Project contract no. 036851

ESONET

European Seas Observatory Network

Instrument: **Network of Excellence (NoE)**

Thematic Priority: **1.1.6.3 – Climate Change and Ecosystems**

Sub Priority: **III – Global Change and Ecosystems**

Project deliverable D9

DATA MANAGEMENT PLAN

Due date of deliverable: month 6
Actual submission date: month 12

Start date of project: **March 2007** Duration: **48 months**

Organisation name of lead contractor for this deliverable: KDM
Lead authors for this deliverable: Michaël DIEPENBROEK

Revision [FINAL]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services	x
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

TABLE OF CONTENTS

ESONET Data and Information Management Plan

Introduction

The European Seas Observatory Network (ESONET) was established as a European Commission FP6 Network of Excellence in March 2006. ESONET's central mission is to create an organisation capable of implementing, operating and maintaining a network of multidisciplinary long-term ocean observatories in deep waters around Europe from the Arctic Ocean to the Black Sea.

As this Network is supposed to generate huge amounts of scientific valuable information, the management of data – from acquisition to dissemination of information for further analysis and decision making – was declared as one key element of ESONET. The development of a Data and Information Management Plan corresponds to Deliverable D-9.

The overall purpose of ESONET data management is to treat data as a valuable resource. It encompasses the development and execution of architectures, policies, practices and procedures that properly cover the full data lifecycle. A general goal is to integrate ESONET's data management into the emerging Global Spatial Data Infrastructures (GSDI, Nebert, 2004), finding convergence with and linking to the wider network of observatories on the international level, and implementing common services. A specific goal is to supply data management guidance for the ESONET demonstration cases, ensuring consistent processing of data thus allowing for serving homogeneous, quality assessed data and compilation of corresponding data products.

This document consists of three parts: (1) an overview of policies with respect to relevant bordering international initiatives and programs; (2) general aspects of sustainable data management; and (3) an ESONET specific part giving guidelines for the so-called demonstration cases and the subsequent long-term network operation.

As ESONET is strongly related with the Global Ocean Observing System (GOOS), the framework and guidelines of the establishment of the ESONET data management takes into account the data management and communication requirements of GOOS, and in particular the coastal module of GOOS. The data management activities cover the observation data, the information on the data collection (meta-data) and integrated products prepared on routine mode. It includes warnings related to sensors status provided in real time, and some warnings related to environmental events in real time and delayed mode, depending on the considered time scales. It does not include higher level information products prepared from the qualified data by scientific analysis activities.

This data management plan is conceived as a generic and dynamic document meeting today's requirements. However, the definition of what is correct is far from straightforward, often being a matter of opinion and opinions are subject to change (Lowry and Loch 1995). With increasing integration of observatories over

time, in particular through the demonstration cases, specifications have to be added and the document must be adapted. The ESONET observatories network follows an evolutionary design. With the ongoing technical and infrastructural development, the type of document chosen is a wiki-based structure, open for the incorporation of further development in the context of data capture, data flow, and data migration. This Data and Information Management Plan will be available at any time in its respectively updated form (http://wiki.pangaea.de/wiki/Data_and_information_management_plan and http:www.esonet-emso.org).

Part I: Overview of policies with respect to relevant bordering international initiatives and programs

ESONET and EMSO[1] deep sea-floor observatories are deployed on specific sites off the European coastline to allow continuous monitoring for environment and security. They will be organised in a unique management structure at European level (and part of a global endeavour), for long-term monitoring of environmental processes related to ecosystem life and evolution, global changes and geo-hazards. EMSO will be a key component of GMES[2] and GEOSS[3]. ESONET was set up to consider the feasibility of such a system.

GMES plans to implement a number of public information services to aid European policies. This initiative is a concerted effort to create co-operation between data providers and users to enhance the availability of relevant information through advanced or new services. GMES is a collaborative initiative by the European Union and the ESA (European Space Agency) and aims at expanding the European capacities in global monitoring. GMES will interact closely with INSPIRE[4] to provide data in a harmonised manner.

The INSPIRE directive intends to make harmonised sources of geographical information available to support the formulation, implementation and evaluation of European Union policies. It intends to force the creation of a European spatial information infrastructure that delivers integrated spatial information services.

Both GMES and INSPIRE will contribute strongly as the European Input to GEOSS by the intergovernmental Group on Earth Observations (GEO; http://www.earthobservations.org/).

ESONET is also strongly related to GOOS[5] as the oceanographic component of GEOSS. GOOS is a permanent global system for observatories, modelling and analysis of marine and ocean variables to support operational ocean services worldwide and is sponsored by IOC[6], UNEP[7], WMO[8], and ICSU[9].

