DATA MANAGEMENT FOR THE OMEX I PROJECT. A CASE STUDY.

Roy K. Lowry

British Oceanographic Data Centre, Bidston Observatory,
Birkenhead, Wirral, Merseyside L43 7RA, UK.

co-authors Zeljko Loncar and Richard Downer

SUMMARY

The British Oceanographic Data Centre (BODC) was contracted in February 1994 to manage the field data collected during the EU Ocean Margin Exchange (OMEX) project. The OMEX I field programme consisted of 47 research cruises undertaken by 17 ships in the period April 1993 to December 1995. Scientists participating in the data collection came from laboratories in Belgium, France, Germany, Ireland, Netherlands, Norway, Portugal, Spain and the UK. The data were highly multidisciplinary covering physical oceanography, marine biology, biogeochemistry, benthic measurements and air/sea interaction studies. Some 612 data sets were collected over the north west European continental shelf edge, primarily in the region of the Goban Spur, south of Porcupine Bank.

The OMEX data management protocol was to first assemble a complete, integrated and documented database for use by the project's scientists and to then make the database available through electronic publication on CD-ROM. The techniques used were developed from those pioneered by BODC for the management of the data from the UK North Sea Project and the Biogeochemical Ocean Flux Study. The project was a success, bringing in some 95% of the data collected within the OMEX I field programme. This paper is a case study describing how this was achieved.

1. INTRODUCTION

The aim of the Ocean Margin EXchange (OMEX) project is to gain a better understanding of the physical, chemical and biological processes occurring at the European ocean margins in order to quantify fluxes of energy and matter across this boundary.

The objective is to provide a more accurate picture of the biogeochemical interactions between the coastal zone and the open ocean. This information is essential for the development of predictive models required to evaluate the response of the shelf and slope area to global environmental changes.

The coastal area, with its enhanced productivity and strong influence from continental input, is an important source of dissolved and particulate matter for the open ocean. On the other hand, deep ocean waters, rich in nutrients and high in dissolved trace elements, are transferred across the shelf edge and help to sustain the high productivity of biota in the coastal zone and shelf seas. The quantification of fluxes across the ocean margins is a fundamental requirement for

the evaluation of the budgets of carbon, nutrients and trace elements between the continents, the coastal zone and the open ocean.

The OMEX project is a major multinational research endeavour carried out within the Marine Science and Technology (MAST) programme of the European Commission. It has been organised into three phases, OMEX I, II/I and II/II, over a seven year period from June 1993 to June 2000 and involves scientists from ten European countries. OMEX I ran from 1 June 1993 to 31 May 1996 and included a major field programme focused primarily on the shelf edge margin south west of Ireland. OMEX II/I ran from 1 June 1996 until 31 May 1997 and was designed to give scientists a funded opportunity to interpret and publish the masses of data collected during OMEX I free from the pressures of additional field campaigns. The only data generated from OMEX II/I were the results of analyses on samples collected during OMEX I that could not be analysed within the resources available for the first phase of the project. The final phase of the project, OMEX II/II, will run for three years from 1 June 1997 and will include an intensive study of the Iberian Margin.

OMEX I involved over 40 Principal Investigators from Belgium, Denmark, France, Germany, Ireland, the Netherlands, Norway, Portugal, Spain and the United Kingdom. The scientists were based in either government research laboratories or universities. The OMEX I field programme was focused on three areas of the European shelf margin in northern Norway, to the south west of Ireland and off Iberia; the main area of the study was around the Goban Spur (see Figure 1). It involved a total of 47 research cruises on ships ranging from fishing boats to large research vessels from 9 European nations. Cruises ranged in duration from a couple of days to several weeks. A list of the cruises is given in Table 1. A wide range of oceanographic measurements was made on these cruises, including meteorology, atmospheric chemistry, hydrography, water column biogeochemistry, water column biology, benthic biology, benthic biogeochemistry and sedimentology. These measurements were made using a vast array of oceanographic hardware including shipboard instrumentation, moored instruments, CTD rigs with water sampling rosettes, nets, plankton samplers, corers, towed fish and benthic landers.

Responsibility of managing the data collected during the OMEX project has been given to the British Oceanographic Data Centre (BODC). This was funded as a supporting initiative (commenced on 1 February 1994) during OMEX I and as a scientific partner in OMEX II. The role of BODC is to provide integrated data management support to OMEX, both for the benefit of the Project's scientists and to ensure the publication of a comprehensive, high quality, fully documented data set at the end of the Project.

BODC has specialised over the past decade in handling the complex data sets collected by multidisciplinary oceanographic research projects. However, we were barely prepared for the scale and complexity of the OMEX I data set. Some considerable thought was given on how to communicate concisely an accurate impression of the scale of the problem to the reader but no suitable words could be found. Instead, the reader is asked to browse the summary of the data set given in Appendix I to comprehend the task that faced the data management team. This paper describes our experiences in bringing together the OMEX I data set.

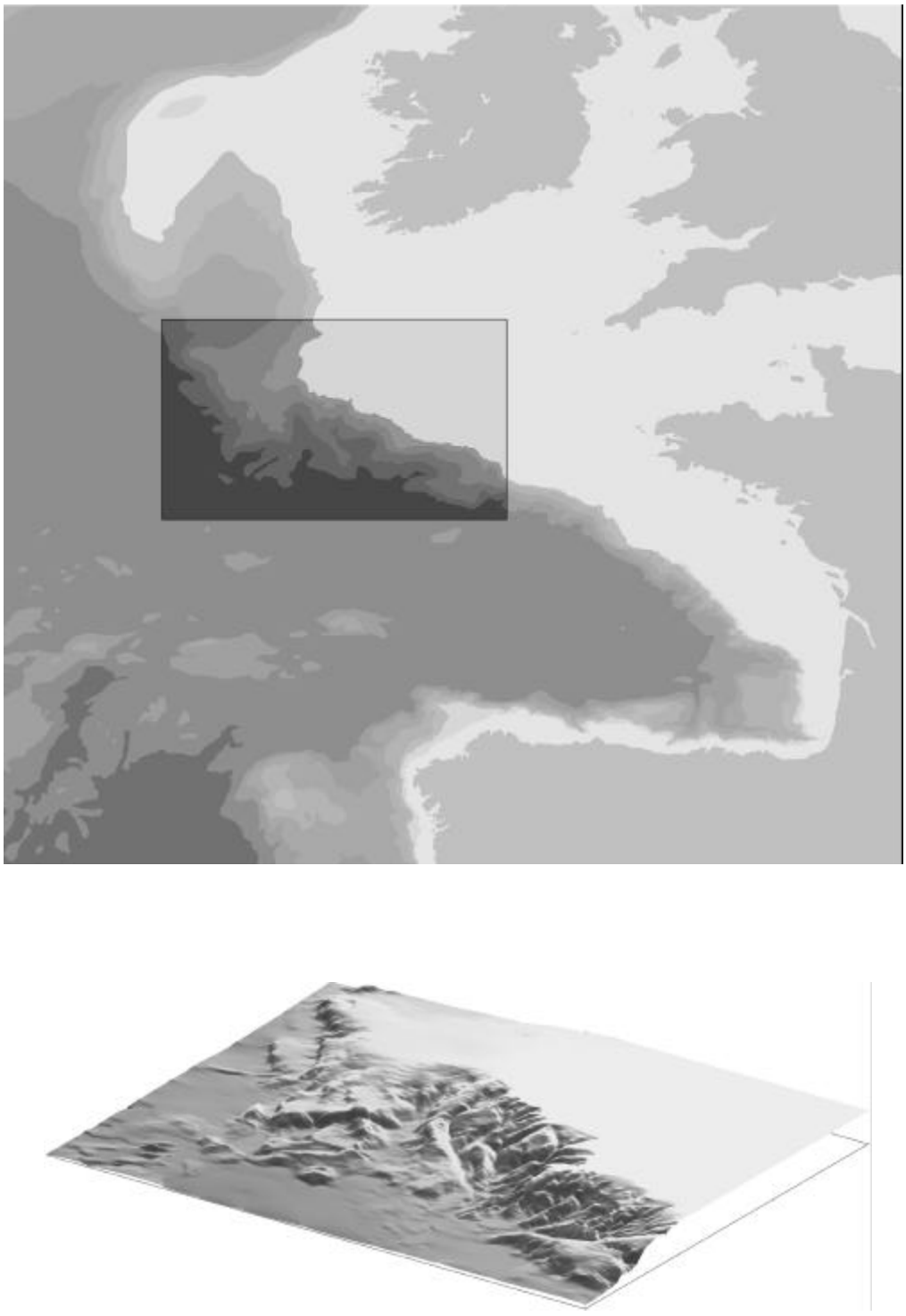

Figure 1. The main area of study for the OMEX I field programme was the region of the Goban Spur south west of Ireland (the illuminated 3D view of the region is at a vertical exaggeration of 35:1)

Ship	Cruise	Chief Scientist	Country	When	Where
Belgica	BG9309	R. Wollast	Belgium	19/04/93 - 06/05/93	Channel/Iberian slope 2E-11W 42-52N
Poseidon	PS200-7	B. von Bodungen	Germany	23/06/93 - 04/07/93	NE Atlantic, 5-15 W 45-50 N
Valdivia	VLD137	T. Raabe	Germany	23/06/93 - 16/07/93	NE Atlantic, 5-15 W 45-50 N
Auriga	Plutur I	J M A Dias	Portugal	26/07/93 - 30/07/93	Shelf off the Tejo submarine delta, 8-10W, 38-39N
Cote d'Aquitaine	NAOX1	J.M. Jouanneau	France	02/09/93 - 05/09/93	La Chapelle Bank, Bay of Biscay, 5-10 W 47-48N
Belgica	BG9322	R. Wollast	Belgium	21/09/93 - 06/10/93	NE Atlantic, 5-15 W 45-50 N
Pelagia	PLG93	W. Helder	Netherlands	11/10/93 - 31/10/93	NE Atlantic, 5-15 W 45-50 N
Auriga	Plutur II	J M A Dias	Portugal	22/11/93 - 03/12/93	Shelf off Tagos & Sado rivers, 8-10W, 38-39N
Charles Darwin	CD83	R.D. Pingree	UK	13/12/93 - 13/01/94	NE Atlantic, 2-35 W 23-52N
Meteor	M27-1	W. Balzer	Germany	29/12/93 - 17/01/94	NE Atlantic, 5-15 W 45-50 N
Charles Darwin	CD84	P.J. Statham	UK	18/01/94 - 02/02/94	NE Atlantic, 5-15 W 45-50 N
Jan Mayen	JM1	K. Tande	Norway	12/03/94 - 16/03/94	NE Norwegian Sea, 15-19 E 69-71N
Charles Darwin	CD85	P. Pugh	UK	08/04/94 - 05/05/94	NE Atlantic, 5-15 W 45-50 N
An Cappall Ban	CAPB1	P. Bowyer	Eire	11/04/94 - 16/04/94	Off Erris Head, 10-12W 54-56N
Jan Mayen	JM2	K. Tande	Norway	11/04/94 - 16/04/94	NE Norwegian Sea, 15-19 E 69-71N
Belgica	BG9412	R. Wollast	Belgium	20/04/94 - 05/05/94	NE Atlantic, 5-15 W 45-50 N
Jan Mayen	JM3	K. Tande	Norway	16/05/94 - 20/05/94	NE Norwegian Sea, 15-19 E 69-71N
Charles Darwin	CD86	T.C.E. van Weering	Netherlands	16/05/94 - 17/06/94	NE Atlantic, 5-15 W 45-50 N
Cote d'Aquitaine	NAOX2	J.M. Jouanneau	France	06/06/94 - 13/06/94	La Chapelle Bank, Bay of Biscay 5-9W 46-48 N
Jan Mayen	JM4	K. Tande	Norway	13/06/94 - 18/06/94	NE Norwegian Sea, 15-19 E 69-71N
Auriga	Plutur III	J. M. A. Dias	Portugal	28/06/94 - 07/07/94	Shelf off Tagos & Sado rivers, 8-10W, 38-39N
Jan Mayen	JM5	K. Tande	Norway	15/07/94 - 20/07/94	NE Norwegian Sea, 15-19 E 69-71N
Jan Mayen	JM6	K. Tande	Norway	08/08/94 - 12/08/94	NE Norwegian Sea, 15-19 E 69-71N
Jan Mayen	JM7	K. Tande	Norway	05/09/94 - 10/09/94	NE Norwegian Sea, 15-19 E 69-71N
Madornina	Monitoring	R. Prego	Spain	09/94 - 04-05/95	Ria of Vigo, 8.5-9.5W, 42-43N

Table 1: Cruises supported during OMEX I data management

Ship	Cruise	Chief Scientist	Country	When	Where
Meteor	M30-1	O. Pfannkuche	Germany	06/09/94 - 20/09/94	Goban Spur, Porcupine, Celtic shelf 11-16 W 38-50 N
Cote d'Aquitaine	NAOX3	J.M. Jouanneau	France	05/10/94 - 11/10/94	Bay of Biscay, 1- 5 W 45-47 N
Jan Mayen	JM8	K. Tande	Norway	10/10/94 - 12/10/94	NE Norwegian Sea, 15-19 E 69-71N
Auriga	Plutur IV	A. Rodrigues	Portugal	23/11/94 - 01/12/94	Shelf off Tagos & Sado rivers, 8-10W, 38-39N
Belgica	BG9506	R. Wollast	Belgium	03/03/95 - 17/03/95	NE Atlantic, 5-15 W 45-50 N
Heincke	HEINK68	M. Kloppmann	Germany	27/03/95 - 17/04/95	Porcupine Bank/SeaBight, 5-15 W 50-55 N
Jan Mayen	JM9	K. Tande	Norway	16/05/95 - 17/05/95	NE Norwegian Sea, 15-19 E 69-71N
Charles Darwin	CD94	P.J. Statham	UK	03/06/95 - 20/06/95	NE Atlantic, 5-15 W 45-50 N
Auriga	Plutur V	A. Rodrigues	Portugal	12/06/95 - 22/06/95	Shelf off Tagos & Sado rivers, 8-10W, 38-39N
Valdivia	VLD153	Mohn	Germany	24/06/95 - 13/07/95	Porcupine Bank/SeaBight, 5-15 W 50-55 N
Jan Mayen	JM10	K. Tande	Norway	25/06/95 - 01/07/95	NE Norwegian Sea, 15-19 E 69-71N
Valdivia	VLD154	A. Spitzy	Germany	14/07/95 - 30/07/95	NE Atlantic, 5-15 W 45-50 N
Pelagia	PLG95A	P. de Wilde	Netherlands	14/08/95 - 05/09/95	NE Atlantic, 5-15 W 45-50 N
Discovery	DI216	P.J. Statham	UK	26/08/95 - 12/09/95	NE Atlantic, 5-15 W 45-50 N
Poseidon	PS211	L. Mintrop	Germany	01/09/95 - 10/09/95	NE Atlantic, 5-15 W 45-50 N
Pelagia	PLG95B	T.C.E. van Weering	Netherlands	08/09/95 - 30/09/95	NE Atlantic, 5-15 W 45-50 N
Belgica	BG9521	M. Frankignoulle	Belgium	11/09/95 - 20/09/95	NE Atlantic, 5-15 W 45-50 N
Jan Mayen	JM11	K. Tande	Norway	19/09/95 - 22/09/95	NE Norwegian Sea, 15-19 E 69-71N
Belgica	BG9522	P. Dauby	Belgium	21/09/95 - 30/09/95	Shelf off Tagos river, 8-10W, 38-39N
Discovery	DI217	R. Lampitt	UK	27/09/95 - 22/10/95	NE Atlantic, 5-15 W 45-50 N
Charles Darwin	CD97	R. D. Pingree	UK	12/10/95 - 06/11/95	NE Atlantic, 5-15 W 45-50 N
Andromeda	Plutur VI	A. Rodrigues	Portugal	20/11/95 - 02/12/95	NE Atlantic, 5-15 W 45-50 N

Table 1(continued). Cruises supported during OMEX I data management

2. DATA MANAGEMENT PHILOSOPHY

The philosophy behind the assembly and management of such a large and complex data set as the one generated by OMEX is worthy of some discussion. There are a number of different strategies that a data management operation can adopt and a range of tools are available with which these strategies may be implemented. The first problem that must be addressed by any data management project is therefore strategy selection, closely followed by an implementation plan.

The easiest strategy, from a data manager's point of view, is to take the 'box of floppy disks' approach. Each data set arrives at the data centre as a floppy disk (or some more suitable medium for larger volume data). Upon arrival, it is secured through routine backup procedures and catalogued. The primary function of the data management operation in this case is the cataloguing of the data. Much can be hidden in the meaning of a word and the word 'cataloguing' here is a case in point. The minimal meaning, if the data management operation is to have any value at all, would be to have an entry for each data set specifying what was measured and the cruise (or cruises) on which the data were collected.

Let us consider the application of this approach to a project of the scale and complexity of OMEX. There are two acid tests of any data management operation. First, can a user from outside the project ascertain whether the data they require are contained within the project data set? Secondly, if the data are there, are they of any use? Let us apply these two tests to the simple 'box of floppy disks' approach.

Users in search of data know which parameters they require, an area of interest and a time window of interest. Ascertaining whether the parameters of interest are present in the database should be relatively straightforward, providing, of course, that they have been described in the catalogue in a manner in which the user can recognise them. Determining whether they are in the area and time window of interest is, to say the least, difficult. With the minimal catalogue approach, the user would either have to have detailed knowledge of the project cruises or ask for all data sets containing the parameters of interest and examine each one to see if it was relevant. However, this problem could be addressed by adding a second catalogue containing basic information on the project cruises. Therefore, our user, albeit with considerable effort, could identify the data sets of interest.

