ARC Centre of Excellence for Climate System Science: Publishing Template - Basic DMP

Data Collection

What data will you collect or create?

Guidance:

Questions to consider:

- What type, format and volume of data?

Overview:

Give a brief description of the data, in each case noting its content, type and coverage. Outline your choice of format and give an estimate of data volumes.

Example Answer:

Raw output: output from 2 simulations of the ACCESS model, the output will be binary UM format for the atmospheric component and netcdf for the other model components. The estimated size is 5 Tb for each simulation. Processed output: The final data product will be the post-processing of the following fields: ... Files will have netcdf format and the estimated size is 100 Gb.

How will the data be collected or created?

Guidance:

Questions to consider:

- What standards or methodologies will you use?
- Are you going to use any input data?

Overview:

Outline how the data will be collected/created and which community data standards (if any) will be used. List any dataset you will be using as input.

Online resources:

For an overview of naming conventions, good data practice and metadata standards:

- Climate & Forecast (CF) metadata conventions ;
- other netcdf metadata conventions ;
- ANDS guide on file format ;
- ANDS guide on metadata ;
- Standford University file naming best practice ;

Storage and Backup

How will the data be stored and backed up during the research?

- NCI server: tape archive
- Institutional server: tape archive
- Other institutional archiving service
- Commercial archiving service
- External hard drive
- Other

Guidance:

Questions to consider:

- Do you have sufficient storage both for analysis and archiving?
- How will the data be backed up?
- Who will be responsible for backup and recovery?
- Are you storing enough information to be able to reproduce the data in case of loss?

Overview:

State how often the data will be backed up and to which locations. How many copies are being made? Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable. Similarly, it is normally better to use automatic backup services provided by IT Services than rely on manual processes.

If you are using the NCI facility, you can archive your data for long-term storage in massdata. NCI has four kinds of filesystems and each project/user is allocated some.

1. /home - Intended to be used for source code, executables and irreproducible data (input files etc), it is backed-up.
2. /short/<project-id>/<user-id> - this is a working area, where for example you run your model and the output is first created, it is not backed-up.
3. /g/data/<project-id> - fast disk to store persistent data. The ARCCSS has its own shared allocation under the ua8

project. If /short allocation is insufficient to analyse your data or you need to share data you can get a temporary allocation under this project by contacting climate_help@nf.nci.org.au or/and including details in the technical section of this plan.
4. massdata - tape system, for long-term storage.

Remember to document your data workflow as much as you can to maximise the reproducibility of your data in case of loss. While it might be difficult to back up all your data, it is easier to back up the code. Consider using git and/or subversion to keep track of versions and changes. More information on this is provided in the following section.

Online resources:

- NCI filesystems

Documentation and Metadata

What documentation and metadata will accompany the data?

Guidance:

Questions to consider:

- What information is needed for the data to be to be read and interpreted in the future?
- How will you capture / create this documentation and metadata?
- What metadata standards will you use and why?

Overview:

Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed.

Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded.

Keeping track of your code development using a form of version control, is both useful in ca se of data loss and as documentation of the data analysis process.If you are not familiar with a versioning control below are listed links to git and other version control software. Wherever possible you should identify and use existing community standards.

Online resources:

- Climate and weather community uses the CF metadata conventions.
- An overview of metadata and other guidance is available from the ANDS website.
- online resources for learning git and github ;
- mercurial tutorial ;

Data Sharing

How will you share the data with collaborators?

Guidance:

Questions to consider:

- With whom will you share the data, and under what conditions?
- For how long do you need exclusive use of the data and why?
- Will a data sharing agreement (or equivalent) be required?

Overview:

If working in a group add details of how will you share data with your collaborators, possibilities are: shared directories, managing access through permissions, setting up a VM on the cloud (i.e. a virtual environment that external collaborators can access too), using other collaborative tools.

You only need to give an overview here, you can add details in the technical phase of the dmp if necessary.

Online resources:

- NECTAR research cloud, virtual laboratories and other e-research resources

How will you make the data publicly available?

