OGS oceanographic data archiving and validation system: the IOC/National Oceanographic Data Centre

A. Giorgetti, A. Brosich and R. Mosetti

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale – OGS, Trieste, Italy

(Received: May 19, 2006; accepted: February 20, 2007)

ABSTRACT The oceanographic database, managed by the National Oceanographic Data Centre (NODC)/IOC Research Unit within OGS, archives 166 millions of measurements of physical and biogeochemical parameters, of current, wave motion, sea level and meteorological data. The information in the data archiving system are fully validated and continuously updated. Its relational schema contains all the information needed by the Medar/MedAtlas format, including the complete set of information qualifying the measured data. All the meta-data are public while the access to archived data is subjected to a data policy, defined at data set level in agreement with the data providers. The oceanographic data archiving and validation system is accessible through the NODC Web Portal, available on Internet. A dedicated data portal has been developed, where all users can register and get different access permissions pertinent to different levels of information and to different sections of the data archiving system. The Italian NODC information system is completely integrated in the European distributed network of oceanographic data centres and is able to answer requests for information through the network by means of a standardised interface and making use of the XML technologies to exchange and qualify information.

1. Introduction

The OGS Department of Oceanography has a long and high level experience in marine data acquisition, management and archiving, obtained through participation in several national and international experimental and applied research programmes. For several years the OGS has played the leading role in the rescuing, qualification and management of hydrological and biochemical data, mainly at international level. Intensive collaborative relation have been started with Italian institutions (CNR, ENEA, Universities) and with European NODCs, among others HNODC/NCMR, IFREMER, BODC, RIHMI-WDC. Since 1991, OGS has been continuously participating in numerous EU-funded research projects, carrying out data management activities. In chronological order they are: the EU project **EDMED**, European Directory of Marine Environmental Data, from 1991 to 1993 (BODC, 1993); MAST/**MODB**, Mediterranean Oceanic Data Base, from 1994 to 1997 (Brasseur *et al.*, 1996); MTPII/**MATER**, MAss Transfer and Ecosystem Response, from 1997 to 1999 (Maillard *et al.*, 2002); **MEDAR**/MEDATLAS II, MEditerranean Data Archaeology and Rescue, from 1999 to 2001 (MEDAR Group, 2002); EuroGOOS-**EDIOS**, European DIrectory of the Ocean-observing System, from 2001 to 2003 (Verduin and Fisher, 2003); **SeaSearch**, a Pan-European Network for Ocean & Marine Data and

Information Management, from 2002 to 2005 (Schaap, 2004); **SeaDataNet**, a Pan-European Infrastructure for Ocean & Marine Data Management, from 2006 to 2011. An important communication and exchange network has been developed with these projects between the European data centres, the data providers and the users, hence allowing the establishment of standard protocols for data exchange, data validation and dissemination [the Medar/MedAtlas standard; MEDAR Group, (2005)] recognized at international level. As a result, comprehensive and fully validated data sets were produced by adopting a unified protocol for data archiving and validation (MEDATLAS 2002, MTP II-MATER 1996-1999). The OGS, acting as Regional Data Centre for the Central Mediterranean, is responsible for the safeguarding of new and historical data and meta-data collected in the Adriatic Sea, Ionian Sea and Sicily Channel by Italian institutes as well as from bordering countries.

In June 2002, the OGS was nominated National Oceanographic Data Centre for Italy (Mosetti, 2003) within the International Oceanographic Data Exchange (IODE) system of the UNESCO Intergovernmental Oceanographic Commission (IOC).

The importance of technology has been a critical aspect in oceanographic data management. In 1979, a first relational database model, working on the IBM 4341 system, was implemented at the OGS Informative Centre to manage the meta-data obtained within the international ASCOP programme (Manca, 1989). The oceanographic data were stored as ASCII files on magnetic tapes and hard discs. The development of a marine data bank, based upon an advanced relational database management system (RDBMS) has been a necessary step to give visibility to NODC oceanographic information, accomplishing the institutional role assumed by OGS as Italian NODC, and to operate at international level with other scientific institutions. The storage of marine data on a structured database allows us to centralise the Italian information on the marine environment, to be easily accessed by the users through look up systems via web. At the same time, the new data management system helps scientists and researchers to insert, update and delete data, to share information, to search for and retrieve data, avoiding redundancy, inconsistency and duplication. NODC started the development of the national meteo-marine database, including most of the national time series of wave motion measurements, sea level, current and meteorological data in cooperation with APAT (the Agency for Environmental Protection and Technical Services) within the Italian oceanographic community. The information system developed by OGS and maintained by APAT can manage all meteo-marine national data and can operate as an effective link between the main national data holding centres through a unique portal.