IOC – as a body of UNESCO – has established already in 1961 the International Oceanographic Data and Information Exchange (IODE) programme to facilitate the exchange of oceanographic data and information

[1] European Multidisciplinary Sea-floor Observatory; http://www.ifremer.fr/esonet/emso/
[2] Global Monitoring and Environment System; http://www.gmes.info/
[3] Global Earth Observation System of Systems; http://www.epa.gov/geoss/
[4] Infrastructure for Spatial Information; http://www.ec-gis.org/inspire/
[5] Global Ocean Observing System; http://www.ioc-goos.org/
[6] Intergovernmental Oceanographic Commission; http://ioc.unesco.org/
[7] United Nations Environment Programme; http://www.unep.org/
[8] World Meteorological Organization; http://www.wmo.ch/
[9] International Council for Science; http://www.icsu.org/

between participating member states and by meeting the needs of the users for data and information products. One major commitment of the programme is the long-term accessibility and archival of oceanographic data and metadata, which was more recently taken up by a variety of working groups (e.g., CCSD[10], OECD[11]), which place strong emphasis on the importance of archiving scientific data on a long-term basis and encourage the open access to this data in order to promote an exchange of ideas, information and knowledge. They recognize the crucial role of adequate data centres for the archival of scientific data in the emerging global science system and its infrastructure. In January 2008 the IODE/JCOMM[12] forum on oceanographic data management and exchange standards will hold a first session to gain a broad agreement and commitment to the adoption of standards related to ocean data management and exchange.

Since 2002 ESFRI[13] aims with its Roadmaps to support a coherent approach to policy-making on research infrastructures in Europe.

Imbedded in this network of high ranking international programs and projects an ESONET contribution can only be successful if it integrates and adapts to international standards and protocols, communication requirements, data management and data sharing arrangements, defined by the correspondingly responsible Commissions.

[10] Consultative Committee for Space Data Systems (2002)
[11] Organisation for Economic Co-Operation and Development (2007)
[12] Joint WMO/IOC Technical Commission for Oceanography and Marine Meteorology; http://www.jcomm.info/
[13] The European Strategy Forum on Research Infrastructures; http://cordis.europa.eu/esfri/

Part II: General aspects of sustainable data management

Data are entities and supposed to be unique. Any data entity is supposed to be a digital object. A data entity consists of meta-information and data. Meta-information is any information describing a data set. Data is the pure scientific information, which can be on hand as numbers[14], text[15], graphics[16], audio and video recording and reproduction etc.

If data has been gathered already – may be as part of ancient inventories, antiquated software, unusual data storage medium, etc. – it is called data legacy. One challenge of data management is to keep legacy data accessible as one day they become cultural heritage. Open access sensu BOAI[17] is one mean to prevent legacy data from eventual loss as the old scientific tradition of publication merges with the new technology of the Internet.

For this reason, the whole documentation history of data from acquisition over processing towards archival is crucial to the understanding and the successful long-term use of any data.

In order to guide data producers and Principal Investigators (PIs) a *Template for Project Data Management* is supplied in the annex, which supplies data handlers with a checkbox list of necessary information and working steps needed for a proper data flow that ensures maximum quality of data.

RRR - Roles, rules and responsibilities

It is an ancient customs that scientists publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. After the famous British H.M.S. Challenger deep-sea expedition (1872–1876) returned, an international team of investigators analyzed the staggering body of observations and converted them into records of qualitative and quantitative data. By 1895 the study was

[14] A **number** is an abstract idea used in counting and measuring. The definition of number has been extended over the years to include such numbers as zero, negative numbers, rational numbers, irrational numbers, and complex numbers.

[15] **Text** may refer to character or string data, computer code segment, a portion of memory or of an object file that contains executable computer instructions, an application whose primary input and output are based on text rather than graphics, plain or formatted.

[16] **Graphics** are visual presentations on some surface, such as a wall, canvas, computer screen, paper, or stone to brand, inform, illustrate, or entertain. Examples are photographs, drawings, Line Art, graphs, diagrams, typography, numbers, symbols, geometric designs, maps, engineering drawings, or other images. Graphics often combine text, illustration, and colour. Graphics can be functional or artistic. Graphics can be imaginary or represent something in the real world. The latter can be a recorded version, such as a photograph, or an interpretation by a scientist to highlight essential features, or an artist, in which case the distinction with imaginary graphics may become blurred.

[17] Budapest Open Access Initiative; cf. http://www.soros.org/openaccess/

completed and a 50-volume report published (e.g., Murray and Renard, 1891). Today, as publication space is limited and data tables are not delivered anymore with the publication, data sets are – if at all – provided to the publisher as discrete entities where they are catalogued and stored as supplements.

At the same time, however, any National Science Foundation requires from data producers and Principle Investigators to take over responsibility for the documentation of scientific data.

World Data Centres and other certified data centres then become the responsible bodies for the long-term archiving of the project's data. They take care for proper documentation including information necessary for the discovery and comprehension of the scientific data and for the understanding of the used instrumentation and sensors, such as information on the principle investigator, time and location of the measurements, and methods used. Likewise, they will quality control data independently and fairly, check against historical records, and validate the contributed datasets and associated metadata. Experienced scientists are expected to assist data centres as data evaluators or as builders of specific data collections. Eventually, archived data sets will be assigned a unique persistent identifier (e.g. DOI[18]).

