TCS): TCS provides a standard for taxon names and taxon concepts in the exchange and integration of biodiversity and natural history data.” The majority of elements in TCS are optional to allow for the variety of different approaches to defining and recording taxonomic names and

[61] http://www.opengeospatial.org/standards/om

[62] http://www.csc.noaa.gov/metadata/sprofile.pdf

[63] http://www.doi.org/doi_handbook/1_Introduction.html

[64] http://www.tdwg.org/activities/tnc/

concepts, hence TCS allows more choices if an expert simultaneously authors concepts AND asserts concept relationships.

Ecological Metadata Language[65] (EML): EML is a metadata specification for the ecology discipline. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological data set.

The ebXML Registry Information Model (ebRIM) [45] defines a general metadata model for digital resources. ebRIM can also be used for geographic datasets, e.g. through the ebRIM application profile of CSW [44] using an extension package for EO data [41].

INSPIRE has defined several relevant metadata standards for specific types of geographic data. Table 1 lists some relevant INSPIRE standards, as well as relevant standards for biological data from other international initiatives.

Table 1 Relevant metadata standards and vocabularies for NMDC application domains.

Application domain	Metadata standard(s) and vocabularies used
Geographical names	INSPIRE Data Specification on Geographical Names - Guidelines v 3.0.1 03.05.2010[66]
Sea regions	Data Specification on Sea regions – Draft Technical Guidelines 04.02.2013[67]
Hydrography	INSPIRE Data Specification on Hydrography - Guidelines v 3.0.1 03.05.2010, D2.8.I.8 INSPIRE Data Specification on Hydrography – Guidelines[68]
Elevation	Data Specification on Elevation – Draft Technical Guidelines 04.02.2013[69]
Protected sites	INSPIRE Data Specification on Protected Sites - Guidelines v 3.1.0 03.05.2010[70]
Meteorological geographical features	Data Specification on Atmospheric Conditions- Meteorological geographical features – Draft Technical Guidelines 04.02.2013[71]
Species occurrence data	Darwin Core [26]
Biological data	OBIS extension of Darwin Core[72]
Biogeographical regions	INSPIRE Data Specification on Bio-geographical regions[73]

[65] http://knb.ecoinformatics.org/software/eml/
[66] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GN_v3.0.1.pdf
[67] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SR_v3.0rc3.pdf
[68] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_HY_v3.0.1.pdf
[69] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_EL_v3.0rc3.pdf
[70] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_PS_v3.1.pdf
[71] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_AC-MF_v3.0rc3.pdf
[72] http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10432

Application domain	Metadata standard(s) and vocabularies used
Habitats and biotopes	Data Specification on Habitats and biotopes – Draft Technical Guidelines 04.02.2013[74]
Species distribution	Data Specification on Species distribution – Draft Technical Guidelines 04.02.2013[75]
Marine data	INSPIRE Network Services and Data Specifications pilot[76]
Oceanographical features	Data Specification on Oceanographic geographical features – Draft Technical Guidelines 04.02.2013[77]
Geology	Data Specification on Geology – Draft Technical Guidelines 04.02.2013[78]

The NetCDF Climate Forecast (CF) conventions [46] defines a set of required elements, including descriptive and usage metadata, that should be stored in a NetCDF file to allow users to both discover datasets and decide whether they are fit for a particular purpose. With NetCDF/CF each variable can be described in detail using standard names and units, allowing computer applications to extract, process (e.g. re-grid) and display user selected values. In addition to parameter names and units, the CF also enables definition of, amongst others, time (point or interval), coordinate axis, depth axis and map projection. To describe the data set as a whole, CF includes metadata elements for data set title, name of institution producing the data originally, data source, history (of processing), references (e.g. scientific or technical literature) and comments.

The EOSDIS Core System (ECS) Core metadata [47] is a metadata standard defined for the NASA's Earth Observing System (EOS) Data and Information System. The ECS Core metadata contains, among others, identification information (title, origin, contents), location (position, date and time), interpretation (descriptions, units, scaling factors, accuracy), and documentation (software version, production history, ancillary inputs).

3.2 Pre-project metadata structure

The metadata structure used in the NMDC pre-project [16] is divided into six parts: (1) Main metadata, (2) Geographic data, (3) Creation information, (4) Parameter data, (5) Citation data, and (6) Media.

