Quantifying and Understanding the Earth System (QUEST)
Themes 1 and 2 Data Scoping Study

Kevin Marsh, BADC
2 October 2006 - Revised December 2006

Introduction

This report describes the outcome of the scoping study carried out by the BADC between May and September 2006 to support the current and planned work of the NERC programme *Quantifying and Understanding the Earth System* (QUEST).

It is a requirement from NERC that all Directed Programmes which it funds 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 publicly funded data readily and easily available.
The NERC Data Policy requires that data are offered to the appropriate NERC Designated Data Centre. In the context of the QUEST programme this is the British Atmospheric Data Centre (BADC).

An integral part of any data management plan is an obligation upon the 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. The aim of the scoping study was to collect information to be used in writing the QUEST Data Management Plan that will be the next step in the practical implementation of the QUEST Data Policy. It has also been the occasion to raise awareness of data management issues with the project investigators.

QUEST research

Throughout its duration (2003 to 2009), the QUEST Programme will fund a number of projects to perform research in 7 main subject areas, as follows.

Theme 1. Carbon cycle
Theme 2. Climate regulation over glacial-interglacial timescales
Theme 3. Global change and sustainable resources
Theme 4. Earth system dynamics
Theme 5. Earth system modelling
Theme 6. Biosphere management
Theme 7. Earth system processes and prediction
- 7.1 Assessing and facilitating QUEST
- 7.2 Biogeochemical cycles and feedbacks
- 7.3 Climate-carbon modelling, assimilation and prediction
- 7.4 Dynamics of the Palaeocene-Eocene thermal maximum
- 7.5 Environmental change and fisheries
- 7.6 Fire data assimilation and prediction

In addition, QUEST intends to invest resources in the following two research activities.

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

The QUEST Data Management Scoping Study

The present scoping study covers the already awarded six projects under Themes 1 and 2. Another scoping study will focus in due time on projects yet to be awarded.

Visits were paid to five of the six principal investigators as shown in Table 1 (it was unfortunately not possible to meet with Prof Woodward). This followed on from some useful initial discussions at the QUEST Annual Science Meeting (ASM) in April 2006. A questionnaire was also developed by the BADC and distributed to co-investigators either directly via email or via the project PI. The purpose of the interviews and questionnaire was to determine specific project data issues and possible inter-project relationships of interest to QUEST data management.

Key pieces of information collected during the scoping study about input and output data are summarised in **Tables 2 and 3** and are commented in the next two sections. Data management questions of general interest for the QUEST Programme are addressed in the *Other data management issues* section. More detailed notes from the visits are attached to this report as **Appendix 1**. The questionnaire itself is given in **Appendix 2**.

Third-party data

Table 2 summarises the answers received to the question relative to the datasets required by the researchers to support their work, as well as the outcome of the enquiries made by the BADC about the availability of the requested data. Below are comments on the contents of **Table 2**.

ARGO — http://www.argo.ucsd.edu/ — 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 Global Data Assembly Centres (GDACs) located in Brest (France) and Monterey, California (USA) (website addresses of these two centres are given in **Table 2**). The GDACs synchronize their data holdings to ensure consistent data is available on both sites. The Coriolis Site — http://www.coriolis.eu.org//cdc/argo_rfc.htm — provides advice and guidance on how to use Argo data effectively.

MOZAIC — http://mozaic.aero.obs-mip.fr/web/ — stands for *Measurements of Ozone by Airbus In-Service Aircraft* and is a European collaboration between the Laboratoire d'Aérologie du CNRS (France), Météo-France (France), the KFA Jülich (Germany), the MPI Mainz (Germany), Cambridge University (UK), Tenerife University (Spain) and Airbus Industry. The project developed a portable system for automatic measurement of ozone, water vapour and temperature aboard commercial flights of the A340 aircraft. The apparatus was flown in the context of Mozaic, and data are now collected routinely in the framework of the Iago project which followed on Mozaic. The data are archived at the CNRM at Météo-France (Toulouse). Following the visit to Dr Pyle in May, contact has been made by the BADC with Dr Fernand Karcher (CNRM, Météo-France) about the possibility for the BADC to mirror the data. Although not opposed to the idea, Dr Karcher underlined that the archive was about to be moved to Paris to be maintained at a dedicated data centre, and that the MOZAIC data are given to any researcher on provision of a project description. The BADC has then contacted Dr Marenco (Laboratoire d'Aérologie), at the time the MOZAIC Programme Coordinator, and three of his colleagues of the Mozaic team about the possibility to mirror the Mozaic data, and is awaiting a reply. In the meantime, data can be obtained from the Mozaic database maintained at the Observatoire de Midi-Pyrénées through a request for collaboration with the Mozaic and Iago investigators. The link to the online request form is given in **Table 2**.

