

## ***Suggested Statements for Data & Information Management in future oceanographic programmes***

initially prepared by Bernard AVRIL and Nicolas DITTERT (20 February 2003)  
slightly revised by Bernard AVRIL (October 2003)

a.- Understanding that at the international level, the project Scientific Steering Committee (SSC) makes **data management recommendations**, that assist the **implementation and enforcement at the national level of a data management policy together with the relevant (national and international) funding agencies**,

b.- Considering that the International Council for Science (ICSU) recommends as a general policy, the **fundamental principle of full and open exchange of data and information for scientific and educational purposes**. Scientific data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. Datasets are regarded as a valuable resource in their own right,

c.- Considering that the minimum requirements are that **scientific data** are **complete** (for global integration), **accurate** (through quality control and analysis) and **authentic (trustworthy); timely released; identifiable, documented** thanks to associated metadata records (*i.e.*, “description of data and including method and references”); **ethically retrievable** when needed and **properly secured**, and **preserved** on the long-term,

d.- Noting that raw data, which are undocumented and uncertain in their quality have the least value, whereas **complete, accurate, well documented and organised data, as used in publications, have the highest value**,

→ Recommends **full and open access to well documented and organised data and required information** for all parties involved in the project.

e.- Noting that the **ultimate success** in data retrieval and exchange, and accompanying advancements in knowledge rely on the **scientists’ participation**, which does not consistently occur unless there are **appropriate incentives** to do so.

f.- Understanding that a **data management plan** improve the efforts dedicated to scientific analysis and new discoveries, by easing those devoted to exploration and manipulation of datasets.

g.- Considering that the project also depends on **historical records and knowledge**.

h.- Considering that during the project, it is expected the **development of new concepts and visions**, together with a **great expansion in marine, biogeochemical and ecological data flow of very large, distributed, heterogeneous datasets**, associated with advances in technologies and scientific concepts.

→ Recommends **a major effort for the project data and metadata management, to be internationally coordinated**, based on a full, coherent data management plan, along with first, a attractive and rewarding system sustaining the voluntary participation of scientists, and second, some fair enforcement procedures, including a data management policy, to be prepared and implemented in coordination with the funding agencies

→ Recommends **a multiplicity of approaches within a semi-distributed, scalable and flexible data management system** in order to address the increasingly important and complex issue of data and knowledge management, taking into consideration the multiplicity of the cultures, international work experiences, national particularities, and constrained relationships that scientists are willing or required to work with.

→ Recommends the **establishment and continuous support of an International Information Management Office (IIMO) and of a Data Management Committee (DMC)** designed to assist the SSC and to coordinate in close liaison with the International Project Office (IPO) all information and data management aspects for the entire project duration (evaluation, guidance, support of the various partners in data management; evaluation of the data flows and usages), and to 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 and DMC 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 the data flows, especially for future observations from new sensors, new satellites, new platforms (autonomous underwater or remotely operated vehicles, inexpensive, low-maintenance monitoring systems and sophisticated buoys), for continuous measurements, global survey, ships of opportunity operations, video recordings, and for systems of data delivery in near real-time and delayed modes or for on-time, push data and information delivery; for model output dissemination, etc. The IIMO and DMC 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 follow developments and review existing governmental, commercial, and legal constraints on data access and intellectual property issues (including published articles, CD-ROMs) for purposes of scientific research.

i.- Recognizing that the project core science is specific and different in its goals and strategies from the already existing initiatives and other core projects sponsored by IGBP and SCOR.

→ Recommends the **establishment and the regular revision of a list of project core parameters** (with related methodological standards and appropriate metadata, as required), which related closely to the project core science, in coordination with all relevant project partners and other external initiatives.

j.- Recognizing that to insure the highest possible quality in the project datasets, a system similar to the peer reviewing for the publications should be adopted for the datasets and associated metadata

→ Recommends the **establishment and basic, continuous support of several Data Evaluation Units (DEUs)**, reporting (activity with performance evaluation) to the SSC and the IIMO. Their primary missions are first, to quality control independently and fairly, to check against historical records and to validate the contributed datasets and associated metadata relevant to a specific group of the project core parameters; second, to deliver appropriately official project-labelled certificates for the datasets and to publicly acknowledge the scientists' contributions to the project, and third, to create project-labelled data collections. Each DEU is flexible in its working structure and is community-oriented (more than the national agencies already existing, such as NODCs), and is led by a voluntary 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 both as data producers and as data evaluators or as builders of specific data collections. The DEU use the delivered datasets and associated metadata only for the tasks defined by the SSC and IIMO.