Main metadata has the following fields: (* is mandatory information)

Name	Explanations
Dataset id*	Id given by the metadata author
Dataset name*	The title of the data set described by the metadata.

[73] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_BR_v3.0rc3.pdf
[74] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_HB_v3.0rc3.pdf
[75] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SD_v3.0rc3.pdf
[76] http://www.environ.ie/en/Publications/DevelopmentandHousing/Planning/FileDownLoad,31327,en.pdf
[77] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_OF_v3.0rc3.pdf
[78] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GE_v3.0rc3.pdf

Summary*	Brief description of the data set that allows potential users to determine if the data set is useful for their needs.
Start date*	
End date	
Progress*	Describes the production status of the data set regarding its completeness.
Language*	The language used in the preparation, storage, and description of the data. It does not refer to the language of the metadata.
Project	Name of the scientific program, field campaign, or project from which the data were collected.
Purpose	
Planned revision	Date for planned revision.
Quality	Information about the quality of the data or any quality assurance procedures followed in producing the data described in the metadata.
Reference	Describes key bibliographic citations pertaining to the data set.

Geographic data has the following fields:

Name	Explanations
Spatial coverage*	Specifies the geographic and vertical (altitude, depth) coverage of the data.
Geographical area	Textual description of area
Location*	Specifies the name of a place on Earth, a location within the Earth, a vertical location, or a location outside of Earth.

Location is a searchable drop-down list.

Creation information has the following fields:

Name	Explanations
Originating institutions*	The data center or data producer who originally generated the data set.
Data centers*	The data center, organization, or institution responsible for distributing the data.
Access constraints*	Information about any constraints for accessing the data set.
Usage constraints*	Describe how the data may or may not be used after access is granted to assure the protection of privacy or intellectual property.

Originating institutions and Data centers drop-down lists are searchable.

Choose contact person to add to Origination Institution or Data centers from the drop-down list. If the contact person cannot be found in the list, then a new contact person can be added using the “Create new person” button. The newly create person will then be found at the bottom of the drop-down list. Contact persons are shared between Originating Institution and Data center.

Parameter data has the following fields:

Name	Explanations
Iso categories*	Identify the keywords in the ISO 19115 - Geographic Information Metadata.
Instrument	Name of the instrument used to acquire the data.
Platform	Name of the platform used to acquire the data.
Science keywords*	Specification of Earth science keywords that are representative of the data set being described.
Keywords	Allows authors to provide any words or phrases needed to further describe the data set.
Data resolution	Specifies the resolution of the data, which is the difference between two adjacent geographic, vertical, or temporal values.
Paleo Temporal Coverage	The length of time represented by the data collected. Should be used when the data spans time frames earlier than 01-01-0001.

Iso categories, Instrument, Platform and Science keywords are searchable drop-down lists.

Citation data has the following fields:

Name	Explanations
Citation	Properly cite the data set producer.
Multimedia sample	Provide information that will enable the display of a sample image, movie or sound clip.

Media has the following fields:

Name	Explanations
Related Urls	Specifies links to Internet sites that contain information related to the data, as well as related Internet sites such as project home pages, related data archives/servers, metadata extensions, online software packages, web mapping services, and calibration/validation data.
Distribution	Describes media options, size, data format, and fees involved in distributing the data set.
Personnel	Point of contact for more information about the data set or the metadata.
Parent datasets	Capability to relate generalized aggregated metadata records (parents) to metadata records with highly specific information (children).

3.3 Proposed NMDC metadata structure

During the pre-project, National research infrastructure Norwegian Marine Data centre, a metadata catalogue was developed to give an overview of marine datasets provided by the NMD partners. The developed database is compatible with SeaDataNet's EDMED directory and NASAs Global Change Master Directory (GCMD). This has been chosen to make it possible to make the metadata available through SeaDataNet and NASA's GCMD as done during International Polar Year (IPY).

Many of the NMDC partners, both in the pre-project and in the current project, are obliged to deliver their metadata records to multiple metadata catalogues, such as EDMED and GCMD, as part of their regular data management activities or as partner in other projects. In addition, the data sets themselves were required to be made available through standard web protocols, with access information (URLs of download and/or sub-setting web services). Thus, the pre-project metadata structure was defined to enable easy delivery to multiple metadata catalogues, in addition to being based on established metadata standards and use of GCMD keywords.

