

Data Management Policy for the NERC Marine & Freshwater Microbial Biodiversity (M&FMB) thematic programme

Approved by the M&FMB Steering Committee: February 2003

1. Introduction

1.1 NERC requires all thematic programmes to plan adequately for the management of the data they collect, in accordance with the Council's current Data Policy (www.nerc.ac.uk/data/policy.shtml) and the overall requirement for grant holders to offer to deposit with NERC a copy of datasets resulting from their research. Thematic planning must cover not only the practical arrangements while the programme is running, but also the subsequent maintenance of datasets of longterm importance. Properly managed, data provides a key NERC resource, which will continue to be used long after the formal end of a programme. The scale of effort dedicated to data stewardship should reflect the anticipated longterm value of the data.

1.2 In the context of NERC Data Policy, "data may be held in either analogue or digital form and be stored either on paper or a variety of computer-compatible media... physical specimens in curated collections are outside the usual sense of the word". Arrangements for culture-deposition and culture access are therefore not covered here. Nevertheless, there are important data-culture linkages relevant to the M&FMB programme, and their policy implications are currently being reviewed by the Steering Committee (for resolution by mid-2003).

1.3 NERC emphasises the importance of environmental data that are time-dependent – since, if lost, such information cannot be regenerated. There is also particular value for datasets that bring together complementary information generated by different approaches, with potential for higher-level analyses and improved interpretation of component studies. For those reasons, M&FMB gives highest priority to data management arrangements for:

- data arising from the AMBITION research cruise, and from any future M&FMB sea-going activities
- data arising from the coordinated programme studies carried out at Priest Pot.

1.4 For the above two categories, submission of project data to a NERC Data Centre is an obligatory programme requirement, for longterm data management. The necessary arrangements are in place, as detailed in the remainder of this document. However, for datasets without links to the above, eg for most molecular-based studies, it is the responsibility of M&FMB investigators to determine the optimal arrangements, maximising 'future usability' on the basis of the guidance given in Section 3 below. M&FMB data policy in this area reflects the much greater diversity of types of information that may be obtained from laboratory-based studies (relevant, in some cases, to some relatively-specialised Data Centres, eg GenBank), also the fact that some technology-dependent data may be of little value within 5-10 years, as technologies change.

2. Role of BODC

2.1 A wide range of data (marine and freshwater, environmental and molecular) will be produced by M&FMB thematic. Whilst there is no single NERC Designated Data Centre covering this range, the British Oceanographic Data Centre (BODC) has been appointed as having the 'lead role' – working closely with other groups and data management organisations able to provide complementary services of a more specialist nature. For example, freshwater environmental data will initially be brought together by CEH Windermere. However, BODC will provide metadata information services for the programme, keeping track of other locations where M&FMB data are held.

2.2 Thus BODC will be aware of the totality of data arising from the M&FMB programme (including fieldwork activities not at Priest Pot or via NERC research cruises), and will, as far as is practicable, work to ensure that similar quality standards for data stewardship apply to all data collected through M&FMB support. Dr Gwen Moncoiffé (gmon@bodc.ac.uk) is the BODC-nominated contact person, and is currently directing BODC effort on the integration of data arising from the M&FMB research cruise.

3. Other data management arrangements

3.1 GenBank (www.ncbi.nlm.nih.gov/Genbank/) is the main international sequence databank, currently containing information on more than 50,000 species. Pre-publication of sequence information to GenBank or other similar databases, eg EMBL (European Molecular Biology Network www.embl-heidelberg.de), is now well-established practice for research groups producing sequence data. Following submission, an accession number is provided – than is then used in the publication to refer to the sequence. Sequence data submitted in advance of publication can be kept confidential if requested.

3.2 The NERC Environmental Genomics programme is developing a bioinformatics-based Data Centre at CEH Oxford to meet its own data management needs. The EG Data Centre (<http://envgen.nox.ac.uk/>) will focus on Expressed Sequence Tags (EST) data and microarray (transcriptomics) information, also the provision of bioinformatics training. The computing infrastructure includes both centralised resources and a network of specialised computers in EG award-holders' laboratories. When these systems are fully operational (in late 2003/early 2004), there may be 'buy-in' options available for the M&FMB programme.