2. Data and meta-data management

NODC detains a large quantity of oceanographic data spanning over mass field measurements, bio-chemical parameters, marine currents data, wind waves and air-sea interface parameters. The data have been collected by means of surveys, moorings and operational large ocean buoy systems. Historical data have been transcribed from magnetic tapes to modern storage devices. The data archiving and validation system consists of a RDBMS integrated with statistical and spatial analysis tools.

The relational schema of the database includes all the Medar/MedAtlas format tag elements

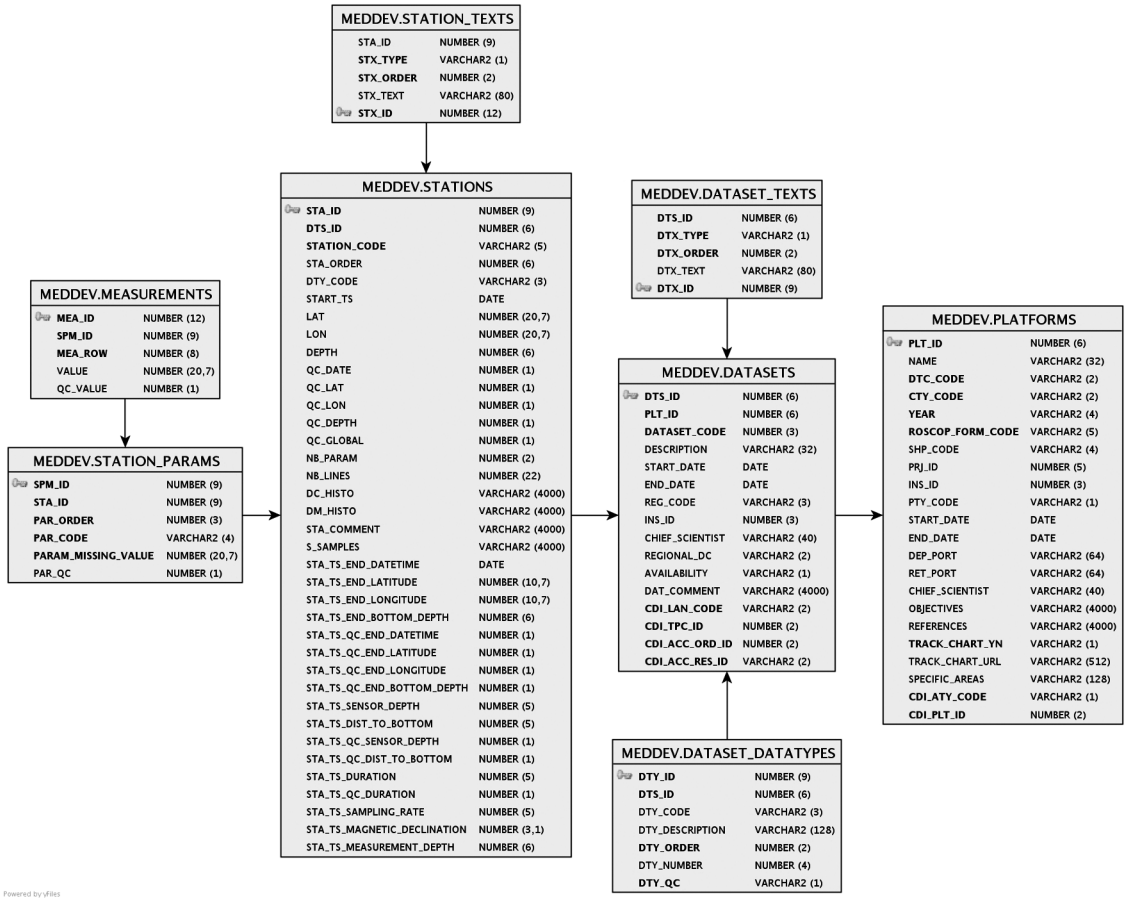

Fig. 1 - Simplified Entity-Relation schema of the NODC data archiving system.