Data infrastructure

From a client's perspective data infrastructure is one of the least important issues in science: Scientific information must be available online wherever you are, whenever you need it, with no restriction at all, comprising any information of interest, supporting your whatsoever operating system, ensuring maximum quality etc. (Dittert et al., 2001). These client's demands require a highly complex data infrastructure that can't be dealt with as a single data centre anymore. A huge body of international initiatives and programs try to keep pace with scientific requirements, technical developments and exponentially increasing amounts of data available in ever shorter time-scales. The aim of the data management system is to create a standard compliant data infrastructure from data capture towards data products based on global standards (Nebert, 2004). Optimizing the use of existing infrastructures will increase compatibility and minimize financial costs. The data management structure can nevertheless be distributed over several centres, benefiting from regional or thematic expertises. Interfacing and communicating will be made possible by the use of common standards.

Standards and Protocols

From a technical and content specific point of view, standards are essential to assure a sustainable high quality data management. As metadata comprise all describing information with respect to a data entity,

[18] Digital Object Identifier; cf. http://www.doi.org/ ; http://www.std-doi.de

different scientific communities will describe different needs for information. Thus, different metadata standards based on standardized XML metadata formats have been developed in order to successfully serve these needs of the different communities. These standards supply corresponding content structures to carry the necessary metadata information.

In 1987 ESDAWG[19] produced the so-called Directory Interchange Format (DIF) for catalogue and data system interoperability. It defined the type of information necessary to describe data. In 1994 the international DCMI[20] initiative started an open forum for the development of interoperable metadata standards for a wide range of purposes. In its simple version, Dublin Core consists of 15 metadata elements. Originally adapted in 1994 and revised in 1998 FGDC[21] promotes the coordinated development, use, sharing, and dissemination of geographic data which eventually led to the Content Standard for Digital Geospatial Metadata (CSDGM). Since its second version the base CSDGM can be customized by adding for example extension elements. Recently, ISO[22] (Kresse and Fadaie, 2004) has developed the most advanced international metadata standard ISO 19115. It is a component of the ISO 191xx series for Geospatial metadata and defines 20 core elements and more than 400 optional metadata elements for describing geographical information and associated services. Of particular interest for ESONET is SensorML[23], a sensor model language, which provides standard models and an XML encoding for describing any process, including the process of measurement by sensors and instructions for deriving higher-level information from observations.

Use of common standards on data and metadata among the communities is one requirement for the deployment of portal technology based on the harvesting of distributed data from data centres. OGC CS-W[24] is one example for a standardized network protocol for distributed searches in the geo-scientific world. Another, HTTP-based protocol is OAI PMH[25] (cf. Van de Sompel et al., 2004) that supports incremental harvesting. Once extracted, metadata can be disseminated in a standardized manner through data portals using Generic Metadata Portal Software Framework (e.g., panFMP, Schindler and Diepenbroek, in press). For technical standards W3C[26] developed appropriate solutions.

Data Processing

A specificity of observatories is real time data. This data should be automatically decoded, reformatted, documented, corrected from measurement artefacts and translated in geophysical parameters before loading in the data management system, in general a relational database system (RDBS). In addition to this processing, quality control procedures should be applied with the following objectives:

[19] NASA Earth Science Data Systems Working Groups; http://www.esdswg.org/
[20] Dublin Core Metadata Initiative; http://dublincore.org/
[21] the US Office of Management and Budget's Federal Geographic Data Committee; http://www.fgdc.gov/
[22] International Organization of Standards; http://www.iso.org/
[23] SensorML; http://vast.uah.edu/SensorML/
[24] Open Geospatial Consortium Catalogue Service for Web; http://www.opengeospatial.org/standards/cat
[25] Open Archives Initiative Protocol for Metadata Harvesting; http://www.openarchives.org/
[26] World Wide Web Consortium; http://www.w3.org/

- to insure a minimum level of quality in data, information and product
- to insure coherence and compatibility between data from different observatories

The quality assurance of the system will insure that all these procedures are fully documented and applied.

Data production and provision

As any project has a clear understanding of what the scientific goals are to be expected at the end of its funding period, the way of how to proceed towards this scientific goal is defined in the *Description of Work*. Based on that fact it should be implicit that the project promotes the development of common core measurements and methods as well as metadata required for individual international experiments particularly those meant to be comparable among different sites. This includes the definition of minimum measurement requirements for the different sites. These specifications should take into account intersections to related international and national projects in order to make results among projects comparable.

Quality assurance (QA)

Quality assurance, or QA for short, is the activity of providing evidence needed to establish quality in work, and that activities that require good quality are being performed effectively. All those planned or systematic actions are necessary to provide enough confidence that a product or service will satisfy the given requirements for quality. For products, quality assurance is a part and consistent pair of quality management offering supposedly fact-based external confidence to customers and other stakeholders that a product meets needs, expectations, and other requirements. QA covers all activities from design, development, production, installation, servicing to documentation. One of the most widely used paradigms for QA management is the PDCA (Plan-Do-Check-Act) approach, also known as the Shewhart cycle (Deming, 1986; Shewhart, 1939).

The creation of several instances that focus on quality of data on different levels of the data- and workflow is one main procedure to assure the quality of the project's data. This requires the strong involvement and co-operation of data providers and data managers.