Envisat satellite data are already available to the PI of QUACC, whose team is involved in several ESA projects. As a UTLS Ozone participant and as a Category-1 ESA grant holder, the PI is entitled to access the MIPAS, SCIAMACHY and MERIS data via the NERC Earth Observation Data Centre (NEODC) if he wishes so. The address of the Envisat website at NEODC is given in **Table 2**.

MTCI — http://www.neodc.rl.ac.uk/?option=displaypage&Itemid=145&op=page&SubMenu=-1 — stands for *MERIS Terrestrial Chlorophyll Index*. The MTCI dataset is a Level 3 product derived by Infoterra from the ESA Envisat MERIS measurements and distributed by the NERC Earth Observation Data Centre (NEODC). Access to the data is granted on application and on provision of a brief research project description. Like other NERC funded researchers, the investigator who has mentioned his interest in this dataset has already access to it via the NEODC. However, the issue is that the data present important gaps, particularly over Year 2002 and over the period January 2005 to May 2006. In order to study ecosystem properties over time (e.g. drought impacts on vegetation phenology, temperature impacts on rate of senescence), it would be a significant asset for the researchers to have at their disposal a long time series. It is suggested that some QUEST funding could be devoted to mandate Infoterra to fill the gaps (or part of these) by processing historical MERIS data.

EPICA — http://www.climate.unibe.ch/clim_recon/epica.html — stands for *European Project for Ice Coring in Antarctica*. The EPICA Dome C Ice Core Data is distributed by the NOAA National Environmental Satellite Data and Information Service (NESDIS) — http://www.nesdis.noaa.gov/ — via the EPICA database webpage, the address of which is given in **Table 2**.

DESIRE (Dynamics of the Earth System and the Ice-core Record) is a French-UK collaboration project that was born in response to a joint NERC-INSU call for proposals to develop a quantitative and predictive understanding of the ice core record of changing atmospheric composition. The proposal focuses on CO2 & CH4. Since the QUEST participant who expressed the wish to access the data to be issued by this project is himself a partner of DESIRE, the data will be readily available to him on acquisition.

Additional third-party data required for the development of the projects and held at the BADC or at the NEODC, such as climate model data sets or data from other NERC projects, will be made available to the participants, subject to current access conditions. If required, the BADC will endeavour to retrieve data sets from other sources at no cost or will negotiate their acquisition at the best possible cost.

Data deliverables

Table 3 summarises the findings of the scoping study regarding data to be produced by the projects. A number of them, particularly in Theme 2, are in their early stage, making detailed estimates of output data volumes difficult. In general, the answers provided were vague and it was difficult to get a definitive picture of the work to be done, including which calculations would be performed and even which models would be used. It is hoped that this situation will improve as the result of iterative interaction between the BADC and the researchers during the projects development.

It is to be noted that it is inherent to a modelling project that the runs of prominent interest only stand forwards towards the end of the study. Therefore, it is understandable that most investigators interviewed said that their data would be ready for archival only at the end of their project. Since no researcher expressed the need of data produced by another QUEST project, the data submission date does not appear to be crucial.

Other data management issues

- ***Communication within QUEST***

  At the Annual Science Meeting of April 2006, many participants expressed the regret that no satisfying communication took place among the QUEST community. To help filling this gap, the BADC has offered to set up the following two tools.