Guidance:

Questions to consider:

- How will potential users find out about your data?
- Will you share data via a repository, handle requests directly or use another mechanism?
- When will you make the data available?
- Will you pursue getting a persistent identifier for your data?
- Will there be any restrictions?

Overview:

If possible all the data output from your research project with acknowledged long term value should be made available to the wider community.

The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data.

Outline any expected difficulties and restrictions in sharing data, along with causes and possible measures to overcome these. Restrictions may be due to confidentiality, lack of consent agreements or IPR, for example. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.

If this dmp is aprt of a grant application, mention earlier examples if possible, to show a track record of effective data sharing.

A default answer covering the long term ARCCSS standard access policy is provided for projects that do not have specific requirements.

Online resources:

- ANDS Open Data overview
- Overview from opendatahandbook.org
- ANDS Research Data Australia
- ANDS guide on DOI

Legal Compliance and Ethics

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Guidance:

Questions to consider:

- Who owns the data?
- How will the data be licensed for reuse?
- Are there any restrictions on the reuse of third-party data?
- Will data sharing be postponed / restricted e.g. to publish or seek patents?

Overview:

In a standard PhD case your institution is the copyright and IPR owner of any data you produce during your PhD. You can find below a list of relevant policies.

For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. The ARCCSS data manager can help you consider any relevant funder, institutional, departmental or group policies on copyright or IPR if this apply to your project. Also consider permissions to reuse third-party data and any restrictions needed on data sharing.

It is good practice to apply a license and having a rights statement accompanying your data, even if the data has open access.

Online resources:

- ANDS guide on copyright

Dataset License

- CC BY Attribution
- CC BY-SA Attribution-ShareAlike
- CC BY-ND Attribution-NoDerivs
- CC BY-NC Attribution-NonCommercial
- CC BY-NC-SA Attribution-NonCommercial-ShareAlike
- CC BY-NC-ND Attribution-NonCommercial-NoDerivs
- Other

Guidance:

Overview:

Only one choice is allowed.

A license is the legal agreement which automatically applies to anyone who wants to use the data. Only the copyright owner, which is usually your institution can enforce a license, so you should make sure that the license you are using does not break any institutional requirement.

Unless your data has access limitations, privacy issues or more than one copyright owner, any of the proposed Creative Commons licenses should be fine for it. A default preferred ARCCSS choice is provided: CC BY-NC-SA.

Online resources:

- You can find more information and the legal code for all the Creative Commons licenses listed here on theCreative Commons website.
- More information on open-access licenses and their application in Australia is available fromAusGOAL and on the ANDS website.

Rights statement

Guidance:

Overview:

A right statement will tell data users what they are allowed to do with the data, which is particularly important if you are not applying a license to the data. If you are applying a license, a right statement is a plain words version of the license itself and it is not necessary to include it here. It should be listed as metadata in RDA/Geonetwork records, README files, in the actual files where it is possible. For example as global attribute in a netCDF file.

A default ARCCSS choice is given and reflects the terms of use in the license CC BY-NC-SA.

Online resources:

- **ANDS guide on rights**

ARC Centre of Excellence for Climate System Science: Publishing Template - Technical DMP

Data detailed description

Detailed description of your input data

Guidance:

Questions to consider:

- Will you need to download new datsets or are they all available locally?
- Is your input data regulated by a license and if yes can you use it the way you intend to?
- Is the input data in an easily usable format?
- Where will you store this data?
- If you are downloading a new dataset could it be useful to other users?

Overview:

You should provide information about datasets you will be using in your analysis or model simulation. Include any details such as: file format, version, size, license and most importantly availability.

You can find a list of dataset available on the NCI server raijin on their geonetwork catalogue.

More data might be available but not be advertised yet, if unsure, check with the ARCCSS data manager by e-mailing climate_help@nf.nci.org.au. It is also a good idea to contact the data manager before downloading a big dataset, if the data can be useful to more than one user, the data manager can provide help with the download and a disk allocation to store it.

Online resources:

- NCI geonetwork catalog
- NCI thredds catalog
- CMS wiki data page

Detailed description of your data output

Guidance:

Questions to consider:

- What will your output represent? Climatology, index, time series, simulation of a particular event or phenomena?
- What is the size and format of your data output, including before and after analysis?
- How is your data structured?

