Limited Distribution

IOC/IODE-XVI/12
Paris, 30 September 2000
Original: English

INTERGOVERNMENTAL OCEANOGRAPHIC COMMISSION
(of UNESCO)

Sixteenth Session of the IOC Committee on International Oceanographic Data and Information Exchange (IODE), Lisbon, Portugal, 30 October – 9 November 2000

Data Formats and Data Unification

Prepared by Nickolay Mikhailov, Evgeny Viazilov, RIHMI-WDC/Russian NODC, Russian Federation

INTRODUCTION

Data formatting is an important component of data Management technologies, which provides the structural basis for data sets ' creation as well as for software interfaces between data sets and procedures of data processing, analysis, visualization and dissemination. Data formatting is of particular importance for data exchange between subject - oriented technologies, between data centres, between a data centre and users.

For a long time the Group of Experts on Technical Aspects of Data Exchange (IODE GETADE) has been responsible for providing the standardization of ocean data exchange. These activities resulted in development of the General Format (GF3) formatting system and, the GF3 subsets for various disciplines (oceanography, currents, geology geophysics) related to marine environment. Subsequently GF-3 was upgraded to meet modern technical requirements and is offered for the International Oceanographic Data Exchange (IODE) as the GETADE format. At the same time many centres and organizations developed data formats for their own needs.

In 1992 GETADE investigated formats, which in that time were in use. But recent years radical changes took place in the field of information technologies. Therefore, examination of the status, use, and utility of currently existing formats should play a positive role for making the decisions concerning the further activity in the field of data formatting and development of the IODE End to End Data Management (IODE E2EDM) system.

To prepare the current presentation the analysis of 21 data formats (Annex 1) used for data management both in the international projects and at a national level was carried out. When analyzing the formats much attention was given to consideration of the following characteristics:
- data type described by the format;
- composition and content of metadata described by the format;
- codes used by the format;
- peculiarities of data structuring.

1. Data formats inventorying and analysis

First it is necessary to note, that the list of the investigated formats includes formats intended for use at various stages of data management: from data collection (TOGA, WOCE formats), to data processing and analysis (OceanAtlas and OceanDataViwer/ODV formats). The formats use by NODC of Argentina, Spain and Russia were considered as examples of national formats.

The analysis of formats allowed the following Format features to be allocated.

1.1.Individual methods of data structure description were used in development of most of the considered formats. Actually, in every large international project on the World ocean study and in the most advanced

IODE NODC/RNODC's the original data formats are applied. In this connection, it is possible to assume, that the amount of formats actively used is not less than 30.

Depending on methods of data structure description the formats can be divided into three groups:

- with a fixed set of parameters: formats of the International Council of Exploration of the Sea (ICES), NODC of the USA (SD2) and also formats of NODC of Argentina, Spain and partly Russia;
- with a variable set of parameters and their own logic of data structure description: formats of the project "Global temperature and salinity data pilot project" (GTSPP) developed by MEDS Canada, of the project "Data Archaeology and Rescue" (WA' 98) developed by WDC-A and the new version of NODC of the USA (P-3);
- with a variable set of parameters and GF-3/GETADE logic of data structure description: the last version of ICES format (ICES Blueprint'86), formats of JGOFS and MEDAR/MEDATLAS projects.

The two last groups of formats are the so-called self-defining formats, which ensure description of both the list of parameters in data records and their specifications such as length of fields, accuracy of numerical values presentation, scaling. The such formats as GTSPP, WA'98 and P-3 formats use the approach when data structure is described through definition directly in the data record.

The formats constructed on GF3/GETADE logic apply special records of parameters definition placed in the data file (usual in the beginning of a file).

In recent years a stable tendency is observed to specialize the formats for supporting different stages of data management process. However most of the considered formats are focused on support of collection and exchange of observation oceanographic data. At the same time characteristic features of such formats as WA' 98 and P-3 are connected with the aim of their application, which is to integrate data from various sources Such formats as OceanAtlas, ODV and a table format used to export data to Ocean PC (CSV-1) should also be noted. They belong to the category of formats used for statistical processing and visualization of data.