(Maillard *et al*., 2001), with the complete set of information qualifying the measured data. It is designed to archive both vertical profiles and time series of oceanographic data. At this stage, the database schema includes 40 tables, containing all meta-data, data and decoding information. The database structure is totally normalised, very flexible for future updating and completely integrates the existing information, that comes from different sectors. Efficient and adapted procedures have been implemented to insert data in the database, extract data, perform the data and meta-data quality check according to the international protocols and guidelines, analyse data and meta-data consistency, disseminate data and meta-data on line.

A simplified Entity-Relation schema is shown in Fig. 1. The upper element of the hierarchy is the table PLATFORMS, directly linked to the table DATASETS, which is linked to STATIONS. All measured data points related to the archived parameters (and the reference parameter, that is to say pressure or time) are saved in the table MEASUREMENTS, containing, at this stage 166,077,531 rows. All measured data and meta-data has its quality control flag. The full set of data and meta-data are contained in five main tables, which are shown in the figure from right to left:

PLATFORMS: containing information about the measurement platform (ships or buoys);
DATASETS: every platform can be linked to one or more sets of collected data;
STATIONS: containing information about a measurement point (time, position, etc.);
STATION_PARAMS: all measured parameters related to the station;
MEASUREMENTS: all measurements for every parameter.

A set of shared tables contain information like data centres code and name, data types code and name, institutes code and name, parameters code, name and unit, projects code and name, sea regions code and name, ships code and name. In particular, eight shared tables were defined, namely COUNTRIES, DATACENTERS, DATATYPES, INSTITUTES, PARAMS, PROJECTS, REGIONS, SHIPS. A set of shared tables contain decoding information and fields necessary to produce the output data formats.

The procedures for inserting and extracting data related to a cruise or a mooring into and from the database consists of a series of Java programs, where all SQL instructions are recalled. The inserting procedure finishes with a commitment statement, if all information in the data set are compatible and consistent and therefore can all be included, or with a rollback statement, should an integrity constraint be violated. In the latter case, no profiles or time series related to the data set are inserted, and a warning message is given to correct the problem. Specific checks are included to avoid the insertion of duplicate profiles or time series, enforcing additional rules not defined using the integrity constraints. The extracting procedure uses web forms as front-end. Users can select single data sets or choose some search criteria on space and time coordinates, data parameters or quality control flags. Data are extracted in an XML intermediate stream and then converted to the format requested by the user.

The **NODC web portal** (http://nodc.ogs.trieste.it) provides general information on the data centre, on the research projects carried out by the data centre, gives free access to the systematically compiled and updated data and meta-data directories. The main page of the web portal is shown in Fig. 2. The meta-data catalogues, listed under the window Marine Information, includes sea cruises information [IOC/ROSCOP Cruise Summary Reports, UNESCO (1991)], research projects reports (EDMERP forms), operational oceanography monitoring site evaluations (EDIOS Forms) and descriptions of collected data [EU/EDMED forms, BODC (1993)]. The window Marine Data gives access to measured data from different sectors: oceanographic data, satellite data, real-time operational data from drifters, Medargo buoys, meteo-oceanographic buoys, seismic data. Oceanographic data retrieval can be obtained through the dedicated web data portal available to internal (Intranet) and external (Internet) users, with different security levels. Data and meta-data can be extracted in comma separated values (CSV) data format, in generic ODV spreadsheet format [Ocean Data View; Schlitzer (2004)], or in Medar/MedAtlas data format. The Medar/MedAtlas data format is the ASCII auto-descriptive text model based on the international IOC/GF3 coding system (UNESCO, 1987), developed and adopted within the Mediterranean oceanographic community. It includes general information on the measuring platform, such as the name of the ship or of the buoy, the responsible institute and scientists, the name of the project and the data availability, general information on the measurements profile, such as the spatial and temporal coordinates, the depth of the measuring point, the measured parameters with the related unit, the data collection and data management history, and the table of the measurements for all the present parameters.