Overview:

You should describe here your data output in more details. Include specific information on what the data is and estimates of unprocessed and processed output. Include information on the data format used for the different kind of output.

If you can, describe also how did you structure your data, a directory structure or a filenaming convention you adopted.

Online resources:

Example Answer:

This project will have as final output a climatology of air temperature derived from ACCESS model simulations. We will run two simulations of the ACCESS1.3 model. Each simulation will produce an estimate of 5 Tb of raw model output. We will analyse air temperature ....{describe the output of your analysis here}. The final processed data output will have an estimated size of 1 Tb for both simulation. Based on the above estimates, this project will need 11 Tb of archive storage and 15 Tb of disk to analyse the data for the duration of the project. The project data output will be in netcdf, this format is self-describing and widely available and used in the Climate community.

Software and data workflow

Hardware and Software

Guidance:

Questions to consider:

- Is any hardware or software you are planning to use well supported and immediately available to you?
- If not, have you considered any alternatives?
- Is the software you are planning to use widely used in your community and/or institution?
- Are you going to use commercial software?
- Would the hardware you are planning to use allow you to share data and code with your collaborators?

Overview:

You should provide information about and the rationale for any hardware or software (including programming languages) which will be used to support your project.

If you're planning to use NCI resources, a link to a list of available software is provided below.

Please indicate if any of the software you require is not yet available to you on the computing facility you're planning to use and the motivation for using such software (i.e. lack of alternative, previous experience, availability of code suite).

You could consider using a cloud or another virtual environment to install any software that could potentially cause issues to your operating system. Snapshots of unix/linux type distribution with pre-installed software are available on the net.

Cloud environments are also good choices if you want to share a working environement with external collaborators.

Online resources:

- NCI software list
- CMS wiki page on NCI cloud
- Getting started with nectar cloud
- Virtualbox

Data Workflow

Guidance:

Questions to consider:

- How much disk space will you need to process your data and for how long?
- How are you going to add some quality control to your data?
- Have you done any testing of your code?

Overview:

You should provide information about the data processing, showing how you will get from the input data to the final product using the hardware and software described in the previous question. You must show that you have considered how you will achieve your final output in practice, including a timeframe estimate for the main workflow phases.

You should also include information on quality assurance, both for the data and the code you are producing to identify possible mistakes early on in the workflow process.

Online resources:

Technical Support and Relevant Experience

Models and/or datasets which you will be using.

- ACCESS
- UM
- MOM
- CABLE
- WRF
- CMIP5
- CMIP6
- ERA Interim
- YOTC
- CABLE data
- Other

Guidance:

Overview:

Include here the models and datasets which you will need assistance with. This helps us correctly resourcing our time and resources to provide extra help and tools where is most needed.

You can use the 'comment' box to specify 'other' non listed resources. If possible add a link to the resource and/or its documentation. Feel free to add as much detail as you like, as well as your reasons for using this resource. The more information, the easier it is for us to evaluate the inclusion of this resource in our supported tools.

Following is a list of the already available resources linked to some of the models and datasets we are already supporting.

Online Resources:

- ACCESS model
- UM model
- WRF model
- MOM model
- CABLE model
- CMIP5 data
- CMIP6 data
- ERA Interim data
- YOTC data
- CABLE data

Which training would be useful for you?

- Introduction to NCI facilities
- python

- NCI basic parallel programming
- Virtual Machines (cloud/virtualbox)
- NCI VDI remote desktop
- NCI CWSlab (includes CMIP5 pipeline workflow tool)
- Other (please specify)

Guidance:

Overview:

You can choose any training options you think could help you with your current project or just generally improving your skill set.

We listed some of the more common options and four of them are courses regularly run by NCI. There are a lot of available options for training: online courses, videos or tutorials, software carpentry courses, courses run by universities. The CMS is trying to increase its involvement in training adn we will use this information to organise new materials, as well as informing NCI and other partners of which training would be more valuable to our community.

Online resources:

- NCI training courses material
- Coursera offers a wide variety of online courses
- Software carpentry

ARC Centre of Excellence for Climate System Science: Publishing Template - Publishing DMP

Publishing plan

How the metadata will be made available?