We now need to consider what the user receives in response to a request for one or more data sets from the database. The simple answer is that whatever was supplied by the data originators is supplied to the user. The critical question is whether this would be of any use and the answer is a definite maybe that reveals one of the fundamental problems of data management.

If the data sets are to be of any value to external users, they need to come with complete header information (location, sampling time and the like) and be documented to the extent that the user may determine whether they are 'fit for purpose'. It would also greatly benefit the user if all the data sets containing the same type of data were in a consistent format. Inevitably, the scientific data collected within a large, diverse community has very high entropy and bringing them into the desired low entropy state requires a significant amount of work. This work either has to be done by the data originator or by those undertaking the data management.

It has been mooted many times that the role of data managers is to define standards that must be met by data originators for their data submissions. Packaging data to external specification is probably the least popular task within any scientific community and in our experience the imposition of such a regime dramatically inhibits the flow of data into a data centre. The only way to maintain any supply of data in these circumstances is by providing powerful motivation to the scientists, usually the threat of funding withdrawal, but this inevitably creates a very unpleasant climate and a very negative perception of data management within the scientific community.

The alternative is for the data centre to take whatever is offered by the data originators and then do the work of bringing the data sets supplied up to the required standard. This certainly enhances the flow of data and creates a more constructive working relationship between scientists and data managers. However, the price paid is a significant increase in the workload at the data centre when compared with a simple cataloguing operation. It must be emphasised that this represents a transfer of workload from the scientists to the data centre and not an increase in the overall workload. Data centres are staffed by specialists equipped with specialised tools to manipulate data efficiently. Consequently, when a project is considered as a whole, the result is a decrease in the resources required for data management.

The above points may be illustrated by examples from the BODC sphere of operations. BODC maintains a collection of what are termed 'Special Data Sets'. These are self contained data packages that are held by BODC for copying and distribution to users as required. A catalogue is maintained containing a description of each data set. This is therefore the minimalist data management scenario described above and it works. However, it only works for a relatively small number of large, usually global or basin scale, data sets that are both self contained and fully documented. Any attempt to use this approach for the 600 or so individual data sets collected during OMEX I would result in a confused mess in which nothing could be found.

BODC maintains a UK National Oceanographic Database (UK-NODB) which has been developed over the past 20 years. This comprises a large number of data files, each containing a 'series' of data (e.g. a CTD cast), that have been converted into a common format, quality controlled and documented by BODC. These files are indexed by a catalogue held in a relational database management system. This is a much more sophisticated entity than a simple data set catalogue. Complete spatial and temporal information, parameter set definitions, linkages to data activities (e.g. cruises), scientific projects and much more are held and may be queried. This database has been built from data supplied by many different originators in hundreds of different formats.

The UK-NODB is an example of an extremely sophisticated 'box of floppy disks' that is powerfully indexed so users can find data. The data are in a consistent format with all necessary header data and data documentation so users can use the product delivered with confidence. Both the acid tests of data management have therefore been passed. Could this be used for OMEX?

The answer in general is no. The reason lies in the concept of the data series that is fundamental to the data model underpinning the system. The bulk of the OMEX data map to an extremely large number of small data series. Each data series handled through the UK-NODB system incurs computing and, more significantly, labour overheads. With the number

of series involved in OMEX, the overheads would build to such a level that available resources would be swamped. Furthermore, the design of the system limits the number of parameters per series to somewhere between 50 and 64 (depending upon the proportion of parameters that have associated quality control flags) and some OMEX samples have over 100 parameters measured on them. However, some of the OMEX data, such as moored instrument data, play to the strengths of the UK-NODB system and it has been used to good effect in the management of these data.

The role computing technology has to play in the enhancement of 'floppy disk box' databases is worthy of some examination. Of particular interest are Object Oriented Databases (OODBs). In simple terms, these package data sets as objects that comprise the data themselves linked with software (termed methods) that gives the data a standardised appearance to the retrieval interface. Object dictionaries are then built that can be developed to the extent that anything contained within the stored objects may be located and retrieved. Formatting differences in the object data are eliminated by the methods and additional information, such as data documentation, may be linked to the data through the method. Our criteria for successful data management are therefore satisfied.

The data for US-JGOFS are managed with a great deal of success using an object oriented data system developed by Massachusetts Institute of Technology and Woods Hole Oceanographic Institution (see URL http://usjgofs.whoi.edu/jgdms_info.html). This provides the user with data access and data manipulation and display tools interfaced on any platform through a Web browser. The user is initially presented with a list of cruises. Once a cruise is selected, the objects for that cruise are listed as a series of hypertext links that cause the data to be served by the appropriate method. For a user to locate the data of interest a degree of knowledge about the cruises is required, but this information is available with the data including a full station listing.

The main advantage of the US-JGOFS system is that it may be implemented as a distributed system that may be accessed from any platform. This allows data originators to post and maintain 'provisional' data sets on their home systems bringing the data into the project shared domain much sooner than would otherwise be possible. However, a strong word of caution is required here. Data distribution must be done in a managed environment with the long term safety of the objects assured. Otherwise, the potential for disaster is enormous.

The main disadvantage of the system is that the access path currently provided to the data is limited. For example, it is not possible to obtain answers to questions like 'Are there any nutrient data in my area of interest?' without visiting a large number of Web pages. The larger the database, the greater this problem becomes. However, the system is under continued development and certainly has the potential for the implementation of more effective query mechanisms.

In the context of OMEX data management, the system was not considered for two reasons. First, BODC did not have the confidence that a community of the size and diversity of OMEX could be educated or motivated to deliver data packaged to the standard that would be required. Secondly, BODC's experience lies in other technologies and it seemed prudent to remain with tried and tested systems when embarking on such a large and complex project.

Commercial OODB systems have recently appeared on the market. The emphasis within these is in the provision of extremely powerful object dictionaries including full spatial and temporal indexing of the data within the objects. The potential for this technology in marine data management is beyond question. These systems are currently under investigation within BODC and are likely to be adopted in the future once the technology has achieved commercial maturity.

There is one final point to make about OODB systems. There is a great danger for them to be considered as electronic laundries where sundry items of data in any old form are thrown to be turned as if by magic into integrated data sets. This is not true. Objects loaded into the database must conform to certain standards. For example, if a data originator forgets to include vital information such as where and when the data were collected, the technology can never save the day. Therefore, whilst formatting becomes irrelevant, beyond the resources required for method development, the information content of objects remains critical.

So far in this section, data management strategies and systems that were not adopted for the OMEX I data management project have been discussed at some length. However, it is now time to focus on the technology and strategy that was adopted and some of the reasons behind this choice.

In 1988, BODC (or MIAS Data Banking Section as it was then known) was asked to provide data management support to the NERC North Sea Project. This represented a radical change in the expectations of the data management operation. Up until this time the group's activities had focused on the acquisition of a very limited range of data types, primarily moored instrument data and CTD profiles, some considerable time after they were collected when the data originator deemed that they were 'finished with' and could be archived. The North Sea Project included many additional types of data including bottle data, benthic data and atmospheric chemistry. Further, BODC were expected to provide the primary vehicle for data exchange within the project whilst it was current. To achieve this, the time scale for data submission to BODC had to be drastically reduced.

These requirements were achieved through the adoption of two novel strategies. The first of these was the development of a close working relationship with the project scientific community through mutual co-operation in the calibration and processing of the data logged by the ship's computers. The working protocols and resulting data management benefits have been described in detail elsewhere (Lowry, 1992; Lowry 1995).

The second strategy was the application of relational database technology to the management of project databases. A database supporting active science in near real time needs to be flexible. Relational databases have developed an unjustified reputation for being large, inflexible, and in need of years of design effort. In fact, modern RDBMS software allows dynamic restructuring of databases even when significant quantities of data are loaded making them an ideal project support tool. The concept of dynamic database development in this context is discussed further in Lowry and Cramer (1995). BODC has been using relational technology in this way for nearly a decade and an interesting observation is that the rate of change in the database structure has decreased dramatically over the years. In other words, the database structure has evolved to what may be regarded as a generalised structure for the storage of multidisciplinary oceanographic data.

During the early 1980s, these strategies were developed and extended to support the Biogeochemical Ocean Flux Study (BOFS), a UK contribution to JGOFS. The BOFS data set was significantly more diverse than the data handled for the North Sea Project, particularly in the scope of the biogeochemical parameters handled. In addition, the increased data volumes associated with deep CTD casts (5,000m as opposed to 100m) and much longer cruises (up to 3 months as opposed to 15 days) presented fresh challenges to the BODC systems.

When the opportunity arose to provide data management support to OMEX I three challenges were presented. First, the success of the North Sea Project and BOFS data management initiatives was largely due to the working relationship established between BODC and the scientific community. Establishing this relationship within the UK was a challenge in itself. Establishing it on a pan-European scale was, to say the least, a daunting prospect. Secondly, the UK projects had been concerned with research vessels operated by a single organisation with common oceanographic hardware and computer systems. OMEX I involved the research vessels of 9 nations, each with their own systems and operating protocols. Thirdly, the data types handled within OMEX were even more diverse than those encountered in BOFS. For example, one partner was measuring up to 100 chemical parameters on a single water sample.

After due consideration, it was concluded that the systems and strategies in place at BODC could either cope, or could be extended to cope, with the challenges of OMEX I data management. History has proved this conclusion to be correct. Detailed descriptions of how this was achieved in practice have either been presented previously (Lowry, 1995) or are included in this paper.

What this decision meant in practice was that BODC was to engage in a venture where the OMEX data supplied would be pulled apart and reassembled into a single integrated database structure, fully covered by data documentation. There are three unavoidable consequences of this. First, the effort required to do the job is at least an order of magnitude greater than would be required for the acquisition and cataloguing of data sets. Secondly, the work has to be done by scientifically qualified staff who must either understand, or develop an understanding, of the data they are handling. The consequences of using technicians or computer scientists to do this kind of work would inevitably be total disaster. Thirdly, there is an enormous potential for error in this kind of work. If there is to be any confidence in the result, the completed database must be carefully audited before release which further adds to the amount of skilled resources required.

Such an investment requires justification and the simple answer is that it lies in the quality of the product that is produced at a small fraction of the cost of data collection and with a minimal burden placed on the scientific community. The data in the resulting product are covered at the spatial and parameter level by a relational index. Therefore, any data may be located with ease despite the size and complexity of the data set. However, the overriding argument is that the data may be retrieved in a fully integrated form using relatively simple delivery mechanisms accompanied by full data documentation. This guarantees their value to oceanographers for years to come and eliminates the need for costly data archaeology exercises in the future.

3. STRUCTURE OF THE DATABASE

The bulk of the OMEX data is stored in an Oracle relational database. However, two other storage strategies, largely based on file storage, are used for storing high volume data. In this section a brief description of these alternative strategies is given together with an example of the relational structures used in the Oracle database. A complete list of the tables in the database, together with an indication of their purpose, is given in Appendix 2.

3.1 Underway data

Many research vessels are equipped with systems that continuously log navigation and oceanographic instrumentation as the ship is underway. These are known to BODC as the 'underway data'. In addition to navigation, the underway data set typically comprises bathymetry, thermosalinograph, fluorometer, transmissometer and meteorological data but may include a number of chemical parameters as well. The data are typically logged at intervals ranging from 30 seconds to 10 minutes.

These are high volume data sets. A typical cruise generates between 40,000 and 80,000 records of data with up to 30 parameters giving a binary data volume of between 2 and 10 megabytes. The problems of handling underway data, including quality control and calibration, were first addressed by BODC in the late 1980s. At this time, our available relational database technology was Oracle running on an IBM 4381 mainframe that delivered less than 0.5 gigabytes for the total site data area. This, combined with the fact that Oracle uses a relatively inefficient storage format (2-4 times less efficient than binary), meant that incorporating the underway data into Oracle was not a viable option.

BODC's UK-NODB was, and still is, based on a binary format known as PXF. This was considered for the storage of underway data, but was rejected because it has a complex structure that imposes a considerable software development overhead. In 1988, BODC were faced with the situation of having to develop a complete underway data processing system to support the North Sea Project on a time scale of between 3 and 6 months. Consequently, it was decided to develop a simple binary format, known as binary merge format, into which underway data from a variety of sources could be merged, processed, quality controlled and calibrated.

The structure of binary merge format is described in the on line documentation that may be found under URL http://www.pol.ac.uk/bodc/omex/omexman.html and therefore is not described here. The format has a number of advantages. For example, it includes internal locks to prevent processing programs (such as calibrations) being run against the file more than once and history flags to indicate what processing has been done to the data in the file.

However, the real strength of the format lies not in its structure, but in the power of the software interface that has been built around it. In addition to data manipulation tools, the interface includes an extremely flexible data retrieval interface that can produce ASCII listings of the complete file or virtually any subset, including dynamically generated averages. Further interface programs provide integrated output from binary merge files and data, such as bottle data and CTD data, held in the Oracle database.

Data management using binary merge format couldn't be simpler. Data from all cruises are held in a single directory under the project master id with filenames of the form binmerge.cruise_mnemonic.

The binary merge format was conceived as a quick fix to allow BODC to provide data processing support to the underway data collected during the North Sea Project. However, it has subsequently been pressed into service for BOFS, OMEX, LOIS and other smaller data management projects supported by BODC. We have now reached a stage where we hold over 0.5 gigabytes of underway data from well over a hundred cruises in binary merge format.

Inevitably, the format is starting to display weaknesses resulting from its usage being extended way beyond the purpose for which it was originally designed. In particular, the single byte parameter codes on which the format is based are becoming strained to the limit with very few of the 200 or so available symbols remaining unused. It is anticipated that the format will be able to cope with OMEX II data management but the time will soon come to consider a replacement. This is not a task to be undertaken lightly due to the large amount of software modification that will be required. Thought is currently being given to this problem and several options, including standard formats like netCDF, relational and object-oriented storage are being considered with implementation proposed by the end of the decade.

3.2 Normalised relational data structures

The normalised three-level hierarchical relational data structure used by BODC has been summarised elsewhere (Lowry 1995). However it is so fundamental to the data management operation that it is worth examining in more detail. The data model is founded on the concept of data collection 'events' that are defined as any activity during a cruise that results in the generation of data. The term event therefore covers anything from deploying a mooring through CTD casts to someone opening a tap to take a surface sea water sample. The structure is implemented in three variants handling bottle (water or air samples), sediment trap and benthic data. The latter, the most complicated case involving a fourth hierarchical level, is described here.

The EVENT table forms the top level of the hierarchy and, in this example, contains a record for each corer deployment with fields giving the sampling location, time of sampling, cruise from which the corer was deployed and the type of corer used. Below this is the COREINDX table. The sole function of this is to implement the one to many relationship between corer deployments and cores. This allows the system to take account of sub-cores taken from box cores and for replicate determinations on the separate core tubes from a multicorer.

Whole core data are held in table CORETOT that has the standard system data structure containing a core reference, a parameter code, parameter value, parameter flag and a reference to the data originator. There is a one to many relationship between CORETOT and COREINDX with one record in CORETOT for each parameter measured on the core represented by the COREINDX record.

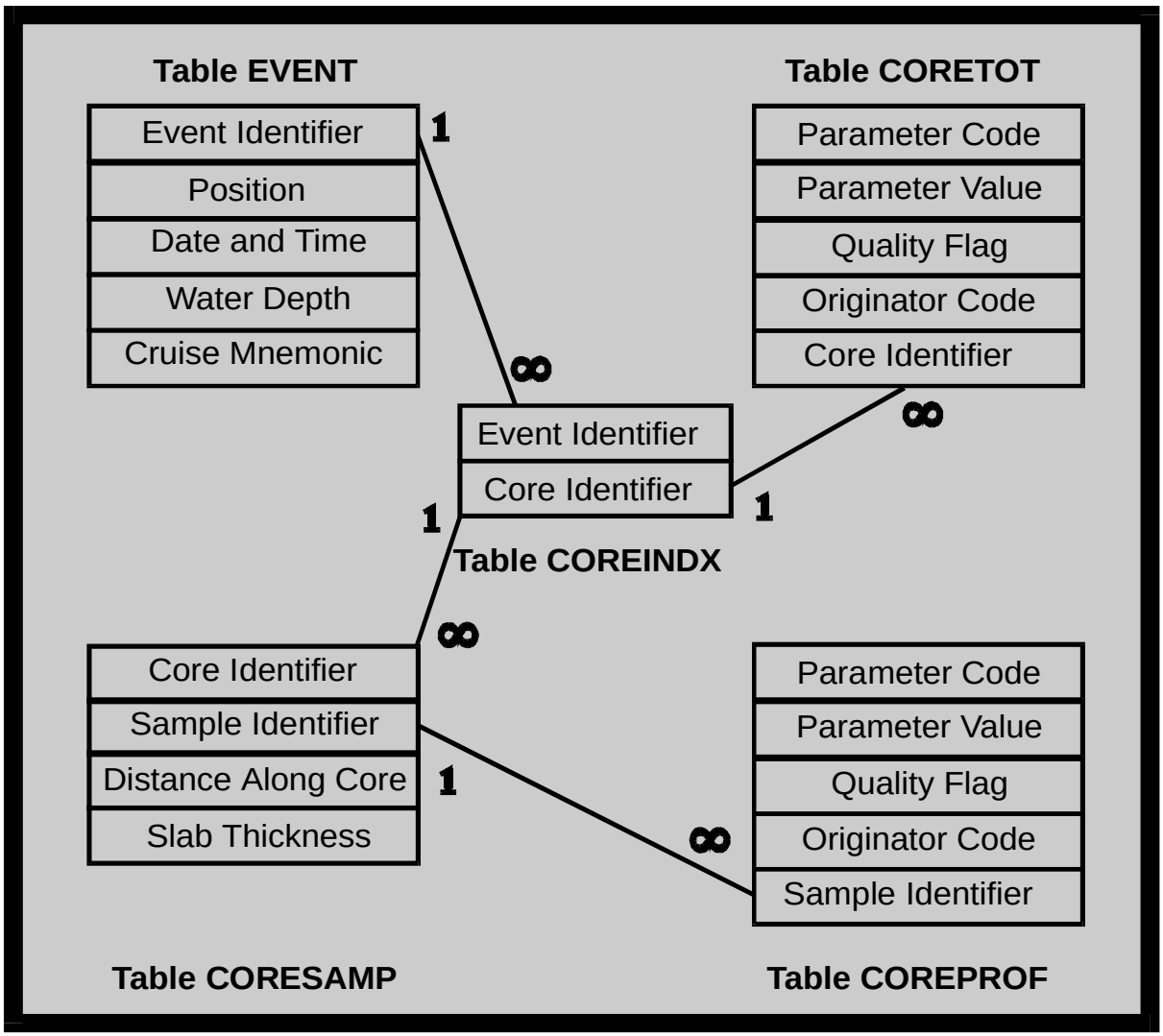

Figure 2. Schematic diagram of the structures used to store benthic data

However, most benthic data are profiles measured along the core and not determinations on the whole core. The profile may be considered as a series of slab samples that have attributes of thickness and distance from the sediment water interface. Point profiles, such as oxygen probe measurements, simply have a thickness attribute of zero. These attributes are held in the table CORESAMP. The parameter determinations along the profile are stored in table COREPROF. This has an identical structure to CORETOT except that it has a linkage to a single profile point in CORESAMP instead of an entire core in COREINDX. The tables and the relationships between them are represented schematically in Figure 2.

