Data Management for Community Research Projects: A JGOFS Case Study

Roy K. Lowry

Abstract

Since the mid 1980s, much of the marine science research effort in the United Kingdom has been focused into large scale collaborative projects involving public sector laboratories and university departments, termed Community Research Projects. Two of these, the Biogeochemical Ocean Flux Study (BOFS) and the North Sea Project incorporated large scale data collection to underpin multidisciplinary modeling efforts.

The challenge of providing project data sets to support the science was met by a small team within the British Oceanographic Data Centre (BODC) operating as a topical data centre. The role of the data centre was to both work up the data from the ship's sensors and to combine these data with sample measurements into online databases.

The working up of the data was achieved by a unique symbiosis between data centre staff and project scientists. The project management, programming and data processing skills of the data centre were combined with the oceanographic experience of the project communities to develop a system which has produced quality controlled, calibrated data sets from 49 research cruises in 3.5 years of operation. The data centre resources required to achieve this were modest and far outweighed by the time liberated in the scientific community by the removal of the data processing burden.

Two online project databases have been assembled containing a very high proportion of the data collected. As these are under the control of BODC their long term availability as part of the UK national data archive is assured.

The success of the topical data centre model for UK Community Research Project data management has been founded upon the strong working relationships forged between the data centre and project scientists. These can only be established by frequent personal contact and hence the relatively small size of the UK has been a critical factor.

However, projects covering a larger, even international scale could be successfully supported by a network of topical data centres managing online databases which are interconnected by object oriented distributed data management systems over wide area networks.

1 Introduction

The primary objective of all scientific data management is to provide the scientist with the data he (or she) needs to support their research. This objective contains three implications for the data manager.

a) The data manager must ensure that the scientist is provided not just with the data but with sufficient information to allow the data to be used with confidence. This includes information on how the data were obtained, how they were processed and on the quality of the data.

b) The data manager must provide the data in such a way that the effort needed to use the data is significantly less than the benefit obtained from the data. The required manner of presentation is obviously user dependent. For example, a biologist may easily obtain the information he needs about a density field from a graphical representation. However, a numerical modeller in marine physics is more likely to require the data in machine readable form in a format which he can easily incorporate into his software.

c) The data manager must ensure that access to the data is maintained until it is universally agreed that the data are no longer of any value to the scientific community. In most cases this means long term maintenance for many decades and beyond.

Consequently, it can be seen that expectations of data managers are high. Not only is a thorough understanding of all types of data collected by the scientists needed but a thorough understanding of the needs of the scientific user community is also required.

This paper describes how this has been achieved in the United Kingdom by a unique partnership between a data centre and the academic scientific community.

2 The British Oceanographic Data Centre (BODC)

The British Oceanographic Data Centre (BODC) was formally created in April 1989. It is located at NERC's Proudman Oceanographic Laboratory at Birkenhead, Merseyside. Whilst managed by the host laboratory on behalf of NERC's Marine and Atmospheric Sciences Directorate (MASD) the BODC mandate is the provision of data management services on a directorate wide basis.

BODC developed from the Data Banking Section of the Marine Information and Advisory Service (MIAS) which was formed in 1976 primarily to provide support to the offshore oil industry by building a central archive of oceanographic data.

MIAS approached this problem by classical data archaeology. A data scout was employed to travel around the data collecting laboratories in the academic, government and commercial sectors persuading them to submit their holdings to the data centre. Once submitted, the data were converted to a common format, screened, documented and loaded onto a database.

Inevitably, the problems associated with data archaeology were encountered. In many cases those collecting the data had lost interest in them and were unwilling to provide the necessary effort to resurrect the data. Much of the data submitted revealed the shortcomings of the parochial viewpoint of local data management and a significant 'data laundry' effort was required to bring the data to a standard where they were of general use.

During the eighties a shift in the source of funding caused the work of the MIAS Data Banking Section to evolve away from supporting the oil industry towards supporting science both in NERC and the universities. During this time the national data archive was further expanded to provide a significant national oceanographic data resource.

During 1988 the concept of BODC was developed with a mandate to:

a) Maintain and operate the national data archive.
b) Provide data management support to major programmes within NERC's Marine and Atmospheric Sciences Directorate (MASD).
c) Make good quality oceanographic data available to UK research scientists, industry, local and central government.
d) Collaborate in international data exchange and data management.

The data centre currently has 8 tenured staff, 5 staff on fixed term contracts and 4 industrial training students. This paper describes the work of BODC in support of two of MASD's Community Research Projects undertaken by 3 of these staff (I tenured and 2 contract) plus 1 or 2 students.

3 BODC and the MASD Community Research Projects

During the mid 1980s MASD developed the concept of Community Research Projects which aimed to target a significant proportion of MASD resources at a small number of specific scientific problems and involving scientists both from NERC's own laboratories and from university departments. Two of the initial projects to be set up, the North Sea Project and the Biogeochemical Ocean Flux Survey (BOFS), involved massive data collection efforts. The BODC remit was to ensure that these data were properly managed.

