Adding Value to Oceanographic Data at the British Oceanographic Data Centre

Roy Lowry, Lesley Rickards and Juan Brown

British Oceanographic Data Centre
Joseph Proudman Building
6 Brownlow Street
Liverpool L3 5DA

Abstract. The British Oceanographic Data Centre (BODC) is the Natural Environment Research Council's (NERC) designated data centre for marine sciences. BODC's core business since its formation in 1979 (under the name Marine Information and Advisory Service Data Banking Section) has been the ingestion of oceanographic data into a national oceanographic data resource, including value enhancement to ensure the data stored may be used with confidence decades after collection without recourse to communication with the original collector of the data.

Quality enhancement is achieved through procedures to:

- Replace *ad hoc* parameter descriptions used by scientists to label data with standardised terms from a parameter dictionary.
- Ensure metadata content satisfies the standards of the BODC data model, including collation of information from multiple sources, resolving contradictory information and pursuing missing information.
- Ensure all available information beyond the scope of the BODC data model is preserved by incorporation into digital documentation linked through the data model to the data.
- Ensure bad data are clearly identified by the addition of data quality flag fields to all data values.

These procedures have been applied to the data centre's acquisitions for over 25 years resulting in an extremely large and valuable marine data resource. Work is starting to extend the BODC data model to fulfil the requirements of distributed data systems such as the NERC DataGrid, migrating information from unstructured plain language documentation suitable for human users into structures more suited to interrogation by software agents.

The Oceanographic Data Management Culture

The marine environment and its inherent physical, chemical and biological processes have no respect for national boundaries. It covers two thirds of the Earth's surface, is 3-dimensional and is highly dynamic. Compared to the scale of the environment, measurements are extremely sparse as no technique is capable of high resolution sampling across all four dimensions. Satellite measurements give good spatial and temporal coverage, but only sample the surface. Ship-based instrumentation gives high resolution vertical profiles, but only at a single point in space and time. Making measurements involves expensive platform operations, such as running research vessels and the results are unrepeatable. Furthermore, these measurements are essential as a baseline for any quantification of change in the marine environment.

The resultant scarcity and high value of oceanographic data has led to the development of a culture where long-term data preservation and data sharing are the norm rather than the exception. Consequently, there is a long history of national and international oceanographic data management and curation. The Intergovernmental Oceanographic Commission (IOC) founded International Oceanographic Data and Information Exchange (IODE)[1] in 1961. This is a worldwide service oriented network of DNAs (Designated National Agencies), NODCs (National Oceanographic Data Centres), RNODCs (Responsible National Oceanographic Data Centres) and WDCs (World Data Centres for Oceanography). During the past 40 years, IOC Member States have established over 60 oceanographic data centres in as many countries. This network has been able to collect, control the quality of, and archive millions of ocean observations, and makes these available to IOC Member States.

In the UK, oceanography is not the only domain with data curation high on its agenda. The Natural Environment Research Council (NERC)[2] operates a network of designated data centres covering geology, atmospheric sciences, hydrology, terrestrial ecology and remotely sensed data in addition to marine science.

The British Oceanographic Data Centre (BODC)

BODC[3] was formed as the Marine Information and Advisory Service Data Banking Section in 1979 and established as a separate organisation under the BODC banner in 1989. It has been the UK NODC in the IODE network since the beginning and as such has a remit to work as a national facility for the marine community as a whole, interacting with Government departments and their agencies, academia, industry and the general public. A significant

element of the data handled by BODC originates from outside the parent organisation, identifying the group as an autonomous data centre and not a value added reseller.

In the late 1980s, BODC's operations diverged from the conventional view of a data centre as simply a repository and distribution point for data due to the problems encountered with working in this way. Scientists rarely submitted their data in a fully worked-up form (let alone with proper documentation) and data were invariably submitted many years after their collection, by which time it was almost impossible to resolve problems identified in the data or to compile the documentation.

To alleviate these problems, BODC pioneered an end-to-end approach to data management, playing a pro-active role within oceanographic project (such as NERC Thematic Programme) field programmes, providing shipboard support and directly involved in the working up, calibration and quality control of the data. The resulting integrated datasets were documented supplied for use by the programme's scientists within the lifetime of the project. This approach is far more cost effective than data rescue/archaeology. It is also of direct benefit to the programme's scientists and facilitates the interchange of data within the programme itself.

In 2001, BODC became established and resourced as NERC's designated data centre for marine science. As such, it is responsible for assuring the long-term stewardship of NERC-funded oceanographic data collected by NERC institutes, collaborative centres and other organisations such as university departments. This work encompasses both an NODB role, assuring stewardship and redistribution of 'user-ready' datasets, and a role providing end-to-end data management support to activities in NERC laboratories and collaborative centres unsupported by BODC's project activities.

