Draft Data[1] Management Policy for Oceanographic data (IGBP/SCOR)

Bernard AVRIL, Nicolas DITTERT

This draft data management plan has been generated to stimulate discussion regarding data management plans for new marine projects. Previous projects, notably the Joint Global Ocean Flux Study (JGOFS), found that the management of diverse and heterogeneous data sets was a significant task and difficult to accomplish. This draft plan is based on experience from JGOFS and other projects. It is not meant to be exclusive.

The purpose of a data management plan is to create an international data system into which national research, observation, and modelling activities can submit data and which will provide services that will increase the value of the data. Such a plan should provide guidance for scientists and funding agencies on national and international levels. A fundamental underlying principle for the policy is full and open exchange of data and information for scientific and educational purposes, which is an obvious necessary condition for an integrated international research project. It is also the policy of the International Council for Science (ICSU), which is the parent of the Scientific Committee on Oceanic Research (SCOR) and the International Geosphere-Biosphere Programme (IGBP).

Benefits of a Data Management Plan

Data management plans are useful because:

- A data management plan will improve the ability of project scientists to analyse and document their research findings and link them to a global context of ocean processes, by easing access to and manipulation of datasets.
- A comprehensive data management plan will make data from field observations, research, laboratory studies, and modelling, available, comparable, and usable by scientists both within and outside a given project.
- Easy availability of data will help in ongoing research planning and synthesis of research results.
- A data management system will provide a regular and consistent procedure for data submissions and a single source of information about how to access data from many investigators working in different countries.
- The value of individual datasets may be increased by the ability of project scientists to compare their data with related datasets from the same region or studying the same processes. This could lead to new scientific collaborations and joint publications.

[1] Scientific data are defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Data do not include preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues.

- Data management, analysis, and visualization tools, developed by data managers and data users, and provided as part of the data management system, will help investigators understand their data.
- Data producers will receive increased recognition when their datasets and related peer-reviewed publications are officially project-labelled and designated with an unique Digital Object Identifier (DOI). Other project scientists will receive professional recognition as data evaluators (within the DEUs) or as data collection builders.

At a minimum, data for a project must be

- complete and comparable (for global integration),
- accurate (through quality control and analysis),
- attributed to the responsible data producer,
- released in a timely manner,
- identifiable,
- documented with associated metadata records, which describe data, including collection and analysis methods used and references,
- retrievable when needed, and
- properly secured and preserved for the long term.

Projects are likely to generate very large, distributed, heterogeneous sets of marine, biogeochemical, and ecological data, associated with advances in technologies and scientific concepts. The range of data will become even larger when other projects participate in the proposed data management system. The project data management system should be developed and implemented as early as possible so that data and information flow are initiated as soon as the first research results are available and maintained until the project's end, to ensure a timely, continuous delivery of the datasets to the community, and also to increase the quality of those datasets and of the ongoing fieldwork and synthesis. The Scientific Steering Committee for a project will be responsible for implementing and enforcing the procedures, and modifying them, as necessary.

Data Policy

- The project will require full and open access to well-documented and organised data and metadata from all parties involved in the project.
- Project data and metadata management will be internationally coordinated.
- The data management system will include incentives to encourage the active participation of project scientists in submitting data and participating in data management activities. It will also include enforcement procedures to respond appropriately to scientists who do not fulfil their obligations. Funding agencies will be involved in preparing and managing the incentive system.
- To ensure the highest possible quality of data, datasets and associated metadata will be peer reviewed before accepted as project data.