3.3 The parameter dictionary

The fundamental principle underpinning the relational data structures is that each measurement in the database is labelled with a parameter code that specifies what has been measured. Within OMEX, over 750 different parameter codes have been used. The definition of what is meant by each of these codes is managed through a set of tables known as the parameter dictionary.

Each parameter code is 8 bytes long and may most easily be considered as having two 4 byte fields, termed the parameter descriptor and the method descriptor. There is a one to many relationship between these fields with each parameter descriptor owning many method descriptors.

This relationship is exploited in more than one way, depending upon the nature of the information that is being coded. In most cases, the parameter descriptor describes the parameter whilst the method descriptor provides information on how that parameter was measured. Consider the following example of some of the codes for phosphate:

PHOSMATX	Manual analysis on unfiltered water
PHOSAATX	Autoanalysis on unfiltered water
PHOSAAD1	Autoanalysis on GF/F filtered water
PHOSAAD3	Autoanalysis on GF/C filtered water

For biological species coding, the one to many relationship between the two elements of the parameter code is exploited in a different way. In this case, the parameter descriptor specifies the genus whilst the method descriptor denotes the species and provides a method code if required. For example the codes for some of the species of Chaetoceros are:

P030M05Z	Chaetoceros breve
P030M21Z	Chaetoceros didymum
P030M56Z	Chaetoceros sociale
P030M91Z	Chaetoceros volans

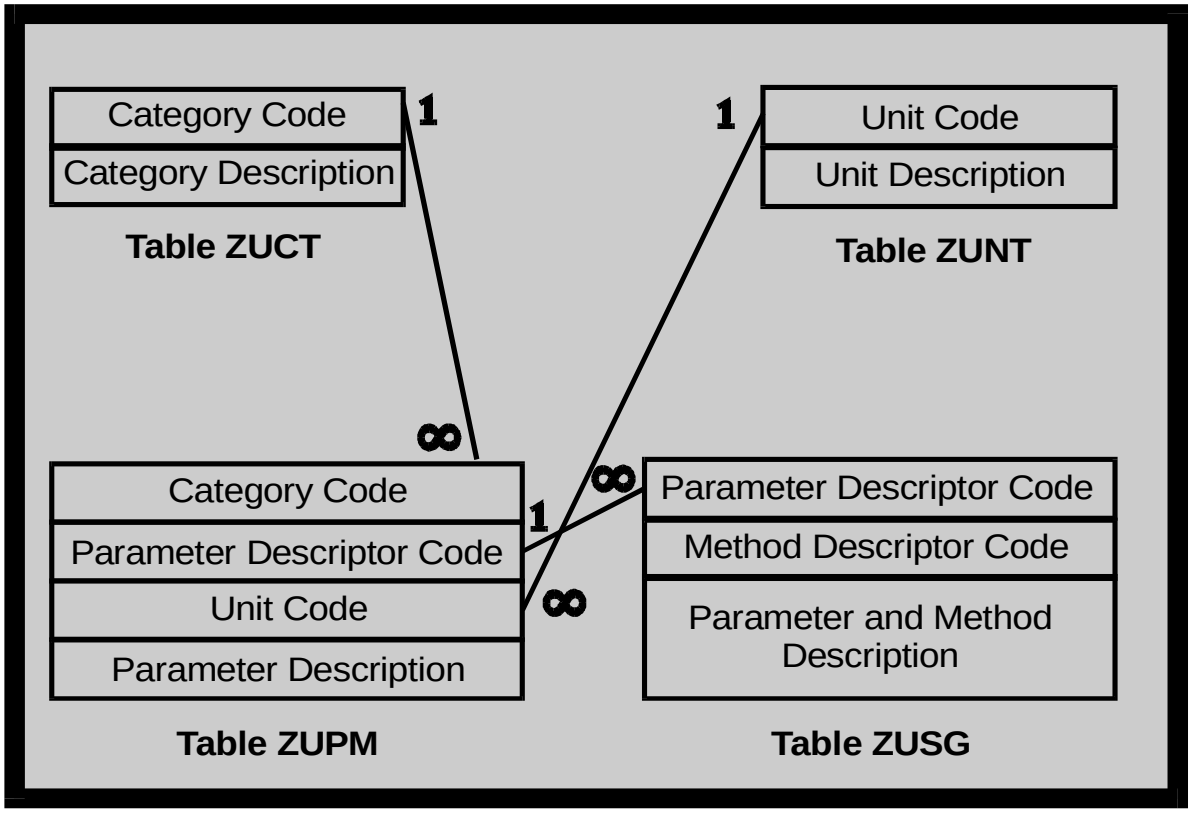

Figure 3. Schematic diagram of the parameter dictionary

The hierarchical nature of the code naturally maps into a relational data structure consisting of two tables (ZUSG and ZUPM). In addition, there are two other tables included in the parameter dictionary. Table ZUCT groups the parameter descriptors into categories and is

simply there to aid dictionary navigation. Table ZUNT is a simple code table for unit definitions. A simplified schematic of the parameter dictionary is shown in Figure 3**.**

It can be seen that each table may be considered as code fields forming primary and foreign keys that implement the logical relationships supported by plain language description fields. The actual table structures are significantly more complex, including management fields such as time stamps, fields to allow implementation of referential integrity checks and fields to support application software such as number of decimal places to be displayed.

The parameter dictionary currently includes over 2700 codes, although not all of these are relevant to OMEX. This level of complexity requires a sophisticated tool to allow the meaning of codes to be ascertained and, more importantly, the code to be located for a given parameter. A bespoke application program (based on wild card searches of either the code fields, the description fields or both) fulfils this purpose with considerable success.

3.4 National Oceanographic Database

The OMEX data set included a number of self-logging instrument deployments, such as current meters and benthic landers. Over the past two decades, the development of the UK National Oceanographic Database (UK-NODB) has been primarily focused on this type of data. In particular, our work on the UK LOIS Shelf Edge Study developed systems that allowed a project database to be 'aware' of a subset of the UK-NODB data holdings. Adopting this technology for OMEX was an easy decision.

The UK-NODB is based on a structure where the bulk of the data are stored as files in BODC's PXF format combined with a relational database that holds the header information and data documentation. This provides a database of unlimited capacity (the files may be kept off line if required) that retains all the search capabilities expected from a relational database. Retrieval utilities are available that generate ASCII listings of the data complete with header information and data documentation.

The project database includes tables that link the available UK-NODB series to entries for mooring deployments in the project EVENT table and provides an inventory of the parameters that have been measured by each instrument deployed. This integration between the UK-NODB and the project databases has proved extremely effective in practice and has obviated the need for considerable additional systems development.

4. DATA TRACKING STRATEGY

4.1 Ascertaining what data have been collected

The data tracking element within oceanographic project data management may be considered as the assembly of a two level information hierarchy. At the top is a list of cruises and below this a list of the data sets collected during each cruise.

The cruise list used by BODC is a simple spreadsheet with fields corresponding to the basic information about the cruise such as the name of the ship, the start and end dates, the principal scientist and the area worked. Compiling this information for the North Sea Project and BOFS was relatively straightforward. In both cases, the field work and science were funded by the

same national organisation. Ship time resources were allocated to the project and this resulted in a series of cruises being clearly labelled as dedicated to the project.

Putting the OMEX I cruise list together was a little more problematical. Some cruises were specifically mobilised for OMEX, designated as such and were relatively easy to identify. However, a significant number of OMEX scientists occupied berths on an opportunistic basis on cruises that did not have an OMEX designation. Finding out about these was much more difficult and in several cases the first BODC knew that a cruise had collected OMEX data was when the data arrived months or even years after the event. Thus, even the most basic data management requirement necessitated a significant intelligence gathering effort to produce.

The data set list used in OMEX was another simple spreadsheet with fields for the cruise identifier, a description of the data set and whether or not the data had been acquired by BODC. In retrospect the tool would have been more effective had fields for the date when the data set was acquired and the name of the responsible principal investigator been included. These refinements will be introduced for OMEX II data management.

The primary tools used for assembling the data set list were the Cruise Summary Report forms and cruise reports. For clearly designated OMEX cruises, there was a requirement that Cruise Summary Reports be completed and this obligation was generally honoured. However, a surprising number of principal scientists had difficulty completing the form properly. In some cases, vital information was missing. In other cases, a form was completed by each principal investigator for their part of the data. Nevertheless, the forms provided a vital overview of the data collected within days or weeks of the cruise docking and their continued use, together with education on how to fill them in, is to be strongly encouraged.

Cruise reports were received for 29 of the 47 cruises covered by the OMEX data management operation. The missing cruises were either of very short duration or cases where OMEX scientists had occupied opportunistic berths. The quality of the reports received varied from a complete description of the activities undertaken during the cruise, including detailed station listings to sketchy and sometimes inaccurate information on the times and positions of the stations worked.

The importance of cruise reports to project data management cannot be overemphasised. Educating principal scientists in what is required is an important task for data managers. This goes beyond simply laying down specifications. If co-operation is to be received, it is essential that scientists understand the importance of the information they provide through demonstrations of how it may be converted into products that are of use to scientists as well as data managers. The other message that data managers need to get through to scientists is that a rough draft report containing full and accurate information arriving a couple of weeks after the cruise is infinitely more valuable than a glossy printed product arriving twelve months later.

Assembling the data set list for the cruises from which cruise reports were not available had to be approached in a different way. During the OMEX project scientists presented their interim results through workshops and annual reports. All workshops were attended by at least one BODC staff member and copies of the annual reports were supplied by OMEX management to BODC. From these sources, information on the data sets was assembled by what can only be described as an intelligence gathering operation. However, the quality of information obtained

in this way often left much to be desired. For example, in several cases we were able to deduce a superset of data sets and a superset of cruises. We then had to make the assumption that all of the data sets were collected on all of the cruises. This was often invalid which led to the generation of 'phantom' data set entries. Significant effort was wasted in pursuit of non-existent data sets as a result.

4.2 Bringing in the data sets

In all data management initiatives, identifying the data sets to be managed is an important element of the work. However, it pales into insignificance in both importance and difficulty when compared with the problem of persuading principal investigators to submit their data to the data managers.

In OMEX I, the principal investigators had a contractual obligation to supply their data to BODC. However, we hold a strong belief that successful data management is founded on winning the co-operation of the scientific community and that the funding weapon should only be used as a last resort. Putting scientists under heavy pressure results in an unpleasant working climate in which only the absolute minimum is delivered and life is difficult for all concerned. The true art of data management is engineering a situation where the scientific community are motivated into supplying their data willingly.

If scientists are to part with their data happily, it is essential that they are convinced that they have absolutely nothing to lose and have something concrete to gain. Scientists can only believe that they have nothing to lose if the data managers are perceived as totally honest data brokers. This level of trust can only be developed if the data managers are known personally to the scientists.

The initial objective of the OMEX data management operation was therefore to become well known throughout the OMEX community. In previous UK projects, this had been achieved through participation in scientific meetings, participation in research cruises and regular visits to scientists in their home laboratories. In OMEX, the latter two strategies were not generally available. Unsolicited applications for valuable cruise berths were deemed likely to lose friends, not win them and laboratory visits were deemed impractical due to the geographic scale involved in a pan-European project. Consequently, BODC efforts were focused on OMEX meetings and workshops. Presentations were given wherever possible to explain who we were, what our work involved, what we required from scientists and, most importantly, what we could deliver in return. In addition, as much personal contact as possible was established between BODC personnel and the project scientists.

A small number of laboratory visits were made. A great deal of planning was dedicated to these to ensure that they were targeted effectively. Top priority was given to visiting cruise principal scientists with whom contact had not been established at meetings. A small number of visits were also made to scientists identified as having particularly large data holdings. Significant quantities of data were collected on these visits as they naturally brought data delivery to the top of the agenda of those visited.

In addition to engendering trust, the BODC effort was directed at convincing the scientific community that there was much to be gained through working with BODC. It was essential that we clearly demonstrated that our system was a source of data as well as a sink. This was

achieved through provision of data to scientists through both a request desk and an on-line service. A significant proportion of the available resources was directed at maximising the standard of these services which did much to enhance the reputation of BODC.

Two other benefits were provided by BODC to the scientific community. First, BODC has considerable experience in the calibration and quality control of scientific data backed up by powerful software tools. In many cases we were able to calibrate raw data, refine data calibrations, improve the data quality through the flagging of spikes or direct the attention of data originators to undetected problems with their data. Secondly, BODC were able to provide additional data security to those who had supplied their data. On a number of occasions, BODC were able to rescue OMEX scientists from disasters caused by the failure of inadequately backed up hard disks.

BODC provided significant benefits to the OMEX community in return for the submission of data. However, it cannot be stressed too strongly that providing benefits will do nothing to enhance the flow of data into data management unless the scientific community is made aware of them through effective, clear and frequent communication.

Establishing a culture within the project where all the scientists have the best of data submission intentions is only part of the problem. It is a fact of life that many researchers today have numerous problems demanding their attention. Some effort is therefore required to ensure that data submission remains high on each scientist's agenda. A number of techniques were adopted to achieve this during OMEX I.

First and foremost, the opportunity provided by personal contact and presentations at meetings and workshops was used to inform the community about how well, or how badly, it was doing in the submission of data. This was backed up by a mail shot to all principal investigators some 18 months into the project (half way through OMEX I) that included detailed information on which data sets had been submitted. In this way a degree of peer pressure was brought to bear on those scientists who were proving a little slower at data submission than their colleagues.

Approximately nine months before the end of OMEX II/I a series of letters and e-mails were sent to specifically targeted individuals who had still not submitted all of their data. These included detailed descriptions of the data that were expected by BODC and yielded both significant quantities of data and valuable information that allowed a number of 'phantom' data sets to be eliminated from our considerations.

Finally, some six months before the end of OMEX II/I all principal investigators were reminded of their contractual obligations by a circular e-mail from the EU Commission official responsible for the OMEX project. It should be emphasised that this revelation of the 'big stick' was only required to persuade, with success, two out of over forty principal investigators to part with their data. All the others had supplied the bulk of the data for which they were responsible before this e-mail was sent.

The success of the strategy described above exceeded our most optimistic projections at the start of the project. A total of 668 data sets were identified as resulting from work supported by OMEX I and OMEX II/I funding. Of these, 580 were submitted to BODC. Of the remainder, 56 were written off as belonging to non-OMEX participants on OMEX cruises, to

samples collected that could not be processed and to analytical disasters. The remaining 32 included a proportion that were possibly unconfirmed 'phantoms'. In other words, the OMEX data management operation achieved the dramatic success of bringing in at least 94.8% of the available data sets.

5. BUILDING THE OMEX I DATABASE

5.1 Building the index framework

The index framework is the top two levels of the relational database hierarchy. This consists of the primary inventory of data collection events (table EVENT) and the gear-specific indices such as the BOTTLE, COREINDX and STINDX tables. EVENT contains basic temporal and spatial information. The index tables fulfil two functions. First, they implement the one to many relationships between events and data. Secondly, they store gear-specific information such as the depths of water bottles and collection times of individual trap samples.

Building the EVENT table for the OMEX database proved to be much more difficult than expected. For the North Sea Project and BOFS we always had an authoritative master navigation file. Scientists had been educated over the years to concentrate on logging accurate times in UT, preferably using the slave clocks linked to the ship's computers that are fitted in the laboratories. BODC then took control of providing positions corresponding to the sampling times using the master navigation file.

In this scenario, BODC had total control of position assignments, including the precise definition of event positions. Consider a deep CTD cast which can take some three hours to execute. During this time, the ship is by no means stationary with respect to the ground. Some workers define the cast position as the ship's position when the CTD left the deck, others as the position of the ship when the downcast was completed and so on. The definition used by BODC is the position of the ship averaged from when the CTD left the deck to when it returned, supported by error bars that can be used to specify a box within which the data were collected. With this level of control the positional information in EVENT could be declared as authoritative with absolute confidence.

In OMEX, exactly the same procedures could be established for the cruises undertaken by the UK vessels. However, OMEX also involved cruises from eight other nations. A number of non-UK cruises also provided automatically logged navigation files. Wherever possible these were used to regenerate positions from event timings extracted from cruise reports. However, event timings were not always provided. In fact, in a significant number of cases event timings were generated by matching the positions supplied with the master navigation. Even when timings were provided, they were often obviously wrong or from an unspecified time zone.