3.1 The North Sea Project: Concept Development

3.1.1 Project description

The proposed fieldwork for the North Sea Project consisted of 30 consecutive research cruises each of approximately two weeks duration: in other words 15 months of continuous sea time. The cruises were subdivided into alternate survey and process cruises. The survey cruises repeatedly worked a network of 123 fixed stations whilst the process cruises were short self contained studies covering a wide range of disciplines within the southern North Sea. In all cases, the primary sampling platforms were the ship's pumped water supply, a heavily instrumented CTD frame with a 12-bottle rosette and atmospheric sampling equipment. In addition, a smaller number of stations were worked using corers and zooplankton nets.

Working on the project were approximately 100 scientists from 4 NERC laboratories and 7 university departments scattered throughout the length and breadth of the United Kingdom. Imposed on this was an irrevocable time limit of 2.5 years after the end of the fieldwork for the completion of the project including a significant modeling effort using the data collected during the fieldwork.

3.1.2 Data management requirements

Viewing this scenario from a data manager's point of view there were two basic requirements:

a) The number of cruises and timescale was such that interpretation would have to be based upon a poor quality unrefined data set unless a highly efficient mechanism for working up the data was installed.
b) Mechanisms had to be devised to give the project scientists easy access to the entire data set from their home laboratories.

3.1.3 Data processing

The solution to the problem of working up the data required a radical rethink of the relationship between the project scientists and the data centre.

Let us first consider what would have happened in practice to the data automatically logged by the ship's systems had the conventional relationship between scientists and data centre been maintained.

The data are taken off the ship on magnetic tape in GF3 format. Each principal investigator (PI) would have required a set of tapes (at least 15 copies of a set of 4 or 5 tapes per cruise). Those responsible for the individual data channels would then have cleaned up and calibrated their part of the data and submitted them to the data centre.

The implications of this for the data centre are horrific. Consider the CTD instrumentation. Pressure temperature and salinity were the responsibility of one laboratory, oxygen a second, light and chlorophyll a third and transmissometer attenuance a fourth. Consequently, the CTD data would have arrived at the data centre as four separate files and almost certainly on four different timescales. The data centre operation to merge these files is a project manager's nightmare even if one ignores the practical problems associated with identifying a common pressure channel.

The scenario would also have placed those PIs who were the users of the automatically logged data in a difficult position. They would have been faced with the difficult choice of dealing with the calibrating PIs direct, waiting for the worked up data to be available at the data centre or using raw data extracted from their copy of the tapes.

To avoid these pitfalls it was decided that BODC should adopt a radically different and far more active role in the data collection exercise. Instead of distributing the tapes to the PIs, only one set of tapes were generated and sent to BODC straight from the ship.

Here, the working up and quality control of the data were centralised. The data from the ship's tapes were first reformatted into disk files which could be handled by the data centre software tools. One of the most powerful of these, the SERPLO data inspection and flagging package (Loch, in prep) running on Silicon Graphics workstations, was used to despike the data. Initially there were frequent consultations with the project scientists as to what constituted credible data. However, these became less frequent as BODC staff gained experience in North Sea oceanography.

The calibration of the data was an exercise in close cooperation between BODC and the project scientists. The way in which this worked is best considered by example. The fluorometer calibrations for chlorophyll were the responsibility of scientists at the Plymouth Marine Laboratory. Samples from the CTD water bottles were filtered and frozen on board ship, sent to Plymouth and the extracted chlorophylls determined. The extracted chlorophyll data were sent to BODC over the UK academic network (JANET) where they were matched to fluorometer voltage readings and the combined file was returned to Plymouth. Here, project scientists identified rogue extracted chlorophyll values and determined the calibration equations. This information was returned to BODC where the equations were applied to obtain the calibrated data set.

The success of this method of working may be judged by the fact that, within three months of the last cruise docking, the data from all 30 project cruises were fully worked up and available for use.

The advantages of this method of working are as follows:

a) Maximum use is made of the skills available. The data centre offers project management, programming and data handling skills whereas the project scientists offer experience in instrument calibration and North Sea oceanography.
b) necessary information flow is not inhibited by the communication problems which inevitably arise between groups working independently.
c) Scientific management is able to obtain progress and status reports from a point source.

d) The maintenance of a common independent variable for all data channels is assured.

e) Data centre staff are integrated into the scientific project team which both provides motivation and ensures that the data management does not become divorced from the research.

3.1.4 Providing access to the data

Having solved the problem of how to work up the data. the mechanism for making the data available to project scientists required consideration. The required data set not only included the automatically logged data worked up by BODC (and therefore present in the data centre) but also the results of laboratory analyses of samples collected on the cruises which were in the possession of the PIs.

The solution adopted was to load the data into an online database managed by BODC. In taking this approach there were a number of problems which had to be addressed.

a) ***Choice of platform and software environment***

The choice of platform and software environment was dictated by circumstances. There were considerable advantages to be gained from hosting the database on the computer facilities available at the Bidston site. It was clear that adequate resources were available here and so the database was implemented on the Bidston computer, an IBM 4381 mainframe under VM/CMS running the ORACLE RDBMS.

b) ***Technical aspects of user access***

Within the UK, all academic institutions (including NERC laboratories) have computer facilities interconnected by the JANET wide area network. Consequently, it was possible for any project scientist to log onto the Bidston IBM from their home laboratory.