The same two main criteria, adherence to established metadata standards and keywords lists, and the ability to deliver metadata records to additional metadata catalogues without a substantial amount of work for restructuring and/or reformatting metadata or data files were used for the current NMDC project. Based on the user requirements identified during the first part of the NMDC project [3], it has been decided to keep the pre-project metadata structure for the first version of NMDC, with the addition of allowing terms from the SeaDataNet-2 controlled vocabulary as keywords in the metadata to mark up parameter names. Appendix A contains an XML representation of this metadata structure, with an example metadata file for a dataset that has been ingested in the NMDC metadata catalogue.

3.4 Recommendations for NMDC

The work in the first year of the NMDC project has determined that the pre-project metadata structure can be kept as long as the system supports additional vocabularies for scientific keywords, notably the SeaDataNet-2 keyword list. Thus, the initial version of the NDMC data catalogue will be based on the metadata structure described in the previous subsection, where the keywords can be chosen from the GCMD science keyword and the SeaDataNet-2 keywords lists. Appendix A shows an example metadata record where GCMD keywords have been used to mark up the parameter names.

By building on previous development, tools for metadata generation, storage and discovery can be reused. Also, through the reuse of a metadata structure that is compliant with SeaDataNet and NASA specifications, new metadata records that are generated in the NMDC project can be submitted to e.g. EDMED and GCMD without additional effort. Furthermore, the metadata structure is based on standards, and can be easily extended to meet changes in user requirements.

For instance, it is recognised that the lists of keywords will need to evolve as new types of data sets are added to the NMDC system. For instance, to register marine biology data sets, using keywords from the WoRMS taxon tree[79], will allow users to search for data sets using scientific names on species and species groups. A more comprehensive nomenclature register is PESI[80] (A Pan-European Species directories Infrastructure). This register for species taxonomy is using WORMS among other sources. PESI will be used when implementing the INSPIRE directive on Species Distribution, and is the taxonomic backbone in the EU BON project. Thus, we recommend using terms from PESI to support search for biodiversity data sets within the NMDC system, and to work with both EU BON and PESI through partner NBIC (Artsdatabanken) to find a way utilize the Norwegian taxon IDs in the European projects and infrastructures for species data.

Extensions to metadata structures in iterations

[79] http://www.marinespecies.org/aphia.php?p=browser&id=2&expand=false#ct

[80] http://www.eu-nomen.eu/pesi/

4 Data formats

4.1 Standard data formats

Table 2 shows a brief summary of standard data formats that are proposed as standard formats for the NMDC system. Data that are stored in other formats or held in databases such as RDBMSs, must be exported in one (or more) of these formats to be integrated in the NMDC system.

Table 2 Proposed standard data formats for geographic data in the NMDC system.

Format	Type of data	Metadata	Reference
NetCDF/CF	Grid (multi-dimensional), time series[81], trajectory[82], profile[83], swath data[84]	CF convention [46]	[14]
GeoTIFF	Grid	TIFF tags for geographic coverage and map projection	[21]
HDF-EOS	Grid, point, swath	ECS Core metadata[85]	[22]
Shapefile	Vector (point, line, polygon)	Metadata can be given in (an optional) .shp.xml.	[23]
CSV (Comma Separated Values)	Vector (point)	Provider specific, may be one of more headers above the data	[24]
GML (Geography Markup Language)	Vector (point, line, polygon)	Can be included in the data file	[25]
SOSI	Text (point, line, polygon)	Limited info in header	[31]
SEG Y	Seismic profile	Can be included in the data file (header) or sent in separate file, e.g. INSPIRE compliant XML	[32]
JSON	Key/value object (string, number, array, object, Boolean, null)	Provider specific, can be embedded as metadata:{object}	[39]
GeoJSON	Geometry (Point, LineString, Polygon, MultiPoint, MultiLineString, MultiPolygon, and GeometryCollection)	Properties object	[40]

4.2 Formats used by NMDC partners

[81] E.g. http://www.epic.noaa.gov/epic/document/conv_time.htm

[82] E.g. https://geo-ide.noaa.gov/wiki/index.php?title=NODC_NetCDF_Trajectory_Template

[83] E.g. http://www.nodc.noaa.gov/data/formats/netcdf/profileOrthogonal.cdl

[84] E.g. http://wiki.esipfed.org/index.php/NetCDF-CF_File_Examples_for_Satellite_Swath_Data

[85] http://edhs1.gsfc.nasa.gov/waisdata/emd/pdf/420emd001.pdf

Table 3 lists the formats offered by the partners for one of more of their NMDC products, i.e. the formats that partners offer to external users.

