RAPID Climate Change Programme

Data Management Plan

NATURAL
ENVIRONMENT
RESEARCH COUNCIL

1. Introduction

NERC requires all Thematic Programmes to plan and implement a data management scheme. The planning must cover the practical arrangements while the programme is running and the subsequent maintenance and long-term curation of the data sets. The latter is increasingly important in view of the Environmental Information Regulations, which place a duty on Government funded bodies to make all publicly funded data readily and easily available.

The NERC Data Policy requires that all data are lodged with the appropriate NERC Designated Data Centre. In the context of RAPID these are the British Oceanographic Data Centre (BODC) and the British Atmospheric Data Centre (BADC), the respective Designated Data Centres for Marine and Atmospheric Sciences. The minimum required standards of stewardship are summarised in section 3.

NERC provides funding to the Data Centres for basic infra-structure support and the long-term maintenance and curation of NERC's data assets. Thematic Programme budgets include the funds necessary for within project data management for the life of the project.

An integral part of the Data Plan is an obligation upon RAPID Programme Principal Investigators (PIs) to ensure that data management is undertaken in a suitable way, and that adequate consideration is given to the "data side" of their work. Individual project 'data management plans' will cover staff responsibilities, data collection policies, data standards, resourcing of data management, data quality and quality assurance.

The data management policy as defined by the RAPID steering group is outlined in Annex I.

This plan has been formulated following a review of the specified resource requirements and outputs set out in the project proposals and a series of discussions between BODC/BADC, the RAPID Science Co-ordinator (Dr Meric Srokosz) and the project PIs in order to assess the scale of data collection/production. These include observational and modelling products, physical samples and the requirements to enable links to third party data sets.

2. The Role of the RAPID Data Centre (RDC)

Submission of and access to data will be through a common 'portal' and for the purposes of RAPID the term RAPID Data Centre (RDC) will refer to BADC and BODC. Data management costs have been allocated in the RAPID budget for RDC services.

Given the complex and broad range of data encompassed by the RAPID programme the nature of the data management will vary between projects. The basis of this has been agreed with the PIs following an initial dialogue with the RDC and RAPID Science Co-ordinator.

The RDC will be the focal point for PIs regarding data issues. The RDC website will contain inventories providing comprehensive up to date information about the status of all project data sets and model runs, so that all RAPID participants can easily request available data. The RDC will service data requests by RAPID participants.

Following the completion of RAPID the RDC will ensure that data are passed to the appropriate International Data Centres, ensuring that NERC meets its international obligations.

3. Minimum Standards of Stewardship for NERC Data

The following minimum standards are expected to apply when (digital) data sets form part of NERC's enduring data resource:

i. The ownership and Intellectual Property Rights to the data set must be established, and NERC's policy towards exploiting and making it available to third parties agreed.
ii. The data set 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 data set must be agreed.
iv. The data must be fully "worked up" (i.e. 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 needed 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.

The above NERC-wide requirements, set out in the NERC Data Policy (www.nerc.ac.uk/data/policy.shtml), will be looked after "automatically" for the RAPID data sets managed by BODC and BADC. Nevertheless, PIs need to be aware of this framework.

4. Data and Sample Acquisition

RAPID data cover a broad subject area, including oceanographic data, sediment cores, palaeo data and the generation of model output. It is not the intention of this document to specify in detail how these data be collected, described and delivered to the data centres, however, a number of generic principles need to be adhered to.

Processed and project-specific data must be provided to the RDC by the Principal Scientist and project teams as they become available, not in the concluding few months or weeks of projects. However, great importance is given, both by the programme and by the RDC, to protecting the interests of data originators, and restrictions on the wider availability of the RDC-held data sets will therefore apply (see the Data Policy at Annex I).

A well structured and user-friendly identification system is essential for cruise-based data collection and sample labelling. Such arrangements are traditionally the responsibility of the cruise Principal Scientist. However, in order to assist the PIs and RDC it is necessary that a representative of the RDC attend the pre-cruise meeting.