However, achieving this in practice required careful consideration if it was to be done without compromising either the data security by releasing it to nonproject participants or the system security by providing an opportunity for hackers.

The platform available for the database made the solution awkward. The obvious method would be to set up a project account on the system with the minimum of privileges required to interrogate the database. The password could then be made known to project participants and modified as necessary to maintain security in an inevitably dynamic user population.

However, VM/CMS only allows one user at a time to be logged onto a specified account. There is also less control over user privileges than with other operating systems such as Unix or DEC's VMS. Consequently, each user had to be given an account on the Bidston system which were configured by BODC to grant the privileges necessary to access the data.

This solution worked well in practice despite the administrative burden and no security compromises have been reported in almost two years of operation.

c) ***User education***

There is no point in creating an online database if the project scientists do not have the knowledge or expertise to obtain the data they require from it. The problem was approached from two directions.

First, the user interface to the data retrieval was made as friendly as possible. This was aided by the user-friendliness of the database interrogation language provided by ORACLE, SQL, which is syntactically similar to plain language and hence easy to learn. However, some queries, particularly those requiring information from several tables can become quite complex. Some of these may be simplified by the creation of database views but situations were identified where commonly required information could not be obtained easily.

In these cases high level language programs implemented as additional CMS commands were written to simplify the user interface. Consider the example of retrieval of a CTD cast from the database. This requires retrieval of station position information from one table, the raw datacycles from a second and the calibration coefficients from a third. In addition, the calibration equations need to be applied to the raw datacycles. Whilst this is possible using SQL, it is extremely cumbersome. However, with a high level language application all that is required of the user is the simple command:

```
CTDLIST <station id>
```

Similar commands were implemented to allow data retrieved from the database to be represented graphically, including contouring software, and output on a range of graphical devices. This gave the user the ability to log onto Bidston, produce a graphical image file, network it to his home computer and generate a plot.

Secondly, steps were taken to ensure that the user was provided with all the information needed to retrieve data from the database. To this end a comprehensive manual describing the database structure and all BODC implemented commands was produced. Several users have successfully mastered the database using this together with system documentation provided by NERC Computer Services and a little interactive help.

Further, the induction of the initial user population. awaiting the database launch, was accelerated by holding a 3-day training course at Bidston. This covered SQL syntax and usage, the database structure and how to run the BODC supplied software with a strong emphasis on hands-on experience.

The course was attended by over 15 scientists, mostly research assistants, and was well received by all. Many of the attendees have subsequently passed on the knowledge gained to their colleagues and have greatly assisted BODC by providing local support to other users.

d) ***Obtaining the sample data set***

When building a database the technical problems associated with design and implementation fade into insignificance when compared to the problems associated with obtaining the data from the scientists who collected them. However, these problems were found to be far less severe in the case of the North Sea

Project database. As BODC had worked up the automatically logged data these were in house and ready for load. Further, a significant proportion of the sample data set (chlorophyll, dissolved oxygen, sediment gravimetry and salinity) had been submitted to BODC during the calibration exercise.

The flow of remaining data into the database were lubricated by five factors which motivated the scientists to submit their data to BODC:

i) The automatically logged data worked up and held by BODC could be used as currency in the sense that they provided a readily identifiable product coming out of BODC in return for the data coming in.

ii) The project scientists soon learnt that once individual sample data sets were loaded into the database they were automatically linked to other measurements on the same samples. Consequently, a merged data set could be obtained from the database with no effort on the part of the scientist.

iii) Value added data products, such as contour plots, could be easily obtained from data held in the database using software supplied by BODC.

iv) The project fostered a genuine team spirit and broke down a lot of the mutual distrust which had previously existed between scientists. BODC's active involvement meant that the scientists viewed the database as a part of their research project and therefore contributed to it willingly.

v) BODC took great care to ensure that they were perceived as honest brokers of the data who would not allow access from outside the project until the data were formally placed in the public domain.

The degree of success in obtaining the data may be judged from the fact that, by February 1990, over 75 per cent of the data collected on the 30 project cruises had been lodged with BODC i.e. within four months of the end of the 30th cruise. At the current time (December 1991) this figure has risen to over 95 per cent.

e) Quality control

The arrangement with the project scientists was that responsibility for quality control should rest with the PI. In the case of the automatically logged data this was achieved by interaction between BODC and the PIs at various stages during the data processing exercise. In the case of the sample data sets BODC assumed that the data were quality controlled prior to submission.

However, each data set was subjected to brief scrutiny to ensure that no gross corruptions of the data had occurred. Further, it was discovered that a relational database is a powerful quality control tool enabling data comparisons to be made over many permutations in a relatively short time. A number of problems were uncovered in this way during routine checks on the database and resolved by consultation with the PI concerned.