- The system will seek to ensure the shortest path of data flow from project scientists to the most appropriate ICSU World Data Centre (WDC). The WDCs represent the best international framework for the long-term preservation and worldwide dissemination of all data.
- A preliminary report will be required within ***XX*** months after the fieldwork or the experiment. This report should include basic information related to the datasets acquired and/or the experiments conducted, such as location and timing of stations; sampling strategy; inventory of all parameters acquired; time frame of subsequent analysis and data delivery; and any requested restrictions of data access.
- Complete data and metadata should be delivered to the relevant Data Evaluation Unit (see below) as soon as possible and within a maximum of two years after the completion of the fieldwork or of the experiment, and before open public dissemination. ***Some fields of research require that data are in a recognized data center before a paper can be published. Is this what we are advocating?*** Some datasets and information are available as soon as the fieldwork or experiment ends (hydrology, meteorology, etc.), or soon after some basic data quality control (dissolved nutrients, etc.). ***Some projects/programs require that the chief scientist of a cruise turn in a preliminary report as he/she is leaving the ship. Should this be advocated?*** The datasets to be delivered later (from 3-month to 2-year delays) and those not acquired in accordance to the initial cruise or experiment plans, should be identified. Data should be available to other scientists participating in the same cruise more quickly. ***Should a specific timetable be set? For some projects, participating scientists are required to share all their data within 6 months to others participating in a cruise and within 2 years to all scientists.***
- The data management system will promote full recognition and acknowledgement of the contributing scientists by providing an project-label for data sets and a DOI for publications that meet project requirements.
- Each national and international project effort will be expected to fully support and promote the project data management system described above.

Data Management Structure

(see Figure 1)

The project data management system will have a semi-distributed, scalable and flexible structure. This will enable data management to involve scientists who generate the data and will accommodate the different research and data management systems found in different countries.

It is important to recognize that the data management system is based on a foundation of committed project scientists and research teams, whose full cooperation is essential for efficient operation of the system and achievement of the full potential of any project research. Other major components of the project data management system will include a International Information Management Office (IIMO) ***(associated with the IPO?),*** a Data Management Committee (DMC), Data Evaluation Units (DEUs), and the project Scientific Steering Committee (SSC) and International Project Office (IPO).

- Project scientists—Project scientists, as the data producers, are the most important component of the data management system. Data producers will be required to submit meta-information and analytical data produced under the project umbrella and to keep the IIMO informed of the current data status. Ultimate success in data retrieval and exchange rely on scientists' participation in data management, which will require the establishment of appropriate incentives.

- International Information Management Office (IIMO)—This office will coordinate all information and data management aspects for the project duration, including data evaluation, support of the various partners in data management, and evaluation of the data flows and usages. The IIMO will also facilitate the integration of the project data management system within the frameworks already existing nationally (NDC, NCP, etc.) and internationally (e.g., WDCs, GCMD/CEOS-IDN, other projects and programmes). The IIMO, assisted by an ad-hoc ***Publications Committee (PC) [Is this necessary?],*** will attribute an project-label and consecutive number for the relevant peer-reviewed publications when they are submitted to a scientific journal, if the related datasets and associated metadata are delivered within the project data management system and certified by a DEU. If the publication acknowledges the specific project appropriately, it will be assigned a unique Digital Object Identifier (DOI). The IIMO will create the project database architecture, based on information from the SSC. The architecture will need to be flexible enough to evolve over the life of the project and as unforeseen data are added. The IIMO will be staffed with paid personnel, whose salaries may be supported fully through the project, may be shared among several projects, or may be provided as an in-kind contribution by a national or international data centre.

- Data Management Committee (DMC)— The Data Management Committee will be a subcommittee of the project SSC. It will consist of scientists from the involved research projects and data managers, serving as volunteers, who will guide the implementation of the project-specific data management system and its continued operation. The DMC will work with the IIMO to track data submissions and tracking.

- Data Evaluation Units (DEUs)—Several Data Evaluation Units (DEUs) will be created to (1) quality control data independently and fairly, check against historical records, and validate the contributed datasets and associated metadata relevant to a specific group of the project core parameters. The DEUs will report to the IIMO. The DEUs will deliver project-labelled certificates for the datasets and will publicly acknowledge scientists' contributions to the project. The DEUs will create project-labelled data collections. Each DEU will be flexible in its working structure, will interact with its data providers, and will be led by a volunteer scientist with strong inter-personal skills and high scientific expertise and recognition, chosen by the SSC, after a call for proposals and some possible direct solicitations. The most experienced scientists are expected to participate as

