Abstract:

This document describes the data policy being developed by IS-ENES and reviews the software which is being deployed to create a distributed archive to manage Earth System Model data in Europe. The Earth System does not respect political boundaries, and the key scientific collaborations in this area also extend far beyond Europe's borders. The delivery a world class data distribution network can only be achieved through an architecture which supports global integration. In the context of climate models, the driving force for global integration comes from the Coupled Model Intercomparison Project phase 5 (CMIP5). IS-ENES is providing the European component of the distributed archive for CMIP5.

Grant Agreement Number:	**228203**	**Proposal Number:**	**FP7-INFRA-2008-1.1.2.21**		
Project Acronym:	**IS-ENES**				
Project Co-ordinator:	**Dr. Sylvie JOUSSAUME**				
Document Title:	Data policy and software review			**Deliverable:**	D 10.2
Document Id N°:		**Version:**	2	**Date:**	29/9/2010
Status:					
Filename:	ISENES_Deliverable_10_2_final.pdf				
Project Classification:	Public				

Approval Status		
Document Manager	**Verification Authority**	**Project Approval**
Martin Juckes		

Revision Table

Version	Date	Modified Pages	Modified Sections	Comments
0.1	25/05/10	2	2	First draft, from M10.2 documents
1	30/06/10			Version 1, submitted for review
2	31/08/10			Version 2, revised after review by Giovanni Aloisio and Christian Pagé

1 Table of Contents

2 Executive Summary

This document describes the data policy being developed by IS-ENES and reviews the software which is being deployed to create a distributed archive to manage Earth System Model data in Europe. The Earth System does not respect political boundaries, and the key scientific collaborations in this area also extend far beyond Europe's borders. The delivery a world class data distribution network can only be achieved through an architecture which supports global integration. In the context of climate models, the driving force for global integration comes from the Coupled Model Inter-comparison Project phase 5 (CMIP5). IS-ENES is providing the European component of the distributed archive for CMIP5. The global collaboration to deliver the CMIP5 data archive is provided by the Earth System Grid Federation.

The objective of the policy is to enable climate data to be evaluated while protecting the interests of the data providers and ensuring that data users are fully aware of the status of the data. Many users will want to gain access as soon as possible, but it is equally important to ensure that users who want a stable and enduring reference dataset are not tempted to use the archive before it has achieved this status. These conflicting requirements are resolved through the definition of a multi-level quality control process. The lowest level allows for rapid dissemination of the data and the highest level provides users with an assurance of high quality data.

The software review provides an overview of the software technology that will be deployed to support the archive. These range from tools associated with reading, checking and manipulating the scientific data and the associated in-file meta-data to web applications to provide users with a flexible and transparent interface to the archive. The archive must support expert users who want to download thousands of files (without thousands of mouse clicks) as well as novice users (perhaps students) who want a quick look at a small number of fields and also need guidance on the different options available.

3 DATA POLICY

3.1 Background

3.1.1 The Earth System Grid federation for CMIP5

The CMIP5 archive is being delivered by a global partnership designated as “The Earth System Grid federation (ESGF)”. ESGF is a non-profit organization formed by participates in the GO-ESSP collaboration to bring their knowledge and experience to bear on critical Earth System federations in the dissemination of climate data and related products. ESGF will be responsible for delivering the CMIP5 data archive. IS-ENES is coordinating the European contribution to ESGF. Worldwide, more than 30 institutions are expected to participate.

3.1.2 Quality control levels

ESGF aims to provide early access to the climate projections in order to facilitate community evaluation of the archive. Early release necessarily implies that there is little opportunity for scientific evaluation of the data prior to release. In order to provide for those who wish to access data which has evaluated by climate scientists, a three step quality control procedure is being deployed. The lowest level allows for rapid publication of data with sufficient quality control to guarantee traceability of the data. Passing the final stage of quality control will trigger the issuing of a permanent Digital Object Identifier (DOI), after which point the data will be held indefinitely.