During the building of the database an unexpected quality control problem emerged. The physical firing of a water bottle can only take place in one position in space and time. This is not the case for the recording of that event in five scientist's log books! The CTD used in the North Sea Project was known to overestimate pressure by 2 db. The water bottles were physically located approximately 2 m above the CTD pressure head. This meant that the depth of a bottle fired with 10 m of wire out in still water could be recorded as 10 m (wire out), 12 m (corrected CTD pressure) or 14 m (uncorrected CTD pressure) depending upon the definition used for 'bottle depth'.

As bottles were fired with a depth separation of as little as 2 m this represented a real problem which only came to light during the assignment of sample identifiers at BODC. Significant BODC effort was expended resolving the problem by designating authoritative bottle firing depths using a consistent definition.

It is worth noting that had a distributed database approach been adopted this problem would only have come to light when users attempted a merge of sample data sets retrieved from more than one source. The result would have been chaotic with much time wasted by duplicated effort.

The North Sea Project database was formally opened to the user community on 1st March 1990 only S months after the main phase of data collection was completed. Subsequently, data from a comparatively small amount of additional fieldwork in 1990 (8 cruises in all) have been added to the database.

It is currently in daily use with over 30 active user accounts which service the data requirements of well over 50 project scientists. Its success, together with that of the BOFS database, will be objectively reviewed later in this paper.

3.2 The BOFS Community Research Project: An Extension of the Established Principles

3.2.1 Project description

The Biogeochemical Ocean Flux Survey (BOFS) represented the initial UK contribution to JGOFS, and commenced fieldwork in 1989 (3 cruises) with further field seasons in 1990 (6 cruises) and 1991 (2 cruises). In each field season, survey work and process studies were combined both during and after the spring bloom centred along the 20W line in the North Atlantic.

Again the project involved between 50 and 100 scientists from 2 NERC laboratories and 8 universities. However, the project community differed markedly from the North Sea Project with only a handful of scientists involved in both projects.

3.2.2 Differences between the North Sea Project and BOFS data management requirements

The data management strategy pioneered for the North Sea Project had worked well and therefore the logical way to proceed was to apply the same principles to BOFS.

Therefore, in general, what has been discussed above in the context of the North Sea Project is equally applicable to BOFS. However, the BOFS project differed. in data management terms, from the North Sea Project in a number of ways.

a) BOFS was essentially a deep ocean study whereas the North Sea Project was totally contained in a shelf sea shallower than 100 m. This required minor modifications to the data processing procedures and a reassessment of the criteria used for quality control decisions. Sufficient deep sea oceanographic experience was available in the BOFS community to guide BODC so the transition was achieved without problem.

b) The BOFS data set was much more diverse both in the range of parameters that were automatically logged and in the range of measurements made on samples collected during the cruise. In addition, these samples were taken using a much wider range of oceanographic hardware. Again this meant that the BODC systems had to be extended and the complexity of the database schema increased. However, the effort required to make these changes was minimal.

c) The North Sea Project was planned in terms of a predefined set of measurements which were intensively and repeatedly taken. The design of the database was therefore relatively easy to establish at the beginning with a fair guarantee that it would not need to change during the project.

However, BOFS was organised in a very different way. After each field season scientific meetings were organised where the results of that season were discussed and used to formulate the fieldwork requirements for the next year.

Consequently, the required database schema inevitably changed as the project progressed. In the past this would have presented a major problem with each change to the database structure requiring a complete dump and reload of the database. However, the ORACLE RDBMS allows additional tables to be specified, columns to be added to tables and even changes to column specifications with the data in place. Consequently. the database was easily able to adapt to the changing requirements of the science.

d) In the North Sea Project principal investigators were designated for the measurement of the basic environmental parameters such as temperature, salinity and chlorophyll. It was these PIs who worked closely with BODC in the instrument calibrations. However, BOFS was directed more towards supporting innovative measurements. Whilst this had clear scientific benefits, it meant that there was no clearly defined responsibility for the calibration of these basic, but nevertheless vital, parameters.

The problem was solved by BODC expanding its role by taking responsibility for the calibration of the CTD sensors and those instruments connected to the ship's pumped water supply for which there was no designated responsibility within BOFS. Nevertheless full advantage was still obtained from the pool of expertise within the BOFS community with acknowledged experts acting as consultants where required.

3.2.3 The importance of sample coordinates

These relatively minor difference apart, the BOFS database was built up in exactly the same way as the North Sea Project database. Due notice was taken of lessons learned during the North Sea Project. For example, it was realised that it was vital to establish authoritative coordinates in time and, particularly, space for all of the samples collected. Consequently, much effort was put into obtaining copies of 'soft' data collected during the cruise such as log sheets. With all of this information collated centrally, discrepancies could be resolved and the space time coordinate framework could be established within the database with confidence.

The importance of resolving any problems with sample coordinates in a multidisciplinary study cannot be overemphasised. Errors can lead to false comparisons and experience handling the two Community Research Project data sets shows that such errors are frequently encountered.

Even such basic procedures as labeling stations can go badly wrong. For example, on one cruise the CTD station numbers given to the computer files got out of step with the station numbers used by the scientists taking samples from the water bottles. The confusion was furthered by the numbering system adopted which allowed overlaps to develop: the sample numbers matched a CTD file but it was for a CTD dip taken 12 hours after the samples were taken.