Long term archiving

Final storage and public access is supposed to be performed by certified data centres only, i.e. data centres that meet the nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification (Dobratz et al., 2006) and the Reference Model for an Open Archival Information System (Consultative

Committee for Space Data Systems, 2002). Criteria for a certification include the data centre's adequate usage of the information on, the access to, and the interpretation of its data on a long-term perspective. One main focus must be the acquisition of metadata for formal and content based description and identification of data. During all phases of submission, storage and usage (access) the integrity and authenticity of the data must be guaranteed. At the same time, the data centre must commit to long-term preservation and permanent maintenance of the usability of the archived data. Data must be uniquely and permanently identifiable through persistent identifiers.

Data products and data publication

There is immanent entropy in data archiving which refers to the degree of organization of data. That means that the overall value of the data increases with its organizational condition. Raw data without meta-data and undocumented quality have the least value. Personal file collections, institutional working data bases, and data centres increase in organizational condition. Peer-reviewed data entities with persistent identifier, which are made available through portals and grids, have the highest value.

The concept of data publication is correspondingly the overall goal for project data products and data publications (e.g., Klump et al., 2006).

The user perspective

As pointed out earlier (cf. data infrastructure), data management is supposed to serve the client's needs. The information system hence must be science driven, and its products must be developed in compliance with the user's needs. This requires a close co-operation between data centre and user taking into account identifying, monitoring, documenting, and reviewing the user's demands in an adequate, efficient and integrative way in order to eventually adapting to them. The system invites user's involvement, input, and recommendations for improvement accompanying the project for advancement and adjustment of its products as the needs of users – like the technical advancement – are subject to change over time. At the same time, data products must serve a broad range of user communities.

On the other hand, appropriate mechanisms for the co-ordination of user requirements need to be created to document and review user demands to improve the delivery of information.

Part III: ESONET guidelines for demonstration cases and the subsequent long-term observatories network operation

Topology

ESONET as the European Seas Observatory network is based on a long-term operation of observatories, starting out with several demonstration cases (Fig. 1).

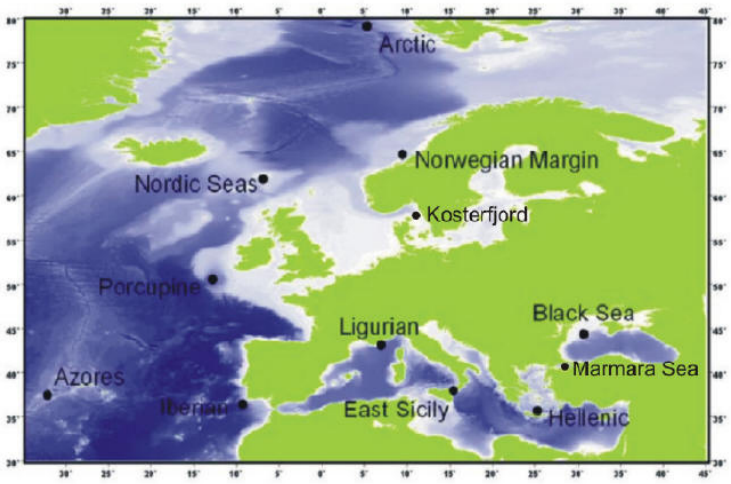


Fig. 1: Twelve permanent sites are initially identified, complemented by a mobile observatory.

These observatories will produce huge amounts of scientific information, which then are subject to further treatment through data management activities. A complement to the data management system is the creation of an interactive topology of existing regional observatories, which itself will bring – through data mining – a complete state of the knowledge on the sites and will eventually impart among regional observatories through a common data infrastructure based on global standards (cf. Part II: General aspects of sustainable data management). This mapping will help the development of strategies for necessary adaptations and extensions of existing nodes.

The data to manage is heterogeneous and consists of observation data with numerical values and digital pictures and video, meta-data, and derived data products.

The operators of observatories have the opportunity to register and update online sensors and scientific equipment through a Sensor Registry Entry Form (Fig. 2) and contribute so to the topicality and further development of the topology. Additionally, the interested public is able to explore the so-visualized project

with an active Google™ Earth based map that serves further and more detailed information on selected items.

Fig. 2: Example for the possible outline of the online Sensor Registry Entry Form.

Staff responsibilities

The data management system is based on a foundation of committed project scientists and research teams, whose full co-operation is essential for the efficient operation of the system. One component of the data management system will be the creation of Data Evaluation Units (DEUs). This diverse group of experienced ESONET scientists takes care for proper documentation including information necessary for the discovery and comprehension of the scientific data and for the understanding of the used instrumentation and sensors, such as information on the principle investigator, time and location of the measurements, and methods used. Likewise, they will quality control data independently and fairly, check against historical records, and validate the contributed datasets and associated metadata. They are expected to participate as data evaluators. On their request, specific ESONET data collections are built.

Project scientist data producers are the most important group within the data management system. They are required to submit metadata and analytical data produced and to keep the data management informed about

the current data status. The eventual success in data retrieval and exchange – though technically solved – relies on scientists' active participation in data management.