BODC currently employs 35 staff. Most of these are 'data scientists' with direct experience of sampling and scientific work, who undertake the task of standardising, quality controlling and integrating data into BODC's systems. These are supported by IT specialists developing and maintaining infrastructure and data distribution systems. The organisation is hosted by NERC's Proudman Oceanographic Laboratory (POL)[4] on the Liverpool University campus.

Physical Security of Data at BODC

Physical security of data is one of BODC's primary concerns and the organisation's approach to this has three components: an 'accession system', physical backup of data storage and long-term preservation of a data archive.

The accession system is a set of operational procedures applied to all data arriving at BODC. The physical integrity of the data is secured by preservation of the original media together with a copy placed in the BODC data archive. Wherever possible, the technological vulnerability of the data in the archive is reduced through creation of an additional version of data supplied in proprietary formats in simple 'lowest common denominator' formats based on ASCII. For example, Microsoft Excel workbooks are output as one or more (if multiple worksheets) ASCII CSV files.

This physical securing of the data is supported in the accession system by the creation of a metadata record providing the data submission with a unique identifier and describing its contents and provenance. Any supporting paper documents that accompany the data are systematically filed, but such submissions are becoming increasingly rare.

The BODC data archive is but a small part of BODC's data storage resources that are spread over PC and UNIX file servers and an Oracle database. This file base holds data in the process of being worked up in addition to the source and final versions of the data. POL provides BODC's computing resources through a service level agreement which includes a systematic backup of all BODC's data storage to DLT. The procedures include daily incremental backups together with regular full backups onto a regularly recycled tape stock to guard against media degradation.

The backup tapes are stored in fire safes in two physical locations on the university campus with a further copy of each full backup located at another NERC facility in Keyworth, Nottinghamshire. This provides reasonable recovery (<1 month data loss) in the event of total destruction of the campus.

Over the past 25 years, data have arrived at BODC in a wide variety of physical formats: paper (cards, listings and tape), magnetic tape (7-track, 9-track, QIC, DLT), floppy disks of many sorts (8", 5.25", 3.5" and 3"), optical disks (magneto-optical, CD-ROM and DVD-ROM), FTP and e-mail. Accession policy until recently was to preserve all physical media received. However, this was developing more into a computing museum than a data resource and much of the redundant media was discarded when BODC moved from Bidston to Liverpool in 2004.

Although source media were discarded, the archive copy of all data received 'as received' has been maintained. The physical implementation of this archive is an area where technological redundancy has been an issue. Initially, the archive consisted of a set of 9-track magnetic tapes with one tape per data accession. These became redundant in 1992 with the demise of BODC's IBM mainframe and were copied to Panasonic phase-change optical disks. It became apparent that these were the optical disk equivalent to 'Betamax' and they were replaced by conventional magneto-optical disks a few years later. Finally, in 2002 'spinning' magnetic disk prices were considered low enough for the archive to be incorporated into the UNIX file system where it currently resides.

The data volumes involved in these migrations have ranged from under 10 gigabytes to several hundred gigabytes, with the archive currently holding some 0.2 terabytes. Interestingly, although the data volumes have increased, the human resources required for the physical copying have tended to decrease due to the increased capacity of individual media.

Adding Value to Data

The considerable effort expended by BODC to ensure the physical security of all data coming through the door would in most cases be totally wasted if that was all that was done. Very few data sets arrive as fully documented entities that could be used with confidence years after their original submission. Most would have degraded by now into lists of meaningless numbers, assuming of course that the list can be deciphered from a binary magnetic tape format.

BODC's primary mission is to ensure that data may be reused with confidence without any need for recourse to the data originator no matter how much time has elapsed since the data were submitted. Ensuring this forms the bulk of the work of BODC's data scientists adding value to the data by application of the procedures described here.

Data Harmonisation

BODC policy is to take data 'as they come' for the originating scientists without any standardisation requirement to maximise the proportion of data reaching the data centre. Consequently, the first requirement is for data harmonisation. 'High' volume (hundreds of measurements plus) data are converted into a common format (NetCDF[5] subset) using a bespoke conversion system[6] requiring minimal code extension (0.5 – 2 days' effort) to

expand its format portfolio. To date this system has been used to harmonise data submitted in over 300 formats.