Station identifiers, navigational information and "basic" oceanographic data (for which the RDC will have quality-control responsibilities) must be provided to the RDC by the Principal Scientist immediately after a cruise. Normal practice, as for other Thematic Programmes, will be for the RDC to meet the ship when it docks in the UK and to take delivery of this material together with a copy of the logs, calibration data and sensor information. If a cruise terminates in a foreign port it will be necessary for the PI and a representative of the RDC to meet immediately on return of the PI to the UK. A copy of the Cruise Summary Report (ROSCOP form) should be provided to the RDC by the Principal Scientist within one working week of the end of the cruise. A copy of the full cruise report should also be sent to the RDC, electronically, as soon as it is completed. The RDC will then assist in making this more widely available (e.g. via a link from the main programme website).

In the case of palaeo digital data, a representative from the RDC needs to be involved prior to and immediately after field campaigns in order to obtain the necessary information to describe the nature of the collected data. Where appropriate, both the raw and derived data will be stored together. For example, data resulting from analyses on an ice core would have raw parameters, such as depth in core, which would be stored alongside the derived parameters, such as age. Accompanying the data will be a description of the method used to arrive at the derived parameters.

For projects collecting physical samples it is the responsibility of the PIs to ensure that appropriate management measures are in place. However, it is important that the necessary collection details are provided to the RDC in order that the information forms part of the overall project information. For deep sea sediment cores, the samples must be deposited with BOSCOR (British Ocean Sediment Core Repository), hosted by the Southampton Oceanography Centre. Those dealing with palaeo samples (e.g. speleothem, bog cores and lake samples) will continue with existing management practices.

In the case of model data, the details for submission and serving will be agreed with individual PIs. Broad principles are given in section 5. In general, information accompanying submitted model data should include the model name and version number and a brief description of the model's general aim. See the metadata protocol (Annex II) for more detail.

Metadata

Metadata are a crucial part of any data archive since they ensure that the data can be understood at a later date. To guarantee the RAPID data archive quality, full documentation on all validated raw and processed data, as well as on models and model results, must be provided to the RDC. It is therefore essential that metadata are submitted at the same time as the data sets to which they pertain. The responsibility for producing the metadata will lie with project PIs and the RDC. A metadata protocol is outlined at Annex II.

In addition to the standard metadata, investigators are encouraged to archive at the RDC all relevant information electronically, including references, papers, reports, etc., unless agreed otherwise between the PIs and the RDC.

5. Data Formats and Data Media

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

In general, model data should be formatted in CF compliant NetCDF files, although there will be exceptions (particularly PP and HDF will also be accepted). Documentation on formats and conventions is available from the RDC (http://www.badc.rl.ac.uk/formats/), which also provides links to downloadable free software packages to support NetCDF access.

Submission of data will generally be via CD-ROM, as a Word/Excel e-mail attachment or by *ftp*. In some instances (e.g. some of the atmospheric model output) an automatic Web based file uploader will be available. At an early stage Investigators should discuss the options with the RDC.

CD-ROMs and or DVDs are currently the preferred means for making integrated data products from thematics available to the wider research community. However, there may be a preference towards a web-based final data product as RAPID progresses. The RAPID Steering Committee will review and decide on this at a later stage in the programme. It is not expected that the choice made will have cost implications.

6. Data Back-up Policy

The consequences of losing data, due to having made insufficient or inappropriate provision for their back-up, are potentially catastrophic in the case of large data collections, and cumulatively serious in the case of smaller data sets. Rigid daily back-up programmes operated at the RDC safeguard major digital databases. Provision and support of back-up strategies for digital data stored locally is the responsibility of individual PIs, or their delegates. Project PIs and Co-Is are responsible for providing appropriate back-up strategies for digital data stored locally and/or via other organisations.

As far as possible, analogue data (such as photographs) should be "disaster proofed" by transferring them into digital form, e.g. 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-ROM or DVD. Note that BODC has considerable experience in managing and publishing image data.

PIs should bear in mind that the timely deposit of data with the RDC will provide additional security for the project data.

7. Protection of Data Originator’s Intellectual Property Rights (IPR)

The Steering Group and the RDC recognise the need to ensure reasonable protection of project scientist IPR. The RAPID Data Policy (see Annex I) addresses this and is intended to provide an appropriate balance between the protection of data originators' IPR and the potential benefits that may arise via data use by the programme, the wider research community and other interested parties.