data evaluators or as builders of specific data collections. The DEUs will use the collected datasets and associated metadata only for the tasks defined by the SSC and IIMO. Datasets that are relevant to the project science but not covered by the expertise of any DEU will be transmitted with their associated metadata preferably directly to the IIMO and possibly to a national institution (preferably national project DMO), in agreement with the SSC or IIMO. ***Another way to organize DEUs would be on a cruise-by-cruise basis.*** DEUs can be individuals, research centres, academic departments, or industrial laboratories, known for their excellence in a particular research topic. They will be recruited through regular calls for proposals. ***Has this concept been tested? Do people respond to calls for proposal that have no funds attached?***

- World Data Centers (WDCs)—Final data storage and public access will be performed by World Data Centers (WDCs). WDCs are institutions that are bound to the rules of, and are supervised by, the Panel on World Data Centers and the International Council for Science.[2]

- Project SSC—The project SSC will promote the development of (any) core measurements, measurement methods, and metadata required for individual international experiments, particularly those meant to be comparable among different sites. The specifications of the project measurements will take into account measurements being conducted by related international and national projects, as well as completed research projects, such as the Joint Global Ocean Flux Study (JGOFS), so as to make it possible to compare results from existing projects as well as completed projects.

- Project IPO—The project IPO will act on behalf of the SSC to ensure that the IIMO functions according to the SSC's instructions.

The IIMO and DMC will investigate or 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 project data flows. This will be especially important for future observations from new sensors, new satellites, new platforms (autonomous underwater or remotely operated vehicles, inexpensive, low-maintenance monitoring systems and sophisticated buoys), as well as for continuous measurements, global surveys, ships of opportunity observations, and for systems of data delivery in near real-time and delayed modes and for model output dissemination. The IIMO and DMC will investigate or adapt new tools and strategies for data rescue, data archeology, data mining, and data integration in relation to the needs of the project. The IIMO and DMC will follow developments and review existing governmental, intergovernmental, commercial, and legal constraints on data access and intellectual property issues (including published articles, CD-ROMs) for purposes of scientific research.

[2] ICSU Panel on World Data Centers. Principles and Responsibilities of ICSU World Data Centers, www.ngdc.noaa.gov/wdc/guide/gdsystema.html, 2002.

The IIMO and DEUs will provide all project scientists with support, recommendations, guidelines, and priorities regarding information and data management plans. They will coordinate the establishment of a specific metadata portal for the project with the help of GCMD – CEOS / IDN. They will promote the project data policy and assist in the development of national project data management efforts (possible national project Data Management Offices, DMOs) and cooperation with national data centres (NDCs) (where these exist).

The project-labelled datasets and related metadata will be placed both on-line (provisionally, with a password-controlled access until the end of the 2-year period) and published as project-specific data products as soon as possible, either by the DEU or another entity (IIMO, WDC…), after full delivery and quality check (with a clear edition date or version number, to avoid confusion). Data may later be revised or reworked in a more convenient, integrated way (e.g., in a "data collection"). This early availability will increase the usefulness of project data.

DEUs and NDCs will be fully involved in the project, 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.

Conventions and Terms

The proposed data management system is based on the guidelines of "Good scientific practice in research",[3, 4] the WIPO copyright treaty[5], and the DOE-NIH Guidelines for Sharing Data and Resources.[6] Data generated by any project will be a substantial resource, which will enable creation of new knowledge and may also be the starting point for further studies.[7,8]

Data—The term "data" embraces the analytical value (number, parameter/variable, unit) and its entire meta-information (i.e., any information that describes the analytical value: reference, method, gear, site, campaign, project, etc.). Data comprise real-time data (e.g., Argo profiling floats) and delayed mode data (e.g., results of biological, chemical,

