Earth Science Academic Archive

2009

Guidelines for Depositors

Rod Bowie

BGS

September 2009

The Earth Science Data and Information Life Cycle

Contents

Preface

The Natural Environment Research Council (NERC) provides grants for scientific research in line with its Science Strategy. All grant holders have to abide by the conditions of the grant including following NERC Data Policy. The policy places a duty on grant holders to offer a copy of their datasets to the appropriate NERC Environmental Data Centre. In the case of the Earth Sciences this is the National Geoscience Data Centre (NGDC), a component of the British Geological Survey (BGS). The Earth Science Academic Archive (ESAA) is the component of the NGDC responsible for working with NERC grant holders.

Earth science data collected under other funding streams can also be deposited with the ESAA, on a voluntary basis, under the same terms and conditions as a NERC funded research.

This document explains the benefits that this deposit offers, describes the process of depositing information, and provides information about other sources of relevant information.

Two sources were drawn on during the preparation of this document. The first is the Research Information Network's document entitled – “Stewardship of digital research data - principles and guidelines”. The second document is the ICPSR Guide to Social Science Data Preparation and Archiving 3rd Edition 2005. We gratefully acknowledge the contribution that both these documents have made to these guidelines.

http://ngdc.nerc.ac.uk/
www.rin.ac.uk/data-principles

About the ESAA

The BGS demonstrated its recognition of the importance of maintaining a major national collection of subsurface information when it established the NGDC in 1984. The NGDC now manages data and information created as part of BGS' own research and that created by the academic community, central government and many commercial sectors. The NGDC is the responsibility of BGS's Head of Information Management.

The NGDC has five principal components:

- BGS-geoIDS (BGS Geoscience Integrated Database System);
- Earth Science Academic Archive (ESAA);
- National Geological Materials Collection (NGMC);
- National Geological Records Centre (NGRC); and
- National Hydrocarbons Data Archive (NHDA).

The ESAA component of the NGDC is responsible for working with NERC grant holders to ensure the long-term preservation, promotion, discovery, re-use and re-purposing of their data. All creators and custodians of analogue and digital earth science data created by NERC grant holders are expected to deposit a copy of the data with the ESAA.

The ESAA seeks to manage information in accordance with NERC and BGS data policies and will:

- manage information of all formats in appropriate environments and use appropriate technologies to ensure its long-term preservation and availability;
- promote the datasets managed with the ESAA via the BGS website;
- ensure that appropriate metadata has been provided by the scientists who deposited the data and that this metadata is disseminated to aid discovery of the resource;
- maintain a system in which all records, in whatever format or media, can be found rapidly;
- encourage the re-use and re-purposing of the data deposited with the ESAA in future scientific research.

The ESAA supports the Research Information Network (RIN) principles for “Stewardship of digital research data”. These RIN principles state:

> *“In order to produce high-quality research, researchers must have access to as wide a range as possible of the data and information produced by other researchers, as well as relevant information produced by other agencies in the UK and overseas. Similarly, successful dissemination and exploitation of research depends on effective flows of information between researchers and other individuals and organisations that have an interest in its results. A successful research and innovation system thus depends on the open exchange of ideas, information and knowledge.*

www.rin.ac.uk
www.rin.ac.uk/data-principles

Why deposit data?

Developments in information and communications technologies are transforming the nature and scale of research, enhancing both quality and productivity. They are facilitating new kinds of research, new organisational models, and collaboration across disciplinary, institutional and national boundaries. But they also demand new ways of thinking about how we manage data and information outputs, so that we can maximise their value, and ensure that precious resources are not lost. In pursuance of those goals, the fundamental policy objective is to ensure that ideas and knowledge derived from publicly-funded research should be made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly and effectively as practicable.

To help achieve that objective, the following five principles provide a broad framework for developing good practice for universities, research institutions, libraries and other information providers, publishers, research funders as well as researchers themselves. The principles are pitched at a high level. The guidance set out in the main document provides pointers as to how policy and practice may need to be changed to ensure that they are to comply with the principles

I. The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with these principles.

II. Digital research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.

III. Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.

IV. The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.

V. Digital research data of long-term value arising from current and future research should be preserved and remain accessible for current and future generations."

The NERC will provide the costs of depositing data with its Environmental Data Centres at the end of the research. NERC grant applicants are required to liaise with the NERC Environmental Data Centres during the development of their proposal so that the costs incurred by the data centres are included within the grant.

ESAA staff are happy to provide advice and support during the course of the research to ensure that the data is in good order and well documented at the end of the research.

Where other organisations maintain similar information, then ESAA will endeavour to collaborate with these bodies to ensure ready access to that information.