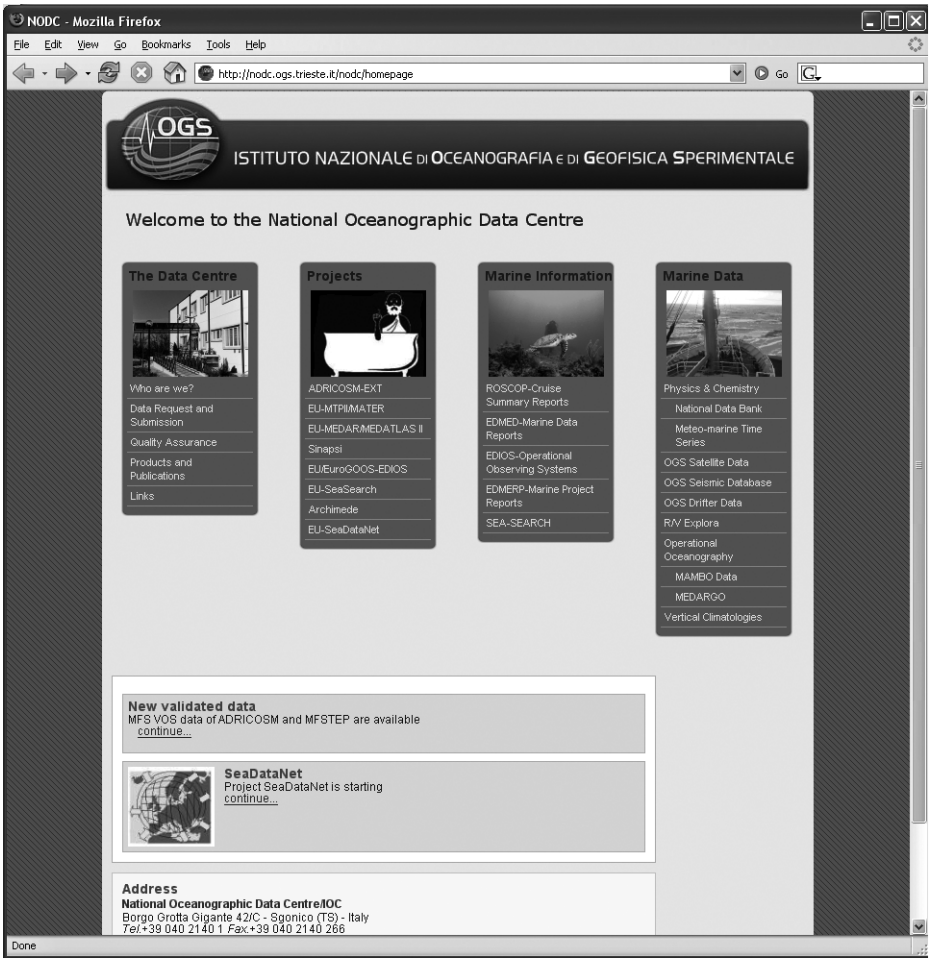

Fig. 2 - NODC web portal.

The **NODC oceanographic database** covers all the Mediterranean Sea, with a particular eye to the Adriatic Sea, the Ionian Sea, the Channel of Sicily, the Levantine basin, and the coastal areas of the Gulf of Trieste, the Emilia-Romagna coastal, the Strait of Messina and the Channel of Otranto. This database is mainly devoted to the marine physical sector (hydrological field, hydrodynamical field and meteorological field), even if information concerning the sectors of marine chemistry, biology, sedimentology, mineralogy are also included. It contains:

- 294,503 hydrological stations (vertical profiles of physical and biogeochemical parameters collected by discrete bottle samples, CTD and XBT profiling probes);
- 889 current time series related to 166 measuring points, collected in the last 30 years;
- 1,353,912 wave motion recordings related to 42 measuring points;
- 10,562,954 sea level recordings related to 45 coastal stations;
- 2,099,770 recordings of meteorological parameters along the Mediterranean coasts;
- several days of surface temperature, salinity and current data collected underway in the