3.1.3 Beyond CMIP5

The project should attempt to determine how the experience from CMIP5 can be turned into a sustainable European capability. Policy on future development will be guided by experience gained during the deployment of the CMIP5 archive. In the CMIP5 experiment, the compliance with the Data Reference Syntax and NetCDF CF protocols is verified by ad hoc software stacks which have these protocols programmed into them: it would be desirable to express the protocols in a more generic form which might allow some separation between the expression of the constraint and the programming language used to test compliance.

3.2 Terms of use

Users registering for access will be required to agree to one of th following “terms of use” statements:

“Unrestricted use” (access to a subset of models)

I understand that the subset of CMIP5 model output that will be

made accessible to me has been designated for "unrestricted" use. I agree in good faith to attempt to understand the limitations of the models used in producing this data. I understand that although the model output has been subjected to a quality control procedure, unrecognized errors remain. I will hold no one responsible for any errors in the models or in their output.

"Non-commercial research and educational purposes" (access to all models)

_I agree to use the CMIP5 model output only for non-commercial research and educational purposes. I agree in good faith to attempt to understand the limitations of the models used in producing this data. I understand that although the model output has been subjected to a quality control procedure, unrecognized errors remain. I will hold no one responsible for any errors in the models or in their output.

”

3.3 Quality control checks

3.3.1 Basic ESGF Conformance Checks (QC L1)

The ingestion process for the CMIP5 archive will use Earth System Grid software, and follow the protocol developed by PCMDI for the replication of data. Data will only be accepted into the archive once it has passed the checks run by the CMOR2 system, ensuring that the metadata in the file complies with the NetCDF CF and Data Reference Syntax protocols.

3.3.2 WDCC Conformance Checks and Subjective Quality Control (QC L2)

Based on the experience of the WDC Climate (WDCC) with IPCC AR4 data, the following data quality checks are currently suggested for the CMIP5/AR5 core data, which fulfill most of the testing properties for the STD-DOI data publication review process (fig. 3).

a) File consistency

- a file exists for each variable for the prescribed time step(s) (e.g. 6hourly, daily, monthly);
- files are not empty and in the end will have the right number of records;
- the layout of each file is consistent to the model design (gridding, filling values);
- strictly regular time steps;
- time bounds are consistent to the time interval specified in the file name;
- no overlap of consecutive time bounds.

b) Data base properties

- each entry in the data base has a counter part in the file system (and vice versa);
- specifications in the meta data of the data base correspond exactly to the layout of the files.

c) Physical properties of variables

- minimum and maximum are checked against specified ranges (default for each grid cell: the magnitude of the current weighted global mean plus twice the standard deviation is smaller than a prescribed threshold (10 to the power of 5), where current weighted global mean is the value from the beginning to the current time step.
- time series are calculated for:
    - min
    - max
    - globally weighted mean

- area weighted mean (reasonable, e.g., for temperature of snow)
- global arithmetic mean
- standard deviation of the globally weighted mean.

A consideration of the CMIP5/AR5 related work and required time on DKRZ's infrastructure has been accomplished. Based on the observed times on a desktop PC, the times required on the HPC IBM Power6 ere estimated, conservatively:

Desktop PC: 50 min per atomic dataset (6hourly interval storage)
IBM Power6 – 1 node (ca. 100 times the performance of a Desktop PC):
0.5 min per atomic dataset (6hourly interval storage),
500 days for all 1.5 Mio. atomic datasets.

The WDCC Conformance checks are completed by subjective quality controls of data and metadata.

3.3.3 DOI Data Publication Process (QC L3)

The results of the quality checks of level 1 and 2 are directly used as testing criteria for the STD-DOI data publication review process of the WDCC (fig. 4). For STD-DOI data publication the data review process is finalized by:

1) Double checks of QC L1 and QC L2 based on log files; discussion and clarification with corresponding data author if necessary.
2) Creation of STD-DOI metadata and assignment of persistent identifiers (DOI / URN) for each experiment / simulation.
3) Data author approval to freeze the data entity in its present version; and update the quality flag to "approved by author".
4) Integration of STD-DOI metadata and persistent identifiers for the frozen version of the data entity into the TIBORDER library catalogue (German National Library of Science and Technology, Hannover).
5) Notification of corresponding data author and ESGF about the finalization of the data publication process.

