Quantifying and Understanding the Earth System (QUEST)
Data Management Plan

Version 0.1 – February-March 2007

Contents

1. Introduction

Throughout its duration (2003 to 2010), the QUEST Programme will fund a number of projects to address 3 main subject areas, or Major Themes (MT). Some cross-cutting projects will also be funded to focus on strategic goals through one of the QUEST three Focused Strategic Activities (SA) listed below. In addition, QUEST will invest resources in a fellowship, as shown below.

MT-1. The carbon cycle and the associated biotic feedbacks
MT-2. Climate regulation over glacial-interglacial timescales - Earth system dynamics
MT-3. Global change and sustainable resources - Biosphere management

SA-1. Earth system modelling
SA-2. The QUEST Earth System Atlas (QESA), a UK-USA collaborative initiative to compile existing data on the Earth System and publish them in the form of maps.
SA-3. Earth system processes and prediction
- SA-3.1 Assessing and facilitating QUEST
- SA-3.2 Biogeochemical cycles and feedbacks
- SA-3.3 Climate-carbon modelling, assimilation and prediction
- SA-3.4 Dynamics of the Palaeocene-Eocene thermal maximum
- SA-3.5 Environmental change and fisheries
- SA-3.6 Fire data assimilation and prediction

QF. QUEST fellowship to carry out accurate atmospheric carbon dioxide and molecular oxygen measurements in the UK in view of supporting studies related to the land and ocean carbon cycle.

This document sets the way to implement principles developed in the QUEST Data Policy (QDP), to which it refers. It is based on information collected at the occasion of visits to the project Principal Investigators (PIs) made during the phase of the scoping study and from the answers to a questionnaire sent to the researchers involved in data production or handling. It is intended to be a work document and some of its sections will be updated in the course of the programme, in particular to take into account new projects as they are awarded.

At the time of writing the DMP V0, a first scoping study(*) has included the currently awarded six projects, under MT-1 and -2, as shown in **Table 1**. In the meantime, two additional activities have been awarded grants, as shown in **Table 2**. Liaison with the investigators of these two activities are underway.

Table 1. Details of projects funded under Major Themes 1 and 2, and dates of initial visit by the BADC.

Theme	Project acronym	Project title	Project duration	PI	PI's university	Date of visit
MT-1	MarQUEST	Marine Biogeochemistry and Ecosystem Initiative in QUEST	May 05 – Sep 09	Andy Watson	UEA	8/8/06
	QUAAC	Modelling of atmospheric oxidants and aerosols: deposition and chemical transformation	Apr 05 – Apr 09	John Pyle	Cambridge	5/5/06
	QUERCC	Quantifying Ecosystem Roles in the Carbon Cycle	Apr 05 – Apr 09	Ian Woodward	Sheffield	/
MT-2	PalaeoQUMP	Using palaeodata to reduce uncertainties in climate prediction	Mar 06 – Aug 09	Sandy Harrison	Bristol	22/5/06
	QUEST Deglaciation	Climate and biogeochemical cycles during the last deglaciation	Apr 06 – Jun 09	Paul Valdes	Bristol	22/5/06
	Quaternary QUEST	Regulation of atmospheric carbon dioxide and climate on glacial-interglacial timescales	Apr 06 – Aug 09	Tim Lenton	UEA	8/8/06

Table 2. Details of newly awarded QUEST activities (February 2007).

Activity	Project acronym	Project title	Project duration	PIs	PIs' institutes
SA-1	DESIRE	Dynamics of the Earth System and the Ice-core Record	Feb 07 - ?	Eric Wolff Pierre Friedlingstein	BAS IPSL
QF	/	High precision measurements of atmospheric CO2 and O2 in the UK	Nov 05 - ?	Andrew Manning	UEA

2. Types of data generated by QUEST

Each QUEST project will be involved in some of the data activities shown in **Table 3**.

Table 3. Data activities of the QUEST projects.

Project		New measurements (1)	Gathering of existing data (2)	Data synthesis (3)	Model output (4)
MT-1	MarQUEST	/	✔	✔	✔
	QUAAC	/	/	/	✔
	QUERCC	✔	?	/	/
MT-2	PalaeoQUMP	/	✔	✔	/
	QUEST Deglaciation	/	/	/	✔
	Quaternary QUEST	/	✔	✔	✔

(*) Kevin Marsh, Quantifying and Understanding the Earth System (QUEST): Themes 1 and 2 Data Scoping Study, October 2006, revised in December 2006 and February 2007 following QUEST Data Management Group meeting and feedbacks.

Data collected under Activities (1) and (2) (see **Table 3**) can present different levels of processing. Raw data, i.e. source measurements in the form that they have when they are first produced, will in general not be archived at the DC (see Section 3) but it is each PI's responsibility to ensure that they are stored safely with the relevant processing software or, alternatively, with documentation on retrieval algorithms. However, raw data must be documented at the BADC.

Processed data, i.e. observation data that have been subject to some treatment or formatting, or data derived from these, will be archived at the designated DC and made available to the QUEST community. However, this does not apply to confidential data collected under Activity (2) — that is, data that may not become public at any time due to their private nature or because they are already subject to some protocol restricting their publication.