Table 3 Data formats provided by partners for the NMDC system.

Partner	Formats provided	Data set / product
IMR	NetCDF NetCDF ODV / NetCDF NetCDF	Fisheries acoustic (historical) Physical oceanography (historical) Chemical data (for EMODNET) Plankton data
FFI	NetCDF/CF	Oceanographic measurements (CTD) Bathymetry data
UiN	CSV	Oceanographic measurements (CTD), some fish data plankton data
METNO	ESRI Shapefile[86]	Ice charts (historical)
	NetCDF/CF	Gridded oceanographic fields
	NetCDF/CF	Time series, profiles and trajectories of scientific data that is not handled in the operational station data system.
	TBD[87]	Operational surface station data.[88]
NERSC	NetCDF/CF, CSV	Oceanographic measurements ((A)XBT)
	NetCDF/CF, CSV	Gridded oceanographic fields
NGU	SEG Y	Seismic data
	ESRI Shapefile, ESRI file geodatabase, ESRI personal geodatabase, CSV, GML, SOSI	Navigation data (sampling points, seismic lines), field observations (seabed sediments), analysis results, geologic interpretations; (point, line, polygon)
	GeoTIFF	Backscatter data; Marine biotopes
	WMS	All geologic interpretations; geochemical data of seabed sediments
NINA	CSV CSV CSV, GML	Fish data Otter database Seabirds
NPI	JSON, GeoJSON, CSV,	Physical oceanography (CTD…), plankton data, sea ice data,

[86] Revision of how these data are handled has started within the scope of NMDC. ESRI Shapefiles will be supported in the future, but other formats may also be supported depending on the decisions made after the ongoing feasibility studies.

[87] This system is under revision, currently ASCII dump with selected field separators are supported. In the future formats may change and expand as well as interfaces are hopefully harmonised.

[88] This is basically surface synoptic weather stations collected in a database with emphasis on climate studies, surface weather data from ships are not stored in this system, but oil rigs are as well as coastal and Arctic stations.

Partner	Formats provided	Data set / product
	NetCDF/CF	other
UiB	Multiple formats supported (export from RDBMS)	Oceanographic measurements (CTD, current meter,...) (historical) Chemical oceanographical data
UiO/CEES	CSV	Plankton
UiT	GeoTIFF, CSV	Rocky bottom fauna picture study Historical hydrography Northern Norway Havmiljødata Nord.Norge TUNU/East Greenland
NIVA	NetCDF/CF Shapefile CSV GML MS EXCEL	Water Chemistry[89] Hydrography Soft bottom fauna Phytoplankton Hard bottom fauna
Uni Research	Multiple formats supported (export from RDBMS)	Antarctic marine time series Current meter data Real time (station M and Site 5)
NBIC (Artsdatabanken)	CSV, WMS, GML, Through an API: XML and JSON, GeoJSON	Species occurrence data (Darwin Core) comprising specimen data from collections and observation data. Over 1.7 mill. records of species living in or by marine habitats from 10 Norwegian source institutions
	CSV, XML,JSON	Habitats in Norway (Naturtyper I Norge) API for online services for search, validate and download habitat classification nomenclature: http://www.naturtyper.artsdatabanken.no/

4.3 Recommendations for NMDC

The data formats in Table 2 are the formats that the NMDC will make available to external users. Internally, the NMDC system will support other formats such as those listed in Table 3. If a data set if provided by in another format, the NMDC will reformat the data set to one of the formats listed in Table 2.

The list of supported data formats can grow as new types of products are ingested into the system.

[89] In principle, each product can be delivered in all formats listed in the previous column, but depending on resources available, delivery may be restricted to fewer formats.

5 Data exchange protocols

The following standard protocols for metadata and data exchange has been identified in the use cases and system requirements documents [3] [4] . These protocols will be further investigated during system design, and are therefore only briefly discussed below.