At the end of the STD-DOI publication process the data entity is accessible within the IPCC AR5 process (WG I – III) and within the wider research community. The STD-DOI data publication process is discussed in detail in a parallel document (Lautenschlager et al., 2010).

3.3.4 Beyond CMIP5

Standards established in the CMIP5 archive should be maintained and improved upon. IS-ENES should promote the establishment of peer reviewed data journals to formalise the quality control and establish an independent review process.

3.3.5 Technical issues

The data files should be confirmed as NetCDF CF compliant and CMOR2 compliant (the latter should imply the former, with additional constraints on required attributes).

3.3.6 Checksums

All files should be check-summed on ingestion into the archive. For data services which do not provide the whole file, a means of providing some check on the data would be useful. It might be possible to convert floating point data to byte data and then use a standard algorithm, but such an approach may produce different results which are dependent on the internal representation of real numbers. A less secure, but more robust, approach would be to provide simple means. E.g. averages over the spatial domain for each time and height level of the data.

3.4 Data management

3.4.1 Data publication

Publication of data occurs in different forms at the different levels of quality control. Level 1 of the quality control is performed at the ESG Data Node, and once this is passed, the data node can publish the data to a gateway, which automatically makes the data available through the gateway discovery services. At the second and third stages of quality control, publication occurs through an independent organisation hosting the DOI repository.

3.4.2 Security

The archive will support both web based access and scripted download through wget. The security architecture will be developed so as to allow access to all archive components through a single sign on procedure. More details are given below.

3.4.3 Removing or deprecating data

Prior to the issue of DOIs, the data will generally be removed if it is replaced. That is, if there are processing errors or other reasons to repeat any stage of the data production cycle, a new data version will be created and old data may be deleted. We do not have resources to store data which is replaced in this way. Once a DOI has been issued, on the other hand, we are committed to finding resources to preserve the data. The careful quality control procedure leading up to the granting of a DOI is designed to minimise the risk that data will have to be replaced once it has reached this stage (though it will, of course, eventually be replaced by more accurate projections in CMIP6 or its equivalent).

a Deprecation of data by owner

The most likely reason to deprecate data is the presence of more recent version of the data submitted by the owners of the original data. In this case, the catalogue entry must be modified to indicate the reasons for deprecation.

b Deprecation of data by others

Models contributing to CMIP5 have all undergone extensive testing and discussion. The archive will act to host the data and will not have a mechanism for judging or rejecting it, so long as it conforms to the file format specifications. If, after publication, data is found to be corrupted (e.g. constant fields) it will be the responsibility of the data owners to specify which files should be deprecated.

3.5 Notifying users of data errors

A procedure should be established for notifying users of data errors. Emails giving notification of errors will be sent to all users registered for access to the CMIP5 archive.

4 Software review

4.1 NetCDF metadata manipulation

4.1.1 NetCDF-4

The netCDF-4 library provides a significant extension to netCDF-3. However, it has been decided that the netCDF-4 library is not sufficiently reliable for use in CMIP5, bearing in mind the need for new software to be incorporated into operational systems at a large number of modeling centres.

4.1.2 CMOR2

The CMOR package is a vital part of the software stack of the data nodes. The package is designed to ensure that all the data files in the archive contain the required attributes in order that the files can be smoothly handled by the archive software. CMOR is developed and maintained at PCMDI. Documentation for CMOR2 is available: http://www2-pcmdi.llnl.gov/cmor/documentation/, and the source can be downloaded: http://www2-pcmdi.llnl.gov/cmor/download/
Requirements include NetCDF-4, udunits2 from UNIDATA (bundled with recent netCDF-4 releases) and uuid.