[3] ICSU/CODATA. Access to databases: A set of principles for science in the Internet era, www.icsu.org/Library/Central/Statem/access.html, 2000.
[4] European Science Foundation. Good scientific practice in research and scholarship, www.esf.org/medias/ESPB10.pdf, 2000.
[5] World Intellectual Property Organization. WIPO copyright treaty, www.wipo.org/eng/diplconf/distrib/94dc.htm, 1996.
[6] HGMIS Staff. NIH, DOE Guidelines Encourage Sharing of Data, Resources. *Human Genome News* 4(5):4, www.ornl.gov/hgmis/publicat/hgn/v4n5/04share.html, 1993.
[7] JOI/U.S. Science Support Program. Acknowledgement and Keyword Policy, www.joi-odp.org/USSSP/acknowledge.htm, 2002.
[8] PAGES. PAGES Data Policy, www.pages.unibe.ch/data/policies.html, 2003.

physical, and geological analyses). Oceanographic data is likely to be highly complex, heterogeneous, and derived from many different sources.

Data Flux—Data produced under any project umbrella should follow the Information Flux Model (Figure 1). This routine will ensure that all project scientists will benefit to the greatest possible extent from each other's results with a minimum of effort. Scientists also may be subject to national data management regulations, but for the reason of data consistency, WDCs will serve as the final archive of project data.

Data Sharing— the projects should promote and encourages the rapid sharing of data that are generated. Such sharing is essential for progress toward the goals of the project, to expedite research in other areas and to avoid unnecessary duplication. Although it should be the policy of any project to maximize data available to the scientific community, it is also necessary to give investigators time to verify the accuracy of their data and to gain some scientific advantage from the effort they have invested. Timely data provision by the project scientists is crucial to the achievement of the project's goals. ***A 6-month period from the time the data or materials are generated to the time they are made available publicly should be a reasonable timeframe in almost all cases. However, more rapid sharing is encouraged. Operational data requiring no (or little) quality control should be available within a few months. – This needs to be more specific.***

Data Publication—The IIMO assisted by a project Publications Committee (PC) will assign a label "PROJECT X Publication" and attribute a Digital Object Identifier (DOI) to publications whose data (1) have been evaluated by a DEU; (2) are available through a WDC; and (3) which acknowledge the project in an appropriate manner.

Data Standards— Any project research must adhere to standards and protocols that have been accepted by the project SSC. International standards and protocols have been (and are still being) developed to (a) increase availability, access, integration, and sharing of digital geographic information; (b) permit inter-operability of geospatially enabled computer systems; (c) contribute to a unified approach to addressing global ecological and humanitarian problems; and (d) ease the establishment of geospatial infrastructures on local, regional and global levels. These standards include the ISO 19100/TC 211 family (standardization of digital geographic information[9]), the FGDC-STD-001-1998 (digital geospatial meta-data[10]), the ANSI/NISO Z39.50 protocol (a unifying interface that allows diverse information systems for coexistence while maintaining a consistent user interface[11]), and the AAP-CNRI DOI (digital object identifier for communication between communities[12]), among many others.

[9] International Organization for Standardization. ISO/TC211 Geographic Information/Geomatics. (2003).
[10] Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata (version 2.0) FGDC-STD-001-1998, www.fgdc.gov/metadata/contstan.html, 1998.
[11] American National Standard Information/National Information Standards Organization, A. N. S. I. N. I. S. American National Standard Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection, www.cni.org/pub/NISO/docs/Z39.50-brochure/50.brochure.part01.html, 1988.
[12] International DOI Foundation. The Digital Object Identifier system, www.doi.org, 2002.

Widespread use of the Internet has revolutionized data distribution and availability in the past few years. For the maximum benefit of the scientific community, WDCs and other archives keep abreast of changes in computing and data distribution and storage and stay at the forefront of new approaches. Any software used by project scientists for data handling and products should be platform-independent.[13]

Data Management Funding—The SSC may require all applicants for affiliation to a project and who expect to generate significant amounts of data or materials, to describe in their application how and when they plan to make such data and materials available to the community. Grant solicitations should/may specify this requirement. These plans in each application will be reviewed in the process of peer review to ensure they are reasonable and in conformity with program philosophy. If a grant is made, the applicant's data sharing plans will become a condition of the award and compliance will be reviewed before continued funding is provided. Scientists and funding agencies should ensure that sufficient funds are included in project proposals to support data management.

