

CPOM 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 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 CPOM 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 CPOM Director
- The CPOM Systems Manager
- CPOM PI's and researchers
- The NEODC

CPOM structure

The Centre for Polar Observation and Modelling (CPOM) is a research centre that studies processes in the Earth's polar latitudes that may affect the Earth's albedo, polar atmosphere and ocean circulation, and global sea level. CPOM uses theoretical and laboratory-derived understanding to form new mesoscale models of interactions between the ice, ocean and atmosphere, and use ground and satellite observations to test the predictions of these and other climate models. CPOM is funded by the Natural Environment Research Council and has research groups at University College London, the University of Bristol and Cambridge University. For more information see <http://www.cpom.org/> .

Scope

The purpose of the CPOM 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.
- CPOM 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 CPOM data archive is created.
- Data and documents can be distributed more widely to the scientific community.

At present there is no funding for NEODC to provide full data support and archival for all Centre of Excellence datasets, and CPOM itself already has some 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, setting up an archive, monitoring of data access, data distribution, publication of results based on CPOM 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 CPOM 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).

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 CPOM scientists, 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 datasets used by CPOM are listed in table 1

3. Sharing CPOM data

Data and model results generated by individual CPOM groups or researchers could be made available to CPOM groups through the NEODC or through existing internal distribution methods.

EO data are already made available through the CPOM archive.

Publication issues are dealt with in Section 6.

CPOM data and model results for internal distribution are listed in tables 3 and 4.

Comment: If useful, NEODC could be used to distribute data, to registered persons. This would be particularly appropriate for datasets which are to be made more widely available once validated.

4. CPOM data archive

4.1 Archive location

The CPOM archive will be located at NEODC and CPOM. CPOM will produce a range of datasets, which will be dealt with in different ways. CPOM handles two types of data:

- *Satellite data* already has adequate distribution, backup, documentation and access control at CPOM. Data is kept in a dedicated storage room and all the stored data is recoverable from the original source distributor. Large parts of the archive are also duplicated online (on hard disk) at MSSL & CPOM/UCL and are available to authorised CPOM users. System/project backups are stored on disk (daily) & SDLT media (monthly) which are kept in a fire safe and off-site at UCL.
- *Derived datasets, and model output to be archived at NEODC.* 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 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: The scope of the archive at NEODC will depend on available resources. The structure described here is a good one to aim for however.

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 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.

CPOM datasets which should be archived are listed in table 2 (appendix).

4.3 Format

All data produced by CPOM should be stored in standard (commonly used by the community) file formats. When deciding on an output format CPOM 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 CPOM 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>.

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 CPOM. 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 CPOM scientists, where appropriate, as soon as possible.

Derived, processed datasets and model output, for which NEODC is used for distribution and/or archival, should be submitted to NEODC as soon as they are ready, and no later than the project end data. Satellite data is already supplied to the CPOM archive.

[Describe NEODC](#) upload method here, e.g. web based file uploader or ftp.
[Describe upload to CPOM archive, if relevant.](#)

5. Data distribution

Different access restrictions are appropriate for different CPOM datasets, although the duration of the “data validation period” during which access is restricted may be a common feature. A password-protected access system can be set up at the NEODC to reflect the defined permissions. Distribution of CPOM data held at the NEODC will take place via the Internet and FTP. During any restricted period, entitled CPOM 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 CPOM 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

CPOM and the data suppliers in any publication using CPOM 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.

Access to data in the CPOM/MSSL archive is provided through the CPOM data management team (A.Muir, J.Gaudelli) by unix group authorisation within CPOM/MSSL as decided by the relevant project PI. Access to data for external users is provided by password protected sftp/http as requested by a PI.

A CPOM Web page will be created at NEODC with links to datasets held at NEODC, CPOM and elsewhere, publications, data access rules etc

6. Publication

Results coming out of CPOM 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 CPOM are used in a CPOM 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 CPOM scientists

The CPOM website at NEODC will be the primary source of information regarding the CPOM data 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 CPOM archive using this website.

8. Support to CPOM scientists

| ??Any other services NEODC could provide?

Appendix 1 – Summary of CPOM data

1. Third party datasets

Dataset	Access
ERA40 reanalysis data	Already have access (BADC) – Bristol use CSAR facility (but would like to avoid) UCL has direct access
NCEP/NCAR reanalysis	Already have access (website)
Upward looking sonar (ULS)	Already have access (website)
Ice draft/thickness data from the National Snow and Ice Data Center	Already have access (website)
NSIDC Arctic sea ice concentration data	Already have access (website)
NSIDC Arctic sea ice velocity data	Already have access (website)
Satellite data	Access through Dundee uni
Met Office data	Already have access, BADC
ERS RA geodetic phase data	Already have access (from?)
IceSAT GLAS data	Already have access (from?)
ERS RA WAP altimeter data (1.7TB)	Already have access from ESA (ftp, Exabyte tape and DVD-ROM)
Envisat RA2 L2 GDR, SGDR (21GB per month) ;	Already have access from ESA (ftp, Exabyte tape and DVD-ROM)
Envisat RA2 L1b (14 GB per month) ;	Already have access from ESA (ftp, Exabyte tape and DVD-ROM)
Cryosat L2 data (140GB/month).	Already have access from ESA (ftp, Exabyte tape and DVD-ROM)