4.1.3 CDO

CDO is a collection of command line Operators to manipulate and analyse Climate model Data. Supported data formats are GRIB, NetCDF, SERVICE, EXTRA and IEG. With necessary additions a conversion from ascii or binary data sets to one of the supported data formats is possible. CDO also allows export of data sets for GrADS, GMT or as Text file. In order to support NetCDF-4 classic model files CDO must be compiled against the NetCDF-4 libraries.
There are some limitations for GRIB and NetCDF datasets. A GRIB dataset must be consistent with NetCDF. This means that all time steps must have the same variables, and within a time step each variable may occur only once. NetCDF datasets are supported only with 1 to 4 dimensional variables and the attributes should follow the GDT, COARDS or CF Conventions.
There are more than 400 operators available. There are operators for file information, regional or time selections, conditional selections and interpolation. CDO contain arithmetical functions and allows transformation between spectral and gridded data. Various statistical operators compute sums, means, minima or maxima for fields or time ranges.

4.1.4 CDAT

CDAT[1] is a large suite of open source tools for the management and analysis of climate data based around the Python programming language. It includes several visualisation components and the graphical user interface VCDAT.

[1] www2-pcmdi.llnl.gov/cdat

4.1.5 CDAT-lite

In situations where a full CDAT installation is unnecessary or undesirable CDAT-lite offers a lightweight installation option for CDAT’s core data management functionality. All software developed at BADC that requires CDAT features will be compatible with CDAT-lite.

4.1.6 CSMLSCAN

Csmlscan generates a “csml” document (an XML document) which is required by the BADC OGC services. It is being developed at BADC, and should be ready for testing at other data nodes late in 2010.

4.2 Access Control

4.2.1 OpenID and MyProxy

Single sign on is supported within the ESG federation so that a user can use a single set of credentials to access secured resources across the organisations making up the federation.
The ESG project has adopted OpenID as the system for browser based single sign-on and MyProxy for non-browser based clients such as wget. A MyProxy service may be deployed alongside an OpenID Provider service at Gateways. Clients may access this service to obtain a short lived certificate using their username/password credentials and use this credential as a token to authenticate when accessing data and services.
In order to access MyProxy, specialised client software is required. This may be difficult to install and configure for non-expert users. As an alternative, an intermediary service is being developed which simply fronts a MyProxy logon call with a HTTPS interface. A client such as wget sends a HTTP call to the service which itself fronts a MyProxy server translating the HTTP request into a MyProxy logon call and returning the user certificate back via the HTTPS interface.

4.2.2 Authorisation and Attribute Management Software

Role based access control is used to secure resources. User role (or *attribute*) information is pushed via OpenID’s Attribute Exchange (AX) mechanism. For the non-browser use case, MyProxy embeds attribute information in the user certificates it issues. In addition, attribute information may be *pulled* from an *Attribute Service*. The ESG Attribute Service uses a SAML based interface. Attribute Services may be hosted by a Gateway or any other organisation which has responsibility for registering users to access resources. Attribute information is queried from an Attribute Service based on a user’s OpenID URL. Attributes may be specific to a given organisation or globally recognised across the federation. PCMDI has a special status in this regard. It will host a Registration Service enabling users to register for a globally recognised *CMIP5* access attribute. Authorisation services deployed at organisations across the federation may define policies securing the resources they protecting using the CMIP5 attribute. When a given user requests access, PCMDI’s Attribute Service may be queried to find out whether the user is registered for this attribute and therefore has the required access privilege.

4.3 Federated archive

The federated archive will use and extend software components developed by the Earth system Grid team. ESG components are divided into Data Nodes and Gateways, the main features of which are described below. In this project ESG services will be extended and augmented using additional software described below

4.3.1 ESG Data Node

The ESG data node software package allows data stored at distributed nodes to be accessed through the ESG gateways (see below). The data node software requires multiple ports to be open. Where possible, the port numbers should be standardised to facilitate exchange of software.

Ports required:
HTTP: port 80
SSL: port 443
GridFTP: ports 2811, a range of ephemeral ports, typically 50000 – 51000

a THREDDS Data Server

Each data node will provide an HTTP interface for downloading and subsetting data via the THREDDS Data Server (TDS). Sub-setting will be supported via the OPeNDAP protocol. TDS is a Java servlet application typically installed inside Apache Tomcat.

b ESG publisher

The ESG publisher tool maintains an internal metadata database for all data on the node and publishes this metadata to an ESG gateway via THREDDS catalogues. In the IS-ENES data portal a more detailed metadata schema will be used: the METAFOR Common Information Model (CIM). Metadata maintained by the publisher metadata will used to generate CIM.