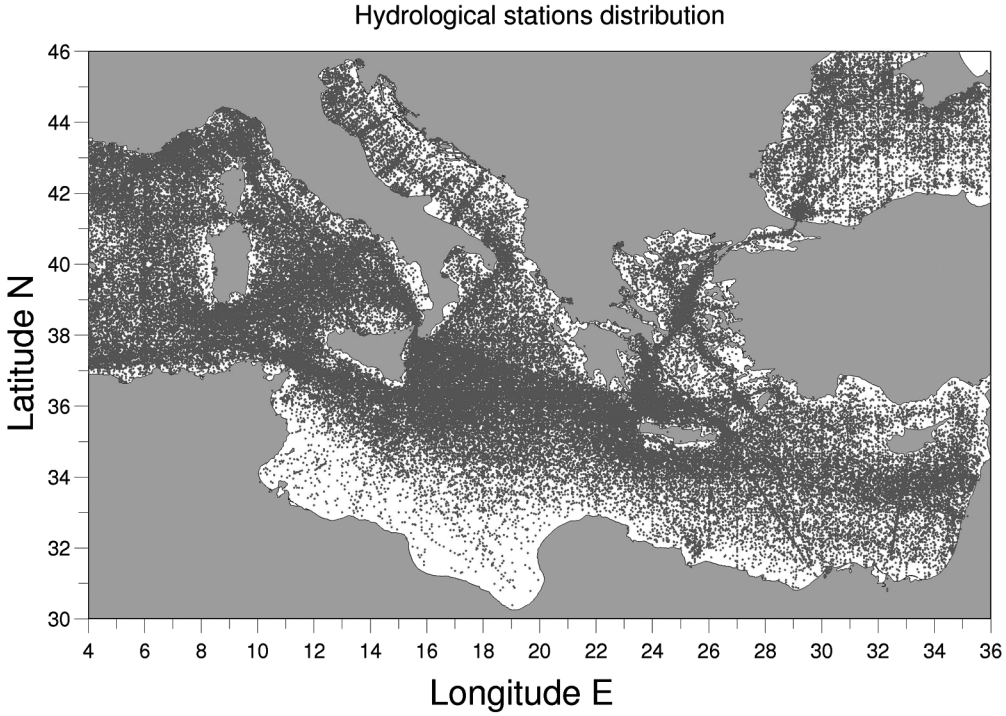

Fig. 3 - Distribution of hydrological stations in the Mediterranean Sea as extracted from the NODC data archiving system.

Mediterranean and Atlantic Sea;

- 45,160 recordings of meteo-marine buoys collected in the northern and central Adriatic and in the Channel of Sicily (three off-shore buoys).

The spatial distribution of hydrological stations (red circles) is shown in Fig. 3, while the spatial distribution of fixed moorings, including wave motion measuring points (blue triangles), current meter moorings (azure squares), sea level stations (red circles) and

Table 1 - Total number of measurements and profiles included in the NODC data archiving system per data type.

DATA_TYPE	DESCRIPTION	RECORDS	PROFILES
H09	Water bottle stations	3,742,824	88,357
H10	CTD stations	31,915,349	40,512
H13	Bathythermograph drops	9,881,942	165,605
H72	Thermistor chain	349	29
D01	Current meters	30,213,769	889
D09	Sea level measurements	42,615,035	128
D72	Instrumented wave measurements	13,867,404	75
M03	Near surface meteorology	618,728	61
M06	Routine standard measurements	16,003,912	244