The Data Evaluation Unit and the long-term archiving centres are fully involved in the data flow from data capture to archiving, including participation in fieldwork, experiments, and science workshops when appropriate, in order to increase the interactions with the scientists and to promote the excellence in data management practises and the utilization of data management, analysis, and visualization tools, as needed.

The purpose is to build a continuously managed data set for ESONET users, also coherent with GEOSS equivalent observing systems operating with real-time and delayed mode. For this institutional levels of responsibilities have to be created and established.

A Principal Investigator (PI) is responsible for a specific observation site. He manages the site activities and provides data and metadata to the DEUs and certified archiving centres. These centres are responsible for setting up data server at a national of regional level according to the specifications approved by the data management group. They guarantee the availability of data, the compliance to the agreed formats, and organize the data processing, transfer and update with the PIs. Each ESONET site should reside only to one centre.

A Global Data Access Service is in charge of providing users a virtual access to the data served by the archiving centres, maintaining the project catalogue, synchronizing this catalogue with a second Service, and implementing viewing services for the global dataset. The creation of a second service is necessary for security reasons.

Data productions

The aim of ESONET is to create an organisation capable of implementing, operating and maintaining a network of multidisciplinary ocean observatories in deep waters around Europe from the Arctic Ocean to the Black Sea. These long-term observatories are crucial for European scientist to maintain world leadership.

In order to achieve this goal from an data management point of view, a first recommendation is the creation of site reports that include any basic information related to the datasets acquired or the experiments conducted such as location and timing of the stations (so-called GEOCODES), sampling strategy and methodology, inventories of all parameters acquired, as well as name and approachability of the responsible scientist. These reports are submitted to the certified data centres, enabling the data management to adequately prepare the prompt availability of preliminary data products or final publications once accordingly available at the centres.

Usable data infrastructure

Today's matter of course using the Internet gained a public face in the 1990s only. Since then, an overwhelming technological evolution happened and still is taking place. For the sake of ESONET, the data management group will investigate and adapt new tools and strategies and appropriate standards (Internet protocols, data and metadata standards, approved protocols for data quality assurance or control) in order to facilitate and promote ESONET data flows. This is particularly important with respect to future observations from new sensors, new platforms (autonomous underwater or remotely operated vehicles), as well as for continuous measurements, and for systems of data delivery in near real-time and delayed modes and for model output dissemination. The certified data centres for the long-term archiving of the project's data are working towards serving common exchange protocols for at least metadata to make use of portal software and make ESONET data available to users at one central online accessible data portal.

Template for project data management

ESONET adheres to the data set[27] philosophy of "Data entities". A data entity is supposed to be a digital object. A data entity is supposed to be unique. A data entity consists of meta-information[28] and data. Meta-information is any information describing a data set. Data is the pure scientific information, which can be on hand as numbers, text, graphics, audio and video recording and reproduction etc.

Meta-information shall contain in any case:

- The citation of the data entity [**Dataset**_Name];
- The Principal Investigator/s[29] (Beveridge and Morris, 2007) or author/s of the data entity and their contact details/name of infrastructure and legal responsible person: in *Citation Table* in ESONET_data_template.xls;
- The event/s[30] concerned: in *Events Table* in ESONET_data_template.xls;
- The variable/s[31]: in *Parameters Table* in ESONET_data_template.xls.

Meta-information shall contain where available and applicable:

- Geo-Code, i.e. latitude, longitude, 3rd spatial dimension, date, time of the event;
- The method/s employed to attain the variable/s;
- Device/s employed to attain the variable/s;

[27] cf. http://wiki.pangaea.de/wiki/Data_set
[28] cf. http://wiki.pangaea.de/wiki/Metadata
[29] **Authorship order** is of increasing importance for scientific careers and the success of collaborations. The first author typically makes the greatest contribution and the last has a leadership role. The process of choosing the order needs to foster understanding and accountability, while recognizing each author's contribution.
[30] **Event** is to be used as synonym for action, activity, episode, occurrence, operation, place, procedure, run, site, etc.
[31] **Variable** is to be used as synonym for determinant, factor, parameter,

- Primary reference towards the data entity;
- Project relevant information;
- Collaborating institute/s, person/s;
- Etc.

In order to gather all necessary information with regard to the data entity, a template is provided (ESONET_data_template.xls).

To give you an idea how the final templates looks like once filled, an example is provided (ESONET_data_example.xls).

The following text boxes give more detailed information on the reasoning of the template. As intuition strongly depends on experience, some will find it appropriate, others too daft, others too little meaningful. Please, let us know where adaptations are required (info@wdc-mare.org).

Likewise, our (geo-scientific) experience may be too constricted to cover all ESONET needs (science, policy, etc.). Please, let us know where extensions are required.

Dataset_Name … is the linking term between *Citation table*, *Events table*, *Parameters table*, and *Data table* (see below). Dataset_Name is plain text prose.

Citation table…includes the project name, description, references (relevant papers, forms or web

sites), quality assessment, and availability of the data set. The citation table is plain text.

It also includes contact details of the Principal Investigator/s (PI) and information to the respective research institution/s for all PIs involved in the work, including all personnel that is responsible for measuring and/or quality checking the parameter, i.e. each parameter is associated with one person (see *Parameters table* below).