Ideally, these errors should be eliminated at source. However, anyone who has experienced the working conditions on a research cruise will realise that this may not be possible. The danger of these errors is not that they are made but that land based data managers assume that such errors cannot be made and design systems accordingly.

3.2.4 Providing access to the BOFS data

Scientists were given access to the database in exactly the same manner as the North Sea Project. Each user, or group of users was provided with a suitably configured account on the Bidston IBM. It should be noted that this was done in such a way that BOFS users were excluded from North Sea Project data and vice versa.

User education again precisely paralleled the North Sea Project with a detailed Users' Guide and a 3-day course held at Bidston. This was run in February 1991 and was more heavily attended than its predecessor with over 20 scientists participating.

One final parallel between the two databases is that the BOFS database is also regularly used by scientists to satisfy their data requirements.

4 An Objective Assessment

There are a number of questions which one may ask of a data management exercise to assess whether it may be deemed a success.
a) What products have been produced?
b) Have the products been produced cost effectively?
c) Have the products been made available on a reasonable timescale?
d) Are the products actively used by the scientific community?
e) Have the objectives of the data manager been satisfied?

4.1 What Products Have Been Produced?

In the case of BOFS and the North Sea Project, the data products are obviously the two on-line databases that have been assembled and made available to the project scientists. The current data holdings in each of these databases are listed in Appendix 1.

4.2 Have the Products Been Produced Cost Effectively?

In order to answer this question, we must first establish some measure of cost effectiveness. From the BODC point of view cost effectiveness is judged by the quantity and quality of the data set assembled costed against the manpower required to produce it.

During the 1980s MIAS (the precursor to BODC) was funded to assemble a data set of CTD data collected by UK laboratories. This was a typical example of conventional data archaeology and therefore provides a useful comparison. The CTD project was operated as follows. First, an inventory was compiled from sources such as ROSCOP forms. The scientists who collected the data were then approached and asked to submit their data to MIAS.

The data that were submitted were converted into a common format, units standardised (including calibration work), screened on a graphics workstation and loaded into the national data archive together with qualifying documentation.

The resources available for this project were 2.4 man years of data centre staff effort. At the end of the project, between 6,000 and 7,000 CTD casts had been added to the national data archive. The figure for resources only includes data centre effort. An unspecified, but significant, amount of scientist's time had been spent working up the data prior to their submission to MIAS.

Compare these figures with the resources required to assemble the BOFS and North Sea Project data sets described in Appendix 1. To date, 3 man years of BODC staff effort have been expended on each project supported by 4 man years of industrial training student (undergraduate students whose course includes one year working in industry or a scientific institute) support. These figures include all project software development and data processing as well as the loading of the data into the databases but make no allowance for the BODC infrastructure, such as existing software systems, which was used as extensively as possible.

These figures clearly show that the data management strategy adopted by BODC for the Community Research Projects considerably outperforms in cost effectiveness terms the data archaeology approach, particularly if the scientist's time saved by BODC's data processing efforts is taken into account.

4.3 Have the Products Been Made Available on a Reasonable Timescale?

Once again a data management project previously undertaken by MIAS may be used for comparison. The JASIN project in 1978, JASIN78, was an intensive multidisciplinary field study involving several research vessels off the NW coast of Scotland. The assembly of the data set for this project was the responsibility of MIAS. Project scientists were given a clear mandate to submit their data once they were worked up. It took between 4 and 5 years for these data submissions to arrive at the data centre.

In contrast to this the BOFS database containing the data from the field seasons in 1989 and 1990 was available 8 months after the completion of the 1990 field season. However, prior to the launch of the database BODC were servicing requests for data from the partially built database and the disk files awaiting load. The underway and CTD data for the 1991 field season (in July) were available for a workshop in early December 1991 and the loading of the 1991 sample data should be completed by February 1992.

The timescale for the North Sea Project was even shorter. The database, containing all the automatically logged data and 75 per cent of the sample data set, was opened to the community only 5 months after the last cruise of the main data collection phase docked. Again access to the data were provided by a request service until the database was released. Indeed, much of the delay in bringing the BOFS database on line was due to resources being tied up servicing North Sea Project requests.

4.4 Are the Products Actively Used by the Scientific Community?

An online database is only of value if it is actively used by the scientific community for which it was provided. Usage of the BOFS and North Sea Project (NSP) databases have been monitored by BODC and the results are presented in the table below:

Month	Number of users		Number of database sessions	
	NSP	BOF	NSP	BOFS
May 90	5		35	
Jun 90	5		37	
Jul 90	5		126	
Aug 90	7		104	
Sep 90	5		69	
Oct 90	4		35	
Nov 90	7		113	
Dec 90	6		97	
Jan 91	9		67	
Feb 91	9	14	79	79
Mar 91	6	6	28	25
Apr 91	7	5	114	24
May 91	7	2	82	3
Jun 91	7	4	54	17
Jul 91	9	4	144	118
Aug 91	6	4	53	31
Sep 91	6	3	37	14
Oct 91	6	1	35	5
Nov 91	7	5	49	28
Dec 91	6	3	72	25

The number of users is the number of different project scientist accounts which have activated the database login macro during the month. The number of database sessions is the number of times the database login macro has been invoked during the month. These figures exclude the activities of BODC staff and accesses to the database using the BODC software interface which cannot be monitored for technical reasons.

