Remarks on Data and Information Management for Science Projects [*August 2003*]

Prepared by Bernard Avril

I. - Advances in oceanographic and atmospheric sciences are mostly based on the timely acquisition of new data either from field studies (*e.g.*, process study-oriented cruises, underway transects, time-series stations, buoys, floats, remote sensing, ships of opportunity) or from model outputs (diagnostic or prognostic). Due to technological development, the amounts and flux of information relevant to those activities are expected to drastically increase in the future (as a general rule, it is admitted that every 2.5 years, the data amount and data flux are multiplied at least by 5), along with the performances of analytical methods and sensors. On the other hand, the relevance, timeliness, accuracy, accessibility and added-value of the outputs from the research on global change and integrated earth system also should also increase as a response to socio-economical challenges and to the requirements of the funding agencies, decision-makers and general audience.

II. - Consequently, in order to optimise especially at the international level the efforts for data acquisition and for the subsequent data quality control, formatting, documentation, access and utilisation, it is crucial to have a defined, organised, yet scalable **data and information management (D&IM) plan** at the beginning of the project and some strategic collaborations, when relevant. The project should thus adhere to a "**vision**" for its D&IM plan, such as the following one:

> *"Provided that all copyright and acknowledgement issues are properly addressed, an identifiable (*authorised*) user of the project D&IM framework has access to relevant, timely, accurate and quality-controlled data and metadata, in the most appropriate, consistent and unambiguous manner, thanks to adequate (*credible*) knowledge interoperability."*

Strictly speaking, D&IM is the process of structuring and controlling the data (*information*) definition to ensure its consistent and optimal use across the project. A workable **data definition** (*i.e.*, textual description, including unit, format and range; accurately maintained and updated sufficiently often to be of maximal value to the project) should be explicitly and commonly agreed upon by the data managers and scientists, also taking advantage of external standardisation efforts (*e.g.*, ISO, W3C, IOC/ICES). In this context and even if the working systems are changing in both technical and operational ways, a science project creates **knowledge**, more than only information, thanks to **metadata** (*i.e.*, data about the data) often

readily available at a very low "cost" during the data acquisition itself. Therefore, **knowledge is an asset with a commonly understood meaning and with added value**.

III. - A data and information management plan makes it possible to create and disseminate this value-added asset (the scientific knowledge) during the project and beyond, based on the **knowledge interoperability**, (*i.e.*, ability to share data (*knowledge*) between systems, combined with the ability to use (*understand*) the data (*knowledge*) so shared), especially in a **semi-distributed and flexible, yet coordinated and robust system** of data and information management. As a basic principle, the project **functioning** (objectives, fieldwork elaboration and expectations, data collection and data use) should be sharply thought and well defined, so that the adequate D&IM **structure** be formed accordingly for its present and future needs, keeping in mind that a "fit for purpose" approach is preferable that an "intellectual perfection" approach.

IV. - In order for the **D&IM efforts to be credible** (*e.g.*, adequate, integrated and focused), and ultimately to do a better science, overall and in each national contribution, a "*D&IM culture*" and a "*spirit of cooperation*" should be developed between the project scientists, project managers, and data & information managers [*see* ***Note A***], and the project should adapt its strategy accordingly, through some pro-active attitudes / actions in the project community, such as:

- involvement of senior data managers in the project design and planning phases, for the improvement of its originality, implementation and fieldwork strategy, and in other relevant activities or meetings of the project,
- organisation of a kick-off D&IM workshop and some later capacity building / knowledge transfer sessions so that all involved data managers agree with and support the D&IM plan, benefit from the expertise of others data managers, and report back to their institution,
- organisation of regular data and model workshops, so that modellers, observationalists and data managers take full advantage of and review / optimize the project D&IM plan,
- continuous reporting to the project steering committee and parent bodies on D&IM issues and activities, especially on operational applications and achievements,

