CTCD NERC EO Centre of Excellence

Draft Data Management Plan

NEODC August 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 was 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.

Discussions have taken place with the CTCD data manager, director and 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:

- The CTCD Data Manager
- The CTCD Director
- CTCD PIs and researchers
- The NEODC

CTCD structure

The Centre for Terrestrial Carbon Dynamics (CTCD) is a NERC Collaborative Centre funded under the Earth Observation Centres of Excellence programme. The aim of CTCD is to solve the equations for the terrestrial carbon balance, at a variety of scales, by a combination of modelling and data. Key to improving current understanding is quantifying the uncertainties associated with such calculations and analysing how best to reduce such uncertainties. The CTCD consists of a consortium involving the Universities of Edinburgh, Sheffield and York, University College London and Forest Research. For more information see http://www.shef.ac.uk/ctcd/ .

Scope

The purpose of the CTCD 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.
- CTCD 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 CTCD 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 and CTCD itself already has existing structures for data management in place. The NEODC can currently provide additional support where there is not a resource issue, but the aim is to identify what the Centres' of Excellence future needs are, in order in a next step to ascertain what funding would be 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 CTCD 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 CTCD 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

2.1 Third-party data external to CTCD

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 CTCD 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.

A comprehensive list of third-party data required by CTCD is provided in CTCD's data management strategy document.

3 Sharing CTCD data and model results

Data and model results generated by individual CTCD groups or researchers are made available to other CTCD groups through individual researchers (e.g. via ftp site at Sheffield), and through UCL for EO data. Metadata for non-EO and EO data are submitted to the Data Manager at Forest Research. Publication issues are dealt with in Section 6.

CTCD data / model results for internal distribution are listed in the CTCD Data Management Strategy document, tables 2.3 and 2.5.

4 CTCD data archive

4.1 Archive location

The CTCD archive will be located at NEODC, but individual datasets may be held at the partner sites. Metadata are currently held at the CTCD metadatabase at Forest Research.

Comment: Metadata will move elsewhere: Sheffield or NEODC but will be maintained by the CTCD Data Manager

CTCD 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, or 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

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 will be archived at the chosen archive location. In addition, investigators are encouraged to submit model results which will have been the basis of theoretical studies or that illustrate the model capabilities.

CTCD datasets for long-term archival are listed in table 1 (annex).

4.3 Format

All data produced by CTCD should be stored in standard (commonly used by the community) file formats. When deciding on an output format CTCD scientists should consider accessibility and future use. If non-standard data formats cannot be avoided, comprehensive format descriptions and read software should 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 documenting the existence of all CTCD 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-specific metadata will be provided by NEODC in due course. 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.

4.5 Supporting collaboration with Collaborative Workspaces

If requested, the NEODC can set up a collaborative workspace dedicated to CTCD. This would be 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 CTCD groups, where appropriate, as soon as possible. This should take place via the CTCD data manager, see CTCD Data Management Strategy document for details.

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.

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

5 Data distribution

Different access restrictions are appropriate for different CTCD datasets, although the duration of the "data validation period" during which access is restricted may be a common feature (if the example data protocol is adopted, it would be 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 CTCD data held at the NEODC will take place via the Internet and FTP. During any restricted period, entitled CTCD 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 CTCD 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 CTCD and the data suppliers in any publication using CTCD 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 CTCD Web page will be created at NEODC with links to datasets at NEODC and elsewhere, publications, data access rules etc. as well as CTCD's own web site.

6 Publication

Results coming out of CTCD 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 CTCD are used

in a CTCD 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 CTCD scientists

The CTCD web page at NEODC will be the primary source of information regarding the CTCD 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 CTCD archive using this website.

8 Support to CTCD scientists

Any other services NEODC could provide?

Annex 1 – CTCD datasets for long term archival

Dataset	Size & format	Data Producer	Where archive	When available to archive
ERS coherence data and forest ages estimates	160 MB (ERDAS Imagine format)	P.Drezet (Sheffield)	NEODC	Now
Derived products from MODIS data: vegetation continuous fields product: MODIS NDVI and EVI over UK ; LAI/fAPAR data over UK ; land cover data ; nadir BRDF-adjusted reflectance and albedo		M. Disney, UCL	UCL (and NEODC?)	Details tbc with M Disney (email exchange started)
Phenology data derived from FASIR		M. Disney, UCL	UCL (and NEODC?)	Details tbc with M Disney (email exchange started)
Fieldwork data: Reflectance, LAI, canopy cover, atmospheric optical depth at Harwood, Barton Bendish, East Anglia and San Rossore		M. Disney, T.Quaife, UCL	UCL (and NEODC?)	Details tbc with M Disney (email exchange started)
Processed data (value-added EO products, MODIS, CHRIS-PROBA, SPOT, SeaWIFS, ERBE)		T.Quaife, UCL	UCL (and NEODC?)	Details tbc with T Quaife (email exchange started)
Products of ForestETp	~10's Mb .txt files	C.Coudun, Forest Research	Forest Research but NEODC for 'reference model output' used in comparisons with Sheffield	From April 06
Products of SDGVM		M. Lomas, Sheffield	*Often have requests for this type of data, usually global npp maps or data, handle this through anonymnous ftp site. I don't think I need the use of a data centre for SDGVM input/output data at the moment. Though, if the size of the data sets or the number of requests significantly increases then a data centre may well make life easier.*	
Products of SPA	20 x ~ 1MB (20 European sites)	M.Williams, Edinburgh	*We are generating output of C an balances for numerous European where intensive measurements ar underway, and these model outputs could be usefully stored and used by other groups. I think we would use a data*	
Products of ACM		M. Williams, Edinburgh		

Comment: In this case it may be appropriate for NEODC to hold metadata and a link to the data – is it currently visible to all who may be interested?

			centre to store these outputs, and that they would become available over the next few months as we work through the European sites.	
Sampling of soils data: Wheldrake Forest, near York. focus on soil respiration and partitioning it into the fluxes: Root+Mycorrhiza, mycorrhiza and soil only	~10Mb, .txt files	A. Heinemeyer, York	CTCD database (and NEODC?)	After June 06
Harwood eddy covariance		C. Nichol, Edinburgh		Contacted re. archival 14/07/05 – reply promised
Hyperspectral data collection		C. Nichol, Edinburgh		Contacted re. archival 14/07/05 – reply promised
Leaf biochemical data		C. Nichol, Edinburgh		Contacted re. archival 14/07/05 – reply promised

Comment: Are metadata produced?

Annex 2 - CTCD Draft Data Protocol

The aims of the Data Protocol are

- to encourage rapid dissemination of scientific results from CTCD;
- to protect the rights of the individual scientists funded by CTCD;
- to have all the involved researchers treated equitably;
- to ensure the quality of the data in the CTCD 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 CTCD Centre. Therefore, to try to meet these aims, all PIs involved in CTCD, in accordance with and on behalf of their co-investigators, must agree to abide by the following conditions:

1. CTCD data and model results produced during the lifetime of the Centre will be made available to all CTCD researchers, and CTCD researchers only, during a dataset-dependent *restricted access period* ending no more than 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 CTCD 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 CTCD 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 CTCD projects or using data acquired during CTCD can be made available `via` the NEODC.
7. Data submitted to the NEODC must be in the data format agreed between CTCD principal investigators and the NEODC. 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 CTCD research groups are used in a publication by a CTCD 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 measurement or calculation, 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 CTCD Steering Committee.