We were therefore confronted by a situation of contradictory information from multiple sources that is all too common in data management. In one example we had two tables in a single cruise report giving the positions of the CTD casts plus positions in the CTD data file headers. Inevitably, they were all different. Significant effort was required in such cases to ascertain what was credible and what was not. However, if logged navigation was available, there is reasonable confidence that the ships were only ever in one place at one time and that the transition between event positions was accomplished at credible speeds.

In cases where we had no logged navigation, we had to depend upon cruise reports for all the data entered into the EVENT table. If we had no cruise reports, timing and positional information were supplied with the data. In general, this information has had to be taken at face value. Obviously, if the information was both present with the data and in a cruise report then it was cross checked and any conflicts resolved. However, information supplied in this way is inevitably of variable quality and in some cases the best we have are nominal station positions.

Similar problems were encountered whilst building the subsidiary index tables. For the UK ships we had a system that automatically assigned water bottle firing depths based on analysis of characteristic markers placed in the data stream by the data logging system in use on the UK ships. These were checked against detailed log sheets, designed in consultation with BODC, and delivered to us with the data. This system provided us with information in which we could have absolute confidence.

Information availability, both in terms of content and quality, again varied significantly for the non-UK vessels. In some cases, cruise reports provided excellent information on bottle firing depths. However, even here the odd problem, such as pressures incorrectly labelled as depths, was encountered. However, for most non-UK cruises we were totally dependent on the information included with the data for bottle depth assignment. This meant that considerable effort was required to resolve conflicts between different data sets from the same cast.

The OMEX index framework is as accurate and internally consistent as we have been able to achieve through the input of considerable BODC effort. However, the result is variable both in terms of accuracy and precision. It can clearly be seen that our experiences have revealed problems with shipboard data management that need to be addressed through education. Over the years, BODC has achieved this through participation in UK research cruises and plans to extend this policy to non-UK cruises during OMEX II.

5.2 Logging data accessions

BODC operates a data accession system that both assures the security of all data submitted to the data centre and allows us to keep track of the hundreds of data packages that arrive each year. Data arrive at BODC on floppy disk, high capacity tapes or electronically through ftp or as e-mail attachments. After virus checking, data in commercial formats (e.g. Excel spreadsheets) are converted into ASCII using the appropriate packages.

All data, including the ASCII translations, are backed up using three separate technologies: optical disk, digital audio tape (DAT) and a metrohm robot tape mass store. The optical disks and mass store are located in different buildings at the Bidston site and an off site DAT copy is kept in a controlled environment at the British Geological Survey in Keyworth, some 100 miles away. All original media supplied are kept in indexed storage.

Each accession is given a unique identifier based on the originating laboratory, the year of the accession and a sequence number within the year. For example accession SOC960133 was the 133rd accession received in 1996 which happened to originate from the Southampton Oceanography Centre. Full details of each accession are logged on a paper form and then entered into an Oracle database. Any paperwork, including hard copy of e-mail messages and

any documentation files, that accompanied the data are stored in a filing system indexed by the accession number.

In addition to this standard system for data supplied to BODC, an additional accession tracking system was implemented for the work on OMEX. This was a simple spreadsheet with fields for the accession number, a summary description of the data content, descriptions of work done or to be done on the accession and flag fields to indicate when all work on the accession was completed and whether there were outstanding queries addressed to the data originator.

A total of 229 OMEX accessions were received, each of which required reformatting and integration into the OMEX database structures. Accessions frequently arrived together which meant a delay before they could receive the necessary attention. If there was a problem with an accession that needed sorting out with the data originator, work on that accession was put on hold until the information required was received. Without an accession tracking tool, there was a very real danger that accessions would be overlooked or left with the work on them only partially completed. The tool described above was simple, but proved extremely effective in practice.

5.3 The nature of incoming data

During OMEX, by far the most popular data delivery mechanism to BODC, particularly for sample data, was the spreadsheet. These were either in internal formats such as XLS and WK4 files or as ASCII dumps such as CSV files. All of these variants could be handled with equal ease through the relatively inexpensive measure of ensuring that at least one copy of the latest version of the more popular packages was available within the group.

The popularity of spreadsheets comes as no surprise because the associated software provides powerful and easy to use data manipulation tools that satisfy the entire data handling requirements of many scientists. However, like any tool, spreadsheet software is open to abuse as well as use.

When receiving a sample data set as a spreadsheet, there are three things that BODC needs to know about its content. First, we need to know the position, depth and sampling time (or event identifier) for each row in the spreadsheet to allow us to link the data to the appropriate collection event in the database. Secondly, we need to know precisely what parameter is contained in each column of the spreadsheet, including its units and methodology, so we can code it correctly. Thirdly, we need full information on the measurement protocols so we can produce the necessary data documentation.

This may not sound a lot to ask, but let us look at what happens in practice. One would think that obtaining the position, depth and sampling time couldn't possibly be a problem. Sadly, this is not the case. One of the worst examples encountered by BODC was a spreadsheet containing three columns: latitude, longitude and parameter value (the parameter involved will remain anonymous). Fortunately, logged navigation was available for the cruise concerned and by poring through a listing of the navigation data it was possible to assign a time to most of the rows. However, there was still a problem. Groups of about a dozen rows had identical positions assigned, together with dramatic differences in the parameter value. It was only when a remarkable coincidence was noticed between these values and profile plots presented

in a year end report that we realised that these groups of rows represented depth profiles and not a series of surface samples taken from a fixed location.

A number of other underway data sets had positions but no times, forcing us into the time consuming task of working through navigation listings. Even when all the information is present, our problems are not necessarily over. Another example underway data set had all the necessary information present, but when the times and positions were checked against the master navigation file it was found that the positions in the data file corresponded to a time approximately 40 minutes after the time given in the data file. Investigation revealed that these data came from an automated analysis system that took some 40 minutes to digest a sample. The data time stamp was taken when the cycle was initiated (i.e. when the water sample was taken), but the position was obtained by interrogation of the ship's navigation log at the end of the cycle.

It was not only continuous surface data sets that provided us with problems. Depth profiles had also set a number of traps for us. One particularly frustrating problem was the practice of labelling spreadsheet rows with just the station identifier and the sampling depth. This was fine if the station had been sampled by a single cast but had significant potential for confusion when multiple casts were made at the station, particularly if the same depths were sampled on more than one of the casts. In such cases there was usually no alternative but to contact the data originator for more detailed information.

Another problem encountered that had the potential for causing a great deal of confusion in deep waters was the failure to accurately distinguish between pressures and depths. Depths were received labelled as pressures and vice versa. Even worse, a column labelled as depth frequently contained depths for some casts and pressures for others. BODC's solution to this problem was to build an authoritative list of the depths sampled, expressed both as depths and pressures, at the earliest possible stage in the data management cycle and then use this to cross reference the sample data sets supplied. This was successful in preventing embarrassing problems such as twenty depths sampled by a twelve bottle CTD rosette that could so easily have otherwise happened.

The confusion between pressures and depths brings us to possibly the most abused feature in spreadsheets, namely the column heading. In the worst cases, we could simply not understand what some abbreviated column headings meant, particularly for data types with which we had had limited experience. One particularly amusing manifestation of this problem was encountered by a colleague in a sister data management organisation. A data set was received with a column headed 'COULTIC'. Our colleague was puzzled by the data values that did not look like any Coulter Counter data he had encountered previously. This was because the column in fact contained coulometrically determined total inorganic carbon data. Gross misunderstandings such as this were usually resolved at an early stage through e-mail exchanges with data originators. The only consequence of this was the time required. However, there were other abuses of the column heading encountered during OMEX that were potentially much more damaging.

The most abused column heading in oceanography is 'nitrate'. The reason for the problem with this particular parameter is that the almost universally adopted method for determining nitrate involves reducing the nitrate to nitrite. Consequently, the parameter determined is nitrate+nitrite and not nitrate. Some data originators determine nitrite through a separate

channel and use the result to calculate nitrate from nitrate+nitrite. Others don't apply this correction, but then supply the nitrate+nitrite data in a column labelled 'nitrate'. This confusion is well known to BODC and we usually take care to clarify the situation with the data originator. However, even if personnel are aware of the problem it is only too easy to take a column heading at face value. In fact, errors in two data sets resulting from this confusion were identified and corrected during the OMEX database audit.

In practice, the vast majority of spreadsheets received by BODC had rows and columns that could be reliably identified and linked into the database by experienced personnel without the need to consult the originator. However, a very small proportion included adequate information to allow the data documentation to be written. Much of this was acquired from other sources, including OMEX scientific reports and the literature, but in a significant number of cases the originators had to be cajoled into providing what we required.

Many of the problems we encounter are the result of errors made when logging information at sea. Anyone who is critical of this should be sent on a research cruise, preferably in winter. Errors inevitably result from the fatigue caused by working shifts on a platform that never seems to stay still. It is also extremely difficult for individual scientists to detect these errors from the limited viewpoint of their own data sets. It is only when one is provided with the overview of several data sets plus reliable automatically logged information that a clear picture is obtained. In other words, there is a clear role for data managers in the quality control of spatial and temporal parameters associated with data.

However, there is significant work to be done to educate the scientific community in what is required for a data submission. When asked what is needed, we have a stock reply thus:

> 'A spreadsheet in which the origin of each row is clearly specified accompanied by a word processor document describing in detail what is contained in each column and a full description of the data collection protocols.'

It is a sad reflection that only one of the hundreds of spreadsheet accessions we have received has fully satisfied our dream specification, although tens of others have come close. Queries from data suppliers indicate that there is a fundamental misunderstanding of the problem. We are frequently asked what format should be used for supplying spreadsheet data. This is irrelevant as format compatibility is provided by the commercial software that we all use. It is the information content of the spreadsheet that needs to be addressed.

5.4 Data loading protocols

The data handling protocols adopted by BODC varied according to the type of data and to how much processing and quality assurance had been carried out on the data by the data originator. The BODC system was designed to fit in with the requirements of project scientists and can accept anything from raw voltages to fully processed data which are worked up as much or as little as required to produce a database containing data to a known common standard. The following section describes the main protocols adopted for handling the OMEX I data.

Underway data

Handling underway data is based upon the binary merge file. This is a binary data file that contains a record for each time step throughout the cruise, set to a sampling interval that matches the data supplied. Data channels are then merged into the file through bespoke programs using time as the primary key.

Quality control is based primarily on the Series Plotting (SERPLO) graphical editor developed in house (Lowry and Loch 1995). This allows rapid visual inspection of the data point by point combined with the ability to set quality control flags at the click of a mouse button. Parameters may be overlaid and a map of the cruise track, incorporating the data cursor, is available. Consequently, comparative screening and spatial context are introduced into the quality control procedure. It should be noted that no data values are modified during this procedure, only the values of the status flags, making it totally non-destructive.

A wide range of software tools is available for checking and calibrating the data held in binary merge files. Data may be checked against CTD or sample data held in the database, check algorithms (e.g. the computation of speed from adjacent navigation points) may be applied and calibration functions may be determined and, as is frequently required, applied. The details of the processing protocol vary from cruise to cruise and result from negotiations between BODC and the data originators.

CTD and XBT data

CTD data are accepted by BODC in the user's native format. In the case of OMEX, 16 formats, or more correctly format variants, have been handled and been converted into the BODC binary PXF format. This potentially onerous task was greatly eased by BODC's generic reformatting system, the Transfer System (Lowry and Loch 1995), that minimises the amount of code that has to be written.

Once in PXF, the CTD and XBT data were quality controlled using SERPLO. All channels were visually inspected, spikes flagged suspect and notes made on any features of concern observed in the form of the profiles. These features were corrected where possible, flagged out after the data had been loaded into Oracle or were noted in the data documentation. If upcasts were present in the data file, these were used in the assessment of the downcast data, but were not subjected to detailed quality control as only the downcasts were stored in the database.

The downcasts were delimited by tagging, simply a case of pressing the appropriate key once the cursor has been correctly located, to limit the data loaded into Oracle. If upcasts were also present in the data file, features associated with bottle firing (usually a clustering of data points around a given depth) were tagged to provide an accurate record of the bottle firing depths for the database.

Once screened, the data were loaded into the Oracle database using a bespoke loader program. Like all BODC bespoke software, this included a wide range of integrity checks with particular attention paid to ensuring that the data time channel was consistent with the station timing information held in the database. As a matter of course, the data were compared with any available sample data and calibrations applied if any systematic differences were observed. Any data channels that were in raw form were worked up, again assuming that either the necessary calibration algorithms or sample data were available.

The CTD data were initially loaded at full resolution into holding tables. Calibration coefficients were progressively assembled in a second table with one row corresponding to each CTD cast. During this time, calibrated CTD data could only be retrieved using a bespoke listing program that dynamically applied the current calibration or warned the user that the data were uncalibrated.

Once the calibrations had been thoroughly checked out, including making the data available to the user community, the final, calibrated version of the CTD data was prepared. This was done by binning all data determined as good to two decibars (one decibar if the cast was shallower than 100m) with limited linear interpolation to fill gaps. This produced a version of the data with minimal dependence on quality control flags as the data were either good or null. The reason for doing this is that in BODC's experience very few users take proper account of status flags. Naturally, the full resolution version of the data with detailed status flags have been safely archived should they ever be required in the future.

Finally, a water bottle data set, comprising CTD downcast values at the bottle firing depths, was generated using a utility program. This is stored in the bottle data structure (table BOTDATA) in the Oracle data base and may be retrieved in a fully integrated manner with other bottle parameters. It is appreciated that this approach has an inherent flaw because water bottle samples are generally collected on the upcast and not the downcast and the water column is dynamic. However, in a scenario where upcast data were unavailable for many of the cruises this lowest common denominator approach was deemed to be the optimal solution, providing consistency at the expense of absolute accuracy.

SeaSoar data

The SeaSoar is a CTD mounted in a towed fish that oscillates between the surface and depths of up to 500m. Data are logged at 1 Hz producing high volume data sets. The data handling technique developed by BODC to make SeaSoar data more manageable is to grid the data with a vertical resolution of between 4 and 8 decibars and a horizontal resolution designed to match a single oscillation of the fish, approximately 4 km for a 500m dive. Each column of the grid is then considered to be a 'pseudo CTD' and is managed as if it were a true CTD cast.

In OMEX, the SeaSoar data were supplied to BODC fully processed by Southampton Oceanography Centre (SOC) as both 1 Hz time series and as appropriately gridded data. The 1 Hz files were passed to the UK-NODB for long term archival and played no further part in OMEX data management.

The gridded files were split into individual 'pseudo CTDs' using a bespoke program, converted to PXF using the Transfer System, screened using SERPLO and loaded into Oracle using a bespoke loader program. The bespoke software included a series of checks to ensure that the navigation data in the SeaSoar file matched the master data in the underway binary merge file.

ADCP Data

Handling underway ADCP data, particularly from water that is too deep for the instrumentation to operate in bottom tracking mode, is a considerable problem if high quality data are to be produced. Problems such as asynchronous determination of ship and water velocities, misalignment between the ADCP and the ship's gyro and sound velocity variations

due to variations in sea water density all need to be addressed. BODC is currently developing systems to undertake this work for raw ADCP data but these were not ready in time for OMEX I.

However, fully processed data were available from two cruises thanks to the efforts of scientists at SOC. The databasing of ADCP data is very similar to SeaSoar data. Each discrete current profile, usually integrated over a 10 minute period, is treated as an event. A secondary index table holds information such as the calibrations applied and the data are held in a single flat table indexed by event identifier.

The data supplied were first split into discrete profiles then converted into PXF using the Transfer System. Each individual profile (over 3000 per cruise) was then screened using SERPLO and any spikes flagged. The main problem encountered was the presence of garbage in bins deeper than the water depth which had not been cleaned out during the SOC processing procedures. Once screened, the data were loaded into Oracle using a bespoke loader program.

Sample Data

The term sample data includes water bottle data, stand-alone pump data, benthic data, sediment trap data, net hauls and the results from tracer incorporation experiments. Invariably, these data were received as spreadsheets or as ASCII files with a simple table structure. Upon receiving an accession of sample data, the first task was to assess whether the data could be mapped to one of the normalised data structures within the database. At present, these cover independent parameter measurements on water, SAP and trap samples, whole core properties and core profiles for parameters other than species level benthic biomass data.

Assuming that the data could be mapped to one of the standard structures, a standard data loading protocol was followed. The description given specifically describes our protocol for handling water bottle sample data. However, the protocols for benthic and trap data were broadly similar and documenting differences of detail is beyond the scope of this paper.

The first stage of the loading procedure was to replace the heading of each data column with a code from the parameter dictionary. During OMEX, this frequently involved coding extensions to the dictionary as additional parameters expanded our horizons. The next stage was to ensure that each row of data was labelled with an event identifier (a syntactically exact match for the OID field in the EVENT table) and either a pressure or a depth.

A visual inspection of the data followed. Particular attention was paid to cells set to zero (that are all too easily generated from empty cells when formulae are applied) and to indications of data below detection. The latter were often present with cell contents such as '<0.3' which then result in data loader errors. In these cases, separate flag columns, named using an extended parameter code (e.g. the flag channel for NTRZAATX is NTRZAATX_F), were created to contain the below detection flag. Another area that frequently required attention was the units used for the data. Each database parameter code has its units defined and if a different unit had been used by the data originator then conversion into the database standard units was required.

Frequently, data originators included plain language notes indicating values about which they were uncertain. These comments were translated to quality flags ('L' meaning value reported suspect by the data originator) in the flag column associated with the data column. If the visual inspection revealed any data values that gave BODC personnel cause for concern, they were flagged 'M' (meaning value considered suspect by BODC).

