DDC Assessment Criteria for Linking External Datasets

Document ownership and history		
Owner	TGICA	
Location	www.ipcc-data.org/documents/ddcLinkingCriteriaV1.pdf	
Author team:	Martin Juckes, Bob Chen, Bryan Lawrence Michael Lautenschlager	
Version	1.0	
Date	February 8th, 2012	
Version history		
Date	**Version**	**Comment**
March 30th, 2011	Draft	Submitted to TGICA, and discussed during April 13th telephone conference.
May 26th, 2011	0.5	Circulated to TGICA mailing list.
February 2nd, 2012	0.9	Revised and circulated for discussion at TGICA 17.
February 8th, 2012	0.95	Revisions after discussion in meeting, introductory text contributed by Volker
February 8th, 2012	1.0	Approved.

1   Introduction

The IPCC Data Distribution Centre (DDC) provides a repository for data from IPCC sources, quality controlled and carefully vetted, and operating within the IPCC mandate. In addition to hosting datasets, it also provides pointers to other centres and groups for data beyond its current holdings that may be useful for impacts, adaptation, vulnerability and mitigation assessments. This document outlines the criteria for assessing the eligibility of such external datasets for being linked to from the IPCC Data Distribution Centre web site. A diverse set of criteria is relevant for the decision to link to a specific dataset or not, including relevance of the dataset to IPCC assessments, availability of appropriate documentation, quality control, institutional stability of the dataset and provider, and accessibility.

A request to link a specific dataset will be handled by the DDC Work Group and DDC staff. The final decision about whether to link to a specific dataset is made by the full membership of the TGICA.

2 Evaluation procedure

2.1 Linking of new datasets

In case of linking to a dataset that the DDC did not point to previously, a three stage procedure will be carried out, the details of which are explained in the following sections:

1. The procedure is initiated by completion of a questionnaire (Section 3), which can be carried out by any interested party (e.g., owners of a data archive who would like to have it linked to, members of TGICA wanting to improve the DDC by adding a link, or members of the public who feel that an additional link is needed). The completed questionnaire should be passed to the DDC by emailing to martin.juckes@stfc.ac.uk.
2. A structured assessment of the questionnaire is carried out by the DDC Work Group and DDC staff, using the criteria set out in Section 4 below.
3. The evaluation is then passed on to the full TGICA membership. TGICA will review the evaluation; once the evaluation is completed (either at a full meeting or through a telephone conference organised between meetings) the decision follows automatically from the decision process as laid out in Section 5 (see also decision tree in Figure 1).

After completion of the process, full details will be passed to the provider and a record of the decision process for successful datasets will be posted on the DDC.

2.2 Review of datasets linked to by the DDC

Once datasets have been linked to, there will be a regular review of the basis on which the links were added by the DDC. Where there has been significant change from the information recorded in the questionnaire, it will be completed with the revised information and passed to the DDC work group for re-evaluation, with the final decision by TGICA as above. If a dataset no longer meets the criteria, the link will either be removed or marked with a warning.

3 Questionnaire

No.	Question	Response	Level or Supplementary Information	Documentation
1	**Description**			
1.1	**Author**			
1.1a	Intellectual Property Right holder		Provide name and address	
1.2	**Dataset**			
1.2a	Title		Provide a long (up to 120 characters) and short version (less than 32 characters) of the title.	

No.	Question	Response	Level or Supplementary Information	Documentation
1.2b	Description		Provide 200-500 word description, highlighting the key distinguishing features of the dataset. This text will be used in the DDC entry.	
2	**Authority**			
2.1	**Use by the IPCC**			
2.1a	Has the dataset been used in an IPCC Assessment or Special Report, e.g., in a figure or table or discussed in text?	Yes/no/not known	1=minimal, 2=one of many datasets, 3=primary source	List report name(s), table/figure/page number(s)
2.1b	Is dataset used in any other IPCC-related documents or materials?	Yes/no/not known	1=minimal, 2=one of many datasets, 3=primary source	List document name(s), table/figure/page number(s)
2.2	**Documentation**			
2.2a	Is the dataset documented in detail in a peer-reviewed journal article or as a peer-reviewed dataset?	Yes/no/not known	List Thomson-Reuter impact factor (or other standard influence factors) for journal if available.	Provide full citation
2.2b	Is the dataset documented in detail in a peer-reviewed book chapter, report or technical document?	Yes/no/not known	List type of peer review if known; list type of book, report, or document, e.g., if part of a series; provide evidence of credibility of authors or publisher (e.g., UN organization)	Provide full citation
2.2c	Is the dataset documented in detail in a non-peer reviewed document, web site, or other resource?	Yes/no/not known	Describe quality control or other review processes used (e.g., crowd sourcing); provide evidence of credibility of authors or publisher (e.g., UN organization)	Provide citation or links
2.2d	Has there been significant discussion of the dataset in the scientific literature?	Yes/no/not known	Were significant questions about the scientific quality of the data raised?	Provide citations to criticisms and responses where relevant.
2.2e	Are the uncertainties associated with the data documented.	Yes/No	What sort of uncertainty information is provided?	Provide link(s) or reference(s).

