CLASSIC NERC EO Centre of Excellence

Draft Data Management Plan

NEODC

July 2005

Introduction

NERC's Data Policy requires the curation of data generated by the research they fund. This means ensuring the long-term archiving and widespread use of the data, and ensuring best practice to achieve this. NERC are implementing this policy through a set of designated data centres, which in the case of Earth Observation, is the NEODC.

A survey of NERC EO Centres of excellence has been carried out (Jan – March 2005) in order to establish: (i) what data is used within the NERC EO Centres and whether there are common requirements best organised centrally, and (ii) to develop each Centre's plan and policy for data management.

Questionnaires were sent to all CLASSIC researchers to determine their needs in terms of data support (provision of third-party data sets or other services). The enquiry also addressed issues related to the data generated by the projects (nature, volume, flow, etc.). The main purpose is to consider data with long term importance and/or use to the wider scientific community.

This draft Data Management Plan is the result of discussions between and response to data questionnaires from:

- The CLASSIC Director
- Classic PIs and researchers
- The NEODC

CLASSIC structure

The Climate and Land-Surface Systems Interaction Centre (CLASSIC) is a NERC Collaborative Centre funded under the Earth Observation Centres of Excellence programme. Hosted at the University of Wales Swansea, CLASSIC brings together a consortium of researchers from the NERC Centre for Ecology and Hydrology at Wallingford and Monks Wood, the University of Durham, and the Hadley Centre for Climate Prediction and Research. See classic.nerc.ac.uk

Scope

The purpose of the CLASSIC data management plan is to set up a coherent approach to data issues for the Centre. Its objective is to ensure that

- Appropriate data support is provided to the scientists within the Centre.
- CLASSIC datasets are archived and distributed in a suitable manner
- Distribution conditions and data usage do not infringe on the individuals' rights to publish their own work.

- Potentially scientifically valuable data are kept for the long-term.
- A high quality documented CLASSIC data archive is created.
- Data and documents can be distributed more widely to the scientific community.

At present there is no funding to provide full data support and archival for all Centre of Excellence datasets. The NEODC can currently only provide these where there is not a resource issue. However, the aim is to identify what the Centres of Excellence's needs are, in order, in a next step, to ascertain what funding is required to meet them.

The following sections cover the main data management issues: provision of a data management plan and a data protocol, potentially setting up an archive, monitoring of data access, data distribution, publication of results based on CLASSIC data and support offered to data providers.

1. Data management plan and data protocol

The present draft data management plan should lead, after discussion with CLASSIC PIs, to a final Data Management Plan. It is suggested that a data protocol be adopted for the Centre (a proposed draft is attached to this document as an Annex).

2. Third-party data

Third-party data required for the development of the projects and held at the NEODC or BADC (e.g. Met Office data, Landsat images), will be made available to CLASSIC researchers, subject to current access conditions. If required, NEODC will endeavour to retrieve data sets from other sources at no cost or will negotiate their acquisition at the best possible cost.

Third-party data sets in use or required by CLASSIC are listed in table 1 of the Appendix. Some of these data sets are already freely available on the web, in which case the URL is given.

3. Sharing data

Data sets and model results generated by individual CLASSIC researchers could be made available to others in CLASSIC through the NEODC.. NEODC could provide a restricted access area where datasets could be placed for use across CLASSIC. Smaller datasets could also be placed on the CLASSIC data archive to be based at Swansea, and would be constructed around the current CLASSIC website. Currently large datasets are passed between the CLASSIC institutions on CDs, DVDs or for very large volumes of data on hard-disks, and through anonymous ftp sites

Comment: This is certainly something to aim towards – sharing datasets via NEODC will help prepare for archiving in the longer term. How much we can do will depend on resources however.

Publication issues are dealt with in Section 5.

CLASSIC data sets and model results for internal distribution are listed in table 2 (appendix).

4. CLASSIC data archive

4.1 Archive location

CLASSIC will produce a range of datasets, which may dealt with in different ways. Where it is considered that data are of wider interest to the community and a long-term archive is appropriate the data should be located at the NEODC/the chosen archive location (provided that it is set up to deal with backups, access control, documentation, dissemination, etc). The data provider is also responsible for providing documentation, metadata and possibly software to decode, interpret and visualise the data. The data provider may also be expected to field some user queries: science questions should be directly addressed to the responsible scientist, and questions about the data availability, format, etc. to the NEODC helpdesk

Comment: This is also a resource issue, leave open for now

4.2 Archiving policy

In recognition that validated raw data (i.e. QA/QC'ed data prior to additional processing) potentially represent an invaluable source of information for the future, the Centre's scientists will archive them in a way that guarantees longevity and accessibility. Although not necessarily located at NEODC, validated raw databases and their access should be fully documented at the NEODC. Processed (final) data should be archived at the NEODC. In addition, investigators are encouraged to submit model results which will have been the basis of theoretical studies or that illustrate the model capabilities.