2. CPOM datasets for long term archival

Dataset	Size	Data Producer	Where archive	When available to archive
Results from numerical models. (a few standard results referred to in papers)	100 Gb per expt	Tony Payne, Bristol	NEODC?	When validated
Radio-echo sounding (RES) layer transects (spreadsheet) Raw data: collected in 70s, only digital copy of the data	< 1Mb 10 Gb	Martin Siegert	NEODC?	When validated, at end of project
UM 4.5 high res data for case studies over Antarctic Peninsula and Greenland. Data include u,v,p,theta,omega fields in UM pp format, as well as UM startdumps	3 Gb	Andrew Orr	Not archived	?
A 1-km Digital Elevation Model of Antarctica plus error map and ancillary grids and support documentation (metadata)	~120 Mb	Jose Luis-Dans, Bristol	NEODC?	?
Velocity maps for individual glaciers and basins of ice sheets ; Bed and internal layer maps from ice penetrating radar data	?	Jonathan Bamber, Bristol	NEODC?	?
Radar Altimeter Sea Ice Freeboard / Thickness Laser Altimeter Sea Ice Freeboard Radar Altimeter Sea Surface Topography	< 100 Gb < 100 Gb < 100 Gb	Seymour Laxon et al, UCL	NEODC ?	?
Antarctic averaged timeseries data from ERS & Envisat altimeter crossovers, Antarctic averaged timeseries data from Cryosat altimeter crossovers	200 Mb 200 Mb	D.Wingham et al	NEODC?	When validated, at end of project

3. CPOM datasets (satellite) archived and distributed by CPOM

Mission	Products	Media	Dates	Amount	Source	Distrib
Envisat	RA-2 L1b, GDR	CD, DVD, disk	1/10/2002->	1200 GB	ESA	ftp/nfs
ERS-2	WAP	Exabyte/ DLT tape	1995->	1400 GB	UK-PAF	ftp/nfs
ERS-1	WAP	Exabyte tape	1992 - 1996	600 GB	UK-PAF	ftp/nfs
ERS-1/2/Env	SAR/ASAR	CDs	1993-1994	7.0 TB	ESA	ftp/nfs
Topex/Jason	GDR/M	CDs	1992->	120 GB	AVISO	ftp/nfs
ERS-1, ERS-2	ALT-OPR	CDs	1993-1998	60 GB	ESA	ftp/nfs
ERS-1, ERS-2	ATSR	Exabyte tape	Not known	900 GB	N/A	ftp/nfs
SIR-C	SAR	CD	Not known	160 GB	N/A	ftp/nfs
Cryosat	L1b	disk	(2005->)	(5TB+)	ESA	ftp/nfs

4. CPOM produced data shared between groups/researchers

Dataset		Share with whom?	How
Results from numerical models. (T.Payne)	100 Gb per expt	Inside CPOM + BAS, other CEs	NEODC ?
Radio-echo sounding (RES) layer transects (Siebert)	< 1 Mb	Inside CPOM + BAS	CPOM – could do through NEODC?
Theoretical work (Wilchinsky)		Inside & outside CPOM	CPOM – could do through NEODC?
Antarctica DEM (Luis-Dans)	~120 Mb	Inside & outside CPOM, large no of institutions in US and Europe	CPOM – could do through NEODC?
Velocity maps for individual glaciers and basins of ice sheets ; Bed and internal layer maps from ice penetrating radar data (Bamber)		Inside & outside CPOM, other institutes nationally and internationally	CPOM – could do through NEODC?
Radar Altimeter Sea Ice Freeboard/Thickness; Laser Altimeter Sea Ice Freeboard ; Radar Altimeter Sea Surface Topography	< 100 Gb < 100 Gb < 100 Gb	Inside & outside CPOM, Hadley Centre, SOC & part of international collaborations	CPOM – could do through NEODC?
Antarctic elevation data	5 Gb	Inside & outside CPOM,	CPOM – could

from crossover analysis from ERS-1 RA ; Antarctic elevation data from crossover analysis from ERS-2 RA ; Antarctic elevation data from crossover analysis from Envisat RA2 ; Antarctic elevation data from crossover analysis from Cryosat ; Antarctic averaged timeseries data from ERS & Envisat altimeter crossovers ; Antarctic averaged timeseries data from Cryosat altimeter crossovers	55GB 5Gb/month 5Gb/month 200Mb 200 Mb	SPRI Cambridge, University of Colorado, US.	do through NEODC?

Appendix 2 - CPOM Draft Data Protocol

The aims of the Data Protocol are

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

1. CPOM data and model results produced during the lifetime of the Centre will be made available to all CPOM scientists, and CPOM scientists only, during a dataset specific *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 CPOM 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 CPOM 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 CPOM projects or using data acquired during CPOM can be made available via the NEODC.
7. Data submitted to the NEODC must be in the data format agreed between CPOM 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 CPOM research groups are used in a publication by a CPOM 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 CPOM Steering Committee.