5.1 Metadata search and retrieval protocols

OGC CSW [8] is a protocol for discovery, browsing, and querying metadata about geographical data (sets), services and other potential resources. CSW (Catalogue Service for the Web) is developed by the Open Geospatial Consortium (OGC)[90], and uses a profile of Dublin Core for discovery metadata.

OAI-PMH [27] is a protocol for harvesting metadata descriptions from a metadata catalogue so that a common entry information system or search service can be built on top of many metadata archives. OAI-PMH Is developed by the Open Archives Initiative[91]. With OAI-PMH data providers expose their metadata and service providers consume these metadata through a client application known as a harvester that issues OAI-PMH service requests over HTTP.

5.2 Data access protocols

OPeNDAP [15] is a protocol for accessing a data set stored at a remote server, and retrieve a description of its structure, a description of the parameters contained (i.e. metadata), or a subset or the whole data set itself as a file.

OGC WMS [28] is a protocol for generating a graphical presentation (raster map) of a geographic data set. The raster map is typically displayed in a browser. It is possible to obtain the map for only a subset (e.g. a limited area) and in a different map projection than the original data set.

OGC WFS [29] is a protocol for obtaining a set of geographical features (points, lines, etc.) from a data set stored on a remote server. The downloaded file is typically in GML format.

OCG WCS [30] is a protocol for obtaining a subsetted or full gridded data set from a remote server. The downloaded file can be in different formats depending on the tool implementing WCS; e.g. GeoTIFF.

The Hypertext Transfer Protocol (HTTP) [37] is an application-level protocol for distributed, collaborative, hypermedia information systems. It is a generic, stateless, protocol which can be used for many tasks beyond its use for hypertext. A feature of HTTP is the typing and negotiation of data representation, allowing systems to be built independently of the data being transferred.

5.3 Recommendations for NMDC

The metadata and data exchange protocols identified in this section are all standards that are widely used in the marine and environmental science community. Together they provide the required functionality as identified in use cases [3], and it is therefore recommended to support them in the NMDC system.

[90] http://www.opengeospatial.org/

[91] http://www.openarchives.org/

6 Summary and conclusions

This report has reviewed some twenty ongoing initiatives and projects addressing marine data infrastructures, data management and service development. Standard metadata and data formats used in these initiatives and projects have also been reviewed, as have the standards used in the NMDC pre-project (which also used standards from several of these initiatives/projects). The metadata structure used in the pre-project was found to be applicable also to the initial user requirements of NMDC, and is therefore kept until new requirements emerge that necessitate future extensions.

The initiatives and projects described in the preceding subsections are all relevant to NMDC development, and NMDC partners will keep abreast with their work and achievements, and seek to adopt upcoming standards and/or tools that can be used to fulfil NMDC objectives. For consistent naming of parameters, units, etc., SeaDataNet, GCMD, WoRMS and INSPIRE are particularly important. For maintaining a standards compliant metadata structure INSPIRE, SeaDataNet and EMODNET will be followed closely, as well as general recommendations stemming from IODE, ICES and ODS.

Best practices for data management, including metadata and data formatting and delivery, will draw upon different projects for different domains. For instance, representation of ocean physical, chemical and biological data sets will be aligned with e.g. SeaDataNet and EMODNET, while for species information key initiatives/projects include OBIS and EU-BON/PESI. For system architecture and infrastructure development, valuable lessons can be learnt from initiatives/projects such as DataONE, EU BON/PESI, EUDAT, EMODNet, GOOS, OBIS, RCN:OceanObsNetwork, SeaDataNet, SIOS and WIS.

The report also describes the formats used by NMDC partners for products to be delivered through the NMDC system. Based on the data sets identified as candidates for the NMDC system, the following data formats will be offered to external users:

- NetCDF/CF
- GeoTIFF
- HDF-EOS
- Shapefile
- CSV
- GML
- SOSI
- SEG Y
- JSON
- GeoJSON

The NMDC metadata structure will be based on ISO 19115, the EDMED metadata structure and GCMD DIF, as in the pre-project. Keywords can be chosen from the GCMD Science Keywords list and the SeaDataNet-2 keyword lists.