Activity (3) (see **Table 3**) consists in the compilation and harmonisation of existing observation data from a variety of origins. This usually involves some kind of modelling and the resulting datasets are more likely to be superseded by new versions than datasets from one single source are. For these two reasons, archival and curation of data produced under Activities (3) and (4) present similarities, and these data will be subject to the *BADC Policy and Guidelines for Archiving Simulation Data* (SDP) (see Annex 1).

While being developed, datasets generated by Activity (3) may be made available to the QUEST community by the PI's team if the appropriate infrastructure allowing this is in place, but they must eventually be archived at the designated Data Centre (DC), in the required format and together with the required metadata.

Selected model output (Activity (4)), together with the required metadata, will be archived at a NERC designated Data Centre (DC). Selection of model output worth to be archived will be done jointly by the project PIs and the BADC, based on the selection criteria listed in the SDP.

3. Data centres

The British Atmospheric Data Centre (BADC) is by default the NERC designated DC for QUEST. Among the currently awarded projects, none is expected to deliver datasets that would be better archived at another NERC DC, but the situation may arise for future QUEST projects. Although MarQUEST deals with Ocean Science, the data produced will be theoretical, and the British Oceanic Data Centre (BODC) feels that the BADC has more expertise in handling such data (especially when issued by coupled ocean-atmosphere models). The fate of palæoclimatological data produced with NERC funding is currently under examination by NERC. In the meantime, a number of locations may be appropriate depending on the nature of the data. While modelled palæo data on the past Earth atmosphere find a natural place at the BADC, BAS has the expertise in analysing ice-core data; if the ability to care for the long-term preservation and distribution of the data exists at BAS or in places other than the BADC, these functions can be fulfilled locally but all the relevant metadata (including how to access the data) must be provided to the BADC, which will set up a central inventory of the QUEST datasets, for future integration into the NERC Data Grid (NDG) (see Section 4).

A website has been set up for QUEST at the BADC, which will be the gateway to all QUEST data: http://badc.nerc.ac.uk/data/quest/ (see Annex 3). This page will include links to all relevant documentation and external sites. It will be updated as the programme unfolds.

For the duration of the Programme, the BADC will
- liaise with the QUEST researchers to get updates on their data deliverables and needs, as these will become better known, and develop the present data management plan accordingly;
- provide support to QUEST investigators in issues related to format, metadata and submission;
- answer data related queries;
- maintain and update the QUEST archive, data portal and uploader;
- integrate the QUEST data into the NERC Data Grid (NDG);
- monitor access applications;
- release the data to the public in due course;
- advise QUEST participants of new developments of the BADC settings related to the Programme.

4. Dataset catalogue and metadata search

An entry will be created for QUEST in the BADC Data Catalogue. As data populate the archive, this will be completed by input (about the instruments, the models, etc.) into the underlying metadata scheme (MOLES), allowing metadata search. The MOLES records will be integrated into the NDG, which will make possible the search of metadata pertaining to datasets throughout the network of all NERC DCs.

At a later stage, part of the MOLES records will be completed by information retrieved automatically from the files, which assumes (and underlines the importance of) properly formatted and CF-compliant data — see Sections 8 and 9).

5. Data archive contents and structure

Table 4 shows the information collected so far from the projects about their data deliverables. Many investigators were not fixed as to which model runs would be performed or which datasets would be collected at the time when they were interviewed. The resulting information hence presents wide gaps and will have to be completed through further interaction with the researchers. The selection of model output to be archived will have to be done on a case-per-case basis in collaboration with the investigators as the expected delivery date is nearing by, based on the principles stated in the SDP (see Annex 1).

Table 4. Data deliverables from the QUEST Themes 1 and 2 projects.

Project	Dataset	Expected volume	Expected date of delivery	Data provider
MarQUEST	Synthesized data + Climate model data	10 GB	October 2008	Erik Buitenhuis, UEA
	Climate model data	250 GB	April 2008	Bablu Sinha, NOC
	Ocean model data	500 GB	2009	Andy Watson, UEA
QUACC	Climate model data	500 GB	2009	John Pyle, Cambridge
QUERCC	Soil nutrient data	~1 GB	2008-09	Matt Aitkenhead, Aberdeen
PalaeoQUMP	Sensitivtiy tables Seasonal fields Synthesis maps	< 100 GB	2009	Sandy Harrison, Bristol
QUEST Deglaciation	Climate model data + Model metadata for which the output is not archived at BADC	< 500 GB	2009	Paul Valdes, Bristol
Quaternary QUEST	Climate model data	10 TB	May 2009	Robin Smith, NOC
	Climate model data	Small	2009	Agatha De Boer, UEA
	Synthesized data	Small	2009	Babette Hoogakker, NOC
	Climate model data	Small	June 2009	Tony Payne, Bristol
	Synthesized data + Climate model data	Several TBs	April 2009	Tim Lenton, UEA
	Synthesized data + Climate model data	100 MB	October 2007	Neil Edwards, Open University
Estimated total volume		**~ 15 TB**		

In addition to the observation and model data, the QUEST archive will hold all documentation that would be too extensive to be archived as metadata within the data files themselves (in particular, source codes of

models used to generate some model output). The QUEST archive at BADC will be divided into folders as described in **Table 5**. Subfolders will be created as needed.