Smaller volume data, such as water bottle data, are integrated into a unified schema in an Oracle database. Water samples are generally collected from a common set of bottles containing water from different depths by many scientists, each of whom undertake analyses and produce a spreadsheet containing their particular set of parameters. Each spreadsheet is labelled with metadata (station, sampling depth etc.) independently logged at sea in notebooks by tired, wet and seasick individuals. Consequently, it contains a significant proportion of errors. The harmonisation process involves reconciling inconsistencies in this information before loading the data for all parameters into a common data structure. As a result metadata quality is enhanced in addition to the production of a merged data set.

Data harmonisation and integration greatly ease the process of data synthesis. High volume data originally held in a multitude of formats become available to the synthesiser in a single format. Bottle data sets are merged once and only once saving significant amounts of scientific time.

Quality control

Once harmonised, BODC quality control procedures are applied to the data through visual inspections of graphical presentations by scientifically skilled staff. Problem data are marked by manipulating quality control flags tagging each data value. This is currently accomplished manually using a bespoke interactive graphical editor[7] but augmentation by semi-automated algorithmic procedures is planned. Undertaking this procedure considerably adds to the value of the data set, particularly if it is to be used for purposes not envisaged by its originator. Consider the case of a data channel with 10,000 values containing a single large spike. The presence of the spike may be safely ignored in an application requiring the channel mean, but it cannot be ignored should the value range for the channel be required. Such an approach adds to the integrity of the data for all subsequent users.

Explicit quality control exposes ‘hidden’ problems with the data. It is not uncommon for a data originator to know that a particular subset of the data is wrong. However, it is uncommon for any indication of this to be included with the data. The originator is simply aware of the problem and avoids doing anything with that particular bit of the data. However, without intervention by the data centre a secondary user would be unaware that a problem existed.

Data Access

No matter how much effort goes into adding value through quality assurance and metadata enhancement, the results are worthless if the data cannot be found, accessed and used. Data retained by their collectors are notorious for becoming hidden or even permanently lost. Through its efforts BODC have added to the value of NERC's marine data by making them available not just to the UK science community but internationally through the IODE and SEA-SEARCH[8] networks.

Even more value may be obtained from data if they are synthesised with data from other sources. BODC is working to facilitate and automate this process across NERC through the NERC DataGrid[9] project and across Europe through the recently funded SeaDataNet[10] project.

Metadata Enhancement

The most significant value addition to data by BODC is through the enhancement of both the content/coverage and quality of the metadata accompanying the data. Attention to both of these areas is essential if data are to be recycled on a decadal timescale without consultation with the originator.

Metadata content in the BODC National Oceanographic Database (NODB) is based upon a data model developed during the late 1970s specifying a minimum requirement for description of oceanographic data. This is remarkably close to the core geospatial metadata requirements specified in ISO19115 considering the BODC model's early origin.

The model is populated for each data entity held in the NODB by a standardised information collation procedure. Any metadata included in the source data files is captured by the software system used to harmonise the data format and used to build a skeleton metadata record. Data scientists then populate the unassigned fields from whatever sources they can obtain, including correspondence with the data originator. In cases where the information for a metadata field is available from multiple sources great care is taken to resolve any discrepancies thus maximising the metadata quality. A number of fields in the model, such as the spatio-temporal co-ordinates, are mandatory and originators are pursued vigorously should any of this vital information not be available.

The result of these efforts is that each data entity held by BODC is covered by a minimum metadata content standard populated with the best quality information available. However, the procedure doesn't stop there. The collation procedure inevitably obtains information that has no designated field

in the data model. This information is written into XHTML documents that are stored in the database with linkages to the appropriate data entities.

One aspect of metadata that has received particular attention in BODC is the description of the measured phenomena. The labelling of these has been identified as a particular area of weakness in the scientific community. Columns of data are often labelled using abbreviated strings such as 'temp', 'T1' or 'chl' whose semantic significance is either locked away in an inaccessible notebook or even forgotten completely.

BODC practice is to replace these labels with access keys to a parameter dictionary. This process is surprisingly time consuming, often requiring research into the methodologies used before sufficient information is available for reliable matching to a dictionary entry. A large amount of effort over 25 years has been put into the development of the parameter dictionary used (the BODC Parameter Usage Vocabulary (PUV)), including a funded dictionary development project (EnParDis)[11] during 2003-2004. The result contains over 17,000 parameter descriptions that have each been built in a standardised manner from discrete information elements that are in turn managed by controlled vocabularies. Thus the syntactic structure and spellings are guaranteed to be consistent throughout. The names of the biological entities have been standardised against the Integrated Taxonomic Information System (ITIS)[12] that further enriches the metadata through access to a biological taxonomy. The full PUV is available on-line[13] together with two parameter discovery vocabularies and associated mappings. It is being developed through the auspices of IOC into an international standard for the oceanographic domain.