ANNEX I

Rapid Climate Change (RAPID) Data Management Policy

This document describes the data management policy for the RAPID programme as drawn by the Steering Committee. The primary aim of the RAPID data policy is

- To encourage rapid dissemination of scientific results.
- To protect the rights of the individual scientists.
- To have all the involved researchers treated equitably.
- To ensure the quality of the data in the RAPID data archive.

These aims conflict at times, and it is hoped that the provisions of the protocol resolve these conflicts fairly. It is recognised that this cannot always be achieved to everyone's complete satisfaction; there are bound to be cases where individual interests clash with those of the RAPID programme. Therefore to try to meet these aims, all PIs involved in RAPID, in accordance with and on behalf of their co-investigators, have agreed to abide by the following conditions as part of the acceptance of the grant award:

Data management

Data collected within the RAPID programme will comply with NERC's policy on data management (www.nerc.ac.uk/data/policy.shtml). The main objective of this policy is to ensure that the data will contribute to a key NERC resource, which will continue to be exploited both scientifically and commercially long after the formal end of the programme. The management of the data collected within the RAPID programme will be the responsibility of the relevant NERC Designated Data Centres (e.g. BADC, BODC), and funds have been made available from the RAPID budget to support this activity. In the absence of a NERC Designated Data Centre for palaeo data, special provision will be made for such data acquired within the programme (while ensuring appropriate links to international projects, such as HOLIVAR). To ensure proper data management the Science Coordinator will work together with a small data sub-group appointed by the Steering Committee (see appendix below).

Recommended RAPID data policy (in line with other Thematics)

The data subgroup proposed the following data policy for the RAPID programme, which has been ratified by the full Steering Committee and will apply to all projects funded through RAPID:

a) Data[1] should be lodged with the appropriate data centre on acquisition[2], together with such metadata as are defined under the RAPID data management plan.

[1] **Data:** includes palaeo data, present-day observations, model output, data syntheses, data-model syntheses, model codes and physical samples.

[2] **On acquisition:** the time-scale may vary between data types (for example, real-time data could go directly to a data centre) but the overall aim is to keep the time-scale as short as possible and certainly less than 6 months. This is to ensure that data acquired during RAPID are available to the RAPID community within the lifetime of the programme.

b) Data will be embargoed for 1 year from acquisition, allowing the PI and co-workers to exploit it in the first instance. The metadata will not be embargoed, to allow the wider community to be aware of work being carried out under RAPID and facilitate community building.
c) Data will be made available to the RAPID community after 1 year, and to everyone after 2 years.
d) Anyone making use of RAPID data within 3 years of it being lodged at the data centre will be required to include the PI and/or co-workers (as appropriate) as co-author/s on any resulting papers, if the PI and/or co-workers so desire.
e) Any corrections, improvements or amendments to data must be lodged with the appropriate data centre as soon as possible.
f) PIs making use of RAPID data are responsible for ensuring that the data used in publications are the best available at the time.
g) Data submitted to the data centre must be in the data format agreed between the data centre and PI. In addition, all agreed metadata must be supplied to the data centre.
h) While data are restricted from the public domain, no data will be transferred to parties outside the programme without the explicit agreement of the originator. In addition, guidance will need to be sought from the Science Coordinator and the Steering Committee if major data transfers are involved, to avoid compromising the interests of other programme participants.
i) In the event of dispute, the final decision rests with the RAPID Science Coordinator and the Steering Committee.
j) PIs and/or co-workers failing to comply with the RAPID data policy would be subject to appropriate sanctions.

Appendix: Data subgroup membership

K. Briffa, P. Challenor, S. Tett, M. Srokosz, C. Gommenginger and a RDC representative.

ANNEX II

RAPID Metadata Protocol[3]

1. Introduction

The term *metadata* encompasses all the information necessary to interpret, understand and use a given dataset. *Discovery metadata* more particularly apply to information (keywords) that can be used to identify and locate the data that meet the user's requirements (*via* a Web browser, a Web based catalogue, etc). *Detailed metadata* include the additional information necessary for a user to work with the data without reference back to the data provider. The metadata required by the RAPID Programme include both discovery and detailed metadata.