3.3 It is recognised that neither 3.1 nor 3.2 above (nor taxon-specific databases, eg the Universal Virus Database www.ncbi.nlm.nih.gov/ICTVdb/) meet the need for many M&FMB researchers to provide wider access to much of the laboratory data collected through project support, yet not publishable *per se*. PIs are therefore strongly encouraged to set up online databases of their own, or hosted by a journal, that may be publication-specific; eg providing supplementary data on methods, or the full results from which the summaries are provided in the paper as tables or figures. Such practice of having additional information online is increasingly being used by Nature and Science. BODC should be informed of these arrangements (where they relate to publications arising from M&FMB support), and sufficient keywords provided so that a central, searchable listing can be developed, with the necessary links. A minimum lifetime of 5 year for independent hosting is suggested, after which BODC may take on responsibility for longterm access arrangements (via inclusion into the M&FMB database).

4. Minimum standards of stewardship for NERC corporate data

4.1 The following NERC-wide minimum standards are expected to apply when (digital) datasets form part of NERC's enduring data resource. These requirements will be looked after 'automatically' for the M&FMB datasets managed by BODC. Nevertheless, PIs need to be aware of this framework, particularly if alternative means of longterm data stewardship are envisaged.

- i) The ownership and Intellectual Property Rights [see Section 8 below] to the dataset must be established, and policy towards exploiting and making it available to third parties agreed
- ii) The dataset must be catalogued to the level of detail required by a NERC Designated Data Centre, so that it can be mentioned in web-based NERC data catalogues
- iii) Formal responsibility for the custody of the dataset must be agreed
- iv) The data must be fully 'worked up' (ie calibrated, quality-controlled etc) with sufficient associated documentation to be of use to third parties without reference to the original collector

- v) The technical details of how the data are to be stored, managed and accessed must be agreed and suitably documented
- vi) The technological implications must be established (digital data stewardship implies the need for an underlying infrastructure of IT equipment and support)
- vii) The resources need to carry out these intentions over the planned life of the data, in terms of staff (whether in project teams or the Data Centre) and IT equipment/infrastructure must be estimated and sources identified.
- viii) A review mechanism must exist to reconsider periodically the costs and benefits of continuing to maintain the data. The intention to destroy or put at risk data should be publicised in advance, allowing time for response by interested parties.

5. Data acquisition (for a NERC Data Centre)

5.1 A well-structured and user-friendly identification system is essential for data collection, particularly when combining information of different types, from different sources. For NERC research cruises, it is initially the responsibility of the cruise Principal Scientist to achieve a commonality of approach, using station and/or observation numbers to achieve a consistent record of gear deployments, environmental observations and analytical calibrations in the context of associated on-board experiments. The protocols followed for the M&FMB AMBITION research cruise (*RRS Charles Darwin* 132) are detailed in the Cruise Report, published in hard copy in December 2002 and soon to be available on the web (via BODC).

5.2 Processed and project-specific cruise data must be provided to BODC by the Principal Scientist and project teams as they becomes available, not in the concluding few months of projects.

5.3 Similar arrangements for 'on-going' data collation are being developed by CEH Windermere to obtain an internally-consistent observational record for Priest Pot, to be integrated as rapidly as possible with data arising from experimental studies of a project-specific nature.

6. Data formats and data media (for NERC Data Centres)

6.1 Digital data should be collected and stored using widely-available software products and their related data formats. Whilst BODC has experience in handling a very wide range of software, formats and media, PIs should discuss with them the proposed use of any data-handling or storage protocols that might be regarded as non-standard.

6.2 CD-ROMs are currently the preferred means for making integrated data products from thematic programmes available to the wider research community. The M&FMB Steering Committee will advise on the number of CDs expected to be produced by BODC, and set target times for their release.