→ Recommends that **the IIMO, DMC and DEUs provide all project scientists with support, recommendations, guidelines and priorities regarding information and data management plans**. They coordinate the establishment of a specific metadata portal for the project or programme with the help of GCMD – CEOS / IDN. They promote the project data policy and assist in the development of national project data management efforts (possible national Data Management Offices, DMO) and the cooperation with national data centres (NDCs).

→ Recommends that **the IIMO**, assisted by an ad-hoc Publications Committee (PC), **attributes a project label and consecutive number for the relevant peer-reviewed publications** when they are submitted to a science journal, and provided that the related datasets and associated metadata are delivered within the project data management system, and project-certified by a DEU, identifiable through a unique Digital Object Identifier (DOI), and the publication acknowledged the project in an appropriate manner.

→ Recommends that **the shortest data flow from the scientists to the most adequate ICSU's World Data Centre (WDC) through the DEUs be promoted** (especially for the countries with a less advanced, developed infrastructure). The WDCs represent the best international framework for the long-term preservation and the worldwide, continuous dissemination of all data. → **Figure 1**

→ Recommends that **data and information flow be initiated as soon as possible and maintained until the project's end**, in order to insure a timely, continuous delivery of the datasets to the community, and also to increase the quality of those datasets and subsequently of the fieldwork strategy and the synthesis themselves.

→ Recommends that **a preliminary report be established as soon as possible after the fieldwork or the experiment**, that includes preliminary, basic information related to the datasets acquired or the experiences conducted, such as location and timing of stations; sampling strategy; inventory of all parameters acquired; time frame and specifications for data delivery and for restricted and open data accessibilities, etc.

→ Recommends that the **data and metadata be delivered to the relevant DEUs 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 datasets and information are available as soon as the fieldwork or experience ends (hydrology, meteorology, etc.), or soon after some basic data quality control (dissolved nutrients, etc.). The datasets to be delivered later (from 3-month to 2-year delays) and those not acquired in accordance to the initial cruise or experience plans, are also identified.

→ Recommends that **the project-labelled datasets and related metadata be both placed on-line** (provisionally, with a password-controlled access until the end of the 2-year period) **and published as project data products as soon as possible**, either by the DEU or another entity (IIMO, WDC...), after full delivery and quality check, and as appropriate or required (with a clear edition date or version number, to avoid confusion), even if they are later revised or reworked in a more convenient, integrated way (*e.g.*, in a “data collection”), in order to increase the work flexibility and the visibility of each achievement.

k.- acknowledging that **a new or specific science field relevant to the project is, exceptionally, not properly covered by any DEU**.

→ Recommends that the **datasets relevant to the project science but not covered by the expertise of any DEU be transmitted with their associated metadata** preferably directly to **the IIMO** and possibly to a national institution (preferably, national DMO), in agreement with the SSC or IIMO.

l.- recognizing that “**education**” is one of the most useful strategies to ensure data delivery and metadata generation

→ Recommends that **the DEUs and NDCs be fully involved in the project, including participations to 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 / data analysis / data visualization tools as needed.

m.- noting that “**full recognition and acknowledgement**” of the contributing scientists is another of the most useful strategies to ensure data delivery and metadata generation

→ Recommends that **all “data publishing” and “data citation index” initiatives and mechanisms be promoted.**

n.- Recognising that the **project is a part of a larger science system,**

→ Recommends that **each national and international project efforts fully support and promote the project data management system above-described** and that **those above-mentioned project recommendations be fully endorsed and implemented nationally and internationally** by all relevant funding agencies, research institutions and international bodies.

### ***Final Remarks***

The scientists should clearly benefit of the project data management system, because:

- They receive **privileged and rapid access** facilities to new datasets, through the project data management system
- They can **increase the added values** of their datasets, either through new scientific collaborations / co-authorship in publications.
- They access faster and more efficiently the datasets and hidden patterns thanks to **data management / data analysis / data visualization tools**, as developed by the project data managers / data users.
- Their recognition and professional status are promoted by their participations to the project, **as data producers**, when their datasets and related peer-reviewed publications are officially project-labelled and designated with an unique Digital Object Identifier (DOI), or **as data evaluators** (within the DEUs) or as **data collections builders**.

## ***Technical Annex – Conventions and terms***

The data management design refers to the guidelines of “Good scientific practice in research”<sup>1,2</sup>, the WIPO copyright treaty<sup>3</sup>, and the DOE-NIH Guidelines for Sharing Data and Resources<sup>4</sup>. Accordingly, data generated by the project are a substantial resource, which enable later verification of scientific interpretation and conclusions. They may also be the starting point for further studies<sup>5,6</sup>.