  - BSCW Workspace — A QUEST online workspace (http://bscw.badc.nerc.ac.uk/pub/bscw.cgi/0/091007?op=login) has been created on the BADC BSCW (Basic Support for Cooperative Work) server. The workspace is visible and accessible only by the QUEST participants and is intended to ease the exchange of ideas, documents and preliminary data between the members of the programme. It is not an alternative to data submission to the BADC but must rather be considered as a discussion forum, a workshop and a temporary repository for data in the validation phase, draft papers, reports, etc.

  - Mailing Lists — If requested, the BADC will set up and run mailing lists for QUEST projects and the core team.

- ***Data management advice and support***

  At the occasion of discussions with participants at the ASM and interviews with the PIs, several additional data management issues have already been tackled, namely
  - the need of providing quality metadata;
  - the question of the data format — the proposed adoption of NetCDF met the participants' agreement;
  - the archival of model output, a major issue for QUEST and the object of a tumultuous debate, concluded by the adoption of the BADC Policy and Guidelines for the Archival of Simulation Data (see **Appendix 3**), endorsed by the QUEST Data Policy.

  Support in these three areas and in any other data management issue will be provided by the BADC to the QUEST researchers throughout the Programme development, via the QUEST website at BADC — http://badc.nerc.ac.uk/data/quest/ — and through individual interactions with the researchers.

Next steps

- Data Management Plan — Based on the current scoping study findings, the BADC will issue a draft Data Management Plan (DMP) to be submitted to the QUEST Core Team for discussion and approval. This will outline the technical aspects of the implementation of the QUEST Data Policy, keeping in mind the specificities of the QUEST projects.
- Remaining rounds scoping study — A new scoping study will have to be conducted to determine the needs and deliverables of projects yet to be funded (including the QUEST fellowship, which was not awarded yet when this scoping study took place).
- DMP update — Updating the DMP may reveal an ongoing activity, as some open issues may find a solution in the course of the Programme.
- Infrastructure settings at BADC — The infrastructure to be set up at BADC to support the Programme will be described in the DMP. Part of this work has already started.

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

Theme	Project acronym	Project title	Project duration	PI	PI's university	Date of visit
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	/
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. Required or desirable third-party datasets to support QUEST Themes 1 and 2 projects.

Project	Dataset name	Provenance	Availability	Required or requested by
MarQUEST	Argo	Argo Global Data Assembly Centres (GDACs) • Brest (France) http://www.coriolis.eu.org/ • Monterey, CA (USA) http://www.usgodae.org/argo/argo.html	Public	Keith Haines, Reading
QUAAC	MOZAIC	Database maintained at CNRM, Météo-France, Toulouse. Application is made through the Mozaic and Iago site at the Observatoire de Midi-Pyrénées at http://mozaic.aero.obs-mip.fr/web/features/database/access.html	On demand of collaboration	John Pyle, Cambridge
	Envisat	ESA MIPAS, SCIAMACHY and MERIS data also archived at NEODC: http://www.neodc.rl.ac.uk/?option=displaypage&Itemid=128&op=page&SubMenu=128	MERIS, MIPAS & SCIAMACHY data available to Cat-1 grant holders	
QUERCC	MTCI	ESA L3 data distributed by the NEODC at http://www.neodc.rl.ac.uk/?option=displaypage&Itemid=145&op=page&SubMenu=-1	Public (no commercial use) Gaps covering Yr 02 and Jan 05 to May 06	Giles Foody, SOTON
Quaternary QUEST	EPICA ice core	NOAA National Environmental Satellite Data and Information Service (NESDIS) http://www.ncdc.noaa.gov/paleo/icecore/antarctica/domec/domec_epica_data.html	Via the web (see left)	Agatha De Boer, UEA
	DESIRE	No data produced yet (proposal being reviewed)	Data will be readily available to the investigator, as a DESIRE participant	Neil Edwards, Open University

Table 3. 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**		

Appendix 1

Summary of Scoping Study Visits

Theme 1

Marine Biogeochemistry and Ecosystem Initiative in QUEST (MARQUEST).
Professor A Watson
University of East Anglia

Summary
MarQUEST will investigate ocean biogeochemical cycles and ecosystems, and their effect on both the oceans themselves and the Earth's climate (due to their effect on the composition of the atmosphere). Representation of these processes in ocean biogeochemical models has been simplified until relatively recently. The increase in complexity has led to problems with identifying the best way to validate such models. The group of projects with MarQUEST will aim to collaborate closely with each other to compare their model outputs and identify the sources of commonality and differences, and to examine the planktonic system.