The above work was done using Excel on a copy of the original spreadsheet. Once completed, the data were saved out as an ASCII comma-separated value (CSV) file and ported from the PC into the UNIX environment. A temporary table was created under Oracle containing the data columns, flag columns, columns for the independent variables and columns for the database primary keys. For example, the temporary table required for nitrate+nitrite data contained the following columns:

OID	Independent variable (event identifier)
DEPTH	Independent variable
NTRZAATX	Dependent variable
NTRZAATX_F	Dependent variable flag
BEN	EVENT table primary key and BOTTLE table foreign key
IBTTLE	BOTTLE table primary key

The independent and dependent variable data were loaded into the temporary table using the standard Oracle utility (SQLLOAD under Oracle 6 and SQLLDR under Oracle 7). Once loaded, values were assigned to the primary keys through SQL UPDATE statements. First, the BEN values were set up which were then used in the assignation of the IBTTLE values. Once a range of simple checks had been completed on the temporary table (e.g. to ensure that all the IBTTLE values were unique) a bespoke loader program was used to copy the data from the temporary table into BOTDATA including linkage of the data to its originator.

Some data could not be mapped into the generalised structures. For example, an in-situ production rig experiment result is meaningless unless the depth of incubation and duration of the experiment are known. Any data storage structure has to ensure that these items of information are maintained in intimate association and the generalised structures fail to do this. Other data sets, for example the benthic species biomass data, have not been mapped into the generalised structures because resources have yet to be found for implementing the massive extensions to the parameter dictionary that would be required.

In these cases, the flexibility provided by relational database technology has been utilised to provide table structures that match the requirements of these data types exactly. Such structures are easy to implement and may generally be populated, either directly or through temporary tables, using the Oracle loader and UPDATE statements to assign primary keys. The price paid for this is database complexity, involving additional documentation and interface forms, but the resources required are significantly less than those required for a generalised solution. However, the long term aim is to work towards structural simplification whenever resources for database restructuring are available and these custom structures should be regarded as temporary entities.

Moored Instrument Data

Moored instrument data were handled using the standard procedures developed for the UK-NODB. The data were converted into PXF format using the Transfer System and screened using SERPLO. If any additional processing of the data was required, for example data from one rig had a 180 degree compass error, then the data were loaded into Oracle, manipulated and dumped back into PXF.

Header information and data documentation were compiled into holding tables. After thorough checking, including many automated systems, the data were loaded into the UK-NODB and the PXF files placed in the standard holding directories. Finally, the index entries showing where the data could be found were set up in the OMEX database.

5.5 Data Documentation

One of the major tasks in the assembly of the OMEX database was the preparation of the data documentation. The documentation was structured as follows:

Underway data	One document per cruise
CTD data	One document per cruise
Moored instrument data	UK-NODB documentation report
Other data	One document per table or group of related tables

The underway data documents describe the channels present in the data, the instrumentation used and a description of the processing history of the data. In addition, a detailed data quality report on each channel, based on BODC's examination of the data plus information from the data originators, is included. The CTD data documents are similar in structure. These documents are written using information in cruise reports, material supplied by data originators and from BODC's experiences working with the data. In some cases, obtaining some of the necessary information was either extremely difficult or even impossible. For example, our efforts to discover what type of thermosalinograph was used on Meteor and Poseidon failed completely.

UK-NODB documentation is constructed by linking together a number of sub-documents. Many of these are standard documents describing the mechanics of the instrument and such like. Others describe the mooring configuration whilst others provide information, such as data warnings, on individual instrument deployments. These are retrieved from the database in the form of a report with duplicated sub-documents eliminated. A single report covering all OMEX moored instrument deployments can easily be created. As many of the OMEX moored instrument deployments used standard instrumentation, such as Aanderaa current meters, many of the sub-documents were already written and stored in the database. The remaining documents were written from information in cruise reports and on the basis of information supplied in response to queries to the data originators.

One of the greatest challenges presented by OMEX was the documentation of the non-CTD data held in the relational database. Producing the data documentation for the normalised data tables such as BOTDATA (bottle sample data), COREPROF (core profile data) and STDATA (sediment trap data) was particularly difficult. Consider the problem of documenting BOTDATA. Over 400 different parameters were measured. Some parameters were measured on different cruises by different originators using different protocols.

After some considerable thought, the following documentation strategy was developed. First, the parameters measured were subdivided into subgroups, such as nutrients, hydrography, dissolved trace metals, etc.. A documentation chapter was written for each subgroup containing four sections. First, a list of the parameter codes used and their meanings was given. Secondly, a cruise by cruise list of the data originators who had provided one or more of the parameters in the first section was prepared. Thirdly, the protocols used by each data originator were described. Finally, a quality control report presenting the results of BODC quality assurance procedures plus any warning information provided by the data originators, was produced. An example section for one of the most commonly measured parameters, nutrients, is given in Appendix 3.

The information for the first two sections came from the database and the information for the fourth section was either offered to us or was reporting on work that we had carried out. Obtaining this information was therefore quite straightforward. Obtaining the information for writing the individual protocol descriptions was also relatively easy. In many cases, the methodology sections from the OMEX I final report could be incorporated directly. Where these were absent or considered too sketchy, the originators were contacted and responded by sending methodology sections from papers, reports or theses. However, even though the information was readily available, the compilation of the documentation was a major undertaking. The document covering the BOTDATA table alone contains over a hundred A4 pages.

Other tables in the database were much easier to document as they generally contained data from a relatively small number of determinations. Again, the information required was drawn from OMEX I reports, particularly the final report, together with methodology sections supplied on request by the data originators.

5.6 Database Audit

It can be seen from the above protocol descriptions that the incorporation of data into the database is a complex undertaking. Consequently, the work is prone to error, particularly as the support staff for the OMEX project were recruited at the start of the project. Whilst some quality control could be achieved through day to day supervision of the work by experienced personnel, it was not possible to check everything in this way. In particular, it was not possible to detect if areas of work had not been completed through being left 'on hold' and then forgotten. It is also possible to undertake much more effective checking that reveals both data loading errors and undetected errors in the original data through comparison of the data held in a unified, integrated structure.

It was therefore decided to undertake a thorough audit of the database between its creation and its electronic publication as the definitive project data set. The entire audit was undertaken by the project manager, the most experienced member of the OMEX team.

The audit of the underway data was based primarily on a re-examination of the data using the SERPLO editor. This revealed a range of problems including data in incorrect units and data sets that had only been partially loaded. A significant number of problems in the original data that had gone undetected in the original screening were also picked up. All calibrations were

checked and the data documentation was audited to ensure that the calibrations described matched the data in the system calibration files.

The audit of the CTD, XBT, ADCP and SeaSoar data was basically similar. The data were re-examined using SERPLO, calibrations were checked and the documentation audited. Relatively few problems were revealed by this beyond a small number of missing entries in the calibration coefficient table. Once everything had been checked out, the utilities were run that binned the data and generated the CTD data set for the bottle firing depths followed by the final archiving of the full resolution data.

The audit of the sample data set was simple in principle but a major undertaking in practice. Essentially, the complete sample data set was retrieved a cruise at a time from the database. The result was then compared with the original data submissions and subjected to a quality control check. Particular attention was paid to the identification of bottle samples contaminated through leakage and to the identification and correction of common problems such as rosette sequence errors. Wherever possible an intercalibration of parameters measured on the same samples by more than one method or by more than one originator was undertaken.

Once the audit on a cruise had been completed, the data documentation for the sample data was written. Reports on the results of any intercalibrations or data quality problems identified during the audit were included.

5.7 Time Scales

An essential component of any case study is an examination of the timing of events. The first area that we may examine is the rate of progress of data delivery into BODC. Individual data set submission time stamps were not available. However, the BODC accession system provided a source of time stamped information and the progress of OMEX data accessions during OMEX I and OMEX II/I is shown on a month by month basis in Figure 4.

This figure tells an interesting story and is worthy of some analysis. In doing this, there are a number of critical dates to consider.

1 June 1993	OMEX I commenced
February 1994	OMEX data management commenced
August 1995	Circular to OMEX Principal Investigators
January 1996	Data submission status revealed at OMEX workshops
31 May 1996	End of OMEX I
September 1996	Targeted e-mail and letter communications
31 December 1996	Advertised final deadline for OMEX I data submissions
31 May 1997	End of OMEX II/I

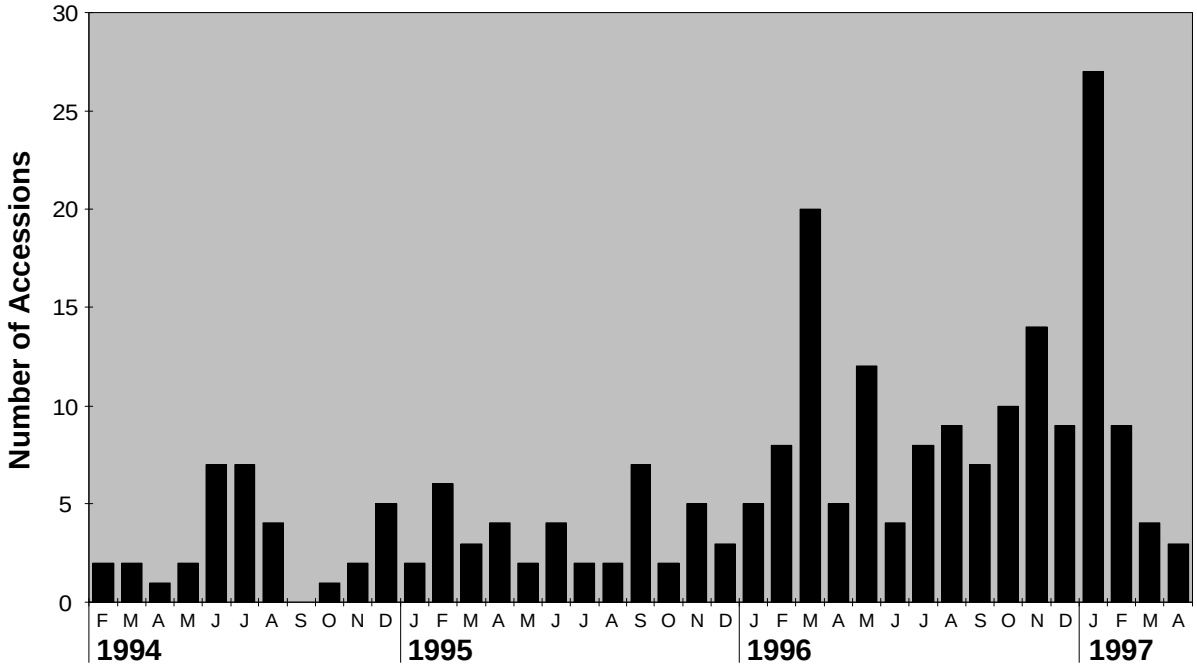

Figure 4. Progress of OMEX Data into BODC

It can be seen that the flow of data into BODC started as soon as the data management operation began. However, this was to be expected as the scientific programme had already been operational for over six months. For the next two years, there is a steady trickle of data. Further analysis of the data reveals that a high proportion of these early submissions were from UK partners with whom BODC had already established a good working relationship during the North Sea Project and BOFS.

During 1994 and early 1995, BODC made intensive efforts to raise the profile of data management in OMEX largely through the vehicle of presentations to project workshops. Whilst these didn't produce any noticeable blips in the rate of data submission, it undoubtedly laid the foundation for our later success. The first attempt to increase the pressure on the OMEX scientific community was the circular letter sent out in August 1995. This was almost certainly responsible for the significant increase in accessions received in September 1995.

It can be seen quite clearly that we received more data during 1996 than at any other time during the project. This was the result of a number of factors. In January 1996, a presentation was given to the biogeochemistry subgroup workshop in Plymouth. A series of overheads were presented there that showed the proportion of data submitted for each cruise and by each sub-project. The latter was particularly significant as all sub-project leaders were at the meeting and was largely responsible for the dramatic peak in data accessions for March 1996.

This was followed by the end of OMEX I which naturally brought data submission to the top of the agenda both in May 1996 and throughout the summer as final reports were prepared and people cleared their desks before turning their attention elsewhere. The fact that this trend continued until the end of the year was the result of intense targeted pressure from BODC. It should be noted that the huge peak in January 1997 really belongs in December 1996 as the holiday season delayed the logging of accessions rushing to beat the December 31st deadline.

From the BODC perspective the trend shown in Figure 4 gives some cause for concern. From this paper it can be seen that we do a significant amount of work on each data set we receive. Approximately half of the data sets we received arrived during 1996 which meant that our workload during the project was seriously unbalanced.

Fortunately, the problem was moderated by two factors. First, many of the data sets received towards the end of the project were sample data that generally required less BODC effort per accession than the underway and CTD data sets received earlier in the project. Secondly, the support staff for OMEX were new recruits and therefore far more effective at the end of the project than they were at the beginning.

From the above comments one might get the impression that BODC were idle for two years and then did some work. However, OMEX I data management was resourced on an initial estimate of 10-15 cruises. In fact, we handled the data for 47 cruises. The true situation was that we were kept adequately occupied for two years followed by a year of intense activity then 6 months of mayhem supported by a project overspend and large quantities of unpaid overtime.

The scale and timing of the workload may be appreciated by studying Figure 5**.** This shows the rate and timing of growth of the three major sample data tables for bottle (BOTDATA), benthic (COREPROF) and sediment trap (TRAPDATA) sample data. The normalised data structures were implemented in December 1994 after it was realised that the more simplistic structures used for the North Sea Project and BOFS would not cope with the additional complexity of OMEX data. Very few data were added to this table for some 9 months. This was partly due to a lack of data to load but the most significant cause was that we were kept fully occupied dealing with CTD and underway data.

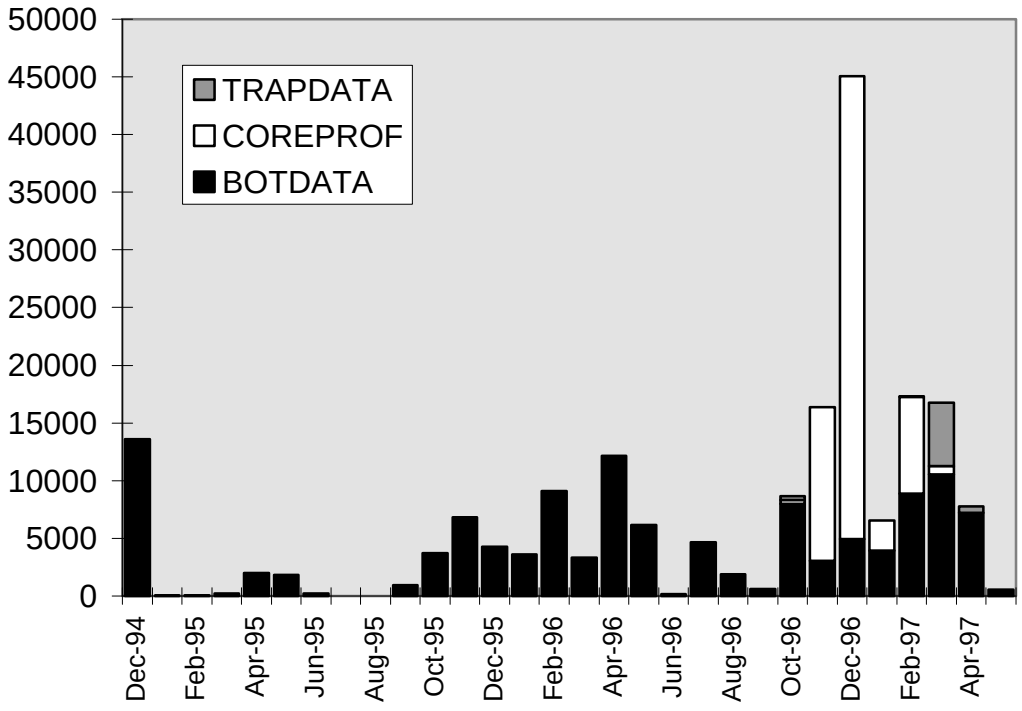

Figure 5. Number of Records Added to the Sample Data Tables by Month.

Until March 1996, BODC had only received a trickle of benthic data. These data started to arrive in significant quantities after the benthic sub-group workshop at Strenglin in that month followed by a steady flow throughout the month. The arrival of the initial rush of data again revealed that the storage strategies we had used in the past were inadequate for the scale and level of complexity of the OMEX data. More sophisticated systems were built during the summer and autumn of 1996 followed by a period of intense data loading activity.

Sediment trap data always arrive later than other sample data due to the large amount of labour intensive sample processing that is required. However, the delivery was not quite as it might appear from Figure 5. In fact, most of the trap data were supplied in the late summer of 1996 but could not be loaded into the system until we had finished dealing with the benthic data.

The pressure of work from November 1996 onwards was further compounded by the database audit that was also underway by this time. It is important that this level of crisis is not allowed to arise again. In OMEX II, all partners have a contractual obligation to deliver their data to BODC six months before the end of the project. This, combined with the fact that many of the OMEX II partners have worked with BODC during OMEX I and that we have improved systems in place, will hopefully provide us with a more even workload.

6. ACCESSING THE DATABASE

One of the commonest criticisms of data management is that there are constant demands for data to be supplied, but nothing is ever received in return. One of BODC's primary objectives is to ensure that the user community we support is able to obtain the data they require to do their science. This is achieved through two data delivery mechanisms, an on-line interface and a request service, that are available from the beginning of the data management operation. In this way we are able to provide an effective vehicle for data exchange within the life of the project we are supporting.

6.1 The on-line user interface

Any OMEX scientist is able to register as a user on the Bidston computer system. This provides access to the UNIX operating system, Oracle's SQLPLUS database interface and BODC application software. Using these tools it is possible to retrieve any of the data held within the system into ASCII that may then be networked to the user's home system.