No.	Question	Response	Level or Supplementary Information	Documentation
2.3	**Lineage**			
2.3a	Is the dataset produced by or under the direction of a national or international body or group?	Yes/no/not known	List scientific body or group	Provide link
3	**Significance Relative to the IPCC Community**			
3.1	**Interest in the data**			
3.1a	Has an IPCC Working Group or the TFI used or expressed an intention of using this data?	Yes/not known	Indicate which group(s) and/or other body.	Provide reference and quote relative passage.
3.1b	Have DDC users expressed interest in these data?	Yes/no/not known	Indicate numbers and/or types of users	Provide user metrics or examples of queries
3.1c	Are there strong reasons for considering the data relevant to the DDC user community?	Yes/no	Give reasons, backed by references.	References from gray literature should have accompanying justification, as for IPCC reports.
3.2	**Uniqueness**			
3.2	Are other datasets available with the same or overlapping variables?	Yes/no/not known	Indicate other datasets and degree of overlap	Give links
4	**Stability of Data and Data Provider**			
4.1	**Curation**			
4.1a	Does the provider have a published data policy?	Yes/no/not known		Give link
4.1b	Does the provider have a succession plan for this dataset?	Yes/no/not known	List organization(s) with long-term responsibility for the dataset (e.g., government agency, library, archive)	Give link to plan or other documentation of agreement
4.1c	Is there an explicit funding model for the dataset or data provider?	Yes/no/not known	Indicate type (e.g., subscription-based, government-supported, submitter fees)	Give link

No.	Question	Response	Level or Supplementary Information	Documentation
5	**Quality Control**			
5.1	**Meta-data and quality control**			
5.1a	Does detailed meta-data exist for this data, in accordance with relevant national or international meta-data standards	Yes/no/not known	Indicate relevant standards (e.g. NASA's Directory Interchange Format).	Link to metadata
5.1b	Is there a stated quality assurance process or procedure for the dataset?	Yes/no/not known	Indicate if relevant quality assurance standards are met (e.g., ISO9000)	Give links
5.1c	Is there a regular validation or calibration process or procedure for the data?	Yes/no/not known	Indicate frequency and/or most recent date	Give links
5.1d	Is technical/usage guidance available?	Yes/no/not known	1=online, 2=offline	Give links
5.2	**Data updates and version control**			
5.2a	Is the data subject to updates?	Yes/no/not known	List frequency	Give link
5.2b	Is there a clear version control process and tracking of data provenance?	Yes/no/not known	Describe version control and data provenance procedures	Give links
5.2c	Are previous versions of the dataset accessible?	Yes/no/not known	List previous versions	Give links
6	**Accessibility**			
6.1	**Access requirements**			
6.1a	Are the data available on-line for download?	Yes/no	1=all data, 2=subsets only	Give links
6.1b	Are data made available in one or more standard (preferably open) formats?	Yes/no	List number of formats available (at least one standard format is required);	Give links