Table 5. Outlined structure of the QUEST archive at BADC.

/badc/quest/	doc/	
	software/	
	data/	marquest/
		quacc/
		quercc
		palaeoqump/
		qdeglaciation/
		quaternaryq/
		desire/
		highres_co2_&_o2/

6. Data versions, curation and backup

The data held at the BADC will be preserved for the long-term unless the dataset review (see SDP in Annex 1) concludes that the data have been superseded by new or better versions. If new versions are submitted, the option of keeping the old ones will be envisaged. In this case, it will be made very clear which one is the most recent version.

Data will be backed up at regular intervals and duplicates will be saved on tape.

7. Data handling

In addition to allowing metadata search through the NERC DC data holdings and giving access to the data files, the NDG will include a data manipulation component. This tool will allow a variety of operations to be performed on data formatted in NetCDF and, to the extent that resources allow, in NASA Ames. These operations include

- extracting subsets of data from the files,
- comparing data from various origins,
- performing basic operations such as differences, averages, etc.,
- plotting and visualising data.

This new tool will replace the current BADC Data Extractor.

8. Formats

QUEST data will be formatted in NetCDF, which is widely used and particularly adapted to the storage of large geophysical model output. A large range of existing software to produce, read and handle NetCDF files is available. If preferred, NASA Ames may also be used for the storage of less voluminous sets of measurements. As underlined in Sections 4 and 7, these two formats will underlie the BADC catalogue search engine and the NERC Data Grid (NDG), so that QUEST datasets will be integrated swiftly into these two instruments, provided that files do include CF-compliant metadata in the formatted fields (see Section 9).

Documentation, tools, templates, examples and help on these formats are provided at the following web pages.

- NetCDF format (binary) — http://badc.nerc.ac.uk/help/formats/netcdf/
- NASA Ames format (ASCII) — http://badc.nerc.ac.uk/help/formats/NASA-Ames/

The BADC will endeavour to provide support to data suppliers using NetCDF or NASA Ames.

Raw data (if any), software and images will be archived in their original formats. Text documents should be archived in PDF.

9. Metadata

Metadata ("data about the data") contain the necessary information to find ("discovery" metadata), read, understand, interpret and use the data.

The Climate and Forecast (CF) Metadata Convention — developed for NetCDF but applicable to any geophysical dataset — will be used to select and format the metadata related to data recorded in the above types of formatted files. For example, as far as possible, CF standard names should be used to name recorded variables, even in NASA Ames files. The set of CF standard names is an evolving nomenclature, open to new soundly founded proposals. Proposals can be submitted to the CF Convention through the CF mailing list — and soon via the web. If requested, the BADC will provide advice to researchers in forming new name candidates.

The metadata should be integrated into NetCDF files through the use of local and global attributes. The NASA Ames (ASCII) data files include a header which contains, in a predefined display, basic information on the data recorded in the file, so that these metadata are inseparable from the data to which they pertain. Both file types provide room for comments, which may include any information that cannot be provided in the formatted fields, although this information, very valuable for the user, will be less useful in terms of automated search: in the NDG perspective, it should be stressed that the formatted fields should be used to their optimal capacity.

The CF Metadata Convention is available from http://www.cfconventions.org/ and the BADC provides additional guidelines on CF at http://badc.nerc.ac.uk/help/formats/netcdf/index_cf.html

Metadata should be as specific, explicit, accurate and complete as possible. They should be formulated in a transparent way, avoiding unexplained assumptions and implicit references to unavailable or undocumented conventions.

The checklist provided by BADC at http://badc.nerc.ac.uk/help/metadata/#Elem includes the following main metadata elements.

- Information on the (physical or theoretical) experiment.
  Date when experiment or model simulation started. Site or trajectory bounding box or domain limits. Platform, instrumentation. Model name.
- Information on the data originator(s).
  Names, affiliation, contact address including e-mail, telephone number. Research programme name, research project code.
- Information on the independent variables (in geophysical datasets, usually a spatio-temporal grid).
  Names, units, domain of definition of independent variables. Interval values when appropriate.
- Information on the data.
  Version number. Date of last revision. Processing level (nature of raw data, derivation method). Nature, name, units, scaling factors, accuracy of dependent variables.
- Information on the format.
  Type of format + reference of format documentation. File structure. Number of lines in file header if any. Record structure.
- Any additional relevant information, such as instrument description and specifications, essential model features and configuration, conditions in which the data were collected, algorithms used to derive the recorded data from the raw data, reference publications, etc.

