ESSC 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 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.

Questionnaires were sent to all ESSC 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 ESSC Director
- ESSC PI's and researchers
- The NEODC

ESSC structure

The Environmental Systems Science Centre (ESSC), at the University of Reading, is the NERC group responsible for research into new ways of handling spatial data, particularly remotely sensed data within computer models, in the environmental sciences. For more information see http://www.nerc-essc.ac.uk/.

Scope

The purpose of the ESSC 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.
- ESSC 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 ESSC 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 ESSC 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 ESSC 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 ESSC 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

2.1 Third-party data external to ESSC

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 ESSC 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.

3. ESSC data archive

3.1 Archive location

ESSC archives will be located at NEODC and ESSC.

Comment: Data of wider interest to the community in the longer term should ideally be archived at NEODC. How much we can archive, and when, depends on available resources. Some ESSC data are already archived elsewhere – duplication may be unnecessary

ESSC will produce a range of datasets, which will be 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 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.

3.2 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.

3.3 Format

All data produced by ESSC should be stored in standard (commonly used by the community) file formats. When deciding on an output format ESSC scientists should consider accessibility and future use. If non-standard data formats cannot be avoided, comprehensive format descriptions and read software should be provided.

3.4 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 ESSC 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. NEODC will produce guidelines for EO-specific metadata but in the meantime, questions can 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.

3.5 Supporting collaboration with Collaborative Workspaces

If requested, the NEODC can set up a collaborative workspace dedicated to ESSC. This would be a secure web space available to registered users only, where scientists can share results, documents and preliminary data files.

Comment: I suspect this is not required

3.6 Data submission

Preliminary data should be made available to other ESSC scientists, 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 or 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.

Comment: Details to be finalised at a later stage

4. Data distribution

Different access restrictions are appropriate for different ESSC 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 ESSC data held at the NEODC will take place via the Internet and FTP. During any restricted period, entitled ESSC 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 ESSC 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 ESSC and the data suppliers in any publication using ESSC 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.

An ESSC Web page at NEODC will contain links to datasets at NEODC and elsewhere, the ESSC web site, data access rules etc.

5. Publication

Results coming out of ESSC 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 ESSC are used in a ESSC 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.

6. Liaison between NEODC and ESSC scientists

An ESSC website will be set up at NEODC with links to ESSC's own web pages, and relevant datasets. This will be the primary source of information regarding the ESSC 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 ESSC archive using this website.

7. Support to ESSC scientists

Any other services NEODC could provide?

Appendix 1 – ESSC data survey responses

ESSC data survey responses[1]

Name	Status in CE	Research	questionnaire response ?
Geoff Wadge	Professorial Research Fellow	Team Leader: Solid Earth Science	Yes, 10 Jan
Dave. Mason & Tania Scott	Reader ; PostDoc	1. Improving river flood models using remotely sensed data (LiDAR, satellite SAR, airborne SAR); EPSRC project 2. Improving coastal morphodynamic models using satellite SAR data (EU FP5 TIDE project)	Yes, 17 Jan
Tony Slingo; Richard Allan ; Margaret Woodage ; Jeff Settle	Deputy Director & Atmospheric Team Leader ; Research Fellow, Research Fellow ; Research Fellow	Evaluation of climate and weather forecast models using satellite, reanalysis and other data SINERGEE: Project to compare Met Office NWP model data with GERB data Studies related to the Surface Radiation Budget Development of the high-resolution version of the UK Met Office Unified Model (HiGEM) to investigate the radiation budget, clouds and diurnal cycle.	Yes, 17 Jan (Allan), 8 Feb (Settle), 8 Mar (Woodage) plus conversation with R Gurney 10th March
Robert Gurney	Professor and Director	Land surface processes	yes – 26 Jan
Jan Pentreath	Prof	Radioactivity in the environment	yes – 30 Jan
Keith Haines	Prof. Marine Informatics	Marine modeling and assimilation ; E-science	yes - 7 Mar

[1] Text in italics are updates following meeting V.Jay/A. de Rudder/R.Gurney 9/3/05