Other Useful Background Documents

ANZLIC Policy Statement on Spatial Data Management – April 1999 – awww.anzlic.org.au/policy/data_mgt.htm

International Argo Data Management, User's Manual – July 2002 – www.coriolis.eu.org/coriolis/cdc/argo/argo-dm-user-manual.pdf

BENEFIT Data Policy – March 2002 – www.benefit.org.na/datamgmt/policy.html

Canada-JGOFS Data Submission Policy – February 1997 – www.meds-sdmm.dfo-mpo.gc.ca/jgofs/Docs/management/policy.html

CoOP Data Policy – January 2002 – www.skio.peachnet.edu/coop/datapol2.html

Digital Object Identifier Handbook, Metadata – December 2002 – www.doi.org/handbook_2000/metadata.html

Dublin Core Metadata Initiative, Guidelines for implementing Dublin Core in XML – September 2002 -- dublincore.org/documents/2002/09/09/dc-xml-guidelines/

GLOBEC Data Policy – February 2001 – www.pml.ac.uk/globec/Data/DataPolicy_June21st.PDF

Committee on Data for Science and Technology (CODATA), scientific access to data and information – February 2002 – www.codata.org/data_access/index.html

Joint Committee on Antarctic Data Management (JCADM) documents and reports – August 2002 – www.jcadm.scar.org/docs1.html

MAST: A Guideline for Project Data Management – March 1997 – www.sea-search.net/guidelines-practices/welcome.html

NSF-Margins Data Policy – January 2002 – www.ldeo.columbia.edu/margins/MARGINSnet4/DataPolicy.html

OSDM Commonwealth spatial dataset access policy – February 2003 – www.osdm.gov.au/osdm/data_acc_policy.html

US-GCRIO Policy Statements on Data Management for Global Change Research (US-GCRIO) – July 1991 – www.gcrio.org/USGCRP/DataPolicy.html

[13] WOCE Synthesis and Modeling Working Group. WOCE observations, data analysis and data products, www.cms.udel.edu/woce/wocedocs/aims/aims3.htm, 1997.

US-JGOFS Data Policy – ioc.unesco.org/oceanteacher/resourcekit/Module1/DataPolicy/u_s_jgofs.htm
WOCE Data Policy and Practices 1995 – December 2002 – www.cms.udel.edu/woce/WOCEDOCS/datapol.htm

Acronyms Used

CEOS	Committee on Earth Observation Satellite – www.ceos.org/
DEU	Project Data Evaluation Unit
DMC	Project Data Management Committee
DMO	Project Data Management Office (national or regional level)
DOE-NIH	Department of Energy - National Institutes of Health
DOI	Digital Object Identifier – www.doi.org/
GCMD	NASA's Global Change Master Directory – gcmd.nasa.gov/
GEOHAB	Global Ecology and Oceanography of Harmful Algal Blooms programme
GSDI	Geospatial Data Infrastructure
ICES	International Council for the Exploration of the Sea – www.ices.dk/
ICSU	International Council for Science – www.icsu.org/
IDN	CEOS' International Directory Network – idn.ceos.org/
IDMU	International Data Management Unit
IGBP	International Geosphere-Biosphere Programme – www.igbp.kva.se/
IIMO	Project International Information Management Office
IOC	UNESCO's Intergovernmental Oceanographic Commission – www.ioc.unesco.org
IPO	International Project Office
JGOFS	Joint Global Ocean Flux Study
LOICZ	Land-Ocean Interaction in the Coastal Zone project
NCP	Project National Contact Points
NODC	National Oceanographic Data Centre – ioc.unesco.org/iode/contents.php?id=97
PC	Project Publications Committee
PI	Principal Investigator
SCOR	Scientific Committee on Oceanic Research – www.jhu.edu/scor/
SOLAS	Surface Ocean – Lower Atmosphere Study
SSC	Scientific Steering Committee
WDC	ICSU's World Data Center System – www.ngdc.noaa.gov/wdc/wdcmain.html
WIPO	World Intellectual Property Organization – www.wipo.org/