CF standards exist for part of the above and should be applied. NASA Ames provides formatting rules for some of the elements above. Any information that cannot be integrated into the data files as local attributes (NetCDF) or in the file header top section (NASA Ames) should either be inserted in the files as global attributes (NetCDF) or comments (NASA Ames), or (if substantially large) be provided in separate supporting documentation. This additional documentation may include collection methods, algorithms, model parameterisations, references, advice to the users, plots, pictures, etc. and will be archived alongside the data. Texts should be submitted as PDF files. Source codes can be archived if they help the

understanding of the archived model output (see Annex 1); in this case, it is preferable to archive also a set of standard model input.

10. Data file names

Data file names will follow the BADC file name convention documented at
http://badc.nerc.ac.uk/help/file_naming.html

File names are composed of three (optionally four) fields separated by underscore signs, and an extension separated from the rest by a dot. Each field may only include lower case letters, digits and the hyphen sign. The file name template is

instr_loc_YYYYMMDD[hh[mm[ss]]][_extra].ext

where fields inside square brackets are optional and where

instr is a standard name for one of the following
- an instrument — sometimes denoted by the quantity it measures (e.g. "uea-peroxides");
- a set of instruments (e.g. "core-cloud-phy" for the set of FAAM core instruments measuring cloud physics data);
- a production method or its result (e.g. "syn-o3" for ozone data synthesis);
- a model (e.g. "mcm-short" for the Master Chemical Mechanism of short lived species).

When the instruments or model are operated by one group, it should be composed of two parts separated by a hyphen, with the first part being the name of the university or institution owning the instrument or model (e.g. "uea-doas" for the Differential Optical Absorption Spectrometer operated at the University of East Anglia).

loc is a standard name for one of the following
- a location or a region (where the data was collected, where the modelled phenomena take place, where the computed trajectory starts or ends, etc.); if the model does not reproduce the conditions at a given particular place, it can be the type of landscape which is simulated; if the data cover the globe, *loc* can be set to "globe"; if none of this applies, it can be the location where the model was run;
- an itinerant platform (such as a van, a balloon, a ship, an aircraft, a satellite, etc.).

YYYYMMDD is one of the following
- the date when the first data record was collected (for observation);
- if the data are made of monthly or yearly averages, *DD* [resp. *MM*] can be replaced by the first or last day in the month [resp. in the year];
- the start date of the recorded scenario (for a realistic time-dependent simulation);
- the date when the first recorded model result was computed;
- if none of the above applies, the date when the dataset was produced.

hh, mm and *ss* (hours, minutes, seconds) can be added if necessary, so that this field can have one of the forms *YYYYMMDD*, *YYYYMMDDhh, YYYYMMDDhhmm* or *YYYYMMDDhhmmss*.

extra is an optional field with a free content.

ext is the format extension; for example,
- *ext* = na for NASA Ames,
- *ext* = nc for NetCDF.

Accepted values for the first two fields and for the extension are listed at
http://badc.nerc.ac.uk/cgi-bin/filespec_doc?id=INSTRUMENTAL&hfile=1

Data providers should submit new values of *instr* and *loc* to the BADC before uploading their data, in order to ensure that their files are not rejected by the web uploader (see Section 11).

11. Data submission and ingestion

Instructions on data submission will be made available from the BADC at
http://badc.nerc.ac.uk/data/quest/

Before submitting their data to the BADC, investigators should

1. Form file names as described in Section 10; if adequate field names cannot be found in the BADC lists of standard instrument, location and extension names, contact the BADC with suggestions or requests for new names.
2. Choose one of the accepted formats described in Section 8 and format their data according to this format rules.
3. Ensure that all required metadata (see Section 9) are included in the data files and are formulated in a clear and unambiguous way.
4. Check sample files against the online tools provided by the BADC at
   - http://titania.badc.rl.ac.uk/cgi-bin/cf-checker.pl for NetCDF files;
   - http://badc.nerc.ac.uk/cgi-bin/dataex_file.cgi.pl for NASA Ames files.

Data are uploaded to the BADC incoming area. To access the incoming area, data suppliers have to register with the BADC and apply for access to QUEST restricted data as explained on the QUEST welcome webpage shown in Annex 3.

Data should be uploaded through the QUEST web uploader which will be set up for that purpose. In three steps (selection of a folder, selection of a format, selection of a file on the home computer), this online tool will enable participants to upload any data file or supporting documentation. Directions will be provided online and will be added to the present document.

Alternatively, if the number of files to be submitted is very large, data and metadata can be uploaded via FTP by connecting to the BADC FTP server at *ftp.badc.rl.ac.uk*

Appropriate folders will be created at the top level of the QUEST incoming area. Subfolders will be created as needed.

Checkings will be performed on files submitted to the BADC via the web file uploader, to spot errors of file name or format. Files with wrong names or format errors will be rejected. Such checkings cannot be made on files submitted by FTP.

An e-mail will be sent automatically to a member of the BADC staff every time new files appear in the incoming area.

The files will be transferred from the BADC incoming area to the QUEST archive. When the structure of the archive will be better known (see Section 5) and where relevant, this ingestion may be automated based on the file names and the incoming folders.