Name	Data sets		Which worth archiving	Times of data
Geoff Wadge	1. Radar Interferometry processed data (several small projects)	100-200 Gb	Both are worth archiving and maintaining for about 5 years (techniques then likely to be superseded)	Continuous 2005-2008
	2. Etna atmospheric flow models using Unified Model	100 Gb		
Dave. Mason ; Tania Scott	DTMs produced from LiDAR	10Gb	None	
Tony Slingo; Richard Allan ; Margaret Woodage ; Jeff Settle	Processed GERB-Met Office model comparisons	50 Gb	GERB-Model comparison dataset – 20 Gb.	2003-2007
	SEVIRI-GERB-Met Office comparisons/simulations	50Gb?		2003-2007
	Local maps of albedo, temperature and emissivity derived from SEVIRI images.			Soon – n/k
	Time series of cloudiness, irradiance, downwelling LWR, etc		Possibly the surface flux measurements	Soon – n/k
	Model diagnostics from various runs ; Processed datasets derived from these runs ; Processed datasets derived from other HiGEM runs	Various sizes ~ 20 Gb ~ 20 Gb	All	Mar 05 – Dec 06
	Future: ARM site in Niger underneath GERB for intercomparisons of data from many sources		All	?
Robert Gurney	some small field data sets which have been gathered in NERC grants, and which will be placed in NERC data centres when worked up		will archive Sonning expt field data (includes passive microwave radiometer data - questions to resolve before making generally available).	2001-2003
Jan Pentreath	Data sets on DPUC values and various dose rate calculations and their results	A few Gb	None (*RG thinks NERC data centre should archive – international treaty*)	
Keith Haines	Model assimilation data sets ; In future possibly model ensembles; Currently we store and serve around 1.5 Tb of data to marine and e-science communities using web services. (includes data served for Met office)	100-200Gb each	All Met office data is archived long term ; All of our 1.5 Tb data stored we would expect to be needed on 5 year timescale at least	Met office data received daily

Name	Of use to other ESSC projects	Share data with other institutes ?	When could data be submitted to a NEODC archive?
Geoff Wadge	No	Yes. Ad hoc arrangements for (1). INGV Catania for (2). Currently distributed themselves, but could distribute through NEODC	When? *Archive at NEODC, but review after number of years.*
Dave. Mason ; Tania Scott	No	DTMs from LiDAR are sent to Bristol, Nottingham and Heriot-Watt Universities from project 1. (Bristol Archive) Data are sent to Padua University archive from project 2. Data is distributed on project websites.	*NEODC should link to data a these archives*
Tony Slingo; Richard Allan ; Margaret Woodage ; Jeff Settle	Yes	yes, inside and outside ESSC: Met Office Other projects within ESSC (e.g. diurnal cycle). Collaborators on the RADAGAST and AMMA projects. Distribute data ourselves.	(When validated) *Gerb - Met Office intercomparison data already made available through ESSC website (link to from NEODC)* *Clarify Met Office agreement and long-term archival.* *AMMA data: through BADC*
Robert Gurney	Yes	We share field data with collaborators elsewhere; the data sets are small and can be FTP'd or distributed on DVD. Distribute data ourselves.	*When validated.* *Could be made available sooner if users restricted.* *Other data archived at USDA and NSIDC: link to these from NEODC*
Jan Pentreath	No	Yes, outside ESSC. Data shared with the EA and labs in Sweden and Norway. Distribute ourselves.	n/a *[NRPB fund UK side and the distribution and archiving are covered under international agreement. It would be in NERC's interest to make this dataset available through NEODC – RG]*
Keith Haines	Yes	yes, see www.nerc-essc.ac.uk/godiva. Distribute data ourselves	*Should link to GODIVA from NEODC pages, but also consider long term archival o (probably a subset) of the datasets*

Name	Third party data? Have access ?	IPR Issues?	Anything else?
Geoff Wadge	ENVISAT ASAR/MERIS ; Met Office UM forecast data ; already have access (through ESA and Met Office)		
Dave. Mason ; Tania Scott	ERS SAR, Envisat ASAR, NERC ARSF ATM data ; already have access	NO (though for the TIDE project data will remain the property of the PIs for a limited time after the end of the project)	
Richard Allan ; Margaret Woodage ; Tony Slingo ; Jeff Settle	CERES radiation budget and other (NASA Langley DAAC) ; SSM/I water vapour, winds, etc (ssmi.com) ; Reanalysis vertical motion data (ECMWF, NCEP) ; SEVIRI data (EUMETSAT) ; BSRN data (bsrn.ethz.ch/) ; HIRS data (Darren Jackson, TBD) ; Met Office Climate model data (TBD) ; TOMS data (http://toms.gsfc.nasa.gov/). Already have access. Archived most at ESSC. Acquisition of SSM/I, BSRN, HIRS, NCEP, TOMS ongoing or TBD SEVIRI images. GERB data. Ground flux data collected by CEH. etc.; already have access; CLAUS (Cloud Archive data), already have access	Met Office owns model data.	*Would be helpful to back-up all these data already archived. Data archived at ESSC is generally not backed-up. (Check this – is this really a problem?)*
Robert Gurney	Airborne altimetry data from the EA (agreements in place). SSR/SSMI/AMSR snow volume data (ex NASA) and related field data from the CLPX experiment (via NSIDC). I generally have access to these data. However, there may be an issue on SMMR data (1979-1988) , SSMI and AMSR data if I need the raw data reprocessing.	No: we make data publicly available as we are a research group. ;	
Jan Pentreath	No		
Keith Haines	Receive Met Office NCOF data daily ; direct from Met Office	Met office data requires agreement that access if for non-commercial use. We have authority to decide that	

Appendix 2 - ESSC Draft Data Protocol

The aims of the Data Protocol are

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

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