- elaboration and promotion of a data policy (guidelines, procedures or recommendations for data acquisition, quality control, documentation, assembly, evaluation, analysis, visualisation, exchange and archival, with freeware, public-domain software), of a limited set of data

definitions, including the description of parameters related to the core science of the project (*core parameters*) and of the expected output specifications (what platforms, sensors, data generators; expected data amount and flow),

- establishment of a moderated project mailing list (for an active, open forum), of a project newsletter (with strong input from the project community), and of a knowledge resource centre (e.g., linked to the IPO and IIMO) for the project community and interested parties (especially, documenting activities and outputs), of a comprehensive "catalogue of project metadata records" (possibly, through a long-term collaboration with GCMD),
- establishment of a direct record (metadata and data) submission feature for the involved parties, with a central control from a moderator body (*e.g.*, IPO and IIMO) before the record is actually included in the project (meta)database and is placed on line,

- involvement of data managers in the data acquisition during the fieldwork (*e.g.*, support for the collection of the basic parameters and the data quality control and documentation),
- promotion of an unique data acquisition and storage, wherever possible, to avoid duplicates and multiple versions (*e.g.*, with different, improper naming or outdated quality control date stamp), to improve the use of communication capability and to favour the transfer of those data really needed within a specific timeframe / objective,
- promotion of *data citations*, especially in peer-reviewed journals promoting / requesting data and metadata submission, and the attribution of a Digital Object Identifier (DOI), for *data evaluations* and for the development of *data collections*,

- avoidance of any duplication work at all stages related the D&IM plan, thanks to interdisciplinarity, cooperation, intercomparison exercises (models and analytical measurement methods), timely publications, or data mining, archaeology and rescue,
- both at the individual and institutional levels, exchange of experiences, expertise, ideas regarding relevant data and information management plans (*e.g.*, "independent" data repository, such as NDC or WDC) and other related initiatives (GCMD, CEOS, IDN) or projects (in ESSP), to launch common activities and produce commons outputs, within a larger science community framework, according to the project, specific historical and multi-national contexts,
- monitoring and recommendation for the implementation of new D&IM technologies / practices, promotion of new, original initiatives (even outsourcing opportunities), so that ideas

and concepts emerge rapidly and develop if valuable, yet without compromising the core project science and achievements,

- development of a project "*corporate*" image and enhancement of the overall visibility of the project community (especially, at major events of relevance) and of the outside recognition for the project deliverables, especially through some outstanding, "high-profile" achievements,
- enhancement of the individual scientists' recognition and professional status, as data **originators**, when peer-reviewed publications and related datasets (*data citations*) are labelled by the project, designated with an unique DOI in the literature, as **data evaluators** (in DEU-like bodies) or **data collections builders** (either on spontaneous or organised basis),

- continuous advocacy for the proper funding of the project data and information management plan, since D&IM only works optimally when the top management and funding agencies of the various elements of the data management system commit the adequate resources to undertake the project D&IM activities, and maintain its enhanced operational effectiveness,
- improvement of the usage of the national and international funding, by optimising the operability and outputs based on the available resources (including the already accumulated scientific knowledge and technical know-how) and the existing infrastructures, including the flexible, open Internet (*e.g.*, searchable gateways and organised portals),
- monitoring of some performance criteria for the D&IM system, such as the tracking of data use and of on-line, available dataset citations, of metadata records, of portal web hits, of published works, the continuous compilation of user feedback; and evaluation of an overall "cost / benefit" ratio in a medium-term perspective,
- address of the issues beyond the project science, relevant to the funding agencies, decision-makers and general audience and also to technological / business applications,

Note A: The SCAR-COMNAP Joint Committee on Antarctic Data Management, which included representatives from 15 contributing countries (*N.B.: JGOFS DMTT represented only 8 countries*) has already established that, in order to improve data availability and operability, an adequate, credible support is needed and a ethically rewarding system is needed (with bottom-up, incentives, proactive approaches through education, tool development, added value to datasets and voluntary behaviour) and some fair enforcement is needed (with top-down, mandatory behaviour, enforcement of data availability and exchange implementation mechanisms).