7. Data back-up policy

7.1 Daily back-up programmes apply at BODC (and other NERC Designated Data Centres) to safeguard major digital databases. Project PIs are responsible for providing appropriate back-up strategies for unique digital data stored locally and/or via other organisations.

7.2 As far as possible, analogue data (such as photographs) should be "disaster proofed" by transferring them into digital form, eg by scanning. Such duplication is not a waste of effort, even though the original, analogue version may have a longer lifetime than the format/media used for the digital transcription. Such data may then be included on a programme CD. Note that BODC has considerable experience in managing and publishing image data.

8. Protection of data originators' Intellectual Property (IP) rights

8.1 NERC policy on IP ownership is given as Annex 1. The following arrangements have been developed by BODC to ensure an appropriate balance between the protection of data originators' intellectual property rights and the potential benefits that may arise via data use by the programme, the wider research community and other interested parties.

- i) All data collected in the M&FMB programme through NERC funding and provided to BODC should be freely available to all programme participants (PIs and CoIs) for M&FMB purposes on the condition that the originator is kept informed about how the data are being used and is duly acknowledged in any exploitation of that data
- ii) Due acknowledgement is considered to be co-authorship, specific reference to the data source or a share of any financial reward. The form of this should be negotiated between the data originator and the data exploiter. If a dispute should arise, then the problem will be referred to the Steering Committee for resolution.
- iii) Until M&FMB data enter the public domain, BODC will not transfer them to parties outside the programme without the explicit agreement of the originator. Steering Committee advice will also need to be sought if major data transfers are involved, to avoid compromising the interests of other programme participants.
- iv) The mechanism for entry into the public domain is expected to be the release of M&FMB CD-ROM(s) at the conclusion of the programme.
- v) A condition of CD-ROM usage is that it is regarded as a data publication and all usage of the data contained therein should acknowledge the data originator through citation

9. Data availability

9.1 It is NERC policy to ensure that "individual scientists, principal investigator teams and participants in programmes will be permitted a reasonable period to work exclusively on, and publish the results of, the data collected by such individuals and teams". Nevertheless, as the M&FMB programme develops, there is necessarily a sequential widening of access to data (and materials, eg isolates and cultures). This process has already been outlined in (8) above. It can be generalised with reference to three access levels:

Level 1 (Project). Availability limited to the investigators responsible for data collection (for M&FMB cruises, most data collection is expected to be a shared responsibility; thus group ownership applies, under overall control of the Principal Scientist); any wider sharing at the discretion of the investigators

Level 2 (Programme). When data are transferred to BODC [or equivalent, with links to BODC] their availability is automatically extended to other investigators within the M&FMB programme. Nevertheless, their further use is still under the control of the data originator, and any wider sharing is at the discretion of the M&FMB Steering Committee.

Level 3 (Public). Data publication, at or near the end of the programme. Availability extended to external users, either openly (for academic use) or at the discretion of BODC/NERC (for commercial exploitation, in consultation with data-originators).

9.2 It is to the benefit of the programme as a whole that inter-project collaborations are developed under Level 1, and that the transition between Levels 1 and 2 is made as rapidly as possible.

10. Identifying data for management purposes

10.1 It is important that the M&FMB programme maintains an awareness of all data collected through its support, including outputs from partnership arrangements (eg from fieldwork using non-NERC vessels, or through collaborations involving other funding agencies). A reporting system will be established to gain information on such data, and their stewardship arrangements, whether or not via BODC.

Annex 1: NERC Policy on Ownership of Intellectual Property

"Ownership of Intellectual Property (IP) and all IPRs arising from standard grants lies with the organisation in receipt of the grant. In NERC research and collaborative centres the ownership normally remains with NERC as the legal entity, but the originating institute can be identified. NERC Council (NERC's governing body) reserves the right for NERC to retain, for a limited period, an exclusive right to exploit IP in partnership with collaborative centres to the benefit of the UK. This is to avoid cases where separate pieces of IP (eg datasets) could reduce the likelihood of exploitation. These must be discussed and agreed prior to the start of any work."