c GridFTP

GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks. It is based upon the Internet FTP protocol, and it implements extensions for high-performance operations that were either already specified in the FTP specification but not commonly implemented or that were proposed as extensions by the GLOBUS Alliance. The current GridFTP protocol specification is now a "proposed recommendation" document in the Global Grid Forum (GFD-R-P.020). Maintained by the GLOBUS Alliance. GridFTP is essential in order to deliver the data transfer speeds which will be needed when the data is being loaded into the archive. It is a mature and reliable software package, already in wide use in the climate data community.
Obtaining the high transfer rates required for the ingestion of data into the distributed archive requires careful attention to proper configuration. GridFTP will by default transfer one file at a time, which can be inefficient when datasets contain many small files.

d PyDAP

The data nodes should also support an OpenDAP[2] through pydap[3].

[2] Open-source Project for a Network Data Access Protocol (opendap.org)
[3] http://pydap.org/

4.3.2 ESG gateway

The gateway will include the ESG portal with search and discovery services. As it is not likely that this can be integrated into the vERC, IS-ENES will be developing separate search and discovery services. The ESG gateway will also include software to handle archive replication, and for this reason it will need to be deployed at DKRZ and BADC. These gateway deployments will also contribute to the testing and evaluation of ESG Data Nodes deployed within the IS-ENES federation.

4.3.3 COWS (CEDA OGC Web Services)

IS-ENES is committed to providing OGC services in order to broaden the accesibility of the data – these may be incorporated in the ESG Data Node or provided as an additional set of software. The services will be built on a Climate Science Modelling Language (CSML)[4] representation of the data. A Web Map Service and a Web Coverage Service will be implemented.

4.4 Portal

4.4.1 Plone/zope

The portal will be built within plone, a highly flexible content management system with an integrated web framework (zope). Plone/zope supports modular development and integration of externally developed services/functionality (plone/zope products). Thus portal services will be developed as independent modules embedded in plone[5].

4.4.2 User interface: javascript

The user interface will rely heavily on javascript to provide the flexibility to deal with complex choices. There is some commonality in objectives with METAFOR, in terms of the kinds of options being presented to users. Several JavaScript libraries are in common use within the METAFOR and IS-ENES JRA4/SA2 development groups. IS-ENES and METAFOR will attempt to converge on a common javascript approach.

a Jquery

Jquery[6] is recommended as the primary library for HTML DOM manipulation and AJAX request processing. Experience within the groups has shown it can work effectively with all other JavaScript libraries considered here.

b Prototype

Prototype[7] is an alternative to Jquery particularly suitable for legacy code.

c YUI (Yahoo)

YUI is a library of javascript functions "*for building richly interactive web applications*" (http://developer.yahoo.com/yui/). Although there are overlaps, its feature set is orthogonal to

[4] csml.badc.rl.ac.uk
[5] plone.org
[6] jquery.com
[7] www.prototypejs.org

Jquery or prototype as it concentrates on user interface widgets.

d Openlayers

Openlayers[8] is a javascript library for presenting interactive maps. It is the basis of the COWS visualisation application developed at the BADC that will be used in JRA4 services.

4.4.3 5.3 Visualisation

a Web Map Service (WMS), Web Coverage Service (WCS) and Web Processing Service (WPS)

The services will be based on the COWS service stack developed at BADC, using CDAT-lite for data access and manipulation, and matplotlib (see below) for creation of images.

b matplotlib

Imaging will use the python matplotlib package, a python package emulating the functionality of Matlab[9].

c WMS, WCS, WFS Clients

The WMS client which allows the user to interact with a WMS in a browser will be primarily in javascript, with some server-side python helpers which interact with the client through AJAX. The javascript components are discussed in section 5.2 above.

4.4.4 Controlled vocabulary server

To support consistent naming and search of entities, the maintenance and integration of controlled vocabularies into the portal is necessary. A concrete example of such an IS-ENES vocabulary is an extension of the vocabulary defined for the CMIP5 data reference syntax.

a ISOcat

The ISOcat (http://www.isocat.org) controlled vocabulary server and management system is being evaluated for interfacing with the portal.

b BODC vocabulary server

METAFOR is going to use the BODC vocabulary server for access to CIM controlled vocabularies.