12. Data submission schedule

The six projects currently investigated do not present any significant synergy, so that there is currently no requirement imposed on submission deadlines by the projects themselves. This may change as cross-cutting projects are awarded.

However, the QUEST Data Policy requires that

> *Data must be lodged with the BADC as soon as they have been validated and no later than 3 months after acquisition. […] Responsibility for this rests with the PI of the individual QUEST funded project, the QUEST fellowship holder, or the individual core team member. Exceptions to the 3-month deadline can be made in the case of model output data where*

after 3 months it cannot yet be determined whether the data will be suitable for long-term, post project curation..

The current submission schedule will conform to the information recorded in Column 4 of Table 4 (see Section 5). This will be completed by more detailed information collected by BADC as the programme unfolds.

13. Data access and use

In accordance with the QUEST Data Policy, access to data will be restricted to QUEST participants during one year after creation, apart from the data produced for QESA, which will be immediately available to the public. Restricted data will be released to the public domain after their retention period.

However, conditions of use will apply to QUEST data at any time:
- During the first year following their production, the data may only be used by the data originators, unless prior consent has been given by them to the candidate user(s).
- During 2 years after a project end, co-authorship on any paper based on the data generated by the project must be offered to the data originator(s).
- At any time, QUEST data distributed via the BADC may not be used for commercial purposes. Requests for commercial use must be addressed to the owner of the IPRs, whether NERC or the data originator(s) (see QUEST Data Policy, Section 8).

Access to QUEST data archived at BADC (whether public or restricted) will require prior agreement with the above Conditions of Use, as shown in Annex 2.

Access to public data will be granted automatically to applicants who will have abided by the QUEST Conditions of Use. Applications for access to restricted QUEST data will be forwarded to the QUEST Project Manager, who will decide whether access can be granted. Anyone who has been granted access to restricted data will have automatically access to public data as well.

14. Third-party datasets

Participants have been consulted on desirable third-party datasets at the time of the scoping study. Investigations have been made about the availability of these datasets and links to the relevant websites are given on the NEU welcome page (see Annex 3). These currently include Argo; MIPAS, SCIAMACHY, MERIS (Envisat); EPICA Dome ice core data; MOZAIC; MTCI.

Contact has been made with the managers of the MOZAIC data about the possibility of mirroring the data at the BADC (discussions are ongoing). Action has been taken to ask the company *Infoterra* to fill the gaps of the MTCI data set (work currently underway).

Every effort will be made to ease the access of QUEST participants to any additional third-party dataset required.

15. Acronyms

BADC	British Atmospheric Data Centre
BAS	British Antarctic Survey
BODC	British Oceanic Data Centre
CF	Climate and Forecast (Metadata Convention)
DC	Data Centre
DMP	Data Management Plan
IPSL	Institut Pierre-Simon Laplace
MOLES	Metadata Objects for Links in Environmental Science
MT	QUEST Major Theme
NDG	NERC Data Grid
PI	Principal Investigator

QDP	QUEST Data Policy
QESA	QUEST Earth System Atlas
QF	QUEST Fellowship
QUEST	Quantifying and Understanding the Earth System
SA	QUEST Focused Strategic Activity
SDP	BADC Simulated Data Policy
UEA	University of East Anglia

Annex 1.

Archiving of Simulations within the NERC Data Management Framework: BADC Policy and Guidelines

Introduction

1. Issues associated with archiving information about the environment made by measurement are relatively well understood. This document outlines a general policy for archiving simulated and/or statistically predicted data[1] within NERC and provides specific policy and guidelines for the activities of the British Atmospheric Data Centre.

2. In the remainder of this document we use the term simulation to cover deterministic predictions (or hindcasts) based on algorithmic models as well as statistical analyses or composites of either or both of simulations and real data.

3. This policy has been developed in response to external legislative drivers (e.g. Freedom of Information Act and Environmental Information Regulations), external policy drivers (e.g. the RCUK promulgation on open access to the products of publicly funded research), as well as the existing NERC data management policy which is based around ensuring that NERC funded research is exploited in the most efficient manner possible.

4. The major question to be answered when considering simulated data is whether the data products are objects that should be preserved in the same way as measured products. In general the answer to this question is non-trivial, and it will be seen that guidelines are required to implement a practicable policy.

Data Management and Simulated Data

5. In general the information provided by models and the information provided by measurements are of a different nature. Simulations are analogues of the "real" world that may provide insights on physical causal relationships, while measured data are the observed symptoms of these relationships.

6. Simulations are generated by either deterministic or statistical models (or a combination of both). Such modelling activity does not generate definitive knowledge. Models are continuously developed and hopefully (but not necessarily) provide improved or more adequate representations of the physical system as time progresses. This is to be contrasted with measurements of the earth system, which by definition, cannot be repeated with the system in the same state and are therefore unique in a rather different way to simulated data.

7. Simulated data is usually produced by individuals, teams, or projects, and may have limited applicability, and/or potential for exploitation, in the wider community. However, the role for data management is not limited to making data more widely available, there is also a recognised role for data management to minimise duplication of activities between individuals, teams and projects, and to facilitate research programmes and collaboration. It is therefore important to develop criteria by which the scope for programme facilitation or wider applicability or exploitability can be recognised.