No.	Question	Response	Level or Supplementary Information	Documentation
6.1c	Are there restrictions (beyond user registration and acceptance of terms of use) on data use, re-dissemination, or reuse?	Yes/no/not known	1=no such restrictions, 2=non-commercial use only, 3=custom restrictions	Give links
6.1d	Is the data available for free or for a charge?	Yes/no/not known	1=free, 2=charges	Give link to price structure
6.2	**Additional information**			
6.2a	Is user registration required or requested?	Yes/no/not known	1=required, 2=requested	Give links
6.2b	Is attribution required or requested?	Yes/no/not known	1=required, 2=requested	Give links
6.2c	Are versions of the data available through open interfaces (e.g., OGC Web Services, REST, SOAP)?	Yes/no/not known	List interface specifications	Give links
6.2d	Are levels of service (e.g., bandwidth, up time) adequate?	Yes/no/not known	List issues	Give links, e.g., to up time metrics
6.2e	Is user support available (e.g., help desk, frequently asked questions)?	Yes/no/not known	List user support services	Give links
6.2f	Is documentation available in English?	Yes/no/not known		Give links
6.2g	Is documentation available in other languages	Yes/no/not known	List languages	Give links
6.2h	What is the spatial domain of the dataset		Specify bounding latitudes and longitudes	

Table 1: Questions to gather information for the decision.

4 <u>Assessment</u>

This assessment should be filled in by the DMWG. Where value judgements are required these should be backed with the level of justification typical of the academic peer review process.

Stage	Issue	Criteria	Explanation	Decisions

A	1	(for information only – clarity required)	These entries are for information only – the text will be used in the DDC entry and should be expressed with sufficient clarity for that purpose.	OK/needs revision
	6.2	(for information only)	Collected for information only – this information will be reflected in the DDC entry.	
B	2.1	Use by IPCC (decision relevant)	Datasets which score 2 or 3 in 2.1a or 3 in 2.1b will be scored high. Those scoring lower will only be included if they have a clean sheet on critical questions.	High/low
C	2.2	Document-ation (critical)	If clear documentation in peer-review literature is not available, acceptance will be subject to TGICA assessment of the reliability of on-line documentation. Documentation must be available in a stable form. If peer-review literature indicates there are significant problems with the data, the data will not be accepted until suitable responses are available in the peer-review literature.	Good/poor/bad
	4.1	Curation (critical)	In the absence of a published data policy, written assurance of the intention to preserve the data should be obtained. Data for which there is no curation policy should either be copied to the DDC or only linked to after a copy has been taken for curation.	
	5.1	Metadata (critical)	The meta-data must be sufficient to provide clear provenance information. DDC managers should review the information described in this section and provide a recommendation to TGICA.	
	5.2	Version control (critical)	If data is subject to updates, clear version control is essential.	

	6.1	Access (critical)	Data must be available in a standard format. DDC managers should advise TGICA. Data which is not available free of charge will be clearly marked as such on the DDC, and preference will be given to free alternatives where available.	
D	3.1	Interest (decision relevant)	If an IPCC group has not expressed interest, TGICA should judge the significance of the interest expressed by other groups. This criterion becomes critical if (2.1) is not satisfied.	High/Moderate/Low
E	3.2	Uniqueness (secondary)	In the case of expression of interest only by non-IPCC groups, the degree or uniqueness will be taken into account.	High/low
F	2.3	Lineage (decision relevant)	Datasets from authoritative international agencies will be scored high.	High/low

Table 2: Assessment criteria.

5 **Decision**

Given the decisions in above table 2, the final decision is determined by the process outlined in figure 1. The key decision points are listed below:

- A) Description and ancillary information: if not adequate, seek revision; if appropriate, look ahead to see if a revision is likely to lead to acceptance;
- B) Use by IPCC: determined by assessment criteria AC 2.1;
- C) DCMVA (Documentation, Curation, Metadata, Version control, and Access): this decision is "Good" if AC 2.2, 4.1, 5.1, 5.2 and 6.1 are all good; "Bad" if there are serious deficiencies, "Poor" if there are problems which can be fixed;
- D) Interest expressed: determined by AC 3.1;
- E) Uniqueness: determined by AC 3.2;
- F) Lineage: determined by AC 2.3;
  The possible outcomes:
- G) Publish;
- H) Reject;
- I) Refer users to alternative data: where availability of better data is a reason for not linking to a particular dataset, the DDC should consider linking to the better data;
- J) If the interest level in a dataset proposed for linking is low, the DDC will not invest effort in fixing any problems, but feedback in the form of the evaluation will be provided to the dataset providers.
- K) DDC looks for solution: If the "DCMVA" material is not OK, in some circumstances the

DDC staff will seek to remedy the deficiencies by, for instance, collecting additional meta-data;

Figure 1: Decision tree.