Significant attention is paid to system security. Each user must identify themselves to the system through two levels of passwords. All session transcripts are logged and regularly inspected by BODC personnel. In this way, we are able to maintain awareness of precisely which data have been accessed and by whom. An indirect benefit of this monitoring capability is that we are able to assess how users are getting on with the system and offer unsolicited user support where appropriate.

The user interface we are currently using was developed for an IBM mainframe nearly a decade ago. It requires users to type commands and have reasonable proficiency in SQL and

basic UNIX. BODC has run a number of training courses for scientists from OMEX and other projects to enable them to get the most out of the system. The interface is highly effective, particularly in the hands of a skilled user, but it has a distinctly dated look about it. With the tools currently available, developing a GUI front end to the database is a trivial task requiring a few weeks work at most. One might therefore ask why this has not been done. The answer is that we have so far been unable to find a way of producing a GUI front end that provides us with the same level of security and usage information as the current system. Until this problem can be adequately addressed our on-line system will maintain its current appearance.

In spite of its antiquated appearance, the system is heavily used. To date, a total of 354 sessions by 18 users from 11 different organisations have been logged.

6.2 Help desk and requests

The on-line system requires a significant learning overhead that is excessive for users who only require infrequent access to the system. Users of the on-line system also sometimes run into problems obtaining the data they require. BODC offers a help desk and request service to cater for these needs.

Users may contact BODC personnel at any time by phone or e-mail for help or data. We give a high priority to this service. Trivial requests are usually dealt with immediately and we strive to respond to anything more demanding within 48 hours.

Considerable resources are required to support this service. Over the 3.5 years of BODC's involvement in OMEX, a total of 184 requests have been serviced.

7. ELECTRONIC PUBLICATION OF THE DATABASE

7.1 Structure of the CD-ROM

The final product of the OMEX I data management operation will be the electronic publication of the complete data set on CD-ROM. At the time of writing, this phase of the project is fully planned with implementation due to commence imminently. It is anticipated that this will take some 4-5 months.

In many ways, the structure of the CD-ROM will reflect the structure of the project database. For example, non-standard storage strategies will be used for underway and moored instrument data. The relational database will be present in a number of different formats both in its entirety and as subsets. Let us look at this in more detail.

The underway data will be included on the CD-ROM in the same binary merge format that is used to store the data under UNIX. This has been used in the past without complaint. PC users are provided with interface software and UNIX users are offered BODC source code, although this offer has yet to be taken up.

Moored instrument and self-logging lander (e.g. BOBO and STABLE) data will be on the CD-ROM in a standard ASCII format generated by the UK-NODB interface software accompanied by a documentation report. This is generated as flat ASCII but it is anticipated

that a PDF version (see below) will be included to assure compatibility with other documentation on the CD-ROM.

BODC practice in the past has been to output relational databases in ASCII 'kit form' format. In this, each table of the database is dumped, in comma-delimited format, as an ASCII file. From these, the project relational database may be recreated under any relational database system on any platform. This concept was introduced over five years ago and still fulfils a need. However, since then relational database software has become much more commonplace on PC platforms with Microsoft *Access* by far the dominant product. Whilst the ASCII kit form may be loaded into *Access* relatively easily, it still requires some effort on behalf of the user. It was therefore decided that the OMEX product should include a second copy of the data in an *Access* compatible format. After considerable thought, Microsoft JET 2.5 format was chosen which is fully compatible with *Access* 2.0 and may be read (or converted into later JET variants) by *Access* 7.0 and *Access-97*.

The inclusion of a version of the data in JET on the CD-ROM has a couple of hidden advantages. First, previous versions of the BODC software interface have used the ASCII kit as a data source. This is very slow compared to software running against a database with an in-built indexing capability. Secondly, the JET container may include objects other than data tables, such as forms and application macros. These are quick and easy to develop and provide a means for BODC to provide enhanced interfaces to the data at little cost.

Curious users of the CD-ROM will notice a number of the larger data tables are present on the CD-ROM as Borland *Paradox* table files. This undocumented feature of the product is present to allow the BODC interface programs, developed under Borland *Delphi*, to run even faster: the Borland database engine significantly out performs the ODBC interface required for JET.

In addition to the complete database, a new relational database concept is being developed for the OMEX database that has been christened the 'Melting Pot Database (MPDB)'. The relational structures used by BODC partially document the data with information on how the samples were collected and analysed. This has obvious advantages but is achieved at the cost of having the same parameter measured in different ways stored in different places in the database. The concept behind the MPDB is to strip the attributes held with the data down to the bare essentials, namely (for water column data) latitude, longitude, date and time, depth and parameter code. The MPDB tables will include data from both the project relational database, underway data files and moored instruments. It is important to note that only data that are believed to be good quality and only parameters that may be integrated with confidence will be included. The MPDB table files will be present on the CD-ROM in ASCII, JET 2.5 and *Paradox* formats and will be primarily interfaced through BODC software.

In addition to the data, the CD-ROM will include extensive documentation, including both the data documentation and the Users' Guide to the product. The North Sea Project and BOFS CD-ROM products both included the Users' Guide on paper. This was because no platform independent delivery system other than flat ASCII was available. However, Adobe's *Acrobat* software has now overcome this problem. Adobe provide freeware reader software for Microsoft Windows (3.n and 95), Apple Macintosh and UNIX platforms that allows users to display and print documents in Adobe's PDF format. This is an active format, equally well suited to text or graphics, that has been set up as a competitor to HTML and includes the facility for extensive cross referencing through hot links. With such a powerful tool available,

it was decided that soft documentation was a feasible option. Further, the hot linking capabilities of PDF will allow the data documentation to be integrated within the Users' Guide in a way that was impossible with paper documentation.

7.2 CD-ROM software interface

The CD-ROM will include three interface programs written by BODC. All of the programs will be restricted to PC platforms running Microsoft Windows 3.n or Windows 95. The first program, the Underway Explorer, provides an interface to the underway data held in binary merge format. This either allows the user to display the data graphically as a time series plot (including a window that shows the segment of the cruise track from which the data were taken) or as digital information in a spreadsheet grid. All graphics are underpinned by Windows printing functionality and the digital information may either be saved in ASCII or passed to other packages via the Clipboard.

The second program provides an interface to the sample data held in the fully normalised database structures: i.e. water bottle, core and sediment trap data. The user first selects the stations of interest, using a combination of dialog boxes or a map. The software then guides the user through a hierarchical selection procedure to choose the parameters of interest. The data will be loaded into a spreadsheet grid that may be saved, printed or copied onto the Clipboard. Subsequent developments, for products due out next year, may incorporate a profile plotting capability. It should be noted that software included on the CD-ROM may be easily upgraded over the Internet.

The third program provides the user with an interface to the Melting Pot Database. This is primarily designed to allow the data to be retrieved in a manner suitable for contouring using packages such as *Surfer*. The interface is therefore biased towards allowing the user to retrieve data from spatial sections or from point time series. Once again, the data are delivered via a spreadsheet grid that may be saved, printed or copied onto the Clipboard.

In addition to the BODC software, the CD-ROM will also include the freeware Adobe *Acrobat* installations for all platforms supported by Adobe. This is included purely for user convenience: the software is readily available from many sources such as magazine cover CD-ROMs and, of course, numerous sites on the Web.

The final component to the interface will be a series of *Access* forms packaged in the JET database container. These will be designed to provide easy access to data stored in the table structures that are not covered by the primary interface software. They will, of course, only be available to Microsoft *Access* users.

7.3 Product enhancements

CD-ROM is a medium associated with multimedia data. In the case of the OMEX CD-ROM, sound and video did not seem appropriate. Nevertheless, pictures and images are appropriate and are used to significantly enhance the product. Pictures provided by the OMEX community include photographs of the sea floor, scanning electron micrographs of SPM samples and X-ray photographs of Kasten cores. The images made available are either satellite images, including composites, and the contoured sections and time series that have been generated from the data held in the database for the Web-based OMEX Nutrient Atlas.

A benthic database was built from data presented in the literature as part of the OMEX benthic modelling work at the Netherlands Institute of Ecology (NIOO). This has been provided for the CD-ROM, together with the source code of the models developed for OMEX. These are included as supplied in a separate partition (in other words a directory) of the product.

It can therefore be seen that the OMEX CD-ROM will deliver significantly more than the OMEX database. The benefits gained from the inclusion of these additional elements go beyond product enhancement. A safe archive has been provided for valuable scientific information that might otherwise have been lost once interest in OMEX had waned.

8. CONCLUSIONS

There are a number of observations that may be made on our OMEX I operation. The first of these concerns the overall cost. The data management service provided by BODC sets high standards. At the end of the project, a fully integrated, fully documented data set is produced in a form where it may be distributed as a neat, highly visible package. The data contained therein have considerable added value over what was initially supplied both through their integration and documentation and through the additional quality control and calibration refinement undertaken by BODC. In addition, throughout the project, a highly effective data exchange vehicle is provided for project participants.

The personnel resources allocated at BODC to OMEX I were one third of the time of a principal investigator plus two support staff. It is difficult to produce accurate costs for the OMEX I project as a whole due to the problem of accurately quantifying the additional national contributions. However, by making a few assumptions, it is estimated that approximately 2.5 per cent of the total spend on OMEX I science was allocated at BODC for data management. As previously discussed, the data management was under resourced and was only completed successfully by drawing from BODC's infrastructure resources, using the breathing space provided by OMEX II/I, over running completion deadlines and by placing pressures on personnel that may only be endured for limited periods of time. The actual spend on data management at BODC was more like 3.5 per cent of the total project budget than 2.5 per cent. Had this been raised to 5 per cent, the work could have been managed comfortably.

Our experiences in OMEX I provided us with the basis for the accurate estimation of the staffing levels required for multidisciplinary oceanographic data management. As a rule of thumb, a fully supported (in terms of systems requirements) data scientist can comfortably handle the data from between three and four major cruises per year. In this context, a major cruise may be defined as one of at least 2 weeks duration with at least ten scientific berths.

Our second observation concerns the staffing of a data management operation dealing with a data set of the scale and complexity of OMEX I. It cannot be emphasised too strongly that this is a job for scientists who understand the data that are being collected. Such are the skill levels required to do the job properly that at least two years on the job training is required before full effectiveness is reached. Any attempt to manage data of this complexity with a team of clerks or technicians backed up by systems developers is a recipe for disaster.

Our third observation concerns the timing of the data management operation in relation to the science. OMEX I data management began over six months after the field work started. In retrospect, it is felt that this was unfortunate for the particular circumstances of OMEX I for two reasons. First, BODC was faced with the massive task of introducing our style of data management to a much wider community. The process of making ourselves known would have been eased had we not suddenly appeared part way through the project. Secondly, our OMEX I work was staffed by two new personnel whose introduction to the job would have been eased had it not started by clearing a data backlog.

It can be clearly seen that the pressures on the data management operation increase alarmingly towards the end of a project. Not only does the data submission rate increase significantly, but additional data management activities such as audit and product development occur at this time. There are two ways in which the adverse effects of this may be reduced. First, through resourcing the data management adequately, even to the extent of over-resourcing the anticipated load during the first half of the project by as much as 50 per cent. During OMEX I, considerable systems enhancement was required and systems development needs to continue if the services we offer are to keep pace with the developing requirements of the science. This is not just a case of software development requiring programmers. Considerable data restructuring is also required that can keep data scientists fully occupied during the quieter times during a project.

Secondly, the bulge in the data submission rate needs to be forced away from the time when the data management operation is due to finish. This may either be done by scheduling the data management operation to run for a specified period beyond the scientific project or by ensuring that all data are delivered well before the end of the project. The latter approach is being adopted for OMEX II/II data management with a contractual obligation placed on all partners to deliver their data six months before the end of the project. BODC has mixed feelings on whether this will succeed. Optimism is fuelled by the fact that many of the OMEX II/II partners were involved in OMEX I and will therefore have a better understanding of what is expected of them. However, pessimism is fuelled by the fact that much of the late delivery in OMEX I was caused by scientists being over committed and simply unable to deliver on time. It will be interesting to observe what happens in practice.

Overall, the data management of OMEX I can only be described as a total success. Some 95 per cent of the data sets collected during the field programme have been assembled into an integrated, fully documented database that is scheduled for electronic publication less than two years after the last cruise docked. No other multinational, multidisciplinary oceanographic data management operation known to the authors has matched this achievement.

9. ACKNOWLEDGEMENTS

The authors give their heartfelt thanks to the OMEX project scientists for providing the data that has made the OMEX data management project a success. In many cases this has involved help and co-operation beyond the call of duty. Useful assistance in our efforts to round up the data have been provided by the OMEX Scientific Steering Committee and our colleagues in the MAST Office at the European Commission.

Dr. Meirion Jones, director of BODC, earned the eternal gratitude of the authors (RKL in particular!) for looking after the administrative side of the project such as contract negotiations, interim and final reports. Meirion also provided a detailed, constructive criticism of this manuscript. Ray Cramer developed the CD-ROM software interface. Other colleagues in BODC have supported the project through development of systems infrastructure, particularly Steve Loch, John Hughes and Dave Neave. Andy Tabor's computer graphics skills provided invaluable assistance with the manuscript illustrations.

Financial support for the work has been provided by the European Union MAST programme and by the Natural Environment Research Council through BODC core funding.

10. REFERENCES

Lowry, R.K., 1992. Data Management for Community Research Projects: A JGOFS Case Study. *In* Churgin, J. (ed.) Proceedings of the Ocean Climate Data Workshop, Goddard Space Flight Centre, Greenbelt, Maryland, USA, pp 251-273.

Lowry, R.K., 1995. OMEX Data Management (ODAM). *In* Weydert, M., Lipiatou, E., Goñi, R., Fragakis, C., Bohle-Carbonell, M. and Barthel, K-G (eds.) Second MAST Days and EUROMAR Market, Office for Official Publications of the European Communities, Luxenbourg, pp 1307-1318.

Lowry, R.K. and Cramer, R.N., 1995. Database applications supporting Community Research Projects in NERC marine sciences. *In* Giles J.R.A. (ed.) Geological Data Management, Geological Society Special Publication No 97, pp 103-107.

Lowry, R.K. and Loch, S.G., 1995. Transfer and SERPLO: powerful data quality control tools developed by the British Oceanographic Data Centre. *In* Giles J.R.A. (ed.) Geological Data Management, Geological Society Special Publication No 97, pp. 109-115.

Appendix 1

Summary of the OMEX I Field Data Set

Cruise Belgica BG9309

Underway data

Latitude and longitude
Temperature and salinity
Wind speed and direction
Barometric pressure
pH, alkalinity, pCO_2 and TCO_2.
Bathymetry
Solar radiation
Air temperature
Dissolved oxygen

CTD Data

Bottle Data

Nutrients (measured by 3 groups)
Spectrophotometric pigments
Dissolved oxygen
DOC (measured by 2 groups)
Colloidal carbon and trace metals
HPLC pigments
POC/PON
Alkalinity and pH
Dissolved trace metals (2 groups)

Centrifuged SPM samples

Elemental analyses (metals and carbon)
Spectrophotometric pigments
POC/PON
Carbon and nitrogen isotopes

Stand-alone Pump Data

Trace metals
Organic carbon and nitrogen

Production/Uptake

Carbon uptake (2 workers)
Nitrogen uptake
Phosphorus uptake

Cruise Poseidon PS200-7

CTD Data

Marine Snow Profiler Data

Longhurst-Hardy Plankton Recorder Data

Underway Data

Latitude and longitude	Temperature and salinity

Bottle data

Reversing thermometer temperatures	Microzooplankton biomass

Sediment Trap/Current Meter Moorings and BOBO Lander Deployed

Cruise Valdivia VLD137

Underway Data

Latitude and longitude	Wind speed and direction
Air temperature	Solar radiation
Barometric pressure	

CTD Data

Bottle Data

Microzooplankton biomass and grazing	Phototrophic/heterotrophic nanoflagellates
HPLC pigments	Fluorometric chlorophyll (2 groups)
Spectrophotometric pigments	Nutrients
POC, PON and particulate phosphorous	DOC
Dissolved oxygen	Fatty acids
Total dissolved nitrogen and phosphorous	Phytoplankton species counts
pH	Turbidity
Bottle salinities	Reversing thermometer temperatures
Bacterial production and counts	Amino acids and carbohydrates

Core Data

Porosity profiles	Organic biomarker profiles
Pigment profiles	

Production data (In-situ/on deck incubations)

Carbon uptake	Nitrogen uptake

Cruise Auriga PLUTUR I

Bottle Data SPM gravimetry

CTD Data

Cruise Cote d'Aquitaine NAOMEX1

CTD Data

Core Data

Grain size profiles
Calcium carbonate profiles
Water content profiles

Cruise Belgica BG9322

Underway data

Latitude and longitude
Solar radiation
Air temperature
Dissolved oxygen.
Chlorophyll
Temperature and salinity
Wind speed and direction
Barometric pressure
pH, Alkalinity, pCO_2 and TCO_2
Nutrients (NO_3, Si)

CTD Data

Bottle Data

Microzooplankton biomass
Dissolved oxygen
POC/PON
HPLC pigments
Normalised carbon uptake (2 workers)
Normalised phosphorous uptake
Alkalinity and pH
Nutrients (2 groups)
DOC
Spectrophotometric pigments
Normalised nitrogen uptake
Dissolved Trace Metals

Centrifuged SPM samples

Elemental analyses
POC/PON
Spectrophotometric pigments
Carbon and nitrogen isotopes