MarQUEST will lead to

- The development of new methods of validating models, making use of remote sensing ocean colour data, in-situ data sets (with strong links to European programmes such as Carbo-Ocean and Euroceans)
- Comparison of different ecosystem models run in the same circulation codes
- Development of a module to simulate the coastal ecosystems, but useable in global ocean biogeochemical simulations
- Detailed comparison of ecosystem models with observations over recent decades, including estimates of the evolution of the CO2, oxygen and di-methylsuiphide fluxes from ocean to atmosphere over the next 50 and 100 years

MarQUEST has three work packages (WP1: 1d models, WP2 1 degree global models, WP3 NEMO ocean model). Some data assimilation work will be done by Keith Haines at Reading.
Professor Watson was uncertain what the actual data archival requirements of his project would be, but supplied contact details for the data producers within his project, and these researchers were sent copies of the questionnaire.

Outputs for BADC
Produced by 2009:

- Forecasts and hindcasts of CO2 into the oceans
- Volume : ~500GB

3rd Party Data Required
None

Other BADC support
None Required

Modelling of atmospheric oxidants and aerosols: deposition, emission and chemical transformation / QUEST (QUACC)

Professor J Pyle
University of Cambridge

Summary
The aim of this project is to study the role of chemistry/climate system coupling surface processes on atmospheric oxidizing capacity and aerosol loading,

This will build on an initiative already underway between the Met Office and the NERC Centres for Atmospheric Science (NCAS) to develop a new community model, UKCA, to study the interaction between climate and composition (gas phase composition and aerosols).
This project has four main foci:

(1) Development and testing of chemistry and aerosol schemes to include in a climate model
(2) Development and testing of a range of schemes to describe (interactively wherever possible) surface emissions of reactive trace gases,
(3) Development and testing of new surface deposition schemes,
(4) Implementation of these schemes into the climate model which will be used to look at climate-related variability of the model system for the immediate past and the near future.

This will allow the interaction between changing climate and surface emissions with full description of the feedbacks occurring within this system. The studywill allow these chemistry/climate feedback processes to be assessed in studies covering the last century and the coming century.

The project has some interaction with QUERCC, but there are no data exchange/accessibility issues which the BADC could assist. Professor Pyle offered to add Data Management issues to the agenda of his upcoming project meeting (25/5/06).
Not possible to say exactly what will be worth keeping until the project is nearing completion.
Some output information from this project may go into JULES.

Outputs for BADC
Produced by 2009:

- Model metadata (setup, initial conditions, etc)
- Subset of the actual model data (Met Office pp format/NetCDF)
- 'snapshots' of the output used in publications
- 40 year forcing run with SST
- Estimated Volume :~500GB

3rd Party Data Required

- MOSAIC data
- GOME/SCIAMACHY data
- Formaldehyde products for biogenic emissions
- FAAM

Other BADC support
Some data could be shared with MarQUEST via the BADC.

Theme 2

PalaeoQUMP:using palaeodata to reduce uncertainties in climate prediction

Dr S Harrison
University of Bristol

Summary
PalaeoQUMP (Quantifying Uncertainties in Model Prediction) is aimed at reducing the uncertainties in the climate sensitivity of current models by examining how they behave when used for very different periods in the earth's history. For this study, the periods around the last glacial maximum (LGM, 21,000 years ago) and that around 6000 years ago (mid-Holocene, MH), have been chosen. Testing climate models under these very different climates should put stronger limits on climate sensitivity. The intention is to run the HadCM3 climate model, using known changes in solar radiation, ice-sheet distribution, and greenhouse gas concentrations for the LGM and MH, and run the same series of simulations with different values for key processes. They will evaluate these simulations using reconstructions of LGM and MH climate. Lake and bog sediments provide ample evidence for changes in vegetation and precipitation during these periods. These will be used to force vegetation and precipitation models, and these will be one of the targets of the climate model simulations. The project is intended to provide a better estimate of the climate sensitivity to doubling CO2 in time for the IPCC Fifth Assessment Report.