Criteria for Selecting Simulated Data for Management

8. If the answer to one or more of the following questions is yes, then simulated data are candidates for professional data management beyond that provided by the investigating team responsible for producing the data.

[1] The word "data" is often claimed by experimental scientists to exclude simulated information, however, most reputable dictionaries include simulated products within the definition.

a) Is there — or is there likely to be in the future — a community of potential users who might use the data without having one of the original team involved as co-investigators (or authors)?

b) Does some particular simulation have some historical, legal or scientific importance that is likely to persist? (Some simulations may become landmarks, in some way, along the route of scientific knowledge. They may also have been quoted to make a statement that might be challenged – either scientifically or legally – and should therefore be kept for evidential reasons.)

c) Is the management of the data by a project team likely to be onerous or result in duplication of effort with other NERC funded activities?

d) Is it likely that the simulation will be included in future inter-comparisons?

e) Does the simulation integrate observational data in a manner that adds value to the observations?

9. If the answer to any of the following questions is yes, then the simulated data should not be archived.

a) Is the data produced by a trivial algorithm that could be easily regenerated from a published algorithm description?

b) Is the data unlikely to ever be used in a peer-reviewed publication, or as evidence to support any public assertions about the environment?

c) Is the data known to be of poor quality or to have had no scientific validity?

d) Is it impossible to adequately document the methodology used to produce the data?

10. If the answer to any of the following questions is yes, then value judgements will need to be made about how much of the simulated data should be archived. Guidelines to assist in this situation appear below.

a) Would storage of the data be prohibitively expensive?

b) Would storage of statistical summaries rather than individual data items provide adequate evidential information about the simulation? (e.g. while it might normally be desirable to store all ensemble members, would ensemble and/or temporal means be adequate in a situation where storage of the individual members at full time resolution might be prohibitively expensive).

Guidelines for Archiving Simulated Data

11. In some cases, datasets may be archived by the investigating team at a national facility, rather than at a NERC designated data centre.

a) This is most likely to occur when the longevity of the dataset is in some doubt, and the added value of using a designated data centre is not clear.

b) Where datasets will initially have restricted access (see para 16) it should normally be the case that the data archive is held at a designated data centre where procedures are already in place for providing secure access to data.

c) Alternative archives should not be established where the result will be that academic staff will be spending significant amounts of time carrying out professional data management which should be carried out within institutions with more appropriate career structures.

12. Where the intention is that a dataset be held outside of a NERC designated data centre, procedures should be in place to ensure that the data holder (or holders) conform to all the following requirements. It should also be ensured that funding is in place to move the data within a designated data centre when the holder (or holding facility) is no longer able to archive and distribute the data. Such datasets will still be

the responsibility of a designated data centre, but those responsible for the remote archives will be responsible for keeping all metadata required by the designated data centre up to date, and communicating the results of internal reviews (especially those which might involve removing or superseding data holdings).

13. All simulated datasets will be subject to regular lifetime review (described below).

14. Given that a simulation dataset is to be archived, what is involved in archiving such a dataset?

    a) The simulated data itself should be archived in a format that is supported by the designated data centre community (whether or not the data is to be initially archived in a designated data centre. It is recognised that in taking on data, potentially in perpetuity, every new format is a significant ongoing cost.)

    b) Any non-self-describing parameter codes (e.g. stash codes) included within the data should be fully documented.

    c) Discovery metadata conforming to appropriate standards and conventions[2] should be supplied for all datasets to the responsible designated data centre.

    d) Where possible, documented computer codes and parameter selections should also be provided (e.g. the actual Fortran, and full descriptions of any parameter settings chosen[3]).

    e) Where initial conditions and boundary conditions are themselves ancillary datasets, these too should be archived and documented.

    f) Estimates of the difficulty (both practically and financially) of recreating the simulation. (This will be needed to inform the lifetime review).

    g) Where special tools (e.g. diagnostic software codes) are available to help interpret the simulation, these tools themselves should be archived if possible.

    h) All documents and information ("further metadata") should conform to appropriate archival standards (published open formats, suitable metadata structures etc)..

15. Where only a subset of the simulation is to be archived, the following considerations should be assessed in making decisions:

    a) Potential usage (e.g. if the climate impacts community are involved appropriate parameters might include daily min/max temperatures, whereas instantaneous values are more likely to be useful if the simulation is to be used to generate initial conditions for other runs).

    b) Illustrative value (where a simulation is being archived because of it's scientific importance, those parameters relative to the scientific thesis should be the most important).

    c) Physical Relevance (e.g. case studies, one might only store those parameters necessary to make the relevant points, but there are obvious risks in retrospectively identifying key parameters).

    d) Volume and cost of storage.

    e) Standard Parameters used in model-intercomparison exercises. Where possible and appropriate datasets should always seek to keep these, and the designated data centre community will provide guidance on current standard lists of parameters.