The list of supported data formats and the metadata structure are expected to grow during the lifetime of the project, to support new types of data sets and changing user requirements. However, both metadata and data formats will continue to be based on community standards.

The report also identified some key standard data discovery and delivery protocols that the NMDC should support to distribute data sets to users. For data exchange, we propose the following standards:

- Metadata search and discovery: OGC CSW and OAI-PHM
- Map (raster) generation: OGC WMS
- Data download: OGC WFS, OGC WCS, OPeNDAP and HTTP

The design and implementation of the NMDC will be done in iterations. Thus, for details on when the above data formats and data exchange products will be available, we refer to the latest information and services announced on the NMDC home page [38].

A. NMDC metadata structure example

The following XML file is an example of valid metadata for NMDC. Further examples of metadata can be obtained from the NMDC project web site [38].

```
<OAI-PMH xmlns='http://www.openarchives.org/OAI/2.0/'
xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
xsi:schemaLocation='http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd'>
<responseDate>2014-05-15T17:25:18Z</responseDate>
<request>http://arcticdata.met.no/metamod/oai</request>
<ListRecords>
<record>
<header>
<identifier>urn:x-wmo:md:no.met.arcticdata.test3::ADC_radflux-janmayen</identifier>
<datestamp>2010-10-18T12:51:10Z</datestamp>
<setSpec>ADC</setSpec></header>
<metadata><DIF xmlns='http://gcmd.gsfc.nasa.gov/Aboutus/xml/dif/'
xmlns:jif='http://www.met.no/schema/metamod/jif'
xmlns:topic='http://www.met.no/schema/metamod/mm2dif'>
<Entry_ID>ADC_radflux-janmayen</Entry_ID>
<Entry_Title>Downwelling surface radiative fluxes at Jan Mayen</Entry_Title>
<Data_Set_Citation>
<Dataset_Creator>Øystein Godøy</Dataset_Creator>
<Dataset_Title>Downwelling surface radiative fluxes at Jan Mayen</Dataset_Title>
<Dataset_Release_Date>Not Available</Dataset_Release_Date>
<Dataset_Release_Place>Not Available</Dataset_Release_Place>
<Dataset_Publisher>Norwegian Meteorological Institute</Dataset_Publisher>
<Version>Not Available</Version>
</Data_Set_Citation>
<Personnel>
<Role>Technical Contact</Role>
<First_Name>Egil</First_Name>
<Last_Name>Støren</Last_Name>
<Email>Not Available</Email>
<Phone>+4722963000</Phone>
<Contact_Address>
<Address>Norwegian Meteorological Institute P.O. Box 43 Blindern</Address>
<City>Oslo</City>
<Postal_Code>N-0313</Postal_Code>
<Country>Norway</Country>
</Contact_Address>
</Personnel>
<!--start of variables-->
<Parameters>
<Category>EARTH SCIENCE</Category>
<Topic>ATMOSPHERE</Topic>
<Term>ATMOSPHERIC RADIATION</Term>
<Variable_Level_1>SHORTWAVE RADIATION</Variable_Level_1>
```

<Detailed_Variable>Atmosphere > Atmospheric Radiation > Shortwave Radiation</Detailed_Variable>
</Parameters>
<Parameters>
<Category>EARTH SCIENCE</Category>
<Topic>ATMOSPHERE</Topic>
<Term>ATMOSPHERIC RADIATION</Term>
<Variable_Level_1>LONGWAVE RADIATION</Variable_Level_1>
<Detailed_Variable>surface_net_downward_longwave_flux</Detailed_Variable>
</Parameters>
<Parameters>
<Category>EARTH SCIENCE</Category>
<Topic>ATMOSPHERE</Topic>
<Term>ATMOSPHERIC RADIATION</Term>
<Variable_Level_1>LONGWAVE RADIATION</Variable_Level_1>
<Detailed_Variable>Atmosphere > Atmospheric Radiation > Longwave Radiation</Detailed_Variable>
</Parameters>
<Parameters>
<Category>EARTH SCIENCE</Category>
<Topic>ATMOSPHERE</Topic>
<Term>ATMOSPHERIC RADIATION</Term>
<Variable_Level_1>RADIATIVE FLUX</Variable_Level_1>
<Detailed_Variable>surface_downwelling_shortwave_flux</Detailed_Variable>
</Parameters>
<!--end of variables-->
<ISO_Topic_Category>Climatology/Meteorology/Atmosphere</ISO_Topic_Category>
<Keyword>Radiative</Keyword>
<Keyword>Flux</Keyword>
<Temporal_Coverage>
<Start_Date>2010-09-01</Start_Date>
<Stop_Date>2010-09-30</Stop_Date>
</Temporal_Coverage>
<Data_Set_Progress>In Work</Data_Set_Progress>
<Spatial_Coverage>
<Southernmost_Latitude>70.933333</Southernmost_Latitude>
<Northernmost_Latitude>70.933333</Northernmost_Latitude>
<Westernmost_Longitude>-8.6666667</Westernmost_Longitude>
<Easternmost_Longitude>-8.6666667</Easternmost_Longitude>
</Spatial_Coverage>
<Location>
<Location_Category>OCEAN</Location_Category>
<Location_Type>ARCTIC OCEAN</Location_Type>
<Detailed_Location>Greenland Sea</Detailed_Location>
</Location
><Location>
<Location_Category>GEOGRAPHIC REGION</Location_Category>
<Location_Type>POLAR</Location_Type>
</Location>