Metadata pertaining to observational data, for example, include details about **how** (with which instrument or technique), **when** and **where** the data have been collected, by **whom** (including affiliation and contact address or telephone number) and in the framework of which research project.

In the case of all submitted data, the RDC needs to know how the values were arrived at. The derivation process must be stated: all processing and calibration steps should be described and calibration values supplied. The nature and units of the recorded variables are essential, as well as the grid or the reference system. The RDC requests that as much information as possible about fieldwork instrumentation be included, e.g. serial number, copies of manufacturer's calibration sheets, and recent calibrations, if applicable.

Metadata pertaining to model output should include the name of the model, the conditions of the calculation, the nature of its output, the geographical domain over which the output is defined (when applicable). Specific conditions applying to the model or the experiment may be mentioned. Metadata also include information on the format in which the data are stored, and the order of the variables, to allow potential users to read them. Metadata pertaining to software models include the key points of the theory on which the model is based, the techniques and computational language used, and references.

The following lists the minimum metadata required to accompany data files submitted to the RAPID Data Centre (RDC). Since there is a large range of data types within RAPID, the RDC will liaise with project workers submitting data on a case-by-case basis to ensure that metadata formats are appropriate and to gain additional relevant information as necessary.

2. Metadata for RAPID Projects

2.1 Metadata for tables of numbers (observations or model output)

2.1.1 Content

Metadata include the following overall information. Some information in this list may be applicable in specific cases only.

[3] Adapted from URGENT Air Metadata document at http://badc.nerc.ac.uk/data/urgent/Metadata.html

- **Information about the experiment.**
  Date when fieldwork, experiment or model simulation started.
  Site or trajectory bounding box or domain limits.
  Platform (e.g. ship, cruise number), instrumentation (including instrument make, model and serial number).
  Model name.

- **Information about the experimenter(s).**
  Names, affiliation, contact address including e-mail, telephone number.
  Programme name, research project number.

- **Information about the independent variables (spatio-temporal grid).**
  Names, units, domain of definition of independent variables.
  Interval values when appropriate.

- **Information about the data, including processing level.**
  Version number.
  Date of last revision.
  Processing level (nature of raw data, derivation method: processing steps, calibrations applied).
  Nature, name, units, scaling factors of dependent variables.

- **Information about data storage.**
  Number of files of the entire dataset.
  File number of current file.

- **Information about data format.**
  Type of format e.g. ASCII, Excel, Matlab, NetCDF.

- **Additional information.**
  May include particular conditions of experiment or model run, model boundary conditions, article reference, and sources of further information.

2.1.2 Metadata storage

Ideally, each data file should include a header containing the metadata. If there is a large amount of information (e.g. description of many processing steps, calibration techniques), then a separate text file can be used as an alternative.

2.2 Metadata for software

2.2.1 Content

Metadata pertaining to a model should include the following.

- **Information on the model**
  Brief description of model general aim.
  Model structure.
  Physical processes involved, including equation set.
  Algorithmic implementation techniques used.
  Spatio-temporal coverage when applying.
  Boundary conditions, including reference(s).
  Initial conditions, including reference(s).
  Program language.

Input nature and format.
Output nature and format.
Summary of model validation, or appropriate reference(s).
Summary of results from former studies conducted with the model, or appropriate reference(s).

- **Information on the author(s)**
  Names, affiliation, contact address including e-mail, telephone number.
  Programme name, research project number.

2.2.2 Metadata storage

Metadata relative to software can be included as comments in the top section of the source file or can alternatively be provided as a separate text file.

2.2.3 Format

Text. There is no particular requirement regarding software metadata formatting.

3. Additional documentation

Any additional documentation on recorded data or images, whether pertaining to a single data file or a whole dataset, that would not find its place into the structures described above (because it does not fall into any described category or because it is too voluminous) may be submitted to the RDC in the form of a text file that will be stored in the RAPID archive documentation directory. These documents may for example include technique description, possible use of the data, study conclusions, etc.