Stand-alone Pump Data

Trace metals

Cruise Pelagia PLG93

CTD Data

Bottle Data

Microzooplankton biomass	Nutrients
Dissolved oxygen	SPM gravimetry

Core Data

Pore water nutrient and sulphate profiles	Pigment profiles
Carbon isotope profiles	Density and porosity profiles
Organic biomarker profiles (2 groups)	Pore water oxygen profiles
Pore water trace metal profiles	Foraminifera morphology profiles
^{210}Pb profiles	Grain size profiles
Total and organic carbon profiles	Total nitrogen profiles
Benthic macrofauna biomass	Benthic meiofauna biomass

Landers

TROL	Sediment oxygen and resistivity profiles
BOLAS	Sediment oxygen demand
BIOPROBE	Benthic boundary layer currents, suspended load and particle characterisation

Cruise Auriga PLUTUR II

CTD Data

Bottle Data

SPM gravimetry	POC

Core Data

Solid phase trace metal profiles	Calcium carbonate profile
Grain size profiles	Organic carbon profiles
Water content profiles	

Cruise Charles Darwin CD83

Underway Data

Latitude and, longitude
Temperature and salinity
Photosynthetically available irradiance
Bathymetry
Optical attenuance
Chlorophyll

CTD Data

Bottle Data

DMS and DMSP
POC/PON
Dissolved oxygen
Nutrients
Fluorometric pigments
Normalised carbon uptake

XBT Data

Drifting Buoys and Current Meter Moorings Deployed

Cruise Meteor M27-1

Underway Data

Latitude and longitude
Solar and ultra-violet radiation
Air temperature and humidity
Bathymetry
Temperature and salinity
Wind speed and direction
Barometric pressure

CTD Data

Bottle Data

Nutrients
POC, PON and particulate phosphorous
CH_4, $\delta^{13}C$ and TCO_2
Dissolved/atmospheric carbonyl sulphide
pH
Dissolved free amino acids
Turbidity and raw fluorescence
Dissolved total P and N
DOC (2 groups)
Carbonyl sulphide production
Dissolved oxygen
Bottle salinities
Dissolved/particulate carbohydrates

Core Data

Pore water oxygen profiles
Porosity profiles
Organic biomarker profiles
Grain size profiles

Landers

BIOPROBE	Benthic boundary layer currents, suspended load and particle characterisation

Cruise Charles Darwin CD84

Underway Data

Latitude and longitude	Bathymetry
Temperature and salinity	Optical attenuance
Photosynthetically available	Chlorophyll

CTD Data

Bottle Data

Nutrients (2 groups)	Dissolved trace metals (2 groups)
Dissolved aluminium (2 groups)	HPLC pigments
DOC	Dissolved oxygen
Bottle salinities	Reversing thermometer temperatures
SPM gravimetry	Microzooplankton biomass
Radionuclides	

Stand-alone Pump Data

Trace metal	Organic carbon

Core Data

Amino acid profiles	Solid phase trace metal profiles
Calcium carbonate profiles	Mineralogy profiles
Carbon and oxygen isotope profiles	Density and water content profiles
Organic and inorganic carbon profiles	Organic and inorganic nitrogen profiles
^{210}Pb profiles	Grain size profiles

Landers

STABLE	Near-bottom current and turbidity data
Bed-hop Camera	Bottom photography

Cruise Jan Mayen JM1

CTD Data

Bottle Data

Phytoplankton species counts	Nutrients
POC/PON	Chlorophyll and phaeopigments

Cruise Charles Darwin CD85

Underway Data

Latitude and longitude	Bathymetry
Temperature and salinity.	Optical attenuance
Photosynthetically available irradiance	Chlorophyll

Bottle Data

Microzooplankton biomass and grazing	Phototrophic and heterotrophic nanoflagellates
Phytoplankton species counts	HPLC pigments
Spectrophotometric pigments	Fluorometric pigments (2 groups)
Nutrients	Bottle salinities
POC and particulate biogenic silica	

Production Data (In-situ/on deck incubations)

Carbon uptake	Nitrogen uptake
Phosphorous uptake	

CTD and SeaSoar Data **Marine Snow Profiler Data**

Continuous ADCP Data **RMT Net Data**

Longhurst-Hardy Plankton Recorder Data

Cruise An Cappall Ban CAPB1

CTD Data **Current meters deployed**

Cruise Jan Mayen JM2

CTD Data

Bottle Data

Phytoplankton species counts	Nutrients
POC/PON	Chlorophyll and phaeopigments

Drifting Sediment Traps

Phytoplankton fluxes

Cruise Belgica BG9412

Underway data

Latitude and longitude
Temperature and salinity
Wind speed and direction
Barometric pressure
pH, alkalinity, pCO_2 and TCO_2
Nutrients (NO_3, Si)
Bathymetry
Solar radiation
Air temperature
Dissolved oxygen
Chlorophyll

CTD Data

Bottle Data

Nutrients (2 groups)
Spectrophotometric pigments
Normalised carbon uptake (2 workers)
Normalised nitrogen uptake
Dissolved oxygen
Metal uptake and partition data
POC/PON
Fluorometric pigments
Normalised phosphorous uptake
Alkalinity and pH
DMS and DMSP
SPM gravimetry

Centrifuged SPM samples

SPM elemental analyses
Spectrophotometric pigments
SPM gravimetry
POC/PON

Stand-alone Pump Data

Trace metal data

Radiometer Profiles

Cruise Jan Mayen JM3

CTD Data

Bottle Data

Phytoplankton species counts
POC/PON
Nutrients
Chlorophyll and phaeopigments

Drifting Sediment Traps

Chemical fluxes	Plankton fluxes

Cruise Charles Darwin CD86

Underway Data

Latitude and longitude	Bathymetry

CTD Data

Bottle Data

Microzooplankton biomass	Nutrients
Dissolved oxygen	SPM gravimetry

Core Data

Pore water nutrient profiles	Pigment profiles
Nitrogen isotope profiles	Density and porosity profiles
Organic biomarker profiles	Pore water trace metal profiles
Pore water oxygen profiles	Foramenifera morphology profiles
Radiocarbon age profiles	Grain size profiles
Total and organic carbon profiles	Total nitrogen profiles
Benthic macrofauna biomass	Benthic meiofauna biomass

Landers

BOBO recovered	Currents and optical attenuance
TROL	Oxygen and resistivity sediment profiles
BOLAS	Sediment oxygen demand

Cruise Cote d'Aquitaine NAOMEX2

CTD Data

Core Data

Calcium carbonate profiles	Grain size profiles
Water content profiles	

Cruise Jan Mayen JM4

CTD Data

Bottle Data

Phytoplankton species counts	Nutrients
POC/PON	Chlorophyll and phaeopigments

Drifting Sediment Traps

Chemical fluxes	Plankton fluxes

Cruise Auriga PLUTUR III

CTD Data

Bottle Data

SPM gravimetry	POC

Core Data

Solid phase trace metal profiles	Calcium carbonate profiles
Grain size profiles	Organic carbon profiles
Water content profiles	

Cruise Jan Mayen JM5

CTD Data

Bottle Data

Phytoplankton species	Nutrients
POC/PON	Chlorophyll and phaeopigments

Drifting Sediment Traps

Chemical fluxes	Plankton fluxes

Cruise Jan Mayen JM6

CTD Data

Bottle Data

Phytoplankton species counts	Nutrients
POC/PON	Chlorophyll and phaeopigments

Drifting Sediment Traps

Chemical fluxes	Plankton fluxes

Cruise Jan Mayen JM7

CTD Data

Bottle Data

Phytoplankton species counts	Nutrients
POC/PON	Chlorophyll and phaeopigments

Drifting Sediment Traps

Chemical fluxes	Plankton fluxes

MADORNINA IIM Section Monitoring Cruises (7 off)

CTD Data

Bottle Data

Nutrients	Dissolved oxygen

Cruise Meteor M30_1

Underway Data

Latitude and longitude	Temperature and salinity
Solar and ultra-violet radiation	Wind speed and direction
Air temperature and humidity	Barometric pressure
Bathymetry	

CTD Data **Marine Snow Profiler Data**

Bottle Data

Atmospheric/dissolved carbonyl sulphide	Carbonyl sulphide production
Nutrients	POC/PON/particulate phosphorous
DOC (2 groups)	Dissolved oxygen (2 groups)
Total dissolved nitrogen and phosphorous	pH
Turbidity	Raw fluorometer
Bottle salinities	Reversing thermometer data

Amino acids and carbohydrates

Atmospheric Radon and condensation nuclei

DMS

Sediment Traps Recovered

Dry weight fluxes
Pigment fluxes
Phytoplankton fluxes
Trace metal fluxes
Carbon and nitrogen fluxes
Current meter data

Core Data

Solid phase trace metal profiles
Pore water N_2O profiles
Pore water nutrient profiles
Organic biomarker profiles
Pore water DOC and TCO_2 profiles
Pore water oxygen profiles
Pore water trace metal profiles
Grain size profiles

IFREMER Lander Deployment

Benthic Lander with Sediment Traps

Current velocity, nephelometry plus mass, carbon and nitrogen flux data

Cruise Cote d'Aquitaine NAOMEX3

Core Data

Calcium carbonate profiles
Water content profiles
Grain size profiles

Cruise Jan Mayen JM8

CTD Data

Bottle Data

Phytoplankton species counts
POC/PON
Nutrients
Chlorophyll and phaeopigments

Cruise Auriga Plutur IV

CTD Data

Bottle Data

SPM gravimetry
POC

Core Data

Solid phase trace metal profiles

Cruise Belgica BG9506

Underway data

Latitude and longitude
Solar radiation
Air temperature
Chlorophyll
Bathymetry
Wind speed and direction
Barometric pressure

CTD Data

Profiling Radiometer Data

Bottle data

Nutrients (2 groups)
Alkalinity and pH
POC/PON
Normalised phosphorus uptake
Microzooplankton biomass
Dissolved oxygen
Spectrophotometric pigments
Normalised carbon uptake (2 workers)
Normalised nitrogen uptake
Atmospheric and dissolved methane

Centrifuged samples

SPM elemental analyses

Air Sea Flux Determinations

Cruise Heincke 68

CTD Data

Bottle Data

Nutrients

Cruise Jan Mayen JM9

CTD Data

Bottle Data

Nutrients	POC/PON
Chlorophyll and phaeopigments	

Drifting Sediment Traps

Chemical fluxes

Cruise Charles Darwin CD94

Underway Data

Latitude and longitude	Bathymetry
Temperature and salinity	Optical attenuance
Photosynthetically available irradiance	Chlorophyll

CTD Data

Bottle Data

Microzooplankton biomass	HPLC pigments
Nutrients (2 groups)	DOC (2 groups)
Trace metals	Total dissolved phosphorus and nitrogen (2 groups)
Dissolved aluminium	POC/PON/particulate phosphorus
Dissolved and particulate carbohydrates	Dissolved oxygen
Reversing thermometer temperatures	Bottle salinities
SPM gravimetry	

Stand-alone Pump Data

Trace metal data

Core Data

Calcium carbonate profiles	Density and water content profiles
Organic and inorganic carbon profiles	Organic nitrogen profiles
^{210}Pb profiles	

Current Meters Recovered and Deployed Current velocity data

Cruise Auriga Plutur V

CTD Data

Bottle Data

SPM gravimetry
POC

Cruise Valdivia VLD153

Bottle Data

Nutrients

Cruise Jan Mayen JM10

CTD Data

Bottle Data

Nutrients
POC/PON
Chlorophyll and phaeopigments

Drifting Sediment Traps

Chemical fluxes

Cruise Valdivia VLD154

CTD Data

Bottle Data

HPLC pigments
DMS
Bucket temperatures

Air Sea Flux Determinations

Cruise Pelagia PLG95A

CTD Data

Core Data

Pigment profiles
Organic biomarker profiles
Organic and total carbon profiles
Porosity profiles

Benthic macrofauna biomass

Lander Data

BOLAS	Sediment oxygen demand
BIOPROBE	Benthic boundary layer currents, suspended load and particle characterisation

Benthic Trawl Data Benthic megafauna biomass

Cruise Discovery DI216

Underway Data

Latitude and longitude	Bathymetry
Wind velocity	Air temperature and humidity
Photosynthetically available radiance	Solar radiation
Barometric pressure	Temperature and salinity
Chlorophyll	Optical attenuance

CTD Data

Bottle Data

HPLC pigments	Trace Metals (2 groups)
Dissolved aluminium (2 groups)	Bottle salinities
Reversing thermometer temperatures	Nutrients (3 groups)
SPM gravimetry	Dissolved Oxygen (2 groups)
Microzooplankton biomass	

Core Data

Amino acid profiles	Mineralogy profiles
Grain size profiles	Pore water trace metal profiles
Pore water nutrient profiles	Porosity profiles

Current Meters Recovered and Deployed Current velocity data

Cruise Poseidon PS211

Underway

- Latitude and longitude
- Air temperature
- Temperature and salinity
- Wind velocity
- Barometric pressure
- Atmospheric and aqueous pCO_2

Bottle data

- DMS, DMSP and DMSO (2 groups)
- Atmospheric ammonia and methylamines
- Dissolved and atmospheric methane
- HPLC Pigments
- Dissolved methylamines
- Nutrients

Air Sea Flux Determinations

Cruise Pelagia PLG95B

CTD Data

Bottle Data

- Bottle Salinities
- Dissolved oxygen
- Nutrients

Core Data

- Pore water nutrient profiles
- Pore water trace metal profiles
- ^{210}Pb profiles
- Organic and total carbon profiles
- Pigment profiles
- Total nitrogen profiles
- Density and porosity profiles
- Foramenifera morphology profiles
- Grain size profiles
- Radiocarbon age profiles
- Organic biomarker profiles

Lander Data

- **BOBO recovered** — Near bed currents and transmissometer data
- **TROL** — Sediment oxygen and resistivity profiles

Cruise Belgica BG9521

Underway data

- Latitude and longitude
- Temperature and salinity
- Wind speed and direction
- Bathymetry
- Solar radiation
- Air temperature

Barometric pressure
pH, Alkalinity, pCO_2, TCO_2
Dissolved oxygen
Chlorophyll

CTD Data

Radiometer Profiles

Bottle Data

Nutrients (2 groups)
Dissolved oxygen
Normalised carbon uptake (2 workers)
Normalised nitrogen uptake
Fluorometric chlorophyll
POC/PON
Alkalinity and pH
Phosphorus uptake
Spectrophotometric pigments
Metal uptake and partition data

Cruise Jan Mayen JM11

CTD Data

Bottle Data

Nutrients
Chlorophyll and phaeopigments
POC/PON

Drifting Sediment Traps

Chemical fluxes

Cruise Belgica BG9522

Underway data

Latitude and longitude
Temperature and salinity
Wind speed and direction
Barometric pressure
pH, Alkalinity, pCO_2 and TCO_2
Bathymetry
Solar radiation
Air temperature
Dissolved oxygen
Chlorophyll

CTD Data

Radiometer Profiles

Bottle Data

Nutrients (2 groups)
Dissolved oxygen
Spectrophotometric pigments
Normalised carbon uptake
Metal uptake and partition
POC/PON
Alkalinity and pH
Fluorometric chlorophyll
Normalised nitrogen uptake

Cruise Discovery DI217

Underway data

Latitude and longitude	Bathymetry
Wind velocity	Air temperature and humidity
Photosynthetically available radiance	Solar radiation
Barometric pressure	Temperature and salinity
Chlorophyll	Optical attenuance

CTD Data **Marine Snow Profiler Data**

Bottle Data

Bottle salinities	Reversing thermometer temperatures
Microzooplankton biomass and grazing	Photosynthetic and heterotrophic nanoflagellate biomass
Phytoplankton species counts	HPLC pigments
Spectrophotometric pigments	Fluorometric pigments (2 groups)
Thymidine and leucine uptake	Bacterial cell numbers
Dissolved oxygen	pH
Turbidity	Nutrients
Total dissolved nitrogen and phosphorous	POC/PIC
SPM gravimetry	POC/PON/particulate phosphorous
Dissolved and particulate carbohydrates	Normalised nitrogen uptake

Production data (In-situ and on-deck experiments)

Carbon uptake	Nitrogen uptake

Longhurst-Hardy Plankton Recorder Data

SeaSoar Data **Continuous ADCP Data**

Cruise Charles Darwin CD97

Underway Data

Latitude and longitude	Temperature and salinity
Optical attenuance	Photosynthetically available irradiance
Chlorophyll	

Bottle Data

Fluorometric chlorophylls

XBT Data **Drifting Buoys Deployed**

Cruise Andromeda PLUTUR VI

CTD Data

Bottle Data

SPM gravimetry	POC

Core Data

Solid phase trace metal profiles	Calcium carbonate profiles
Radioisotope profiles	Grain size profiles
Organic carbon profiles	Total nitrogen profiles
Water content profiles	

Ships of Opportunity

Continuous Plankton Recorder Merchant ship transects over the Goban Spur area for 1993-95.