Achievements

One has to take care when quantifying BODC's achievements. The data centre specialises in handling data sets that are small in volume, but extremely high in complexity covering a wide range of physical, chemical, biological and geological parameters that need to be supported by very rich metadata if they are to have any value for future reuse. Statistics based on data volume are therefore inappropriate and a suitable alternative metric is difficult to find. A feel for the level of success may be obtained when it is considered that the BODC data holdings include:

- Over 1,500,000 chemical and biological measurements on water samples covering over 5,000 different parameters dating back to 1988.
- Over 60,000 profiles of temperature and salinity dating back to 1968.

- Over 7,000 site years of sea level data dating back to 1842.
- Over 6,000 current meter records dating back to 1967.

All the above have had their value enhanced through harmonisation, quality assurance and metadata assembly procedures that assure their suitability for reuse in decades to come or even longer.

A further measure of the success of a data centre lies in its value to the community and in ensuring that the data are used. In this respect BODC has a wide range of users. For example, from April 2004 to March 2005, 38,021 requests from the scientific community, educational establishments, Government bodies, industry and the general public were serviced.

Future Plans

Whilst BODC's procedures and working practices to add value to data are based on 25 years of experience and currently work well, the desire for improvement is important for the continuing health of the system. Although the data model showed remarkable foresight when it was developed in the 1970s, its content is light for anything other than the minimal requirements of current standards and models. Another difficulty is that the fundamental entity in the data model is too fine-grained to be considered as a 'dataset' in the ISO19115 context. These shortcomings have been particularly exposed through BODC's participation in the NERC DataGrid project, but are equally limiting to any system based on software agent rather than human interaction. Extensions are currently being developed to enrich the metadata content of the model and manage entity aggregation to build true datasets.

A significant proportion of the information required for the metadata enhancement is currently held in the supporting XHTML documentation rather than designated fields. This limits software agent access to free-text searching, which is notoriously unreliable when the syntax and spellings in the target are not standardised. BODC is therefore looking at ways in which the most important items of information may be captured into tagged fields. The document stock is held as XHTML element contents in a relational database rather than documents per se, which means that a significant degree of automation may be incorporated in this procedure.

The goal of adding further value to BODC's data holdings through data interoperability with other organisations is frustrated to some by technical obstacles such as parameter description diversity. Many organisations built their own parameter dictionaries in the past and these differ significantly in granularity and semantic richness. Interoperability requires that these

vocabularies be mapped. Two approaches to this problem are currently being undertaken by BODC. The first is to further develop the BODC PUV with particular emphasis on the handling of synonyms, which is essential for reliable mapping. The second is to investigate the capabilities of the ontology-based vocabulary mapping strategy and supporting tooling being developed by the Marine Metadata Interoperability (MMI)[14] project. If these prove sufficiently robust then work will begin applying them to the construction of operational maps.

It is clear that maximum value from data may only be obtained if they are available on-line. Whilst BODC has some on-line data delivery capability, this does not cover all the datasets held. The launch of a completely revamped website in August 2005 represents the start of major investment to address this. Further database interface systems are under development and nearing completion and a little further into the future involvement in interoperability projects such as NERC DataGrid and SeaDataNet will enhance the value of BODC's data holdings through increased accessibility.

Conclusions

Data curation has been high on the oceanographic agenda for over 40 years and is currently a high priority for the environmental sciences in the UK through NERC's network of designated data centres.

Since its beginnings in 1979 BODC has made a significant contribution to the curation of UK oceanographic data adding significant value to the data it has acquired through harmonisation, quality control and metadata enhancement. The aim is to continue this process whilst raising the standards of quality and service in the future.

Reference URLs

[1] http://ioc3.unesco.org/iode
[2] http://www.nerc.ac.uk
[3] http://www.bodc.ac.uk
[4] http://www.pol.ac.uk
[5] http://my.unidata.ucar.edu/content/software/netcdf/index.html
[6] http://www.bodc.ac.uk/about/information_technology/software_engineering/transfer_system.html
[7] http://www.bodc.ac.uk/about/information_technology/software_engineering/visual_programs.html
[8] http://www.sea-search.net/
[9] http://ndg.nerc.ac.uk
[10] http://www.seadatanet.org
[11] http://www.bodc.ac.uk/projects/uk/enpardis
[12] http://www.itis.usda.gov/
[13] http://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/bodc_para_dict.html
[14] http://marinemetadata.org