The Earth Science Data and Information Life Cycle

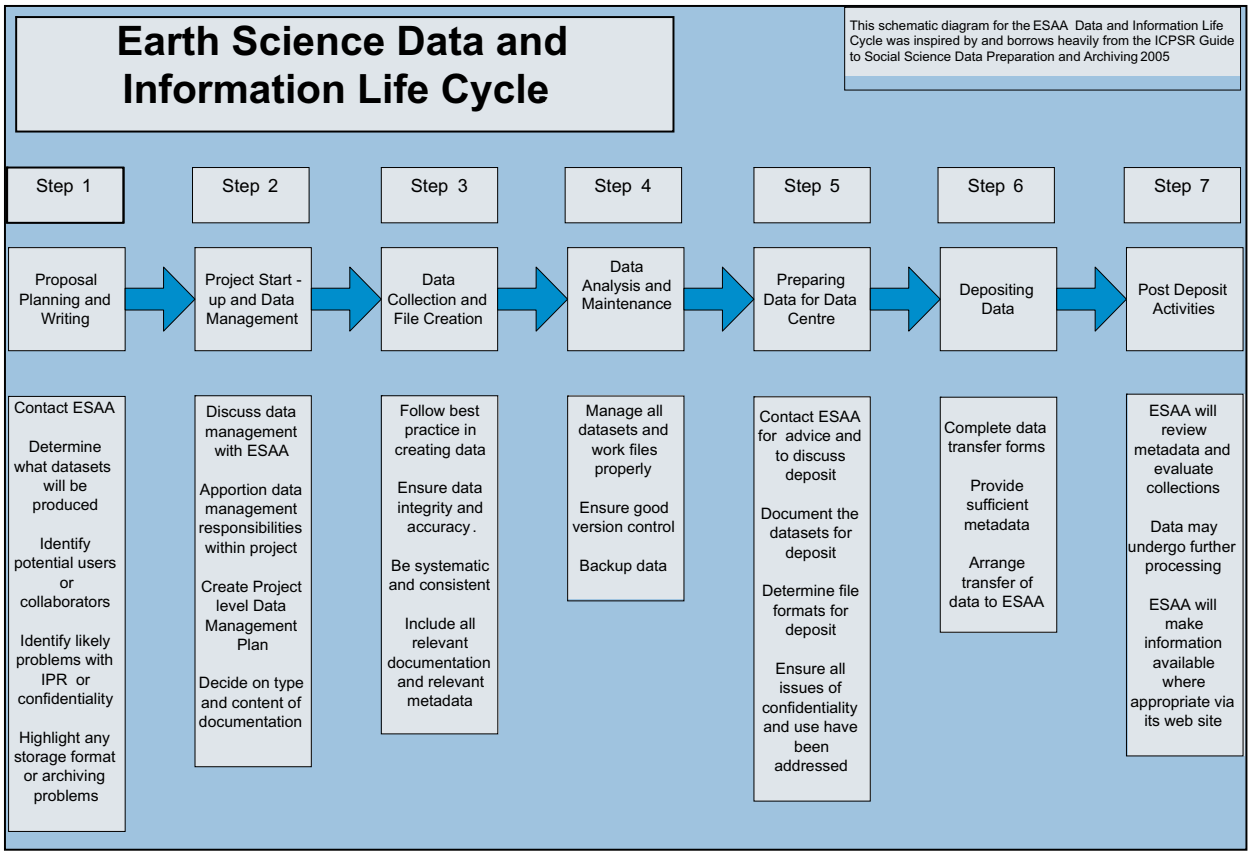
Using these guidelines

.

This document is aimed at supporting potential depositors with the ESAA engaged in any part of their project, from applying for a research grant, through the data collection, and to the preparation of the data for deposit in the archive. These guidelines are constantly evolving as they reflect NERC Data Policy and best practice and the experience gained from other organisations, professionals and specialists. Researchers can contact the ESAA at any time to discuss their plans with respect to the selection and preparation of datasets.

Many investigators are more than willing to make their data available to others, but are unsure of how to go about preparing data for outside use, particularly in terms of complete documentation. The guidelines are intended to help researchers document their datasets and prepare them for deposit with the ESAA.

The main steps are set out below, in an order that reflects the lifecycle through which research data are created, used and made accessible over the long term.



Step 1 - Proposal planning and Writing

It is important to contact ESAA staff as soon as possible. They will provide advice and assistance on the following:

- datasets that already exist that may be of use to the research;
- data management planning for the grant proposal;
- how to document the datasets created and what metadata will be required;
- advise on periods of exclusive use by the creators of the dataset;
- what would be appropriate for final deposit with the ESAA; and
- costs to include in the grant proposal for data management and archiving.

If you are developing a proposal for a research grant to NERC then the proper management of the data is considered to be part of the project and failure to deal adequately with this aspect may contribute to rejection or cause delay in approval. By contacting the ESAA you can gain professional advice and guidance that will help you meet the standards that NERC expects in its grant applications.