Events table – An event (for definition see above) corresponds to the deployment of an instrument (e.g. CTD, fishing net, weather balloon) or a package of instruments (e.g. a "*rosette*" equipped with CTD, bottles and sensors for PAR, fluorescence, oxygen, etc). The events table is plain text, date, time, and plane angle degree.

Requested are name/s of event/s campaign/s, research vessel/s, and original station name/s as appearing in log books. Requested are a short description of the event and a quality assessment comment of the sampling equipment and procedures at each event.

Requested are date/s, time/s, longitude/s, and latitude/s of each event.

Parameters table…includes the name, unit, and description of each parameter or variable (for definition see above); it includes the name of the investigator responsible for each parameter, the reference/s (publication/s, web site/s, …), and a quality assessment comment of the method for this parameter.

Data table

Each datum (as the singular form of data; for definition see above) is associated to one parameter and to one event.

I. Preparing Data Tables

Data are usually of numerical format but can also consist of other formats (cf. chapter 1). Each datum is surrounded by metadata describing the what-, where, who- and howabouts.

Scientists are responsible for the scientific quality of the data and for providing all respective meta-information. Data base managers are responsible to cross-check, evaluate, and dig into data quality from a technical point of view. Scientists and data base managers together review the final version of the data set before it is released to the public.

This task is sometimes tedious but experience shows that it is the only way to ensure sustainable availability, long lasting integration, and conscientious use of data by other scientists.

Whether you submit one large dataset or several smaller ones is up to your convenience.

Example for a *Data Table for Biogeochemistry*

We suggest that biogeochemical data are submitted in table format consisting of a two-rows heading and as many columns as needed, but obligatory including two columns to the left referring to the *Event* Name/s and the respective sampling *Depth/s*.

Dataset_Name is the same through the whole data entity and links between all files related to the data entity. It is plain text prose and describes in a few words the content of the data entity.

Event_Name is (at best) an official label that was given from e.g. the chief scientist of a cruise: It is the label of the event at which one device is used at one single location at one time unit. If the device changes, or if the location changes, or if the time unit changes another event label is used. Profiles of measurements can have a start and an end of the location and the time. Instead of inventing something new, please make sure that you use the official label (cf. chapter 6).

Sampling_Depth are (at best) official numbers that were given from e.g. the chief scientist of a cruise: In water it is the depth, in air it is the altitude, etc. Instead of inventing something new, please make sure that you use the official number.

Parameter names have names, short cuts, and units on which internationally was agreed on. If you use units that are no SI units, please contemplate whether you can calculate your scientific values into SI units.

Data values are the numbers you measured. Please, make sure that particularly there is no typo and that you did not apply wrong factors [e.g.mmol/kg → µmol/l].

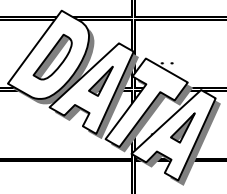
Data table example:

heading

Dataset_Name *Carbon isotopes, hydrography, and nutrients measured at mooring OG5 in the Schelde River in 2003-3004*

Event_Name	**Sampling**_Depth [m]	*Phosphate*	*Silicate*	…

M23_211	*2*	*0.23*	*2.9*	…
M23_211	*20*	*0.16*	*2.4*	…
M23_211	*50*	*0.30*	*3.5*	
M23_211	*75*	*0.41*	*3.9*	…
…	…	…	…	…



II. Preparing Citations Tables

In scientific publications there is an internationally agreed, informal way to give credits to collaborators: The most important contributors become co-authors, suppliers (persons and institutions) get acknowledged in the terminal paragraph. In fact, as we perceive data entities as data publications, credits must be given to the respective persons and institutions likewise.

We ask that the dataset citation is submitted in a table format that consists of a nine-rows heading and seven columns documenting all relevant persons or institutions.

Citation table example:

Dataset_Name	*Carbon isotopes, hydrography, and nutrients measured at mooring OG5 in the Schelde River in 2003-3004*
Dataset_Project_Name	Nordost-Atlantik-Expedition 1971

Dataset_Description	Transatlantic 14C-se[illegible] 49°S N
Dataset_Reference_Papers	Roether, Wolfgang; Münnich, Karl O; Ribbat, B; Sarmiento, Jorge L (1980), Meteor Forschungsergebnisse, Deutsche Forschungsgemeinschaft, Reihe A Allgemeines, Physik und Chemie des Meeres, Gebrüder Bornträger, Berlin, Stuttgart, A21, 57-70
Dataset_Reference_Metadata	EDMED-Form OG5_1_a.xls
Dataset_Reference_URL	http://doi.pangaea.de/10.1594/PANGAEA.602246
Dataset_Quality	Final quality check approved by PI
Dataset_Availability	Online available at www.pangaea.de

Investigator_Author_Rank	**Investigator**_Last_Name	**Investigator**_First_Name	**…**
1st	Roether	Wolfgang	
2nd	Münnich	Karl O.	
…	…	…	…

You can list for example organisations, data managers, scientific directors, scientists, any individual involved in the work and as many individuals as you wish. All scientists associated to a parameter (see 7. *Preparing Parameters Tables*) must be included in the Citation Table, as these have proprietary rights and responsibility for the data.