This project will have strong interaction with Paul Valdes's QUEST project. Dr Harrison will concentrate on the production of the synthesis data, and will also be keeping an archive of the model codes and selected model output at Bristol. Prof Harrison is also concerned that runs should not be archived until they have completed their analysis, and that the data synthesis which are archived should be 'versioned' to dissuade researchers from using old versions of the dataset in their research. It was also noted that the raw data used to produce the synthesis would not be archived.

Outputs for BADC
Produced by 2009:

- Table of sensitivity values
- Global seasonal temperature, precipitation, vegetation maps for each experiment
- Limited model output (NetCDF format)
- Version of the model codes and model metadata
- Synthesis datasets and associated metdata

Estimated Volume :<100GB

3rd Party Data Required
None

Other BADC support
Some support may be useful in extracting the HadCM3 model output from the Met Office, and this is an area in which the BADC could assist.

QUEST Deglaciation: Climate and Biogeochemical Cycles during the last deglaciation

Professor PJ Valdes
University of Bristol

Start: 1/4/06
End: 1/4/09

Summary
This project looks at climate change and the complex interactions involved between the atmosphere and the biogeochemical cycles which have led to abrupt changes, such as the LGM. It will study this by using computer climate models (run over the past 21000 years) with extra features, including a dynamic global vegetation model that can predict changes in wetlands, deserts and forest fires) and will enable the simulation of the land-atmosphere exchanges of many important substances that affect climate, such as carbon dioxide, methane, volatile hydrocarbons, dust and soot. By using efficient versions of the models, many more and longer simulations can be performed than is usually done. To complement this work, there will be close interaction with PalaeoQUMP. A major synthesis of the data from sediment cores around the world will be produced, including pollen counts (an indicator of past vegetation), charcoal counts (an indicator of past fires) and carbon isotope measurements. This work will provide a continuous picture of the state of the Earth's land surface from the last ice age up to recent times. This will be used to assess how well the climate models are working, and examine how the vegetation changes are interacting with the climate and the composition of the atmosphere

Professor Valdes currently maintains his own website at Bristol for making his model output available to the research community. The model data currently on Professor Valdes's website could be transferred to the BADC to ensure it's long term preservation. Publications are seen as being the key output from this project.
The project also has some interaction with Tim Lenton's Quaternary QUEST project, and the outputs from this work may be of interest to researchers working on that project.
The model simulations produced will be from the HadCM3, FAMOUS and GENIE models. The code for these could be one of the archive products, though it was felt unlikely that these would be useful to other researchers. The simulations themselves can be considered to be available for use by the wider community once they have been produced.

Outputs for BADC
Produced by 2009:

- Model metadata (setup, initial conditions, etc)
- Subset of the actual model data (CF compliant NetCDF)
- Estimated Volume :~500GB

3rd Party Data Required
None

Other BADC support
Some support may be very useful in extracting the FAMOUS and GENIE model output from the QUEST machine at the Met Office, and this is an area in which the BADC could assist.

Quaternary QUEST: Regulation of atmospheric carbon dioxide on glacial-interglacial timescales and its coupling to climate change

Dr TM Lenton
University of East Anglia

Summary
The Earth System is a complex system, which has many interlinked parts. These parts are affected by changes in atmospheric circulation, vegetation cover, dust, ice and a host of other factors. Studies indicate that this behaviour is predictable, but that it is highly sensitive to small changes.

Part of the project will involve the production of syntheses of ancient records from ice cores and sediments, and use that information to improve and test the GENIE Earth System model.
This work will also focus on how the changes in atmospheric carbon dioxide interact with the Earth System.

Most of the output from GENIE will be <500GB; output data from the French OCA model will be larger.
The table below shows a summary of the planned data output from this project.