These figures demonstrate that both databases are in regular use. The patterns of usage show a marked contrast between the two project communities. Much of the North Sea database access is by a group of 4 or 5 users who regularly log onto the database indicating that they are using the database as a data analysis tool.

However, in the case of BOFS there is very little overlap in the user community from month to month. This indicates that users log onto the database and siphon the data they require into their home system for analysis. Contact with the BOFS community also indicates that they use the interface software in preference to native SQL and consequently the figures above significantly underestimate the usage of the BOFS database.

One can only speculate as to why this difference in database usage patterns has developed. Possibly it is because the BOFS community contains a higher proportion of scientists whose computing is based on PC packages and hence are uncomfortable in a mainframe environment.

4.5 Have the Objectives of the Data Manager Been Satisfied?

In the introduction to this paper it was stated that the objective of the scientific data manager was the provision of data to the scientist. In this, the exercise has definitely succeeded as all the project scientists know they can look to BODC for the data, and supporting information, they require. In terms of BODC's objectives both projects can be seen to have succeeded. The provision of data management support has been clearly demonstrated. Furthermore, once organised into in house project databases, the data may easily be incorporated into the national data archive.

In summary, by adopting a different approach to data management BODC now has two large multidisciplinary data sets in-house months rather than years after they were collected.

5 The Implications for Large-Scale Data Management

5.1 Data Management in JGOFS

Let us start this discussion by looking at data management within JGOFS for the data collected during the North Atlantic Bloom Experiment (NABE) in 1989. During this experiment research vessels from the USA, Canada, Germany, the Netherlands and the UK undertook a coordinated study along the 20W line northwards from 47N during the 1989 spring bloom.

JGOFS data management representatives for the NABE first met at Kiel in March 1990. Here, a model was adopted whereby project data management would be achieved by mutual exchange between the national data management representatives who would then service the needs of their respective scientific communities.

This has met with limited success. From the UK viewpoint, the American data are available as a comprehensive data report, including full documentation. backed up by data in machine readable form on floppy disk. The Canadian data are available as a set of floppy disks containing a set of Lotus spreadsheets with no accompanying documentation. The Dutch data are available as hardcopy listings without documentation or data in machine readable form. No readily identifiable German data set is available.

The UK data are currently available to the other NABE participants via a request service operated by BODC. Requests for data are submitted via the other national data management representatives and the data supplied on floppy disk, tape or as networked files. Data have been supplied in this way to scientists in the USA. Canada, Germany and the Netherlands with a turnaround of the order of a couple of weeks.

What has gone wrong? The answer is simply that data management in JGOFS has been grossly under resourced. Of the five nations participating in the NABE only the UK and USA had dedicated data management resources available in 1989. The Canadian and Dutch data sets were assembled by scientists taking on the data management responsibility in addition to their research work. In Germany, a data management project was submitted for funding in competition with research projects and failed. However, some resources have subsequently (in 1991) been designated to the assembly of the German NABE data set at the German data centre.

Even where resources have been allocated. they are insufficient to give JGOFS data management the support it deserves. In the case of the UK, the resources were allocated to support BOFS and therefore the processing of the data from the BOFS cruises in 1989. 1990 and 1991 and its assembly into an online database has had to be the first priority. Finding the resources required to support JGOFS i.e. by completing the task and assembling the necessary data documentation for use by scientists not involved in the field programme, has proved problematical.

The question of how the data management of large multi-national projects should be approached must therefore be addressed.

5.2 The Centralised Model

The data management approach adopted by BODC for the North Sea Project and BOFS has worked. The question is whether such an approach could be extended to a large multi-national project such as JGOFS. The answer would have to be no for the following reasons, mainly associated with problems of scale.

a) BODC's success in the UK has resulted largely from the strong working relationship developed between the project scientists and data centre staff. This has been achieved by data centre staff visiting laboratories, attending scientific meetings and participating in research cruises. Whilst this is feasible within the geographical confines of the UK, if attempted on an international scale more time would be spent traveling than working on the data.

b) The NERC research vessels utilise common hardware and data logging software and are operated by a single organisation. The interface between the research vessels and the BODC data processing system was therefore relatively easy to set up and operate. Even so, problems were encountered where individuals adopted different working practices at sea. Such problems would inevitably be magnified in a scenario where the research vessels were independently managed with incompatible hardware and software systems.

c) It would be difficult, if not impossible, to grant access to any database on equal terms for all the contributing nations. In particular, a centralised data centre inevitably requires a host nation whose scientists would be perceived as being in a privileged position.

5.3 The Distributed Model

Let us now consider the opposite extreme of totally distributed data management. In this case, each Principal Investigator throughout the project is responsible for an individual data set. The data sets would usually either be a single set of sample measurements or the data logged by an individual instrument and would be stored under the PIs home hardware and software environment.

These isolated data sets are then linked and opened to other project scientists by a distributed data management software system. Two systems of this type are known to the author: the system set up in the USA for US JGOFS data management (Flierl, 1992) and the system set up at Lamont for managing Earth Science data (Menke et al, 1991).