<Location>
<Location_Category>GEOGRAPHIC REGION</Location_Category>
<Location_Type>ARCTIC</Location_Type>
</Location>
<Project>
<Short_Name>iAOOS-Norway</Short_Name>
<Long_Name>iAOOS-Norway</Long_Name>
</Project>
<Project>
<Short_Name>IPY</Short_Name>
<Long_Name>INTERNATIONAL POLAR YEAR</Long_Name>
</Project>
<Access_Constraints>Restricted to iAOOS-Norway</Access_Constraints>
<Use_Constraints>Not Available</Use_Constraints>
<Data_Set_Language>Not Available</Data_Set_Language>
<Data_Center>
<Data_Center_Name>
<Short_Name>NO/MET</Short_Name>
<Long_Name>Norwegian Meteorological Institute, Norway</Long_Name>
</Data_Center_Name>
<Data_Center_URL>http://met.no/</Data_Center_URL>
<Personnel>
<Role>Data Center Contact</Role>
<First_Name>Egil</First_Name>
<Last_Name>Støren</Last_Name>
<Email>Not Available</Email>
<Phone>+4722963000</Phone>
<Contact_Address>
<Address>Norwegian Meteorological Institute
P.O. Box 43
Blindern</Address>
<City>Oslo</City>
<Postal_Code>N-0313</Postal_Code>
<Country>Norway</Country>
</Contact_Address>
</Personnel>
</Data_Center>
<Reference />
<Summary>
<Abstract>Downwelling surface radiative fluxes observed at the meteorological station at Jan Mayen Island in the Greenland Sea. Measurements are made using Kipp and Zonen CMP21 and CGR4 pyranometers and pyrgeometers. Daily maintenance is performed by the meteorological personnel at the station. Data are averaged over the last minute and the time is set to UTC. This data set has been collected with support from the Norwegian Research Council. This data set has been collected with support from the Norwegian Research Council. The quality control performed is rudimentary and the dataset will be updated.
</Abstract>
</Summary>

```
<Related_URL>
<URL_Content_Type>
<Type>VIEW RELATED INFORMATION</Type>
</URL_Content_Type>
<URL>http://thredds.met.no/thredds/catalog/arcticdata/met.no/radflux-janmayen/</URL>
</Related_URL>
<IDN_Node>
<Short_Name>ARCTIC/NO</Short_Name>
</IDN_Node>
<IDN_Node>
<Short_Name>ARCTIC</Short_Name>
</IDN_Node>
<IDN_Node>
<Short_Name>IPY</Short_Name>
</IDN_Node>
<IDN_Node>
<Short_Name>DOKIPY</Short_Name>
</IDN_Node>
<Metadata_Name>CEOS IDN DIF</Metadata_Name>
<Metadata_Version>9.8.2</Metadata_Version>
<DIF_Creation_Date>2010-10-18</DIF_Creation_Date>
<Last_DIF_Revision_Date>2010-10-18</Last_DIF_Revision_Date>
<Private>False</Private>
</DIF>
</metadata>
</record>
…
</ListRecords>
</OAI-PMH>
```