Only the organisations and/or individuals that are ranked (1, 2, 3 …) in the ***Investigator**_Author_Rank* column will appear as authors in the citation of the dataset. Authors can consist of all individuals involved in the work or can be limited to a few, e.g. organisations, scientific directors, or data managers. Ranking of authors is done as you would normally do in the case of publications submitted to a scientific journal.

III. Preparing Events Tables

Events can have many names and synonyms, whose use strongly depends on the scientific community you work in. For some, the term “event” is a festive evening, pale-oceanographers understand “hole” or “leg” or “site” depending on the size of the event, meteorologists apprehend “station”, and so forth (for definition see above).

We ask that metadata are submitted for each *Event* in a table format that consists of a two rows heading and 21 columns documenting the *Events*. The first column to the left must correspond to the *Event_Name* that

appears in the first column of the Data Table. You are not supposed to fill all cells but as many as you are concerned.

Events table example:

heading

Dataset_Name *Carbon isotopes, hydrography, and nutrients measured at mooring OG5 in the Schelde River in 2003-3004*

Event_Name	**Event**_Cruise_Name	**Event**_Ship_Name	**Event**_Station_Name	…
M23_211	M23	Meteor (1964)	211	
…				

METADATA

An *Event* is an official label that was given from e.g. the chief scientist of a cruise: It is the label of the event at which one device is used at one single location at one time unit. If the device changes, or if the location changes, or if the time unit changes another event label is used. Profiles of measurements can have a start and an end of the location and the time. Instead of inventing something new, please make sure that you use the official label.

An *Event* corresponds to the deployment of an instrument (e.g. CTD, fishing net, weather balloon) or a package of instruments (e.g. a "*rosette*" equipped with CTD, bottles and sensors for PAR, fluorescence, oxygen, etc). Events table is plain text, date, time, and plane angle degree.

Requested are name/s of event/s campaign/s, research vessel/s, and original station name/s as appearing in log books. Requested are a short description of the event and a quality assessment comment of the sampling equipment and procedures at each event.

Requested are date/s, time/s, longitude/s, and latitude/s of each event.

Of course, an instrument can be deployed several times during a research programme and several instruments can be deployed at one location or one time. Nevertheless, each deployment of an instrument is bound in space and time and constitutes a distinct *Event*. For example, two "*rosette"* casts conducted at the same sampling station during the same night off Hawai'i at one distinct time on the same research vessel constitute two (!) distinct *Events*. In case of a *"rosette"* cast, several instruments are screwed to one frame. The data obtained by all instruments included in the *"rosette"* package (e.g. CTD, sensors for PAR, fluorescence, oxygen, etc) as well as measurements/experiments made from the water collected in the bottles (e.g. hydrochemistry, taxonomy, pigments, primary production, etc) are all part of the same *Event* and will share the same ***Event**_Name*. Yes, there is a logic behind: the singularity of the device. If you are unsure how and what to put at which place: please, inquire!

Instruments that continuously measure a parameter over time and/or space (e.g. gliders, argo floats, towed instruments, underway fluorescence, etc) are treated as one single sampling Event that has one distinct start location and time and one distinct end location and time.

Description of columns in the Events Table

Event_Name is (at best) an official label that was given from e.g. the chief scientist of a cruise. Please make sure that you use the official label (see above).

If no Event_Name is available – almost impossible – use simple, unique terms. Data curators at the data centre will assign respective Event_Names based on the *Cruise_Name, Ship_Name*, *Station_Name*, and *Event_Description* that you provide.

Event_Cruise_Name is (at best) an official label that was given from e.g. the chief scientist of a cruise. The term cruise has many synonyms among them "campaign", "excursion", "expedition" and so forth.

Event_Ship_Name is (at best) the official name of the research vessel (ship, aircraft, etc.).

Event_Station_Name is (at best) the official name defined by the research programme or in the log book. However, this is often prose but nevertheless anchored in everybody's minds (e.g. *Waikiki beach mesocosm* instead of *M23_211*).

Event_Description is the description of the sampling procedure, the instruments used, sample handling, treatment, storage, etc. – everything that falls into the intuitive understanding of *Describing the event*. If it is a standard procedure, please refer to the appropriate literature. However, scientists are used to adaptation and invention: in this case create prose as if you describe your neighbour at home in detail what exactly you were doing ! Example: net tow's mouth area, sampling volume, type and duration and speed of tow, calibration, wire length and wire angle. Describe as much as possible… up to analysis or experimentation. If you run out of space, create a new file and refer to it in Event_Description.

Event_Quality_Comment is the comment you would give to your PhD student on the real quality of the sampled event. Did the bottle close correctly, was there leakage, was it sterile, was voltage a little low, etc.? Try to be honest.

Date_Nominal is the date when the event happened. Please, indicate the time-zone or refer to UTC. The format is dd/mm/yyyy. You may add columns to accommodate other formats.

Date_Start & Date_End are the days when the event started and when it finished during a >1 day event or crossing midnight. The format is dd/mm/yyyy. You may add columns to accommodate other formats.