Most of formats provide data representation in ASCII codes with a fixed length of lines in a logical record (usually 80 bytes). Exception is made for one of GTSPP format versions, providing binary presentation of parameter values, and for WMO format (BUFR) intended to record parameter values in the form of bit lines of variable length.

1.3. The important feature of formats is composition and content of metadata placed in the data file. The highest level of metadata is supported by GF-3 formatting system, which has the advantage of providing possibilities for metadata unification in both composition and method of location in data file. But unfortunately GF-3 did not gain wide acceptance, and due to this fact every format developer solves the problem of metadata formatting individually.

The analysed formats use different levels of metadata presentation: from the elementary reference to cruise number (Ocean Atlas, ODV) to detailed information about data QC and data management history (GTSPP, MEDATLAS). In general the characteristic feature of the recently appeared formats is that they support opportunities of wide metadata representation in comparison with formats developed in previous years.

1.4. Only a small part of formats is supported by data management software ensuring the complete set of data processing operations. Most of formats are accompanied by data reading programs. Exception is made for MEDATLAS format which is accompanied by the developed software set ensuring data QC and providing access to data. Creating OceanPC the ICES developed the system programme - converters and these system may be used as a general converter for a number of formats (WAOD98, ICES, ICES Blueprint ' 86, SD2).

2. Data coding

Most of the considered formats use a wide spectrum of data coding methods: codes of WMO, IOC, GF-3, ICES, WDC-A, national codes, codes specially developed at the planning stage of data management within international programmes (e.g. WOCE, GLOSS)

The official IODE standard for the data coding are GF3 codes tables (IOC Manuals and Guides N 17, 1987), including both algorithm of a code parameter formation, and specific codificators. At the same time GF3 codes tables have not been reconsidered for long time. Probably, due to this alongside with GF3 codificators other code tables are used, which partly duplicates each other especially in metadata coding (Annex 2). It should be noted that in code tables which are in use (GF3 and others) ice, chemical pollution and biological parameters are poorly reflected.

In process of expansion of technological opportunities of the IODE centres the supporting and application of additional parameter codes for ocean data management becomes more urgent. For instance if several data sets and data bases are integrated, the final integrated data set should contain references to data sources and their formats. In specialized marine data sets it is necessary to identify time series, standard sections, fixed stations, sizes of data systematization areas and etc. It should be noted that the existing tables of codes do not contain codificators for description of derived and calculated/modelled data.

But the most complicated problem is related to parameter coding. This problem requires careful consideration in connection with: (I) growing amount of codes used for the same parameters; (ii) the additional requirements placed on coding on the part of new information technologies.

The use of various approaches for parameter coding, which actually is not felt when a specific task of data processing is fulfilled, however it will have considerable negative consequences for development of the distributed multi-level oceanographic databases. To illustrate the problems which may appear in the future due to the use of various parameter codes the following example is given. If in *a database A* sea water temperature has the identifier *TEMP7XXD*, and in a *database B* the identifier of sea water temperature is *00001*, the use of these data from two connected databases will require converting at a interface level of between two databases or between databases and a users.

The necessity of development of a technique for parameter coding is connected also with the use of electronic marine maps, Geographical Information Systems (GIS) and Data Base Management Systems (DBMS) in frames of ocean data management. In thus case numerical (sometimes they are referred to as factographical), spatial, textual and graphical data are used jointly under GIS/DBMS technology. GIS/DBMS applications used for processing, analysis and interpretation of data and information on marine environment need to link numerical data (digital sets of climatic fields, results of modelling and etc with corresponding thematic marine spatial data (vector climatic and modelled fields) and GIS cartographic basis.

3. Data unification - problems and proposals.

The data formats continue to remain a serious barrier, which we need to overcome during the marine data management. Despite of the large number of already existing data formats, their amount continues to grow.