- ANDS Research Data Australia (as part of ARCCSS collection)
- NCI geonetwork (as part of ARCCSS collection)
- as above but as part of a different collection (please specify in comment)
- An institutional metadata catalog service (please specify)
- Other web resources (please specify)
- Other (please specify)

Guidance:

Questions to consider:

- Is there any reason why you shouldn't use the default options?
- Is this research part of a bigger project that has a policy on data and metadata?
- Does your institution require you to deposit your metadat in one of their repositories as well?

Overview:

You can choose more than one options, the first two are the default choice for ARCCSS. Metadata has quite a broad definition, so you should include here any web based documentation and/or if you generated your own code for this project any code repository, such as bitbucket or github. The CMS group manages a github repository for ARCCSS users and collaborators, see link below.

Online resources:

- NCI geonetwork record on RDA
- CMS github repository

How will users access your data?

- NCI thredds catalog service
- Institutional thredds or equivalent web dataset repository
- Other thredds or equivalent web dataset repository
- FTP server
- Data will only be accessible upon request
- Other (please specify)

Guidance:

Questions to consider:

- Is the access method suitable for your data format, size, access restrictions and project requirements?
- Is the access method easily available to you?
- Will the access method be maintained for the dataset lifetime?
- Will it be accessible to users outside your institution? and if not, have you thought of an alternative for them?

Overview:

You should list here one or more way to access your data. If they are not the default ARCCSS choice you should use the "Comment" box to detail the chosen service and your reason for this choice.

The default ARCCSS choice is to add the data to the NCI thredds catalog, as a sub category of the ARCCSS data collection. In this case the data will also have a geonetwork record for its metadata which will be the main reference url.

If you choose more that one access please signal which is the preferred one.

Online resources:

- NCI thredds catalog

Publishing details

Dataset title

Guidance:

Overview:

The dataset title or name should be as descriptive as possible. It should include keywords to provide context for non-specialist users, as well as information such as the nature of the data and spatial and temporal coverage. Acronyms should be avoided or resolved.

Online resources:

- ANDS guide on naming collections

Example Answer:

ACCESS1.3b model output from the Amazonian Deforestation (AMZDEF) experiment

Title abbreviation

Example Answer:

ACCESS1.3b_AMZDEF

Guidance:

Overview

The abbreviation for the title will be used in the metadata records but also as the main directory for the dataset files both on the filesystem and on the thredds catalogue.

Dataset version

Guidance:

This is the version to use for the dataset. If this is the first version just use the deafult v1.0 .
Giving a version is important even if you are not planning any update at the moment, in case should you have to republish the data for any reason, ie. to correct a mistake, it is less confusing for the users that all the different versions of your datasets are clearly defined.

Related information

Guidance:

Overview:

Keywords and field of research

Guidance:

Overview:

Preferred keywords or subjects are ANZSRC fields of research, try to include at list one (see link below). You can choose as many as applicable, you can also add your own keywords.

It is not necessary to repeat words already present in the dataset title, since they will be also used by the search engine.

Format:

For FOR codes use the code follow by explanation, example:

040104 Climate Change Processes

For any other keyword just write text, every keyword should be in a separate line

CO2

land surface model

Online resources:

- ANZSRC fields of research
- ANDS guide on subjects

Example Answer:

040104 Climate Change Processes

CO2

land surface model

Spatial coverage

Example Answer:

-90.0, 90.0, -180.0, 180.0

Guidance:

Overview

This is the spatial range covered by the data, we need the minimum and maximum latitude and longitude to create the bounding box requested by the standards followed both by NCI and ANDS. As an alternative you can choose to name one or more regions answering question 7.

Format

Minum an dmaximum latitude followed by minimum na dmaximum longitude, separated by a comma, as:

minLat, maxLat, minLon, maxLon

Latitude is a number between -90 to 90 and longitude between -180 to 180. Do not use E/W/N/S!