Both of these systems are based upon the interrogation of databases fronted by software interfaces using a standardised protocol over a network. The user phrases a request for data. Software on his system consults a directory, locates the required data and submits a request to the system holding the data phrased in the designated protocol. This message is interpreted by the interface software and the requested data returned to the user's system, again in a standardised format.

At first glance, this seems an ideal system. Certainly, from the data requester's point of view it would seem to satisfy all his requirements. However, there are problems associated with such a scenario:

a) Both systems described above are based upon Unix platforms connected via LANs or Internet. In order to provide data management support on an international basis all scientists would require access to such platforms, or at least to platforms operating some analogue of Unix daemons.

b) Somebody has to code the necessary software interfaces. Considering the range of platforms and local data management strategies involved, this is no small task though the work involved would be reduced if PIs could be persuaded to standardise their local data management strategies. The question of who would undertake this work would also need to be addressed. It is doubtful whether all PIs would be willing, or even capable, of undertaking such software development.

As well as software interfaces, a directory needs to be maintained to tell the system where the individual components of the data are located. Presumably this would be undertaken by the PIs using directory maintenance utilities. These would be required for a range of platforms further increasing the software development requirement.

c) Responsibility for the data has been delegated beyond the limits where it can reasonably be expected to be maintained. The primary responsibility of individual PIs is to their own research: services to the wider community are the results of good will.

This has a number of consequences. First, no-one assumes responsibility for forcing related data into a unified space time coordinate system. The discussion above clearly illustrates how this can give rise to problems: when the data are brought together from distributed sources they may not fit together. The scientist requesting data therefore finds himself with an unexpected workload if he is to produce a merged data set from its component parts.

Secondly, no-one assumes responsibility for guaranteeing availability of the data until they are no longer required. It is easy to imagine a PI running short of disk space on his home system and wiping data which he has finished with but are still required by others working on the project.

Thirdly, no-one takes responsibility for ensuring that consistency is maintained within the distributed data set. Thus, it would be perfectly possible for a request for chlorophyll-a data to return HPLC, fluorometric and spectrophotometric data: all would be perceived as chlorophyll-a by the scientist who collected the data.

Finally, no-one takes responsibility to ensure that the requesting scientist receives all the qualifying information required to ensure that the data are fit for his purposes.

5.4 A Way Forward

The previous section may appear damning. However. all of the criticisms result from the extension of distribution to the individual PI level. Consider a scenario of project data management by a number of small topical data centres, each responsible for assembling an online database from a clearly defined subset of the data collected during the project.

Linking these together with an object oriented distributed database management system provides, in the author's view, one the most exciting ways forward in international data management.

6 Conclusions

The main aim of this paper has been to describe the workings of a topical data centre: a small group of data management professionals working symbiotically on a project with the research scientists.

In the case studies presented, a small team within BODC have achieved what can only be described as a major data management success. The scientists have benefited both from the removal of a significant data processing burden and in increased data availability. BODC has benefited by obtaining a far more complete contribution to the national data archive than would otherwise have been possible and has acquired these data on a very short timescale.

The topical data centre model shows great promise for large scale project data management, particularly when interfaced to current developments in distributed database technology. However, if the data management of any project is to be successful, adequate resources specifically targeted at data management must be made available.

Acknowledgments

This work described in this paper was only possible by the dedication and sheer hard work of a number of individuals. Polly Machin (BOFS) and Ray Cramer (North Sea Project) have made significant contributions as have the students (Mike Jones, Jeremy Ashley, Bill Cave, Steve Ng, Pete Brocklehurst, Mark Bell, Gareth Trevor and Andy Spiller) who have spent some of their time at BODC on the projects.

Mairi Marshall accurately keyed in numerous data sets submitted on paper. Other colleagues in BODC, particularly Lesley Rickards and Steve Loch, were always willing to lend a hand when the going got tough. Dave Neave taught SQL on both database courses.

I thank the director of BODC, Meirion Jones, for his sound advice and undying support. Last but not least, I thank the project scientists for their assistance and cooperation. without which the data management initiative would surely have failed.

References

Flierl, G. (1992) Data Management for JGOFS: Theory and Design. These proceedings.

Loch, S.G. (in prep) An Efficient, Generalised Approach to Banking Oceanographic Data.

Menke et al (1991). Sharing Data Over Internet with the Lamont View-Server System. *EOS* 72, pp 409-414.

Appendix 1
Data Holdings in the North Sea and BOFS Project Data Bases

North Sea Project Data Summary

70,000 nautical miles of underway data
3,800 CTD casts
10,000 water bottle samples
168 production experiments
749 net hauls
59 core stations

North Sea Project Data in Detail

Underway Data

Underway data for 38 cruises (30 from the main 1988/1989 field season plus 8 related cruises). These contain the following parameters sampled at 30 second intervals:
Thermosalinograph temperature and salinity
Fluorescence (calibrated as chlorophyll for most cruises)
Optical attenuance (calibrated as suspended matter load for most cruises)
Bathymetry

In addition, 20 of the cruises have dissolved oxygen logged at 15 or 5 minute intervals and 8 of the cruises have underway nutrient data at intervals of 30 minutes, I minute or 30 seconds.