4.5 Monitoring

System usage should be monitored in order to provide details on usage, to ensure continuing interoperability of distributed components, and in order to identify priorities for further development.

4.5.1 Web services

BADC runs a service monitor which checks the responsiveness of a list of URLs and notifies

[8] openlayers.org
[9] www.mathworks.co.uk

responsible parties by email if there is an interruption.

4.5.2 Archive operations

BADC: Processes which are run automatically for routine maintenance report outcomes by email, and Outlook is used to produce a summary overview.

4.5.3 System level

At the operating system level, machine performance statistics need to be monitored.

4.5.4 Energy usage

Green500[10] lists high-performance computers by efficiency, peaking at 536MFLOPS/watt in June 2009. Specpower[11] lists server side Java operations per watt, a measure seen by some as more relevant to single server operation. Their 2009 survey shows variation between 200 and 2000 ssj_ops/watt. The EPA has started a process of establishing a rating system, and Hewlett Packard launched the first servers to qualify for the EPA Energy Star rating in June 2009. These servers allow users to monitor energy usage. General information is provided in a recent ECRIM review[12], but specific options for immediate implementation are not available.

4.6 Virtual servers

Virtual servers allow greater flexibility in keeping multiple services running and in creating replications of installations to parallelise the service. BADC is using XEN[13] for virtualisation. DKRZ will use XEN on new hardware which will be available starting 2010, in parallel vmware[14] is used in the development/test environment.

4.7 Grid software

Grid Superscalar and GRelC software will be developed and deployed as described in the project delivery plan. A GRelC instance has been deployed in Lecce at CMCC for testing purposes and it is used to test access in grid environments to XML-based documents. This task aims at providing a grid-enabled access to CIM based documents.

4.8 Metadata

4.8.1 CMIP5 questionnaire

Inside Metafor WP4 BADC developed a questionnaire to capture model, software and activity metadata from climate models (please refer to: http://cmip5.metafor.ceda.ac.uk/cmip5/). The main goal is to capture these metadata to create the metadata base for the calculations for the CMIP5 project inside the next IPCC assessment report (AR5). It is planned to make Metafor services (like the metadata catalogue search service) accessible through the data portal as soon as stable versions are ready

[10] http://www.green500.org/lists/2009/06/list.php
[11] http://www.spec.org/power_ssj2008/results/power_ssj2008.html
[12] http://ercim-news.ercim.org/en79
[13] www.xenserver5.com
[14] www.VMware.com

4.8.2 GeoNetwork CIM XML review

GeoNetwork is a human GUI based on Ajax technology to create and search XML records. At MPI-M and DKRZ (please refer to: http://anticyclone.dkrz.de:8088/geonetwork/) a version of GeoNetwork is developed and implemented to review and update Metafor intermediate CIM XML records.

5 Annex A. Relevant existing data policies

5.1 PCMDI CMIP3 Terms of use agreement[15]

These data are for use in research projects only. A 'research project' is any project carried out by an individual or organized by a university, a scientific institute, or similar organization(private or public) for non-commercial research purposes only. Results based on these data must be submitted for publication in the open literature without any delay linked to commercial objectives.
An additional restriction is that the UK Met. Office / Hadley Centre requires strict adherence to the conditions set forth in their "License Statement", which applies to their model output:
"These data are licensed for use in Research Projects only. A 'Research Project' is any project organised by a university, a scientific institute, or similar organisation (private or public), for non-commercial research purposes only. A necessary condition of the recognition of non-commercial purposes is that all the results obtained are openly available at delivery costs only, without any delay linked to commercial objectives, and that the research itself is submitted for open publication.
Data provided by the UK Met. Office / Hadley Centre are expected to be acknowledged by:
(c) Crown copyright 2005, Data provided by the Met Office Hadley Centre"

5.2 BADC disclaimer[16]

Terms and Conditions for data and information provided by the NERC/BADC

5.2.1 Exclusion of Liability

Your use of information provided by NERC is at your own risk. Please read any warnings given about the limitations of the information.
NERC gives no warranty as to the quality or accuracy of the information or its suitability for any use. All implied conditions relating to the quality or suitability of the information, and all liabilities arising from the supply of the information (including any liability arising in negligence) are excluded to the fullest extent permitted by law.
NERC gives no warranty as to the accuracy or completeness of data or images in the form in which they are cached or downloaded to your computer, as they may be affected by on-line conditions over which NERC has no control.