It is important to consider Intellectual Property Rights (IPR) and copyright at an early stage in proposal planning and writing. The eventual aim is to ensure that datasets created as part of NERC research will be available to for re-use and re-purposing by future researchers. To this end researchers are asked to specify that the ESAA should have a non-exclusive license to distribute the deposited datasets. The ESAA will of course agree and recognise reasonable periods of exclusive use by the creators of the datasets.

Where researchers are planning to use commercial datasets, owned by third-parties, they are encouraged to set up agreements so that the ESAA may store a copy of the dataset on a Commercial-in-Confidence basis for an agreed period. The ESAA will acknowledge that the copyright of these datasets are vested with the organizations that created them

It is important to contact ESAA staff once the funding has been agreed with NERC so that detailed data management planning my progress.

The first step will be ensuring that all parties are clear about their roles and responsibilities and their expectation of other parties that are involved. This is in line with first of the RIN principles for the stewardship of research data which states that:

The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and that they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with the principles set out in this document.

Planning for the management and storage of data at the beginning of the project is essential if the data is to be well managed throughout the life of the project and archived effectively. Without adequate data management planning data can be lost or corrupted and the costs of archiving increased.

The types of project documentation that will be required need to be considered. How will key decisions be documented during the projects life? What standards will be adopted by the project? How will these standards be implemented? What procedures need to be developed and how will they document, disseminated and preserved? What ontologies will be adopted or created and how will these be managed and preserved? How will change be managed?

Datasets should be well documented at the time of collections. This requires early planning to make sure that consistent documentation is created. Standards are now emerging for data product specification and these should be considered in the process of planning the documenting of datasets. Appropriate metadata will be required at the time of deposit to aid discovery. An appropriate profile of ISO 19115: 2003 Geographic information – Metadata will normally be expected. See example metadata (Appendix 8)l

ISO 19131:2007 Geographic information – Data product specification

Step 3 – Dataset Creation

The second RIN Principle states:

Research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.

“From the archivist's and end user's perspective, a “good” dataset is one that is easy to use. Its documentation is clear and easy to understand, the data contains no surprises, and users are able to access the dataset with relatively little start-up time.” ESAA staff would add that it has sufficiently detailed metadata to ensure it could be easily discovered. To achieve these goals effective data management must begin early.

It should be assumed that datasets deposited with the ESAA will be seen and re-used by other researchers. It is therefore important to adopt appropriate international standards and be able to demonstrate the use of best practice, so that those re-using and re-purposing the information have confidence in the quality of the datasets. This confidence is demonstrated by having a clear understanding of the authenticity, reliable, lineage and completeness of the dataset. This in turn is built by having clear documentation of their origin, including the methods, standards, procedures, instruments and techniques used in their creation or collection?

Where datasets are created in collaboration with other researches it is essential that the IPR and copyrights issues are clearly defined.

During the course of the project appropriate security measures must be in place. For digital data appropriate measures should be used to protect the datasets from virus, theft of equipment, etc. A backup regime should be in place to ensure that datasets and supporting documentation are protected. Analogue information and specimens should also be protected appropriately.

See page 7 of ICPSR Guide to Social Science Data Preparation and Archiving 3rd Edition 2005

Step 4 – Analysing and Maintaining Your Data

The third RIN Principle states:

Research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.

The research community and anyone with an interest in the data should have timely, user-friendly access. This does not mean that all data should be made available immediately to all who may have an interest in it. There is a need to balance conflicting interests and rights. The requirement is for access in a managed environment; and to achieve that, a number of issues must be addressed.

The ability of users to find and re-use data depends on having a full record of information relating to the content, structure, context and source of data that may be relevant to their needs. Such information – data about data – is commonly referred to as metadata. Where appropriate, metadata should be created and made available in accordance with recognised international standards. But the essential requirement is as full a record as possible of the context in which the data was created or collected. Without this information, users cannot be aware of the nature and limitations of the data and will not be able to interpret them properly, or use them effectively.

Data are of no use without the facilities, software applications and other tools required to access and use them. Most users will have access to these applications and tools; but use of the data may only be possible if steps are taken to make the necessary applications and tools available.

All relevant files, particularly datasets under construction, should be backed up frequently to prevent having to re-enter data. It is a good practice to maintain a master version of the dataset that is stored on a “read only” basis. These master datasets should be backed up every time they are changed in any way. Computing environments in most universities and research centres support devices for data backup and storage. Although everyone knows the importance of backing up data, the problem is that few actually follow this through. It is also advisable to maintain a backup copy of the data off-site. One fire can destroy many years of work.

Physical specimens and samples are a vital part of many projects and they should be well labelled, properly documented and kept securely. Doing sophisticated analyses or producing complex theories based on poorly recorded specimens is obviously of little value. The ESAA can provide a secure repository for materials at any time.