The data sets to be archived are listed in table 3 (appendix).

4.3 Format

All data produced by CLASSIC should be stored in standard (commonly used by the community) file formats. When deciding on an output format CLASSIC scientists should consider accessibility and future use.
A preferred format should be agreed between CLASSIC scientists, e.g. raw binary with a separate header for image data, but other formats may also be allowed e.g. fits, ascii, hdf, netCDF and so on if there is good reason**.**
If non-standard data formats cannot be avoided, comprehensive format descriptions and read software must be provided.

4.4 Documentation

Metadata (i.e. information on the data) are a crucial part of any data archive since they ensure the accessibility and readability of the data. It is therefore essential that metadata be submitted at the same time as the data sets to which they pertain. Metadata relating to all raw CLASSIC data not archived at the NEODC should also be supplied to the NEODC.

To guarantee the data archive quality, full documentation on all validated raw and processed data, as well as on models and model results, must be provided to the NEODC. Standard metadata will be archived within data files. For an example of the sort of metadata that should be provided see: http://badc.nerc.ac.uk/help/metadata.

Guidelines for EO metadata will be made available by NEODC soon. In the meantime, questions may be directed to neodc@rl.ac.uk.

In addition to the standard metadata, investigators are encouraged to archive all relevant information, including model and experiment descriptions, references, papers, reports, etc. Copyright information and conditions of use (i.e. for academic use only, or academic and commercial use) should also be included.

4.5 Supporting collaboration with Collaborative Workspaces

The NEODC has set up a collaborative workspace dedicated to CLASSIC at http://bscw.badc.rl.ac.uk/bscw/bscw.cgi/0/62151. This is a secure web space available to registered users only, where scientists can share results, documents and preliminary data files.

4.6 Data submission

Preliminary data should be made available to other CLASSIC groups, where appropriate, as soon as possible.
Via NEODC or internal transfers – state which/how
Processed data and model results should be supplied to the NEODC/chosen data archive location as soon as they are ready, and no later than the project end date (2008).

If using NEODC – describe upload method here, e.g. web based file uploader or ftp.

Comment: Details on this section depend on decisions made on which datasets it is feasible to host at NEODC, etc

5. Data distribution

Different access restrictions are appropriate for different CLASSIC datasets, although the duration of the "data validation period" during which access is restricted may be a common feature (determined by the data provider, but typically one year from the project end date). A password-protected access system can be set up at the NEODC to reflect the defined permissions. Distribution of CLASSIC data held at the NEODC will take place via the Internet and FTP. During any restricted period, entitled CLASSIC scientists who have applied for access to the data will be allocated an account at the NEODC allowing them to directly download the data from the archive. This facility can be extended to external collaborators who will have been personally authorised to access the data by CLASSIC PIs.

At the end of the retention period, the data will be released to the public domain. The Intellectual Property Rights (IPR) to those data need not be transferred. After release, NEODC will make the data available to other bona fide researchers. Anonymous users will be requested not to use the data for commercial purposes; they will be asked to contact the relevant data providers before using the data and to acknowledge CLASSIC and the data suppliers in any publication using CLASSIC data. If required, a system can be put in place by which users will be asked to indicate agreement to these (possibly amended) terms prior to being given access to the data.

A CLASSIC Web page will be set up at NEODC, with links to datasets at NEODC and elsewhere, CLASSIC's own web site, publications, data access rules etc.

6. Publication

Results coming out of CLASSIC projects will be published in the usual way. During the data validation period, each investigator will have the right to refuse the use of his/her results in a publication or a presentation prior to the investigator's own publication of that work. If measurements or model results from other groups within CLASSIC are used in a CLASSIC participant's publication during or after the project, joint authorship must be offered. This will not necessarily have to be accepted, particularly in cases where due credit and acknowledgement can be given in other, possibly more appropriate, ways. References of publications should be communicated to the NEODC where a list of published works will be held.

7. Liaison between NEODC and CLASSIC scientists

The NEODC CLASSIC website will be the primary source of information regarding the CLASSIC archive.

The NEODC will keep in touch with the PIs and their collaborators, e.g. to exchange information on the submission procedure, relevant WWW links, the Data Management Plan and on the population of the CLASSIC archive using this website.

8. Support to CLASSIC scientists

NEODC may be able to provide additional data support to CLASSIC scientists if required. The possibility of remote processing of ATSR data on NEODC computers has already been suggested and will be investigated.