In total, there are 1.37 million datacycles in this set which approximate to some 70,000 nautical track miles of data.

CTD Data

The database contains approximately 3,800 CTD casts. In addition to pressure, temperature and salinity each cast includes fluorescence (calibrated as chlorophyll for most cruises), dissolved oxygen and optical attenuance (calibrated as suspended matter load for most cruises) data. Over 3.000 of the casts also include upwelling and downwelling photosynthetically active irradiance data.

Water Bottle Data

The database contains data from over 10,000 water bottle samples. Of these nutrients were done on over 6,000, extracted chlorophylls and suspended matter on nearly 5,000, trace metals on over 1,700 and sulphur compounds and halocarbons on nearly 800.

Production Data

The database contains 154 in-situ and 806 on-deck 14C primary production measurements, 118 thymidine bacterial production measurements and with 72 oxygen production measurements. These represent 144 14C experiments, 15 thymidine experiments and 9 oxygen experiments. each over a range of depths.

Plankton Species Distribution

The abundances of 15 zooplankton classes are held for 749 net hauls. Phytoplankton species distributions are held for 59 stations.

Core Data

The database contains pore water profiles, sedimentology and some nutrient flux data for 59 core stations.

BOFS Data Summary

45,000 nautical miles of underway data
535 CTD casts
1,089 Seasoar profiles
90 XBT profiles
196 production experiments
105 phytoplankton species distributions
89 zooplankton samples
259 cores

BOFS Data in Detail

Underway Data

Underway data for 11 cruises (3 in 1989, 6 in 1990 and 2 in 1991). These contain the following parameters sampled at 30 second or 1 minute intervals:

Thermosalinograph temperature and salinity
Fluorescence (calibrated as chlorophyll)
Optical attenuance
Bathymetry (4 cruises)
Nutrients (7 cruises)
Photosynthetically active radiation
Solar radiation
Wind velocity
Air temperature
Barometric pressure

In addition, on 5 cruises, dissolved oxygen was logged at 5 minute intervals and CO2 parameters (pCO2, TC02, pH and alkalinity) logged approximately every 10 minutes.

In total, there are 0.8 million datacycles in this set which approximate to over 45,000 nautical track miles of data.

CTD and Related Data

The database contains 535 CTD casts. In addition to pressure, temperature and salinity each cast includes fluorescence (calibrated as chlorophyll), and optical attenuance. Over 75 per cent of the casts include upwelling and downwelling photosynthetically active irradiance data and about 50 percent have dissolved oxygen data.

The database also contains Seasoar (a towed undulator measuring temperature, salinity, dissolved oxygen and chlorophyll) from two cruises. These are held in the database as 1,089 pseudo CTD casts on a 4 km spacing loaded from a gridded data set.

Finally, there are 90 XBT profiles held in the database.

Water Bottle Data

The database contains data from over 5,000 water bottle samples. Of these nutrients were done on over 2,700, extracted chlorophylls on over 2,000, DOC and CO2 parameters on over 400, POC/PON on over 500, and pigments by HPLC on over 300.

Production Data

The database contains 508 14C primary production measurements, 539 thymidine bacterial production measurements 134 (many with associated bacterial counts), oxygen production measurements (many with associated TCO, data) and 117 '5N new production measurements. These represent 76 14C experiments, 68 thymidine experiments, 24 '5N experiments and 28 oxygen/TCO, experiments, each over a range of depths.

Plankton Species Distribution

Phytoplankton species distributions are held for 105 samples collected from 17 stations over a range of depths.

Zooplankton Biomass and Grazing

The database holds data for 16 zooplankton grazing experiments, 89 biomass stations and 43 mesozooplankton gut content determinations.

Core Data

The database contains chemical, sedimentological and bioturbation (by radionuclides) data from 259 cores taken at 118 coring stations.

Management and Assimilation of Satellite Data for JGOFS

Robert Evans

Summary

Mr. Evans described the data flow that has been established and noted that understanding this flow was essential to understanding the data base. He noted that the complete suite of sensors as well as data transfers all need to be considered. This includes the process from initial satellite recordings to the final geophysical values that these represent. Further he illustrated the coordination that needs to be established between the satellite and field programs. Using SEAWIFS as an example, the usages and product delivery needs all must be considered. Presumed geophysical values must also undergo a quality assessment that involves in situ air and sea values. finally you wind up with suites of data and data products that are available in time scales from near-real time to months or even years later. It was noted that changes in algorithms, correction factors and calibration require that data be available for reanalysis at later dates.

Using AVHRR as an example, Mr. Evans showed the level of effort that was required in order to build this high quality time series data set, complete with error bars. This is the type of data that is most useful to climate studies but it required working and reworking buoy data used as sea truth and AVHRR data from satellites.

Finally the some of the critical elements required for a successful system were: Timely access, simple mechanisms which allow one to include all partners in a large project, allowance for metadata so users are aware of how product was derived and distributed system interfaces that readily accessible and easy to use.