Data formatting assumes great importance when “end to end” data management (E2EDM) technology is considered. In frames of such a technology both metadata and data may refer to various disciplines and be presented by different data types (numerical, spatial, textual and graphical data) and in different forms: flat data files, data bases under DBMS/GIS and HTML documents. Currently there are used the data management technologies of two distinctive types:

- specialized subject-oriented technologies which have regulated input and output data flows;
- technologies based on DBMS (in some cases on GIS) for which input/output data flows may be both regulated and non- regulated.

And of the main conditions for E2EDM process establishment is interconnection of technologies implementing various steps (phases) of data management. Interconnection of technologies means that the output data flow of one technology appears to be the “transparent” input data flow of the other ether directly or by means of data structure conversion. By consideration E2EDM the problem of formatting of the data develops into a problem of unification of the data between various functionally oriented technologies.

The task of data unification is difficult in itself, but some peculiarities of IODE E2EDM make it even

more difficult. These peculiarities lie in the fact the IODE concept envisages building and developing the E2EDM system based on existing national and international information systems which are independent at present and will remain independent in the future. Therefore data unification in IODE E2EDM can not be associated with overall and mandatory unification of data structures and information technologies.

Probably the only field, where data unification will not have serious objections, is the realization of the unified information space, which is understood here in the technological sense. (see IOC/IODE- XVI/23) i.e. as a set of tools solving E2EDM integration problems. In this context data unification means the complex solution in the form of methods, rules and tools, which provides information interfaces (or interfaces for input/output data flows) for technologies, without any transformation of their internal peculiarities.

This solution may be based on data description language automatically recognized by components of instrumental software which is used as a media for development of subject - oriented technologies in the IODE system. Considering a rapid progress of Web - technologies for data management on local and global levels XML (eXtensive Makeup Language) seems to be the best selection. On the basis of XML, individual specialized languages may be developed. Creation of XML extension for marine environmental data (marine XML) is the problem to be specially considered at the session.

Alongside with marine XML the ocean data unification requires performance as a minimum of two more conditions. The first condition lies in the fact that data unification may be achieved only when software converting input and output data flows, which are integrated by a technology into marine XML media, be available for the IODE centres to use it.

And the second (may be not the last) condition is related to the necessity to use the unified system of codification at least with respect to the dictionary of parameters. The structure of such a dictionary is not clearly defined, so far but mast likely it should contain a certain universal parameter code, parameter description and a set of parameter codes applied in currently existing technologies, which are most widely used, and in data sets (bases), which are in the greatest demand in IODE system.

The alternative solution in the field of data unification, which is closer to conventional methods of data exchange standardization, is based on GF-3 solutions. The principle of this solution is that for a given discipline (oceanography, coastal hydrometeorology, pollution etc) on the basis of existing data formats and supporting software a set of data structure descriptions for various data presentation (observed - derived - calculated/modelled data) and available software is prepared. This set is recommended as the IODE standard to describe input/output data flows in technologies.

The choice of solution should consider everything - from logical perfection to economical benefits. In any case the problem of unification is the fundamental problem as far as integration of current and development of future technologies of the IODE System is concerned. Now matter how complicated this problem may be it is necessary to make every effort to solve it.