5.2.2 Notes on Limitations

* Scientific observations are made according to the prevailing understanding of the subject at the time. The quality of such observations may be affected by subsequent advances in knowledge, improved methods of interpretation, and better access to sampling locations.
* Raw data may have been transcribed from analogue to digital format, or may have been

[15] http://www-pcmdi.llnl.gov/ipcc/info_for_analysts.php#Terms_of_use
[16] http://badc.nerc.ac.uk/conditions/badc_anon.html

acquired by means of automated measuring techniques. Although such processes are subjected to quality control to ensure reliability where possible, some raw data may have been processed without human intervention and may in consequence contain undetected errors.

* Detail clearly defined and accurately depicted on large-scale maps may be lost when small-scale maps are derived from them.

* Although samples and records are maintained with all reasonable care, there may be some deterioration in the long term.

* The most appropriate techniques for copying original records are used, but there may be some loss of detail and dimensional distortion when such records are copied.

* Data may be compiled from the disparate sources of information at NERC's disposal, including material donated to NERC by third parties, and may not have been subject to any verification or other quality control process.

* Data, information and related records which have been donated to NERC have been produced for a specific purpose, and that may affect the type and completeness of the data recorded and any interpretation. The nature and purpose of data collection, and the age of the resultant material may render it unsuitable for certain applications/uses. You must verify the suitability of the material for your intended usage.

* The data, information and related records supplied by NERC should not be taken as a substitute for specialist interpretations, professional advice and/or detailed site investigations. You must seek professional advice before making technical interpretations on the basis of the materials provided.

* If a report or other output is produced for you on the basis of data you have provided to NERC, or your own data input into a NERC system, please do not rely on it as a source of information about other areas or features.

5.3 IPCC Data Distribution Centre [BADC][17]

5.3.1 Proper use of data

Access to this data is free and unrestricted. If the data is to be used for any publication we encourage registration, so that you may be notified of any changes. When the data is used in publications, proper credit should be given to the data producers.

5.3.2 Privacy

We collect no personal information about you when you visit this Web site, unless otherwise stated or unless you choose to provide this information to us. However, we collect and store certain information automatically. What we collect and store automatically is:

* The Internet Protocol (IP) address of the domain from which you access the Internet (e.g., 123.456.789.012) whether yours individually or provided as a proxy by your Internet Service Provider (ISP)

* The date and time you access our site

* The pages you peruse (recorded by the text and graphics files that compose that page)

* And, the Internet address of the Web site from which you linked directly to our site.

We use the summary statistics to help us make our site more useful to visitors, such as assessing what information is of most and least interest to visitors, and for other purposes such as determining the site's technical design specifications and identifying system performance or problem areas. This

[17] http://www.ipcc-data.org/ddc_anon-download_terms.html

information is not shared with anyone beyond the support staff to this Web site, except when required by law enforcement investigation, and is used only as a source of anonymous statistical information.

5.4 IPCC Data Distribution Centre [DKRZ][18]

5.4.1 License Statement

These data are licensed for use in Research Projects only. A 'Research Project' is any project organised by a university, a scientific institute, or similar organisation (private or public), for non-commercial research purposes only. A necessary condition of the recognition of non-commercial purposes is that all the results obtained are openly available at delivery costs only, without any delay linked to commercial objectives, and that the research itself is submitted for open publication.
Data provided by the UK Met. Office/Hadley Centre are expected to be acknowledged by :
(c) Crown copyright 2005, Data provided by the Met Office Hadley Centre