Appendix 1 – Summary of CLASSIC data

1. Third party datasets

Dataset	Access	Who	When
Met Office and other surface climate data ;	Already have access (BADC?)	Durham	n/a
STRM global topographic data ;	Already have access (http://seamless.usgs.gov)	Durham	n/a
MODIS continuous field vegetation	Already have access (http://glcf.umiacs.umd.edu/data/modis/vcf/)	Durham	n/a
AVHRR	NEODC /ISLSCP	Durham	n/a
Landsat TM	Already have data for testing routines, may need to purchase additional data later (NEODC)	Durham	?
Freely available satellite products (e.g. MODIS)	Already have access (NASA on-line)	CEH Monks Wood	n/a
Met Office – reanalysis from Unified model.	BADC (already have access?)	Swansea	n/a
AATSR and ATSR data for land-surface biophysical property retrieval.	NEODC	Swansea	?
LANDSAT and SPOT images of selected test sites	NEODC	Swansea	?
MSG SEVIRI	EUMETSAT	Swansea / Wallingford	
AVHRR level 1b	NASA GSFC	Swansea	
Land surface temp, and Precipitation	Tyndall data centre	Swansea	
Sea Surface Temp	NOAA	Swansea	

Field measurements	HAPEX-SAHEL, AMMA, BOREAS, LBA,	All	
ECOCLIMAP	ECOCLIMAP	Wallingford/Durham, Leicester	

2. CLASSIC produced data shared between groups/researchers

Dataset	Size	Share with whom?	How
Durham data (model output?)	<100 Gb	Within CLASSIC	?
CEH Monks Wood data (model output?)		Within CLASSIC	ftp
FASIR/NDVI	1 Gb	Within CLASSIC & CTCD	NEODC
Aerosol optical depth at 550nm	10's Gb	Remote sensing and atmospheric communities	NEODC

3. CLASSIC datasets for long term archival

Dataset	Size	Data Producer	When available to archive
Model outputs from Dynamic Vegetation Models ?	<100 Gb	Durham	
FASIR/NDVI already available via ISLSCP initiative as one, half and quarter degree squared resolution. Higher resolution (up to 0.1 deg squared) is available, but requires documentation.	1 Gb (10Gb for higher resolution)	Sietse Los, Swansea	1, 0.5 and 0.25 deg sq resolution available now. Higher resolution still needs documentation.
Aerosol opacity at 550 nm time series (ten years) over land and water, with aerosol type and land surface reflectance over Northern Africa and Western Europe using	10's Gb	Will Grey, Swansea	Processing of Northern Africa and Europe will get underway during the summer, probably completed by September

AATSR and ATSR-2. May be extended to other continents			
AATSR and ATSR-2 surface reflectance	A few Tb	William Grey	Will be derived as a by-product of aerosol retrieval

Appendix 2 - CLASSIC Draft Data Protocol

The aims of the Data Protocol are

- to encourage rapid dissemination of scientific results from CLASSIC;
- to protect the rights of the individual scientists funded by CLASSIC;
- to have all the involved researchers treated equitably;
- to ensure the quality of the data in the CLASSIC data archive.

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

1. CLASSIC data and model results produced during the lifetime of the Centre will be made available to all CLASSIC researchers, and CLASSIC researchers only, during a *restricted access period* ending one year after the concerned project end date, after which data and model results will be released to the public domain. At a principal investigator's request, access may be extended to personally authorised collaborators.
2. The designated CLASSIC data centre is the NEODC.
3. The longevity of validated raw data must be ensured in a secure archive, if possible at NEODC. Details pertaining to the validated raw data (i.e. metadata), whether or not archived at NEODC, must be sent to the NEODC, as well as information on how to access the data.
4. When relevant, preliminary data must be made available to CLASSIC collaborators as soon as possible. Any corrections or amendments to the preliminary data should be announced as soon as possible.
5. Validated processed data (i.e. data sets in their final form) must be archived at the NEODC. Archival must take place no later than the end of the concerned project.
6. Results of model studies feeding other CLASSIC projects or using data acquired during CLASSIC can be made available via the NEODC.
7. Data submitted to the NEODC must be in the data format agreed between CLASSIC principal investigators and the NEODC. Read software must be made available where non-standard formats have been used. All agreed metadata describing data, models and model results, regardless of their archival location, must be supplied to NEODC. Format and metadata are documented at NEODC.
8. It is each principal investigator's responsibility to ensure that the data used in publications are the best available at that time.
9. If measurements or model results from other CLASSIC research groups are used in a publication by a CLASSIC participant, joint authorship must be offered. This does not necessarily have to be accepted, particularly in cases where due credit and acknowledgement can be given in other, possibly more appropriate, ways.
10. Whilst the data are restricted from the public domain (see Clause 1), each principal investigator has the right to refuse to allow his/her work, whether direct measurement or derived product, to be used in a publication or presentation prior to the PI's own publication of that work.
11. Whilst the data are restricted from the public domain, no data should be transferred to a third party without the originator's consent.
12. In the event of dispute the final decision rests with the CLASSIC Steering Committee.