Appendix 2

Tables of the Oracle Relational Database

ADCP/ADCPINDX	ADCP profile index and datacycles
ARGOS	Drifting buoy tracks
BINCTD	CTD datacycles
BOTDATA/BOTTLE	Water/air sample index and datacycles
C14DAT/C14HDR	^{14}C uptake long in-situ and on-deck incubation data and ancillary information
COREINDX	Index of cores and sub-cores
COREPROF/CORESAMP	Core profile dependent and independent variables
CORETOT	Whole core property (e.g. benthic flux) data
CPR_COLOUR	Continuous Plankton Recorder silk colour data
CPR_PHYTO	Continuous Plankton Recorder phytoplankton species data
CPR_ZOO	Continuous Plankton Recorder zooplankton species data
CTDCAL	CTD calibration coefficients
CTDINDX	CTD profile index
CTDTYP	CTD type code table (defines a field in CTDINDX)
EVENT	Catalogue of data collection events.
EVENT_COMM	Plain language comments supporting EVENT
FORAMS	Benthic foramenifera species data
G_CODE	Event gear code table
INTBOT	Column integrated size-fractionated chlorophyll data
LHPR	Longhurst-Hardy Plankton Recorder biomass data
MEGADAT/MEGAHEAD	Benthic megafauna species data
MEIODAT/MEIOHDR	Benthic meiofauna species data
MFDAT/MFHEAD	Benthic macrofauna species data
MOORINDX	Catalogue of moored instrument deployments
MOOR_PARAMS	Parameter code table supporting MOORINDX
MSP	Marine Snow Profiler data
MTALDAT/MTALHDR	Trace metal uptake kinetics data
N15DAT/N15HDR	^{15}N uptake long in-situ and on-deck incubation data and ancillary information
NEPH	CTD nephelometer data
NETINDX	Net haul index
ORGCODE	Data originator code table
P33DARK/P33DAT/ P33HDR	$^{32/33}P$ uptake long in-situ and on-deck incubation data. PvI data and ancillary information
PRINDX/PRPROF	Profiling radiometer data
SSINDX	SeaSoar 'pseudo-CTD' profile index
STINDX/TRAPDATA	Sediment trap sample index and data
XBT/XBTINDX	XBT profile index and data
ZUCT/ZUNT/ZUPM/ ZUSG	Parameter dictionary

Appendix 3

Nutrient Data Documentation

Parameter Code Definitions

AMONAAD2 Dissolved ammonium
Colorometric autoanalysis (0.4/0.45 μm pore filtered)
Micromoles/litre

AMONMATX Ammonium (unfiltered)
Manual colorometric analysis (unfiltered)
Micromoles/litre

NTRIAAD2 Dissolved nitrite
Colorometric autoanalysis (0.4/0.45 μm pore filtered)
Micromoles/litre

NTRIAAD5 Dissolved nitrite
Colorometric autoanalysis (0.2 μm pore filtered)
Micromoles/litre

NTRIAATX Nitrite (unfiltered)
Colorometric autoanalysis (unfiltered)
Micromoles/litre

NTRZAAD2 Dissolved nitrate + nitrite
Colorometric autoanalysis (0.4/0.45 μm pore filtered)
Micromoles/litre

NTRZAAD5 Dissolved nitrate + nitrite
Colorometric autoanalysis (0.2 μm pore filtered)
Micromoles/litre

NTRZAATX Nitrate + nitrite (unfiltered)
Colorometric autoanalysis (unfiltered)
Micromoles/litre

PHOSAAD2 Dissolved phosphate
Colorometric autoanalysis (0.4/0.45 μm pore filtered)
Micromoles/litre

PHOSAAD5 Dissolved phosphate
Colorometric autoanalysis (0.2 μm pore filtered)
Micromoles/litre

PHOSAATX Phosphate (unfiltered)
Colorometric autoanalysis (unfiltered)
Micromoles/litre

PHOSMATX Phosphate (unfiltered)
Manual colorometric analysis (unfiltered)
Micromoles/litre

SLCAAAD2 Dissolved silicate
Colorometric autoanalysis (0.4/0.45 µm pore filtered)
Micromoles/litre

SLCAAAD5 Dissolved silicate
Colorometric autoanalysis (0.2 µm pore filtered)
Micromoles/litre

SLCAAATX Silicate (unfiltered)
Colorometric autoanalysis (unfiltered)
Micromoles/litre

SLCAMATX Silicate (unfiltered)
Manual colorometric analysis (unfiltered)
Micromoles/litre

UREAMDTX Urea (unfiltered)
Manual analysis using the diacetylmonoxime method
Micromoles/litre

Originator Code Definitions

Belgica cruise BG9309

10	Ir. Marc Elskens	VUB, Brussels, Belgium
14	Dr. Lei Chou	ULB, Brussels, Belgium
66	Dr. Ricardo Prego	CSIC, Vigo, Spain

Belgica cruises BG9322, BG9412, BG9506, BG9521 and BG9522

10	Ir. Marc Elskens	VUB, Brussels, Belgium
14	Dr. Lei Chou	ULB, Brussels, Belgium

Cruises Pelagia PLG93, Charles Darwin CD86 and Pelagia PLG95B

11	Dr. Wim Helder	NIOZ, Texel, the Netherlands

Meteor cruises M27_1 and M30_1, Valdivia Cruise VLD137 and Discovery cruise DI217

9	Mr. Thomas Raabe	Hamburg University, Germany

Charles Darwin cruise CD83

39	Mr. Bob Head	Plymouth Marine Laboratory, UK

Charles Darwin cruise CD84

12	Dr. David Hydes	Southampton Oceanography Centre, UK
14	Dr. Lei Chou	ULB, Brussels, Belgium

Charles Darwin cruise CD85

3	Dr. Ian Joint	Plymouth Marine Laboratory, UK

Charles Darwin cruise CD94

9	Mr. Thomas Raabe	Hamburg University, Germany
53	Prof. Mike Orren	University College Galway, Eire.

Discovery cruise DI216

12	Dr. David Hydes	Southampton Oceanography Centre, UK
14	Dr. Lei Chou	ULB, Brussels, Belgium
53	Prof. Mike Orren	University College Galway, Eire.

Jan Mayen cruises JM1-JM11

61	Dr. Paul Wassmann	University of Tromsø, Norway

Poseidon cruise PS211

70	Dr. Ludger Mintrop	IfM Kiel, Germany

Heincke cruise HEINK68 and Valdivia cruise VLD153

90	Dr. Pete Bowyer	University College Galway, Eire.

Originator Protocols

Ir. Marc Elskens

Water samples were taken from manually filled bottles deployed from an inflatable boat away from Belgica (ria surveys) or taken from water bottles deployed on a CTD rosette. On two cruises (9322 and 9412) continuous underway measurements were made by drawing discrete

samples at frequent intervals from the continuous seawater supply. Note that these data are stored in the underway binary merge files and not in the BOTDATA table.

Nutrient determinations were carried out on board ship, immediately after sampling. Nitrate plus nitrite and phosphate were determined using a Technicon AA2 autoanalyser as described by Elskens and Elskens (1989).

Ammonia was determined according to the manual method using indophenol blue described in Koroleff (1969) using a Baush and Lomb Spectronic 21 spectrophotometer.

Urea was determined using the diacetymonoxime method of Mulvena and Savidge (1992) modified to allow precise analyses when strict control of the reaction temperature is impossible as described by Goeyens et al (submitted 1996).

Dr. Lei Chou

Manual spectophotometric analysis for phosphate, nitrite and silicate were done using the methods specified in Grasshoff et al (1983). These analyses were usually carried out on board ship as soon after sampling as possible. Samples were kept refrigerated and dark between collection and analysis.

Samples for nutrient determination by autoanalysis were kept frozen until analysed. A separate set of samples were usually taken specifically for silicate analysis. Samples were analysed on a SKALAR autoanalyser.

Dr. Ricardo Prego

Nitrate plus nitrite was determined using a Technicon AAII autoanalyser with the adaptation described in Mourino and Fraga (1985). Phosphate and silicate were determined using a Technicon AAI autoanalyser following the method described by Hansen and Grasshoff in Grasshoff et al (1983).

Dr. Wim Helder

Samples were taken from water bottles deployed on a CTD rosette and analysed at sea using a TRAACS 80 autoanalyser, usually within 4 hours of collection. Samples were stored in cool and dark conditions between collection and analysis.

On cruise Pelagia 93, the samples from the CTD rosette were analysed unfiltered. On Charles Darwin 86 and Pelagia 95, the samples were filtered through a 0.45 micron acrodisc filter to improve the quality of the ammonium results.

The following chemistries were used:

Ammonium:	Phenol method
Phosphate:	Ammonium molybdate / ascorbic acid method
Nitrate / nitrite:	Sulphanylamide / napthylethylenediamine method using a Cu/Cd coil (efficiency >98%) for reduction
Silicate:	Ammonium molybdate / ascorbic acid method

Samples were always analysed from the surface to the bottom to minimise the risk of cross-sample contamination.

Working standards were freshly prepared daily by diluting stock standards to the required concentration with natural, aged, low-nutrient seawater. The nutrient concentrations in this were determined by manual colorometric analysis. The low-nutrient seawater was also used as a wash between samples. A second mixed nutrient stock, poisoned with 0.2% chloroform or 20 mg/l $HgCl_2$, was used as an independent check. Pipettes and volumetric flasks were calibrated before each cruise and standard batches were intercalibrated.

Accuracy of analyses is reported as about 1% of the full scale value for nitrate, nitrite and silicate and 2% of the full scale for phosphate and ammonium.

The data were reported as nitrate and nitrite, the nitrate values having been computed by subtracting nitrite from nitrate plus nitrite. BODC practice is to store nitrate plus nitrite and the values in the database have been determined by summing the nitrate and nitrite values supplied. In cases where multiple bottles were fired at a single depth, nutrient values were reported from each bottle. These have been averaged, excluding any bottles flagged as leaking, to give a single nutrient value for each depth.

Mr. Thomas Raabe

Water samples were taken from bottles deployed on a CTD rosette and analysed immediately on board ship. Samples were analysed unfiltered, provided the particulate content was not considered too high in which case the samples were GF/C filtered. Parameter coding has assumed analysis of unfiltered samples.

Samples were analysed using a Technicon autoanalyser system using the method of Murphy and Riley (1962) as modified by Eberlein and Kattner(1987) for phosphate, the method of Grashoff (1983) for silicate, the method of Koroleff (1969) for ammonia and the methods of Armstrong et al (1967) for nitrate and nitrite.

Dr. David Hydes

Samples were collected from either bottles on the CTD rosette or the continuously pumped surface sea water supply and immediately analysed unfiltered using a Chemlab autoanalyser. Samples were analysed in triplicate and the mean value stored in the database.

Professor Mike Orren

Samples were collected from either bottles on the CTD rosette or the continuously pumped surface sea water supply and analysed using an Alpkem autoanalyser. This machine and the chemistries employed for phosphate and silicate were progressively modified during the project in an attempt to obtain reasonable performance. The following modifications were described:

The length of all tubing was reduced to the absolute minimum.

The instrument was thoroughly cleaned with Decon90 before each procedure.

The phosphate determination wavelength was switched to 760 nm, with wash and sample times switched to 60 and 30 seconds respectively.

The wavelength used for silicate was switched to 795 nm. The ascorbic acid reagent was prepared without the recommended acetone addition, the sulphuric acid concentration was doubled and the ammonium molybdate was filtered prior to each procedure.

Samples were generally analysed on board ship but some samples taken towards the end of a cruise had to be analysed back in the laboratory. These were kept in the dark and as cool as possible between collection and analysis.

Dr. Paul Wassmann

Water samples were taken from the CTD rosette, fixed with 0.2 ml of chloroform and kept cold (4 °C) and dark until analysed by autoanalyser following the protocols of Føyn et al (1981).

Dr. Ian Joint

Standard autoanalyser methods were used as described in Rees et al (1995). Nitrite corrected nitrate data were supplied to BODC. The nitrite corrections were removed and the data stored as nitrate+nitrite in the database.

Mr. Bob Head

Standard autoanalyser methods were used, with a 3-channel (nitrate+nitrite, silicate and phosphate) instrument logged onto chart recorders.

Dr. Ludger Mintrop

Water samples were taken from the ship's 'moon pool' and immediately frozen. The samples were transferred to the Polarstern and analysed several months after collection using standard photometric methods on a four channel autoanalyser.

Nitrate and nitrite were supplied as separate channels. These were summed by BODC to give the nitrate+nitrite channel stored.

Dr. Pete Bowyer

Samples were taken from the CTD rosette bottles, filtered and immediately frozen. Back in the laboratory, the samples were analysed on an Alpkem autoanalyser (the same instrument used by Professor Mike Orren) with four channels (nitrate, nitrate+nitrite, phosphate and silicate).

The nitrate data supplied to BODC had been corrected by subtraction of nitrite. These were restored to nitrate+nitrite for storage in the database.

Comments on Data Quality

Belgica cruise BG9309

The SKALAR autoanalyser phosphate data were supplied with a warning that there may be problems. On a number of stations all three laboratories provided phosphates and for a number of stations there were also manually analysed phosphates from ULB. Comparing these data it can be clearly seen that the SKALAR values are frequently way too high. Consequently, the SKALAR phosphate data set has been flagged ‘L’.

For the stations where intercomparison of NO_3+NO_2 data is possible, the ULB data are generally higher than the VUB data which are, in turn, generally higher than the CSIC data. None of the data have been flagged. Users are advised to retrieve all three data sets and reach their own conclusions about which data to use.

Belgica Cruise BG9412

On this cruise the ULB NO_3+NO_2 data, with the exception of a handful of points, are significantly lower than the VUB data. Differences of 10 per cent and more are predominant throughout the overlapping data set.

The phosphate and nitrite data sets show excellent agreement.

Charles Darwin CD83

Problems with the colorimeter were reported for this cruise, giving rise to variable sensitivity and non-linear calibrations. The problem was circumvented by careful calibration for each individual CTD cast and is not believed to have affected data quality.

Charles Darwin cruise CD84

Both ULB and SOC measured the nitrate+nitrite profile at the Belgica station. The two data sets show very good agreement.

Charles Darwin cruise CD85

The nitrate+nitrite data for station 04_09 exhibited a curious gradient from 7 to 10 µM increasing towards the surface in the top 10m. The data points concerned have been flagged suspect as has a single anomalously high ammonium value. Other than these, no problems could be identified in the data set.

Charles Darwin cruise CD94

A subset of the nutrient channels (NO_3+NO_2, PO_4 and silicate) were measured by both Hamburg and Galway universities. Both data sets included a small number of anomalous data values. These have been flagged suspect ('M') together with data from bottles where there is strong evidence of contamination through leakage.

The nitrate+nitrite and silicate data from the two groups compare extremely well and no systematic difference between the two data sets could be established. On some casts the

Hamburg data were slightly higher whilst on other casts it was the Galway data that were slightly higher.

Regressing the two data sets gave the following results:

Nitrate+nitrite Galway = Hamburg * 0.9591 + 0.4471 (R^2 = 98%)
Silicate Galway = Hamburg * 1.0188 - 0.1091 (R^2 = 99%)

The results for phosphate were not as good. The Galway values were systematically significantly lower than the Hamburg data, sometimes by as much as 50%. The intercalibration plot exhibited much more scatter than the plots for the other two nutrients.

Regressing the two data sets gave the following result:

Phosphate Galway = Hamburg * 0.9234 - 0.0939 (R^2 = 83%)

The Hamburg data compare more favourably with data from other cruises where the phosphate values are believed to be good quality. It is therefore recommended that the Galway phosphates be used with caution, bearing in mind that they are probably low. However, either nitrate+nitrite or silicate data set may be used with confidence.

Discovery Cruise DI216

Nutrients were measured by three groups on this cruise: nitrate+nitrite, silicate and phosphate were measured by SOC; phosphate was determined manually by ULB; nitrate+nitrite and silicate were determined by the Galway group.

The ULB and SOC phosphate data show very good agreement. ULB reported some phosphate samples contaminated and these have been flagged ‘L’ in the database.

The SOC data are believed to be of extremely high quality. Indeed the data were used successfully to identify CTD rosette misfires due to the close proximity of the values from unintentional ‘blind duplicates’. The only problem encountered with the SOC data were the nitrate+nitrite values for one cast (CTD4) which were obviously low. This was attributed to the reduction column being poisoned by mercury in an internal standard and the data have been flagged.

The Galway data from CTD bottles were compared with the SOC data and flagged if they deviated from the SOC values by more than 10 per cent. The same ‘blind duplicates’ described above were analysed by Galway but the replication was very poor. Users are recommended to use the SOC data rather than the Galway data whenever possible.

Samples from the continuous sea water supply were not analysed by SOC. The Galway data are erratic and in many cases incredibly high. With the exception of samples taken on a section up the Channel right up to the Solent, surface nitrate+nitrite values in excess of 0.75 μM and silicate values in excess of 1.0 μM have been flagged suspect by BODC. The remaining data should be used with caution.

Cruise Poseidon PS211

A small number of the nutrient values were obviously anomalously high for oceanic surface sea water. Nitrites in excess of 0.5 μM (plus the associated nitrate+nitrite values), phosphates in excess of 1.5 μM and silicates in excess of 5 μM were flagged suspect. This affected between 1 and 4 data values in each channel.

There is, however, some concern about the remaining data, particularly the silicates and, to a lesser extent, the nitrate+nitrite channel. The pattern of the data is more uneven than one would expect for surface values, particularly in the lower nutrient waters encountered south of 52 °N. Users are advised to examine the data carefully and make their own judgements on whether further data should be rejected before making use of this data set.

Cruise Heincke HEINK68

A small number of the nitrite values were anomalously high. All values in excess of 0.5 μM (four in total) were flagged suspect in the database.

Cruise Valdiva VLD153

A number of isolated values that were obviously anomalous have been flagged suspect in the database.

However, the main problem with the data from this cruise were the nitrites. The values for stations 40-58 and 93-104 were consistently and unrealistically high (0.9-5 μM) whereas the values from the remaining stations, apart from a couple of high spikes, were normal. Consultation with the data originator revealed a calibration scaling problem, by a factor of 10, for these samples. On the basis of this information, the nitrite data in the database for the affected stations have been divided by 10.

Note that the uncorrected nitrites were added to the nitrate data to give nitrate+nitrite so as to accurately reverse the correction made by the data originator.

A number of the silicate profiles, particularly stations 76, 82, 83, 86, 88, 90, exhibit oscillating values rather than a progressive increase from depth to surface. The fact that this phenomenon was confined to consecutive samples from one of the three sections raised a question as to whether this was real and not an analytical artefact. Consequently, the profiles have not been flagged.

A references section containing all relevant citations is present in the original document, but has been excluded from this Appendix.