5.5 ENSEMBLES Data Policy

5.5.1 Terms and conditions of use

- ENSEMBLES data held in the main ENSEMBLES data centres are made available over the internet without charge for use in research, education and commercial work.
- Users of ENSEMBLES data must publish their results based on using these data in open literature without any delay linked to other, e.g. commercial, objectives.
- Users must submit a copy of their results based on these data to the data centre from which the data were obtained and to the partners who produced the data.
- No redistribution of ENSEMBLES data for commercial re-use or reselling, however processed or derived, by any party who receives data from any ENSEMBLES participant is allowed.
- Data must not be supplied as a whole or in part to any third party without the prior authorisation of the data manager of the data centre from which the data was obtained.
- Users should help improve the quality of the data and its delivery by giving feedback where appropriate.
- All data use, however small, derived or embedded, must be acknowledged, as in Section 2 below.

5.5.2 Acknowledgement

- Articles, papers, or written scientific works of any form, based in whole or in part on ENSEMBLES data, will include the following acknowledgement: “The ENSEMBLES data used in this work was funded by the EU FP6 Integrated Project ENSEMBLES (Contract number 505539) whose support is gratefully acknowledged.”
- Subsequent references can refer to the data in terms such as “ENSEMBLES data”, “the ENSEMBLES dataset” or by the data’s generic name which is from the “ENSEMBLES data archive”.

5.5.3 Registration and institutional directives

- Some types of data require registration before downloading. Registration means you can be kept informed of future updates. Registration details and process will be specified before the point of

[18] http://www.mad.zmaw.de/IPCC_DDC/html/SRES_AR4/index.html

data access in an appropriate way.
• This data policy does not replace the data policy of the institute serving the data which always takes precedence, but should be viewed in conjunction with it. ENSEMBLES data and its availability are in accordance with Directive 2003/98/EC* of the European Parliament and of the Council on the re-use of public sector information (the PSI Directive).

5.5.4 Data limitations

• Although every care has been taken in preparing and testing the data, ENSEMBLES cannot guarantee that the data are correct in all circumstances; neither does ENSEMBLES accept any liability whatsoever for any error or omission in the data, its availability, or for any or damage arising from its use.

- Full text available at: http://europa.eu.int/eurlex/pri/en/oj/dat/2003/l_345/l_34520031231en00900096.pdf

6 ANNEX B. Overview of software in use

Table 1 Overview of software. In the last column, "developer" is interpreted broadly to include those leading on deployment and evaluation issues.

Package	Purpose	Providing organisation and reference
CDAT Climate Data Analysis Tools	Python packages to support processing of climate data.	PCMDI, http://www2-pcmdi.llnl.gov/cdat/
CDAT-lite	A reduced version of CDAT to facilitate integration with other packages while retaining core functionality.	BADC, Stephen Pascoe, http://proj.badc.rl.ac.uk/ndg/wiki/CdatLite
CDO Climate Data Operators	A collection of command line Operators to manipulate and analyse Climate model Data	MPI, http://www.mpimet.mpg.de/fileadmin/software/cdo/
Climate Model Output Rewriter (CMOR)	Creation of standardised netcdf files for the CMIP5 archive	PCMDI, Karl Taylor, http://www2-pcmdi.llnl.gov/cmor/
GridFTP	GridFTP is a high-performance,	GLOBUS Alliance http://www.globus.org/grid_software/data/gri

	secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks.	dftp.php
OpenID	Browser based single sign-on.	OpenID foundation http://openid.net/ Java (ESG – Luca Cinquini, NCAR and Neill Miller, ANL) and Python based (NDG Security package - Philip Kershaw, BADC) implementations are available for ESG.
MyProxy	Single-sign on for non-browser based access.	Globus Alliance and a number of independent OpenSource client packages are also available see: http://grid.ncsa.illinois.edu/myproxy/download.html
ESG Attribute Service	Provides user attribute information for access control	Java and Python based implementations are available. Java, as part of the ESG software stack (Luca Cinquini, NCAR); Python, as part of the NDG Security package (Philip Kershaw, BADC)
GRelC	grid access interface to data sources	Grid service GSI-enabled to access to heterogeneous data sources in a grid environment. (Giovanni Aloisio, CMCC)