2 In October 2005 this would be NASA GCMD DIF documents with the Numerical Simulation Extensions.

3 It is hoped that in the near future, the Earley Suite being developed at the University of Reading will provide an appropriate formalism for Unified Model Simulations.

f) Can the temporal or spatial resolution be decremented without losing impact

16. When simulated data is initially archived, it may be possible for access to be embargoed in some way for a defined period[4]. When this occurs the following issues need to be addressed:

    a) To which community should it be restricted and for how long?

    b) Should conditions of use apply to the data during and/or after the retention period (e.g. communication with investigators, offer of co-authorship, acknowledgement in publications)?

17. Where it is known a priori that simulation data will be archived, they should normally be archived at the time they are produced. Where multiple versions are expected within a project, and no other groups are expecting access to the data before a final version is produced, early simulations need not be archived. It should never be assumed that any part of a dataset would be archived after the end of the originating project.

Archive Lifetime

18. As described in the introduction, continuous model improvement/development may make obsolete datasets made with previous versions. All simulated datasets should be subject to more frequent review procedures than measured datasets.

19. Where a dataset is being held for legal reasons, or because of historical interest, such a dataset might be kept indefinitely.

20. Where a dataset has been formally cited and formally published, it should be kept indefinitely, unless it is not possible to migrate the format to future media.

21. A suitable timescale for review of simulation datasets held at designated data centres would be at four-year intervals. Four years should give time for work to be published and follow-up work to be performed, and for an initial assessment of the likely longevity of datasets to be established. Most international programmes (e.g. IPCC) should have exploited datasets on a timescale of eight years, and again, further longevity could then be assessed. More frequent reviews may be appropriate where datasets are held elsewhere.

22. Reviews should involve at the minimum: the data supplier (if available), the custodians (especially if not held inside a designated data centre), representatives of the user community (if it exists), and an external referee.

23. Reviews may recommend removing subsets of a dataset.

24. Reviews may recommend acquiring new datasets to supersede existing datasets (and to keep multiple versions).

25. Reviews should consider the availability of tools to manipulate datasets.

26. In all cases metadata should be kept for datasets which have been removed.

Custodial Responsibilities

27. The custodial responsibilities of designated data centres are described elsewhere. These points are here to provide guidance for the minimum responsibilities of facilities formally archiving simulation data on behalf of one or more designated data centres.

[4] The Freedom of Information Act (2000) and the Environmental Information Regulations (2004) stipulate that an embargo, if any, can only apply for some limited amount of time, to allow for "work in progress".

28. All archived data will be duplicated, either in a formal backup archive, or by complete archive duplication at multiple sites (in which case the remote sites must support all the same metadata structures, and they must advise the designated data centre should they consider removing their copy).

29. All cataloguing and metadata required by the designated data centre must be provided and kept up to date.

30. User support must be provided to include help with any access control, on how to view and interpret the metadata, and on how to obtain and use the data in the archive.

31. Formal dataset reviews must be carried out.

32. Adequate bandwidth to the data holdings must exist.

33. Appropriate tools to use and manipulate the data must be provided.

Annex 2. Conditions of Use of QUEST Data

(http://badc.nerc.ac.uk/cgi-bin/dataset_registration/register.cgi.pl)

BADC Dataset/Service registration

Please read the following conditions carefully and click on the accept button only if you agree to them.

Quantifying and Understanding the Earth System

Conditions of Use of QUEST Data

1. QUEST data distributed by the BADC may not be used for commercial purposes.
2. During a period of one year after the data have been created, any use of the data by researchers other than the data originator(s) requires prior consent of the originator(s).
3. During a period of two years after a project end, any user of the data generated by this project is required to offer the data originator(s) co-authorship on any resulting paper.

I agree to abide by the terms and conditions for usage of QUEST data as stated above.

I accept>>

Annex 3. QUEST Welcome Page at BADC (http://badc.nerc.ac.uk/data/quest/)

Quantifying and Understanding the Earth System

Quantifying and Understanding the Earth System (QUEST)

Introduction

QUEST is a 6-year *Directed Mode* research programme funded by the British Natural Environment Research Council (NERC). The programme, which started in 2003 and will end in 2009, aims at providing a co-ordinated approach to the understanding of the Earth System and the complex feedbacks it involves. The programme is structured into several research themes. Each of the projects funded by QUEST focuses on one of these themes.

Access to data

Please note that in order to be able to apply for access to QUEST data, you should first register with the BADC.

QUEST data	Conditions of access [1]	How to apply for access [2]	Conditions of use	Where to find the data
• More than 1 year old • Earth System Atlas (QESA)	Public access	Submit application [3]	QUEST Conditions of Use[5] apply at any time	QUEST archive
• Other than QESA & less than 1 year old	Restricted to QUEST participants	Submit application [4]		

[1] As defined in Section 8 of the QUEST Data Policy.
[2] Application involves abiding by the QUEST Conditions of Use shown in Column 4.
[3] Access to public data will be granted automatically.
[4] Access to restriced data will be granted to QUEST applicants after approval by the QUEST Programme Co-ordinator. This will also include access to public data, so that only one application is necessary.
[5] The QUEST Conditions of Use include a clause stating that the data distributed by the BADC must not be used for any commercial purpose. Requests for commercial use must be directly addressed to the owner of the IPR, that is NERC or the data originators.