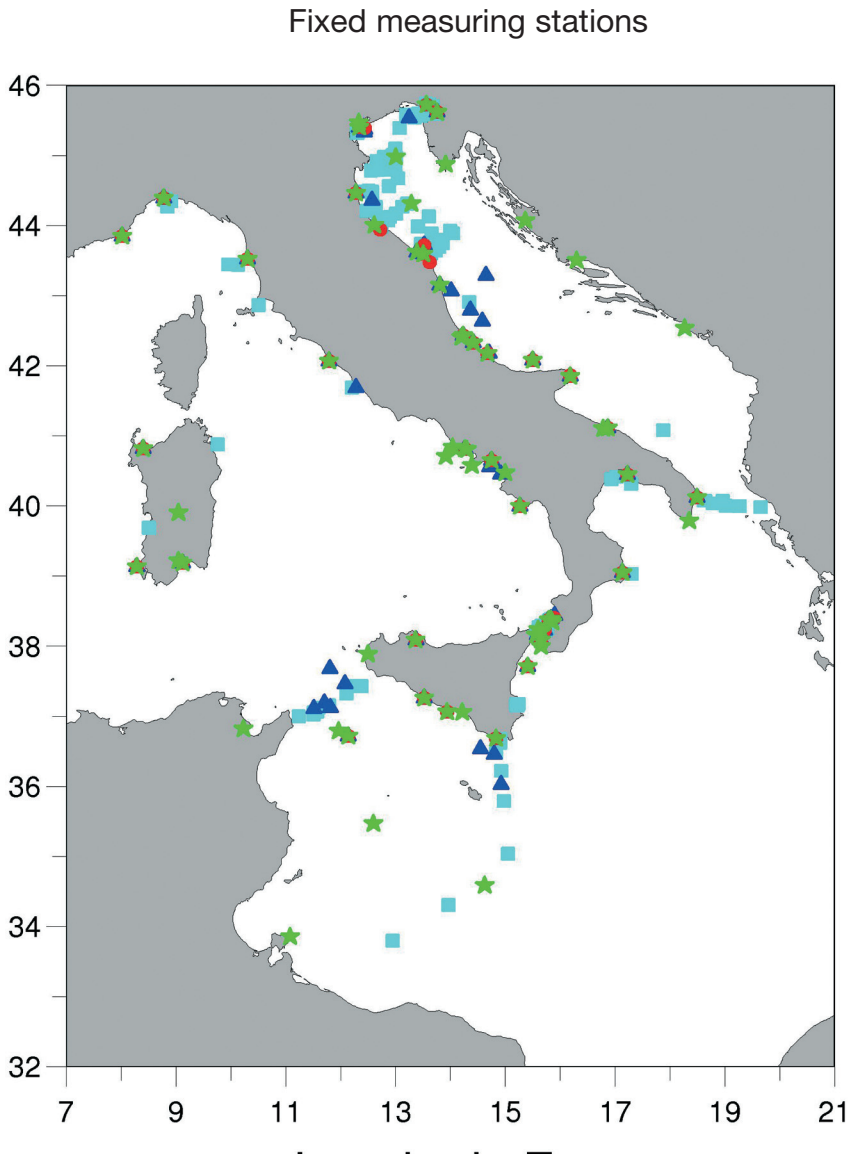

Fig. 4 - Distribution of fixed measuring stations in the Mediterranean Sea as extracted from the NODC data archiving system.

meteorological parameters measuring points (green stars), is shown in Fig. 4. The number of measurements and profiles per data type present in the data archiving system is reported in detail in Table 1.

The marine meta-data and data in the archiving system can be obtained through FTP or SMTP, contacting the institute operator, or through HTTP, with the compilation of a dedicated form for data extraction. Efficient and adapted routines are implemented to query information

in the database and extract data in Medar/MedAtlas format from a dedicated searching mask implemented in the NODC web portal. It is possible to extract an entire data set, related to a cruise or a mooring, as well as a part of one or more data sets satisfying a search request. The search criteria are based on the selection of a spatial and temporal interval, related to the measured data, and on the selection of different or multiple hydrological archived parameters.

3. Data validation

All measured data put in the oceanographic data archiving system are completed with meta-data and documentation. This information concerns the measuring sites, the measuring instruments and parameters, the responsible body, the collection and processing procedures applied to the row data. All this information has been coded and inserted in the relational database structure.

The **data integration** issue is strictly linked to the data standardization issue, which concerns meta-data, documentation and quality control (Giorgetti *et al*., this vol.), and has been tacked and deeply analysed by handling scientifically and technically large multidisciplinary data collections. In fact, the measured values included in the NODC data archiving and validation system covers different sectors, different time periods and originates from different sources. References for a common standard have been adopted. Moreover, the data integration issue has been faced while handling measurements coming from different continuous acquisition systems, related to profiling probes as well as moored measuring instruments.

All measured data put in the oceanographic data archiving system have been validated, according to international protocols recognised in literature for the related measurement sector (UNESCO, 1993). Each data set of vertical profiles or time series is submitted to a set of automatic (objective) and visual (subjective) checks. As a result of the validation process, a quality flag is defined for all information checked (in the data and in the meta-data) without changing or eliminating any data points. The quality control flag [GTSPP flag scale, UNESCO (1990)] is a number associated to each measurement field, whose value grows according to the importance of the failure (0 = not controlled, 1 = correct, 2 = suspect, 3 = dubious, 4 = wrong, 5 = changed, 9 = missing).

The **quality control procedure** for vertical profiles and time series of oceanographic data, defined according to UNESCO/IOC and ICES international guidelines, includes:

- checks for information and data format completeness;
- check of the date and of the measuring position;
- check of duplicate stations or measures.

All these checks are applied to the meta-data, archived in the data header lines. Subsequently, the checks on the data points including a series of automatic checks with visualisation of the quality flag assigned, are applied. These aim at highlighting:

- the presence of casual errors, that may be due to unhappy manual operations, to voltage drop during measurements, to data transmission problems, to sensor calibration problems, to ordinary or extraordinary maintenance operations;
- the presence of systematic errors, that may be due to changing of measuring routine operations, changing of instrumentation, changing of the measuring site, changing of

Table 2 - Statistical results after the application of the data quality control procedure for vertical profile core parameters.

