CASIX 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.

Discussions have taken place with the CASIX data manager and CASIX researchers regarding 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 CASIX Data Manager
- The CASIX Director
- CASIX PI's and researchers
- The NEODC

CASIX structure

CASIX, the Centre for observation of Air-Sea Interactions & fluXes, is a NERC Centre of Excellence in Earth Observation. The scientific focus is on advancing the science of air-sea interactions and reducing the errors in the prediction of climate change. The primary goal is to quantify accurately the global air-sea fluxes of carbon dioxide (CO2). CASIX will accelerate the exploitation of new Earth Observation satellite data to further the understanding of marine biogeochemistry in the Earth System. The CASIX partnership comprises:

Plymouth Marine Laboratory (PML)
Proudman Oceanographic Laboratory (POL)
National Oceanography Centre Southampton (NOCS)
University of East Anglia, School of Environmental Sciences
University of Edinburgh, School of Geosciences

University of Leicester, Earth Observation Science Group
University of Plymouth, School Of Earth, Ocean And Environmental Sciences, Geomatics Group
University of Reading/ESSC, Marine Science
University of Wales Bangor, School of Ocean Sciences, Centre for Applied Oceanography
Met Office
See http://casix.nerc.ac.uk/ for further information.

Scope

The purpose of the CASIX 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.
- CASIX 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 CASIX 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 CASIX 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, setting up an archive, monitoring of data access, data distribution, publication of results based on CASIX 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 CASIX 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 Annex).

2. Third-party data

2.1 Third-party data external to CASIX

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 the 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 data sets in use or required by CASIX 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 CASIX data and model results

Data and model results generated by individual CASIX groups or researchers could be made available to CASIX groups through the NEODC. Internal mechanisms are already in place for several datasets (e.g. NOCS GADS server, University of Plymouth bio-optical database, BODC and ESSC GODIVA server).

Publication issues are dealt with in Section 6.

A list of relevant CASIX data / model results for internal distribution is given in table 2 (appendix).

4. CASIX data archive

4.1 Archive location

The CASIX archive will be located at ???? – can be in several locations : ESSC GODIVA, Southampton GADS server, BODC, PML and NEODC – need to check if GODIVA will be a long-term archive.
CASIX 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 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 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.

CASIX datasets to be archived are listed in table 3 (appendix).

4.3 Format

All data produced by CASIX should be stored in standard (commonly used by the community) file formats. When deciding on an output format CASIX 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 CASIX 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 provide specific guidelines for EO data in due course, but 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 CASIX. 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 CASIX 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 or chosen data archive location as soon as they are ready, and no later than the project end date. (when??)

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

Comment: Details depend on chosen archive location and timescale

5. Data distribution

Different access restrictions are appropriate for different CASIX 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 CASIX data held at the NEODC will take place via the Internet and FTP. During any restricted period, entitled CASIX 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 CASIX 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 CASIX and the data suppliers in any publication using CASIX 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 CASIX Web page will be set up at NEODC with links to datasets at NEODC and elsewhere, publications, data access rules etc, and to the CASIX web site.

6. Publication

Results coming out of CASIX 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 CASIX are used in a CASIX scientist'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 CASIX scientists

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

Appendix 1 – Summary of CASIX data

1. Third party datasets

Dataset	Access	When
Met Office data ;	Already have access (Met Office are a partner)	n/a
Satellite data from ESA (SAR, MERIS, (A)ATSR), NASA, NOAA, and maybe other space agencies (e.g. India).	Already have access NASA Oceancolor: http://oceancolor.gsfc.nasa.gov/ PODAAC: http://podaac-www.jpl.nasa.gov/ ESA: http://earth.esa.int/services/catalogues.html	n/a
pCO2 from other academic institutions, notably Lamont-Doherty	Already have access List relevant URLs here	n/a
SAR data	On order from ESA	n/a

2. CASIX produced data shared between groups/researchers

Dataset	Share with whom?	How
Biooptical database:	Inside CASIX & other institutes (POL, Met Office)	BODC and Plymouth database
EO products	Inside CASIX & other institutes (POL, Met Office)	CASIX, possibly NEODC
Model output, climatologies	Inside CASIX plus other institutes	CASIX
Climatologies of wind speed, sea surface temperature and other parameters ; Climatologies of gas transfer velocity ; Climatologies of CO2 flux	Inside CASIX plus other institutes	NEODC (restricted access initially) for institutes outside the CoE
Retrieved CO2 vertical columns from SCIAMACHY Estimated size: 10Gb	Outside CASIX for validation of results with other CO2 retrievals – IUP Bremen, SRON and University of Heidelberg.	CASIX

3. CASIX datasets for long term archival

Dataset	Size	Data Producer	Where archive	When available to archive
Biooptical database: http://www.research.plymouth.ac.uk/casix/d		U. Plymouth	BODC	*Link to this archive*

atabase/				*from NEODC*
EO products: Initially for model areas (North Atlantic), but possibly global after that. e.g. http://www.research.plymouth.ac.uk/geomatics/casix/datasetv2/index.html The modellers are interested in taking this to 1km resolution (would be swath data, not reprojected) and then globally.	~ 720 Mb (5 yr dataset, N.Atlantic, 9km resolution)	U. Plymouth	Godiva initially. (*link to dataset from NEODC*) Archive at NEODC later?	Version 1 (5 year N Atlantic 9km): to be finished shortly. Version 2: April 2006.
FOAM HADOCC model output : Final products at the end of CASIX: ten year runs assimilating EO (and other) data producing CO2 variables as well as the model state variables. Possibly some shorter runs also.	~20 Tb	Met Office	GODIVA then NEODC?	
POLCOMS model output: Final products at the end of CASIX: ten year runs assimilating EO (and other) data producing CO2 variables as well as the model state variables. Possibly some shorter runs also.	~20 Tb	POL	GODIVA then NEODC? BODC?	
Climatologies of wind speed, sea surface temperature and other parameters (interim products, probably not worth archiving) ; Climatologies of gas transfer velocity ; Climatologies of CO2 flux ;	1-2 Gb	NOCS	NEODC	Within next 6 months
Gridded altimeter data	1-2 Gb	NOCS	NEODC?	Within next 6 months
pCO2 measurements (still to be taken) – sensors funded but not yet processing, archiving etc		PML	NEODC?	
Processed SAR/colour imagery		Bangor		
Retrieved CO2 vertical columns from SCIAMACHY	10 Gb	Leicester	NEODC?	

Deleted: & NOCS