WP/ Institution	Type of data	Nature	Variables	Dimensionality	Volume @BADC	Delivery period
WP7 / Reading (Smith)	Numerical climate model results	numeric	climate fields	4D – 30 vertical levels ~3 degree horizontal resolution	10 Tb	May '09
WP 2/4 / OU (Edwards)	Model results and data synthesis	numeric	Ocean data – physical and biogeochemical	3D model and 1D data synthesis. Low temporal resolution	10-100 Mb	mid-project (end 2007)
WP4 / UEA (DeBoer)	Model results	numeric	Ocean circulation and state variables	4x4 degrees	? (small)	next 12 months
WP6 / Bristol (Payne)	Model results	numeric	3D global grib chemical data	GENIE resolutions	'not large'	throughout project
WP1/5 Cambridge (Hoogakker)	Model results and daat synthesis	Images	Pollen data and Oxygen and Carbon isotope data from cores	-	Respondent suggests none, see below	2007-2009

Quaternary QUEST is committed to the final output of the project being stored in one place (i.e. at BADC) and being fully accessible. In the case of WPs 1 and 2 there already exist ice core and isotope datasets to which data is likely to be contributed. However, in collaboration with BADC we would be keen to also archive a copy of that data with the Quaternary QUEST dataset also.

The GENIE model runs should be reproducible through the versioning and database system used to store and run the model. Metadata on how to reproduce model runs would therefore be a key output of projects using the GENIE models. It would also be sensible to store some model output (1 or 2 key data fields and graphics / video of the parameters of interest?) for each of the model runs, for the purposes of easy data access and checking of model re-runs. We anticipate regular engagement with BADC over the next year to develop a policy for standardising the archiving of GENIE output/ model runs.

Appendix 2

QUEST Data Management Scoping Study Questionnaire

The BADC has been directed by NERC to perform a scoping study to assess the data management issues associated with QUEST. We would be grateful if you could complete the following questionnaire as fully as possible and return it to Kevin Marsh at the BADC (k.marsh@rl.ac.uk) as soon as possible.

Conclusions of the scoping study will be summarised in a document based on the answers to this questionnaire and will provide the elements of a data management plan for the programme. However, a number of data management principles have already been outlined in the QUEST Data Management Policy (http://quest.bris.ac.uk/introduction/policies.html); the Data Management Plan is intended to provide guidelines and tools to implement these principles.

Section 1: Your QUEST project

Your name

Your contact details

QUEST Theme

QUEST Project

QUEST Project start date/duration

QUEST Project team members

QUEST Project collaborators

Section 2: Datasets required

Are there any 3rd party datasets (external to QUEST) which you require which the BADC could attempt to obtain on your behalf?

Will you need data from other QUEST projects? (which ones, at which stage of your project?)

Section 3: Major data output from your project

What type of data will be produced? (observation, model results, synthesis of existing data, etc.)

What is the nature of the data produced? (numeric, images)

What is the nature of the measured/modelled variables?

If applicable, what are the dimensionality and resolution of the space-time grid?

Are your data of such a nature that they should be archived at another NERC Data Centre than the BADC? — If so, which one(s)?

Which dataset produced by your project will require final archival at the BADC? At another data centre? How much volume do you expect each of these to represent?

When will these data be produced?

NetCDF (binary) and NASA Ames (ASCII) are the two data formats recommended by the BADC for QUEST numeric data. Do you feel comfortable using one of these two formats? Will you need help in formatting your data? Do you have arguments to support the use of a different format?

Will there be a requirement to store any "development" data at the BADC during the lifetime of your project?

If so:

How much storage space would be required?

When will this begin?

What format will the data be held in?

Will your data set need to be 'versioned'?

To which quality control procedure will your data be subject before submission to the data centre?

The QUEST Data Policy states that, with the exception of model output that cannot be immediately identified as suitable for long-term archival and for which submission may be delayed towards the end of the project, all QUEST data will become publicly accessible at the latest 1 year after acquisition (and immediately after creation for the data generated for the Earth System Atlas) but that, until 2 years after the project end, users must offer co-authorship to the data originator(s) on any paper based on the data.
Are there confidentiality reasons (individuals' privacy, commercial value) for which access to (part of) the data you will produce should remain restricted? Would you, on the contrary, prefer to make your data publicly available at an earlier stage (e.g. at the time when the data are submitted to the data centre)?

Any other information you feel would be of use?

Thank you for providing us with this valuable information. If you have any further questions, please contact us at badc@rl.ac.uk.

Appendix 3

Archiving 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.

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

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.

    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)..

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.

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.

    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,

[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".

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.

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.