### *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., biology, chemistry, geology). In contrast to previous programs (e.g., WOCE, JGOFS), the project comprises a highly complex data heterogeneity.

Data producer (in a broader sense) is the scientist (Principal Investigator, PI). The PI signs responsible for a data set (i.e. the smallest data entity). Usually, a PI is an individual. In exceptional cases a PI can be an institution.

### *Data flux*

Anytime data are produced under the project umbrella, the PI shall follow the Information Flux Model (Fig. 1). This routine shall insure that any scientist benefits to a maximum from each other with a minimum of effort - besides other positive effects (e.g., data availability).

A project Data Management Committee (DMC) shall track and update any data activity. Data producers are indentured to submit any meta-information and analytical data produced under the project umbrella and to announce the current data status during the entire processing period to the DMC.

Data evaluation shall be carried out by Data Evaluation Units (DEUs). DEU describes a functionality rather than an institution and adopts the responsibility of quality assessment (completeness and trustworthiness of data). DEUs can be individuals, research centers, academic departments, or industrial laboratories standing out by their excellence in a particular research topic. They are recruited by expertise through regular calls for proposals.

Data storage and public access shall be performed by World Data Centers (WDCs). WDCs are institutions that are bound to the rules and are supervised by the Panel on World Data Centers and the International Council for Science<sup>7</sup>.

Scientists may be committed to national data management regulations, too. However, funding received through the project structure indentures the PI to follow the project data policy. However, subordinate data activities (e.g. through national data centers) are welcome. In any case, for the reason of data consistency WDC are preferred to other data centers.

### *Data sharing*

The project promotes and encourages the rapid sharing of data that are generated. Such sharing is essential for progress toward the goals of the project – and to expedite research in other areas - and to avoid unnecessary duplication. Although it shall be the policy of the project to maximize outreach 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. Furthermore, intellectual property protection may be needed for some of the data. Timely data provision by Principal Investigators (PI) involved in the field program is crucial to the achievement of the project goals: Consensus shall be the concept that a 6-month period from the time the data or materials are generated to the time they are made available publicly is a reasonable mean value in almost all cases. However, more rapid sharing is encouraged.

### *Data publication*

The IIMO assisted by a project Publications Committee (PC) shall award a label “Project 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) who acknowledge the project in an appropriate manner.

#### *Data standards*

Any unit shall be committed to implement standards and protocols that have been evaluated by a Project International Information Management Office (IIMO). An entire family of international standards and protocols has been (and is 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; (d) ease the establishment of geospatial infrastructures on local, regional and global levels. These standards comprise the ISO 19100/TC 211 family (standardization of digital geographic information<sup>8</sup>), the FGDC-STD-001-1998 (digital geospatial meta-data<sup>9</sup>), the ANSI/NISO Z39.50 protocol (a unifying interface that allows diverse information systems for coexistence while maintaining a consistent user interface<sup>10</sup>), the AAP-CNRI DOI (digital object identifier for communication between communities<sup>11</sup>), among many others. Widespread use of the Internet has revolutionized data distribution and availability in the past few years. For the maximum benefit of the scientific community data providers (WDC and other archives) shall keep abreast of changes in computing and data distribution and storage and stay at the forefront of new approaches. Software provided publicly for data handling and products should be platform-independent<sup>12</sup>.