	Samples	QC flag 1,2	QC flag 4	QC flag 3	QC flag 2
Temperature	20,007,798	19,778,452 (98,8 %)	28,205	19,977	41,121
Salinity	11,650,700	11,437,784 (98,2 %)	26,557	9,529	20,106
Dissolved Oxygen	1,475,232	1,010,201 (68,5 %)	96,800	365,636	81,041
Phosphate	151,828	147,512 (97,1 %)	1,940	2,375	1,446
PH	150,766	145,247 (96,3 %)	1,215	469	0
Silicate	121,364	119,419 (98,4 %)	1,571	373	892
Nitrate	81,348	77,803 (95,6 %)	1,109	2,436	1,722
Nitrite	73,596	72,553 (98,6 %)	832	211	0
Chlorophyll-a	58,615	54,678 (93,3 %)	29	3,908	4
Ammonium	30,252	29,789 (98,5 %)	89	374	0
Alkalinity	14,737	14,379 (97,6 %)	250	72	0
Total Phosphorus	11,178	10,995 (98,4 %)	182	1	0
Total Nitrogen	6,327	6,320 (99,9 %)	3	4	0

environmental conditions at the site.

In the specific, the check of the data points for both the vertical profiles of marine data as well as for the time series of physical and meteorological parameters, includes:

- comparison with min & max values fixed for each parameter archived,
- comparison with other close measures,
- check with a reference climatology,
- check for spikes,
- check for vertical stability (marine data on the vertical),
- tidal and residual analyses (time series of sea level data).

In the case of vertical profiles of bio-chemical parameters the international protocols (Maillard *et al*., 2001) have been further focused and adapted to the Mediterranean Sea sub-basins (Giorgetti *et al*., 2005). The result of the quality control procedure applied to all vertical profiles included in the NODC database is summarised in Table 2, where the statistics of the occurrences of the quality control flags for the core parameters is presented. These results show that 95% or more of the data are qualified as good (QC = 1, 2), the percentage grows for temperature and salinity with the only exception of oxygen measurements.

It has to be considered that data validation is all the more important the more people that take it into consideration in their analyses. If fact, meta-data gives fundamental details on the measurements collected. Any statistics computed on a dataset can be strongly affected if wrong measurements are included, and errors can be eliminated by excluding high QC values (QC = 3, 4) from the computations.

5. Conclusion

The data archiving and validation system developed by the NODC within OGS contains a large quantity of oceanographic data spanning over the last century and covering the entire Mediterranean Sea. Data have been collected by means of ship surveys, moorings and large ocean operational monitoring systems. The NODC information system consists of a relational database management system, integrated with efficient and adapted procedures to insert data in the data base, extract data in different formats and disseminate data on line.

On one side, the high level of OGS oceanographic activities conducted over the past 40 years, in pure and applied research, both at national and international level, has allowed OGS to become one of the reference points in Italy for physical oceanography and to obtain a large amount of quantitative information on the marine environment. The synthetic description of the available oceanographic data points is one of the most important elements for their evaluation in view of a future dissemination and use.

On the other side, the NODC expanded the information system to new technologies, boosted its marine database, centralising the information heritage of the past 40 years, to be easily accessed by the users through look up systems via the web. Data retrieval can be obtained through the NODC web portal available to internal (Intranet) and external (Internet) users, with different security levels.

The NODC information system currently archives, manages and gives access to environmental marine national data and is ready to operate as an effective link between the main national data holding centres through a unique portal.

The development of the NODC marine information system, both in the hardware and software components, opened the way to the creation of a dedicated informative tool, available for the scientific community. The European distributed network of Mediterranean oceanographic data centres, have created common standards and an infrastructure to give visibility to comprehensive and fully validated data products, to field experiments and to operational oceanography activity. A step further has been done to integrate the Italian NODC information system in the distributed marine databases network of European NODCs by adapting the information infrastructure to XML technologies. A standardised interface is obtained by mapping the database information with Common Data Indexes (CDI), defined at Pan-European level. The implementation of the XML data format and of the XML schema, allows us to answer and validate the requests of objects over the network, addressed by the user interface of the centralised navigator (virtual server). In this way, the NODC information system fully answers Marine XML technology as well. At the Italian level, the NODC gives access to historical and new collected marine environmental data, obtained by national and international monitoring networks. The NODC holds the scientific, technical and physical means to support Italian research activities maintaining and developing high quality oceanographic data and meta-databases, readily available for Italian researchers in academia, government and industry.