Third-party datasets
Links to third-party datasets of interest to QUEST can be found below under the Links section.

Instructions to data providers

Data file names
Data providers are kindly requested to follow the BADC File Name Convention. Note that the first

component of the file name is the instrument (or group of instruments, or model, or technique) name. Please refer to the current list of standard instrument/model/technique names and to the current list of standard location/platform names, and advise the BADC if your instrument, model, technique or location are not in the lists, so that we can ensure that the lists of standard names are up to date before you submit your data. A file name checking based on these lists will be performed at the time of uploading data to the BADC; files not complying with currently accepted standard names will be rejected.

Data file format
Data must be formatted in NetCDF or NASA Ames. A format checking will be performed at the time of uploading. If you wish to check that files are correctly formatted before uploading them please use the BADC NetCDF format checker or the NASA-Ames format checker.

Metadata
When inserting metadata in your data files (in NetCDF attributes or in NASA Ames file headers), please be as specific, explicit, accurate and complete as possible, and avoid references to implicit or undocumented conventions. Metadata inserted in NetCDF files must follow the CF Metadata Convention. When applicable, please follow CF Metadata recommendations even when formatting your data in a format different from NetCDF. For example, it is recommended, if possible, to use CF standard names for variables recorded in NASA Ames files. Additional guidelines on writing CF-compliant metadata are provided by the BADC.
Please also refer to the format standard for the time variable.

Data file submission
In order to upload data files to the BADC, you first need to be granted access to the restricted QUEST archive. To register, please see the "Access to data" section above. The QUEST web uploader *[Link not active yet]* will allow you to submit data files to the BADC. Large amounts of files can also be submitted by ftp (please contact the BADC help desk for further information).

Documentation

- QUEST Data Policy.
- Last version of the QUEST Data Management Plan. This working document will be updated as the QUEST Programme unfolds.
- BADC Metadata help page.
- NetCDF format.
- NASA Ames format.

Links

QUEST websites

- QUEST website at the University of Bristol.
- QUEST website at NERC.

Supporting datasets

- Argo data are available from the following two Argo Global Data Assembly Centres (GDACs).
    - The Coriolis Centre in Brest (France).
    - The Monterey GDAC (California, USA).

Argo is a global array of 2,740 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean. All data are relayed and made publicly available within hours after collection (real-time data). In addition to the real-time data stream, Argo provides salinity/temperature/pressure profiles that approach ship-based data accuracy. Both real-time and delayed-mode data are available from the GDACs, which synchronize their data holdings to ensure consistent data is available on both sites. The Coriolis Site provides advice and guidance on how to use Argo data effectively.

- Envisat data from the MIPAS, SCIAMACHY and MERIS instruments are available to ESA Category-1 grant holders from the NERC Earth Observation Data Centre (NEODC).

- EPICA Dome C Ice Core data are distributed by the NOAA National Environmental Satellite Data and Information Service (NESDIS). More information on the project can be found from the EPICA (European Project for Ice Coring in Antarctica) site at the University of Bern (Switzerland).

- MOZAIC (Measurements of Ozone by Airbus In-Service Aircraft) data are available on application for collaboration with the MOZAIC and IAGO investigators, from the MOZAIC site of the Laboratoire d'Aérologie in Toulouse (France). The data include measurements of ozone, water vapour, carbon monoxide and nitrogen oxides on board commercial flights of the Airbus aircraft. The MOZAIC Data Protocol and the application for collaboration form can be filled online on this site.

- MTCI (MERIS Terrestrial Chlorophyll Index) is a Level 3 product derived by the Company *Infoterra* from the measurements made by the MERIS instrument on board the ESA Envisat satellite. MTCI data come in two resolutions:
  - 4.63 km (global coverage)
  - 1 km (regional coverage of Europe, Australia and North America)

  MTCI data are distributed by the NEODC. Access to these data is granted on application and on provision of a brief research project description.

Other data sources of potential interest to QUEST

- The Southampton Assimilated Oceanographic Data at the BADC.
- The BRIDGE (Bristol Research Initiative for the Dynamic Global Environment) web site makes both model and paleo data available, and is mainatined by Paul Valdes at Bristol University.
- The COAPEC Dataset at the BADC.
- The HadGEM1 Control Run Dataset at the BADC.
- The HiGEM Dataset at the BADC.
- The LINK Dataset at the BADC.
- The OCCAM global model at Southampton.
- The GODIVA web site at Reading.
- The ENACT project.
- The HadISST dataset at the BADC.

Metadata Standard

- UNIDATA CF Metadata Conventions.

Contacts

- The QUEST Leader is Colin Prentice, University of Bristol.

- The QUEST Deputy Leader is Wofgang Knorr, University of Bristol.
- General queries about these pages or browsing the data should be directed to the BADC support line.