Annex 1
Data Formats Summary Table

Format Name	Obs. type	Metadata	Codes	Data Structuring	Software support	Self-defining
ICES	H, CTD	**CI**	WMO, IOC, ICES	FPAR	OceanPC	No
ICES Blueprint '86	H, CTD	CG	GF3	VPAR,	OceanPC	Yes/GF3
GF3 (Subsets)	M, H, MBT/XBT, CTD,…	CR, CG, OM	WMO, GF3, IOC	VPAR,	GF-3 Proc	Yes/GF3
JGOFS	M, H, MBT/ XBT, CTD, Bl, PL	CR, CG, OM	WMO, GF3, added	VPAR,	W/R	Yes/GF3
GTSPP	M, H, MBT, XBT, CTD	CI, OM, SQ, HQ, HM	WMO, IOC, GF3, MEDS, NODC	FPAR	W/R	No
BUFR (subsets)	Bathy, Tesac	CG	WMO	VPAR,	W/R	Yes/BUFR
NODC (SD2)	M, H	CR, CG, OM	WMO, IOC, NODC	FPAR	?	No
NODC (BT)	MBT, XBT	CR, CG, OM	NODC	FPAR	?	No
NODC (h/r STD/CTD)	CTD	CR, CG, OM	WMO, IOC, NODC	FPAR	?	No
OCL (WAOD'98)	H, MBT/XBT, CTD, BL	CI, OM, HM	WMO, NODC/ OCL	VPAR	R	Yes/?
TOGA	M, H, XBT, CTD, …	CR, CG, OM	WMO, IOC, NODC, added	FPAR	?	No
WOCE	M, H, XBT CTD,…	CR, CG, OM	WMO, IOC, NODC, added	FPAR	?	No
GLOSS	Sea level	CR, CG, OM	added	FPAR	R	No
MEDS	M, H, MBT/XBT CTD	CG, OM, SQ, HQ	IOC, WMO, GF3, MEDS	VPAR,	W/R	No
MEDATLAS	H, MBT/XBT, CTD	CR, CG, OM, HQ	IOC, WMO,	VPAR,	QC Medar	Yes/GF3

			GF3, ICES			
P-3 (NODC OPDB)	M, H, MBT/XBT CTD	CR, CG, OM, HQ	WMO, NODC	VPAR,	DBMS	Yes/?
NODC Russia	M,H, MBT/XBT, CTD, PL	CR, CG, OM	WMO, IOC, National	FPAR,	HDDL System	Yes/HDDL
NODC Argentina /RNODC	M, H, MBT/XBT, CTD	CG, OM	WMO, IOC, National	FPAR,	W/R	No
NODC Spain	M, H, MBT/XBT, CTD	CG, OM	WMO, IOC, National	FPAR,	W/R	No
OceanAtlas	H, CTD	**CR**	No	FPAR	OA	No
TSV-O(DVO)	H, CTD	CR	No	FPAR	DVO	No

Observation type (instruments): M - meteo H-bottle cast MBT/XBT – mechanical/expandable bathythermograph CTD – high resolution casting BL – biology PL – chemical pollution	Metadata: CR – reference to metadata CG – general metadata (country, ship, organization, cruise number, …) CI – identifiers (country, ship) OM – observation methods SQ – QC summary HQ – history of QC HM – history of data management
Software (available for users) W/R – examples write/read software	Structuring: FR – fixed length of logical record VR – varying length of logical record FPAR- limited set of parameters VPAR- unlimited set of parameters

Annex 2

Code metadata tables sources for ocean data management

Element	Code Sources
IOC Country	GF3
Platform type	GF3
Organization	WDC-A, MEDAR, MEDS
Project	WDC-A
Ship	WDC-A
Call Sign	WMO, MEDS
Instrument	WMO (for meteo), WDC-A (T, biological parameters), OPDB, ICES (current)
Special platform	GF3
Modified IHB ocean/sea area	GF3, MEDAR
Coastal station index	WMO
IGOSS message identifier	MEDS
Units	GF3, OPDB, ICES (S), WDC-A (chemical)
Date and time within day	GF3
Time and frequency	GF3
Selection of depth levels code	GF3
Method of depth determination	GF3, WDC-A
Position and navigation	GF3, WOCE
Hod of latitude / longitude measurement	GF3
Prime navigation aid fix flag	GF3
Confidentiality/ Availability	MEDAR, MEDS
Media	MEDAR
Ocean weather station	WDC-A
Code table for TS probes	WDC-A
Digitization method	WDC-A, MEDS
Digitization interval	WDC-A
Data treatment and storage	WDC-A
Position precision	WDC-A
Declared national program	WDC-A
XBT probe type	WDC-A
Bottom hit	WDC-A
Time precision	WDC-A
STD-SCAN Condition	WDC-A
Sample Interval	WDC-A
Data type	MEDS
GTSPP status	MEDS
Stream identification	MEDS
Parameter code	GF3, ROSCOP, P-3, WDC-A
Validation flag	GF3, WDC-A, MEDS

(end of document)