Online resources

ANDS documentation on spatial coverage

Alternative spatial coverage: named region

Example Answer:

Australia

Pacific Ocean

Guidance:

Overview

This is and alternative to question 6 on spatial coverage. You can name one or more regions representing the spatial range covered by the data. For example if you have global averages or a clmatology that applies to an entire continent or ocean. You should use wherever possible the available ISO standards for named regions. Links to the standards are in the ANDS documentation provided in online resources.

Format

A named region for each line, as:

Australia

Pacific Ocean

Online resources

ANDS documentation on spatial coverage

Temporal coverage

Example Answer:

2017-01-01 2017-01-01

Guidance:

Overview

This is the time period covered by the data. It should be expressed as from-date and to-date. If the data cover two or more distinct periods you can list them on separate lines. If the data refers to a particular period ypou can use the alternative named period in question 9 or the from and to dates can be the same. Both ANDS and NCI require the dates to follow the W3CDTF conventions.

Format

Each period should be expressed as a couple of dates, one period per line, as

2014-01-01 2017-12-31

1970-01-01 1996-03-04

where each date is YYYY-MM-DD

Online Resources

ANDS documentation on temporal coverage

Alternative temporal coverage: named period

Example Answer:

Last Glacial Maximum

Pre Industrial Era

Guidance:

Overview

This is and alternative option for the time period covered by the data. You can use this when the data refers to a particular period. The named periods should follow some conventions to be understable by all. Sometimes this conventions are different depending on the discipline so please avoid acronyms.

Format

Each period should be expressed as text, one period per line, as

Last Glacial Maximum

Pre Industrial Era

Online Resources

ANDS documentation on temporal coverage

Date created

Example Answer:

2018-01-05

Guidance:

Overview

This is an indicative date for when the dataset was created.

Format

The date should be formatted as YYYY-MM-DD

2014-01-01

Online Resources

ANDS citation dates guides

Dataset abstract

Guidance:

Dataset abstract is different from the project abstract which was focusing on the scientific reasons why the data has been created. This is the abstract that summarise what the dataset is and its main characteristic that will be used in all metadata record. It should be sufficiently detailed for a potential user to make a decision about the utility of your data, so contain inforation about the region and time period covered, what are the fileds made available and how the data was created (model, input data). At the same time it should be easy to read for users coming from a variety of disciplines.

Published records

Dataset citation

Guidance:

Overview:

This is the dataset equivalent of a paper citation. It should allow to refer to a dataset in a bibliography. If a DOI (Digital Object Identifier) will be minted for the dataset, it can be used as reference. An alternative is to use a paper or technical document that describes how the dataset has been generated.

If both a DOI and a paper reference are both available they can be combined.

Online resources:

- ANDS guide on dataset citation
- ANDS guide on DOI and other identifiers

Example Answer:

Lorenz, Ruth, 2014: ACCESS1.3b model output from the Amazonian Deforestation (AMZDEF) experiment v1.0 . NCI National Research Data Collection , doi:10.4225/41/563BEB78E3A93

Lorenz, R., and A. J. Pitman (2014), Effect of land-atmosphere coupling strength on impacts from Amazonian deforestation, Geophys. Res. Lett., 41, 5987–5995, doi:10.1002/2014GL061017

ANDS Research data Australia record

Example Answer:

https://researchdata.ands.org.au/derived-optimal-linear-combination-evapotranspiration/815690

Guidance:

Once your dataset has been published on the ANDS RDA metadata repository, you should use this field to record the url. This will complete the information on your dataset. This information will also be used to create a geonetwork record for NCI and to pass a list of published datasets to the centre administrators. This field will only accept a valid research data australia url.

NCI geonetwork record

Example Answer:

https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f8778_8141_2140_4765

Guidance:

Once your dataset has been published on the NCI geonetwork metadata repository, you should use this field to record the url. This will complete the information on your dataset and it will also be used to pass a list of published datasets to the centre administrators. This field will only accept a valid geonetwork url.

Dataset DOI

Example Answer:

http://dx.doi.org/10.4225/41/563BEB78E3A93

Guidance:

Once a DOI has been minted for your dataset, you should use this field to record the url. This will complete the information on your dataset and it will also be used to pass a list of published datasets to the centre administrators. This field will only accept a valid DOI url.