#### *Data management funding*

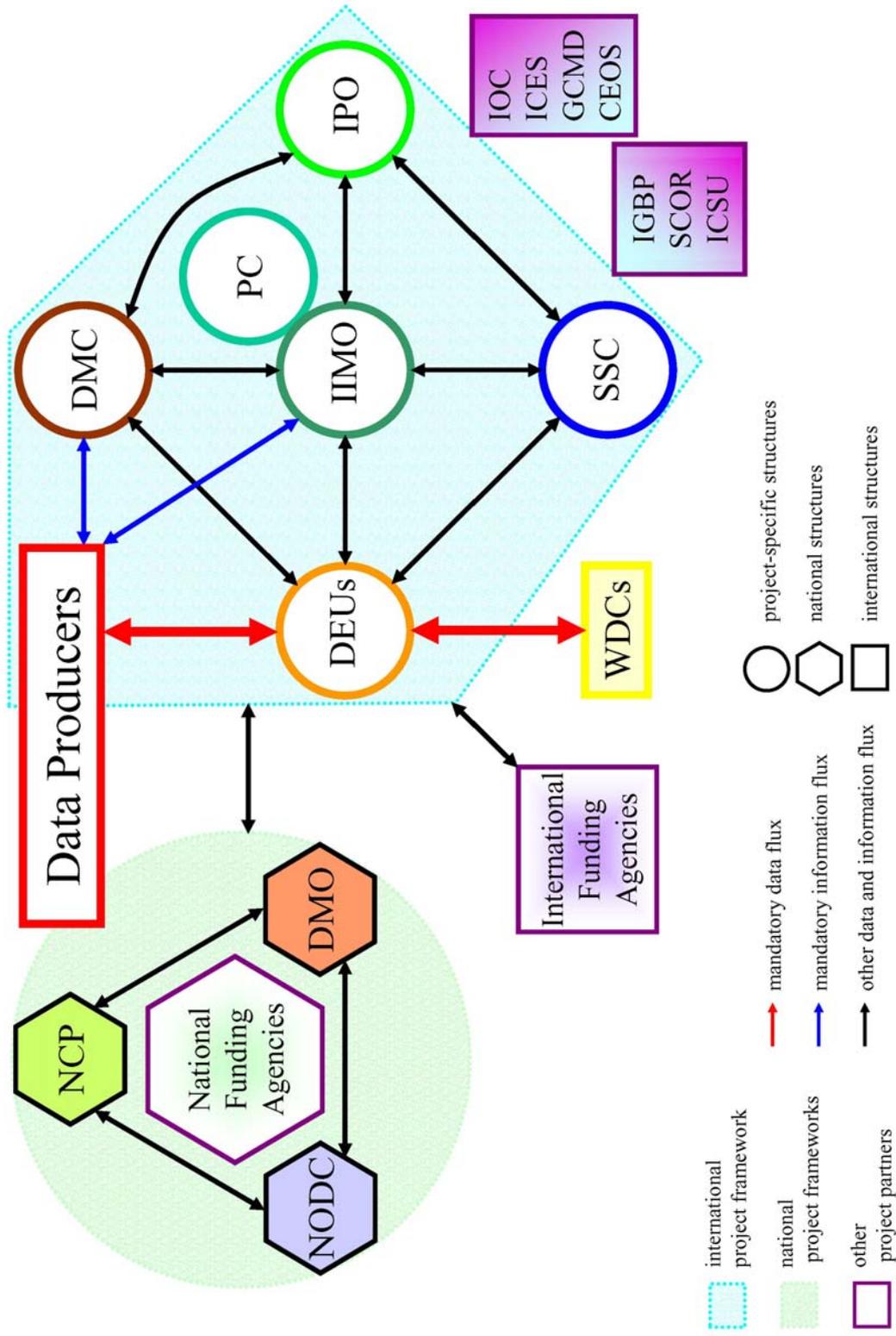
SSC shall decide to require all applicants expecting to generate significant amounts of data or materials to describe in their application how and when they plan to make such data and materials available to the community. Grant solicitations shall specify this requirement. These plans in each application will be reviewed in the course of peer review and by staff to assure they are reasonable and in conformity with program philosophy. If a grant is made, the applicant's sharing plans shall become a condition of the award and compliance will be reviewed before continuation funding is provided. Progress reports shall be asked to address the issue. The project recommends that scientists and funding agencies ensure that sufficient funds are included in project proposals to support data management.

## ***References***

1. ICSU/CODATA. Access to databases: A set of principles for science in the Internet era, [www.icsu.org/Library/Central/Statem/access.html](http://www.icsu.org/Library/Central/Statem/access.html), 2000.
2. European Science Foundation. Good scientific practice in research and scholarship, [www.esf.org/medias/ESP10.pdf](http://www.esf.org/medias/ESP10.pdf), 2000.
3. World Intellectual Property Organization. WIPO copyright treaty, [www.wipo.org/eng/diplconf/distrib/94dc.htm](http://www.wipo.org/eng/diplconf/distrib/94dc.htm), 1996.
4. 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](http://www.ornl.gov/hgmis/publicat/hgn/v4n5/04share.html), 1993.
5. JOI/U.S. Science Support Program. Acknowledgement and Keyword Policy, [www.joi-odp.org/USSSP/acknowledge.htm](http://www.joi-odp.org/USSSP/acknowledge.htm), 2002.
6. PAGES. PAGES Data Policy, [www.pages.unibe.ch/data/policies.html](http://www.pages.unibe.ch/data/policies.html), 2003.
7. ICSU Panel on World Data Centers. Principles and Responsibilities of ICSU World Data Centers, [www.ngdc.noaa.gov/wdc/guide/gdsystema.html](http://www.ngdc.noaa.gov/wdc/guide/gdsystema.html), 2002.
8. International Organization for Standardization. ISO/TC211 Geographic Information/Geomatics. (2003).
9. Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata (version 2.0) FGDC-STD-001-1998, [www.fgdc.gov/metadata/contstan.html](http://www.fgdc.gov/metadata/contstan.html), 1998.
10. 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](http://www.cni.org/pub/NISO/docs/Z39.50-brochure/50.brochure.part01.html), 1988.
11. International DOI Foundation. The Digital Object Identifier system, [www.doi.org](http://www.doi.org), 2002.
12. WOCE Synthesis and Modeling Working Group. WOCE observations, data analysis and data products, [www.cms.udel.edu/woce/wocedocs/aims/aims3.htm](http://www.cms.udel.edu/woce/wocedocs/aims/aims3.htm), 1997.

