Guidance on Preservation of NERC Model Code and Model Output

Version #1 07/07/2017

1. Introduction

This guidance outlines the NERC advice on the preservation of model code and model output data, the primary aim of which is to support the openness, transparency, auditability and integrity of the research process.   Model data and the code that created them are valuable assets of NERC funded research, hence it is vital that measures be taken to preserve them beyond the lifetime of their projects. The preservation of data and code enables other researchers to benefit from the outputs of previous NERC funded research.

2. Requirements

In summary the mimimum requirements for model code must be:

- where possible, developed in an open source environment,
- governed by a source code management tool such as GIT or subversion,
- made available in non proprietry format for storage e.g. ASCII text format,
- adequately documented.

Model configuration files and input or forcing data must be:

- preserved in standard formats (e.g. netCDF) in a versioned repository (e.g. GIT or subversion),
- adequately documented.

Model code and the associated input or output data that are used to generate scientific publications should be provided to and managed by a NERC Data Centre and cited using Digital Object Identifiers (DOI). Where this is not possible, they must be referenced with a URL (preferably a persistent URL).

If licensing constraints exist these can be discussed with the relevant NERC data centre.

Documentation of model code should follow the [NERC metadata standards for models](#) and the application of that model.

3. Notes on Requirements

The guidance is presented in terms of minimum requirements that are expected to be met by NERC funded researchers. NERC acknowledges that in some cases it is not possible to make source code available due to licensing constraints.

It is important to note that making source code available will not necessarily allow for the exact reproduction of particular data and that providing source code should not be seen as an alternative to good documentation.

For further information please contact the relevant [data centre.](#)

4. Documentation and management of code and data

Documentation is essential in order to enable re-use and time must be allocated for writing documentation when developing a data management plan for models. Documentation of model code should use the [NERC metadata standard](#).

Where source code is unavailable or is not able to be shared because of licensing constraints it is essential that the documentation describes what software systems and what version of those were used to create the data. Where only executables are available a declaration must be made about what software was used to generate the data and when the data was created. There is no requirement to keep executable files.

Community modelling programmes must make effective use of version control systems for maintaining and developing code. Individual researchers must also use version control systems to maintain model code, pre- and post-processing software and output analysis software.

Model input data must either be preserved or referenced. Where model input data are not specific to an individual experiment, they must be referenced as independent objects to facilitate reuse and avoid duplication of creation and/or storage.

A URL is sufficient to reference the model code and related data, though formal citation via Digital Object Identifier (DOI) is preferred for both model data and code. For model code to be assigned a DOI it must meet the same criteria for stability, quality of code and metadata and permanence as is required for [DOI-ed datasets](#). DOIs will allow linking back to the source code in its frozen and archived state, meaning that models can be cited in papers in the same manner as other research objects.

Definitions

Adequate Documentation: In this context adequate documentation is comparable to the discovery level metadata in the NERC data catalogue service. It states what model, input data, and any pre- or post-processing software was used along with version information. It includes the date when the model output data were created and the persons and institutes responsible for running the model.

Configuration files: Model specific files that set parameter values. Configuration files are not a substitute for good documentation.

Digital Object Identifier: A digital identifier used to uniquely identify an object such as an electronic document. Metadata about the object is stored in association with the DOI name on a DOI landing page; this metadata may include a location, such as a URL, where the digital object can be found.

Input data: Data used to initialise a simulation and data used to constrain a simulation as it runs.

Metadata Creation Framework: Metadata creation tools based on an underlying data model. Examples include the Metafor questionnaire for CMIP5, CEDA MOLES and the NERC data catalogue.

Model: In the context of this position statement a model is considered to be some software system that simulates single or multiple components of the Earth system including physical, chemical and biological processes. However the statement is intended to apply to all NERC modelling activities such as statistical modelling.

Model output: Data output from a simulation

Open Source: A production and development methodology that promotes free redistribution and access.

Rich Metadata: Documentation that goes beyond discovery level metadata to allow one to answer science questions such as why a particular simulation was performed.

Source code: Computer instructions written in a human-readable computer language.

Source Code Management Tool: a tool to support the collaborative development of software and the tracking of changes to software source code over time (e.g. subversion).