Time_Nominal is the time when the event happened. The time zone is UTC. Any other time must be defined. The format is hhmm. You may add columns to accommodate other formats.

Time_Start & Time_End is the time when the event started and when it finished during a >1 instant event. The time zone is UTC. Any other time must be defined. If applicable and if available, please indicate the

times when the instrument was deployed and recovered. The format is hhmm. You may add columns to accommodate other formats.

Latitude_Nominal & Longitude_Nominal are the two dimensions of the respective spot on the planet. The formats are plane angle degrees and decimal minutes (e.g. 65°45.7'S) OR decimal degrees (-65.1234). Please, be consistent through all events/files when selecting a format.

Latitude_Start, Latitude_End, Longitude_Start & Longitude_End are the two dimensions of the two respective spots where the event started and where it ended (e.g. profile). If applicable and if available, please indicate the latitudes and longitudes where the instrument was deployed and recovered. The formats are plane angle degrees and decimal minutes (e.g. 65°45.7'S) OR decimal degrees (-65.1234). Please, be consistent through all events/files when selecting a format.

IV. Preparing Parameters Tables

The term parameter is ambiguous and has many synonyms as determinant, factor, variable, etc.

A variable is a quantity whose value may vary over the course of an experiment (including simulations), across samples, or during the operation of a system. Variables are generally distinct from parameters, although what is a variable in one context may be a parameter in another. It is a measurable factor, characteristic, or attribute of an individual or a system—in other words, something that might be expected to vary over time or between individuals.

We ask that parameters' metadata are submitted in a table format that consists of a two-rows heading and five columns documenting the *Parameters*. The *Parameter_Names* in the first column to the left must correspond to the *Parameter_Names* in the heading of the Data Table.

Parameters table example:

heading

Dataset_Name *Carbon isotopes, hydrography, and nutrients measured at mooring OG5 in the Schelde River in 2003-3004*

Parameter_Name	**Parameter**_Type	Parameter_Units	Parameter_Investigator	…
Phosphate	measured	µmol/l	Roether, Wolfgang	
Silicate	measured	µmol/l	Roether, Wolfgang	
…				

METADATA

Description of columns in the Parameters Table

Parameter_Name is (at best) the name – including its shortcut – on which internationally was agreed on. In any other case name the parameter to the best of your knowledge.

Parameter_Type specifies whether the values were measured, determined, computed, converted, etc.

Parameter_Units are supposed to indicate the units in which you report the values. Units are (at best) given in SI. If you do not use SI units, please contemplate whether you can calculate your scientific values into SI units. In any other case name the unit to the best of your knowledge.

Parameter_Investigator_Last-Name gives a sort of proprietary right to one person to this particular parameter and thus a responsibility for the quality of the data. Use last names that you report in the citation table.

Parameter_Method_Description describes the analysis or experimental procedure for each parameter. This requires one text for each parameter. Please describe the equipment used, experimental protocols, analyses, detection limit of instruments, confidence intervals, methodological artefacts, etc. Include literature where applicable and appropriate. If you run out of space, create a new file and refer to it in Parameter_Method_Description.

Parameter_Method_Reference In addition to the Parameter_Method_Description, you can provide references that describe the analysis or experimental procedure. This can be a conventional text reference, a DOI reference, a “permanent” web address, or a .pdf file that you wish to archive along with the data.

References

Beveridge, C. and Morris, S., 2007. Order of merit. nature, 448(7152): 508.

Consultative_Committee_for_Space_Data_Systems, 2002. Reference Model for an Open Archival Information System (OAIS). CCSDS Secretariat, Washington, DC, pp. 148.

Deming, W.E., 1986. Out of the crisis. Cambridge University Press, Cambridge, 528 pp.

Dittert, N., Diepenbroek, M. and Grobe, H., 2001. Scientific data must be made available to all. Nature, 414(6862): 393.

Dobratz, S., Schoger, A. and Strathmann, S., 2006. The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification. Texas Digital Library.

Klump, J. et al., 2006. Data publication in the Open Access Initiative. Data Science Journal, 5: 79-83.

Kresse, W. and Fadaie, K., 2004. ISO Standards for Geographic Information. Springer, Berlin.

Murray, J. and Renard, A.F., 1891. Report on deep-sea deposits, based on the specimens collected during the voyage of H.M.S. Challenger. Report on the scientific results of the voyage of H.M.S. Challenger during the years 1872-76, under the command of Captain Sir George S. Nares and Captain Frank Tourle Thomson, 24. Her Majesty's Stationary Office, London, 525 pp.

Nebert, D.D., 2004. Developing Spatial Data Infrastructures: The SDI Cookbook.

Organisation_for_Economic_Co-operation_and_Development, 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD Publishing, Paris, pp. 24.

Schindler, U. and Diepenbroek, M., in press. Generic Framework for Metadata Portals. Computer & Geosciences.

Shewhart, W.A., 1939. Statistical Method from the Viewpoint of Quality Control. Dover, New York.

Van de Sompel, H., Nelson, M., Lagoze, C. and Warner, S., 2004. Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine, 10(12).