## ***Examples of Further Readings***

- ANZLIC Policy Statement on Spatial Data Management – April 1999 – [www.anzlic.org.au/policy/data\\_mgt.htm](http://www.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](http://www.coriolis.eu.org/coriolis/cdc/argo/argo-dm-user-manual.pdf)
- BENEFIT Data Policy – March 2002 – [www.benefit.org.na/datamgmt/policy.html](http://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](http://www.meds-sdmm.dfo-mpo.gc.ca/jgofs/Docs/management/policy.html)
- CoOP Data Policy – January 2002 – [www.skiio.peachnet.edu/coop/datapol2.html](http://www.skiio.peachnet.edu/coop/datapol2.html)
- Digital Object Identifier Handbook, Metadata – December 2002 – [www.doi.org/handbook\\_2000/metadata.html](http://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/](http://dublincore.org/documents/2002/09/09/dc-xml-guidelines/)
- GLOBEC Data Policy – February 2001 – [www.pml.ac.uk/globec/Data/DataPolicy\\_June21st.PDF](http://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](http://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](http://www.jcadm.scar.org/docs1.html)
- MAST: A Guideline for Project Data Management – March 1997 – [www.sea-search.net/guidelines-practices/welcome.html](http://www.sea-search.net/guidelines-practices/welcome.html)
- NSF-Margins Data Policy – January 2002 – [www.ldeo.columbia.edu/margins/MARGINSnet4/DataPolicy.html](http://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](http://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](http://www.gcrio.org/USGCRP/DataPolicy.html)
- US-JGOFS Data Policy – [ioc.unesco.org/oceanteacher/resourcekit/Module1/DataPolicy/u\\_s\\_jgofs.htm](http://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](http://www.cms.udel.edu/woce/WOCEDOCS/datapol.htm)



**Figure 1.** Schematic diagram representing the interaction between the project partners and external parties, and the main fluxes of data and information.

## ***Acronyms Used***

CEOS – Committee on Earth Observation Satellite – [www.ceos.org/](http://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/](http://www.doi.org/)  
GCMD – NASA’s Global Change Master Directory – [gcmd.nasa.gov/](http://gcmd.nasa.gov/)  
GSDI – Spatial Data Infrastructure  
ICES – International Council for the Exploration of the Sea – [www.ices.dk/](http://www.ices.dk/)  
ICSU – International Council for Science – [www.icsu.org/](http://www.icsu.org/)  
IDN – CEOS’ International Directory Network – [idn.ceos.org/](http://idn.ceos.org/)  
IDMU – International Data Management Unit  
IGBP – International Geosphere Biosphere Programme – [www.igbp.kva.se/](http://www.igbp.kva.se/)  
IIMO – project International Information Management Office  
IOC – UNESCO’s Intergovernmental Oceanographic Commission – [ioc.unesco.org](http://ioc.unesco.org)  
IPO – project International Project Office  
NCP – project National Contact Points  
NODC – National Oceanographic Data Centre – [ioc.unesco.org/iode/contents.php?id=97](http://ioc.unesco.org/iode/contents.php?id=97)  
OCEANS – Ocean Biogeochemistry and Ecosystems Analysis – [www.igbp.kva.se/obe/](http://www.igbp.kva.se/obe/)  
(now *Integrated Marine Biogeochemistry and Ecosystem Research*, IMBER, June 2003)  
PC – Publications Committee  
PI – Principal Investigator  
PS – project Publications Committee  
SCOR – Scientific Council of Oceanic Research – [www.jhu.edu/~scor/](http://www.jhu.edu/~scor/)  
SSC – project Scientific Steering Committee  
WDC – ICSU’s World Data Center System – [www.ngdc.noaa.gov/wdc/wdcmain.html](http://www.ngdc.noaa.gov/wdc/wdcmain.html)  
WIPO – World Intellectual Property Organization – [www.wipo.org/](http://www.wipo.org/)

***NB.*** This draft document was initially prepared by Bernard Avril and Nicolas Dittert, 21 Feb. 2003.