Acknowledgements. The oceanographic data base is the result of the continuous effort of a large number of OGS scientists and technicians. Particular thanks go to Dr. Bruno Manca for his data management activity, to Mr. Luciano Perini for his constant contribution in data transcribing and validation and to all colleagues engaged in marine data collection and management.

REFERENCES

Brasseur P., Beckers J.M., Brankart J.M. and Schoenauen R.; 1996: *Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of a historical data set*. Deep-Sea Res., **43**, 159-192.

BODC; 1993: *Completion of the European Directory of Marine Environmental Data (EDMED) for all EC member countries.* Final Rep. to CEC-DG XII, NERC, Proudman Oceanographic Laboratory, Birkenhead (UK), 78 pp.

Giorgetti A., Bruschi A., Inghilesi R., Morucci S. and Orasi A.; 2007: *Maritime data standardization in the Archimede Project.* Boll. Geof. Teor. Appl., this vol..

Giorgetti A., Burca M., Manca B.B. and Tomini I.; 2005: *Compilation of a quality controlled database of biological and chemical oceanographic parameters in the Central Mediterranean Sea*. Boll. Geof. Teor. Appl., **46**, 357-376.

Maillard C., Fichaut M. and MEDAR/MEDATLAS Group; 2001: *MEDAR/MEDATLAS Protocol (V3) Part I: Exchange Format and Quality Checks for Observed Profiles*. Int. Rep. of TMSI/IDM/SISMER/SIS00-084, Institution IFREMER/SISMER, Centre de Brest, France, 49 pp.

Maillard C., Balopoulos E., Giorgetti A., Fichaut M., Iona A., Larour M., Latrouite A., Manca B., Maudire G., Nicolas P. and Sanchez-Cabeza J.-A.; 2002: *An integrated system for managing multidisciplinary oceanographic data collected in the Mediterranean Sea during the basin-scale research project EU/MAST-MATER (1996-2000)*. J. Mar. Sys., **33-34**, 523-538.

Manca B.; 1989: *Una base di dati oceanografici multidisciplinare per il controllo delle qualità delle acque del mare Adriatico.* Boll. Ocean. Teorica ed Appl., Num. Spec., 249-269.

Mosetti R.; 2003: *IODE National Report for Italy*. IOC/IODE-XVII/10.15, 3 pp.

MEDAR Group; 2002: *MEDATLAS 2002 Mediterranean and Black Sea database of temperature, salinity and biochemical parameters. Database Climatological Atlas CD- ROM*. European Commission Marine Science and Technology Programme (MAST), 4.

MEDAR Group; 2005: *A Mediterranean and Black Sea oceanographic database and network*. Boll. Geof. Teor. Appl., **46**, 329-343.

Schaap D. M.A. ; 2004: *Sea-Search Project, First Annual Report,* EVR1-CT-2002-20009, 47 pp.

Schlitzer R. ; 2004: *Ocean Data View,* http://odv.awi-bremerhaven.de.

UNESCO; 1987: *GF3: A General Formatting system for geo-referenced data, Vol. 2: Technical description of the GF3 format and code tables.* IOC Manuals and Guides, **17**, 111 pp.

UNESCO; 1990: *GTSPP Real-Time Quality Control Manual, IOC Manuals and Guides n. 22*. In: Manual of Quality Control Procedures for Validation of Oceanographic Data. CEC: DG XII, MAST and IOC: IODE Manual and Guides, **26**, pp. 305-432.

UNESCO; 1991: *Manual on International Oceanographic Data Exchange*. IOC Manual and Guides,. **9**, Revised Edition, 82 pp.

UNESCO; 1993: *Manual of Quality Control Procedures for Validation of Oceanographic Data.* CEC: DG XII, MAST and IOC: IODE Manual and Guides, **26,** 436 pp.

Verduin J. and Fischer J. ; 2003: EDIOS: *European Directory of the Initial Ocean Observing System*. In: Dahling H. Flemming N.C., Nittis K. and Petersson S.E. (eds), Building the European Capacity in Operational Oceanography, Proceedings of the Third International Conference on EuroGOOS, Elsevier Ocean. Series, **69**, pp. 265-271.

Corresponding author: Alessandra Giorgetti
Istituto Nazionale di Oceanografia e di Geofisica Sperimentale - OGS
Borgo Grotta Gigante 42/c, Sgonico (Trieste), Italy
phone: +39 040 2140391; fax: +39 040 2140266; e-mail: agiorgetti@ogs.trieste.it