There should be a clear process for determining who should have access to what data, and on what terms. Otherwise there is a danger of either unwarranted access or unnecessary restrictions. Access therefore requires research funders, researchers and institutions to decide how the benefits of access to data can be properly managed and optimized.

Copyright and the ownership of datasets and work must be considered and researchers should ensure that they have a right to transfer any data received from other parties or used as part of their research. This means establishing relevant procedures and audit trails. Any agreements should be made clear to ESAA staff who can provide help and advice.

Users will want to access research data alongside publications, in order to assess the evidence on which the research is based. It is important that researchers should receive appropriate credit for the work that has gone into creating or collecting data. Citation of research data (alongside citation of publications) is becoming common in some areas, and is likely to become more important part in evaluating research quality.

Evaluating data for permanent retention

The fifth RIN Principle states:

Research data of long term value arising from current and future research should be preserved and remain accessible for current and future generations.

Not all research data are of long-term value. Many datasets may have little value beyond the life of a specific project. It is essential that researchers put effective procedures in place to determine which data are of continuing value, and to ensure that arrangements for their stewardship are sustainable. Some datasets are of such significance and value that they need to be preserved and made accessible over the very long-term.

As research data are increasingly updated, amended and annotated over their lifecycle both by data creators and by subsequent users, provenance protocols and audit trails are needed to indicate clearly who has annotated or amended data, how and when. Otherwise there is the danger that data will be misinterpreted or used inappropriately.

What needs to be done?

Ideally preparation for the eventual deposit of data should begin at the start of the project. This may initially be best done by establishing a dialogue with the ESAA staff. Preparation of Data Management Plans and properly documenting the information and metadata is essential. Evaluating suitable formats and arranging storage whether for digital or analogue data or materials.

Any documentation should enable a third party to make sense of the data. The ESAA data management questionnaire (Appendix 2) is a good starting point. This makes donors think about the information the project is creating: the ESAA Good Data Management guidelines (Appendix 1) sets out best practice. However, there may be documentation requirements specific to particular types of data or research, so potential depositors should ensure these are recorded. Much of the required basic data will be held or supplied in the accompanying metadata, but additional information helps with the understanding of the data. The following list provides some headings for additional information that could be of value:

- Background to the project
- The purpose of the project
- Information about methodologies
- Sources used
- Related research
- Sampling procedures
- Content and structure of dataset

- List of filenames and description of contents
- Any thesaurus or dictionaries used
- Explanations of codes or anagrams used
- Description of any known errors or weakness in the data
- Documentation of any record or data conversion or format change
- References to publications resulting from or related to the project
- List any information about other Data Centres, Universities, etc. that may hold material related to the dataset
- Indicate of how long archive is to be retained (indefinitely or for fixed period)

Once the ESAA has received your data staff will establish the following:

- Assess their intellectual content and thus the level of potential interest in their re-use. The creation of an electronic data resource requires a significant amount of effort. Assessing the result is difficult, as its form differs from that of traditional research projects.
- Assess how (even whether) they may viably be managed, preserved, and distributed to potential secondary users.
- Assess the presence or absence of another suitable archival home. More information about the criteria for evaluating datasets can be found in the ESAA Collections Policy Framework (Appendix 5).

Step 6 – Depositing the Results of your Research

Potential donators should contact the ESAA Data Collection Officer for information about deposit either by letter or by e-mail or telephone (see Contacts). If there are any queries about policy they will be dealt with by the Manager. There is a booklet available “Guide to the Collection of Geoscientific Collections” which sets out some information about the data and deals with frequently asked questions.

Datasets should be accompanied by written information, which should stipulate how and by whom the data can be reused, and how the data and its documentation will be transferred to the ESAA. The forms are available from the Data Collection Officer and downloadable Web versions are available. See documentation.

Datasets will be assessed by the ESAA Manager to ensure they are appropriate for deposit with the ESAA.

Once accepted, datasets will be accessioned using the standard ESAA workflow (see Appendix 3).

Accessioning procedures include data validation, scanning, indexing and databasing. The size of the dataset will determine how long this process takes, but it is our target to complete this process in less than 3 months. If there is a special reason for needing datasets to be processed more quickly, please contact the ESAA in advance to discuss your requirements.

The fourth RIN Principle states:

The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.

To maximise the benefits and minimise the costs of managing and research data, it is essential that research data must be both efficient and cost-effective in the use of public and other funds.

Although it is costly to maintain this information, managing and providing access to research data brings huge potential benefits in enhancing the overall efficiency of research; avoiding unnecessary duplication of effort and needless costs in knowledge transfer. Effective and efficient management of data requires specialist professional support services, which are provided by the ESAA to ensure that data is properly selected and stored, that it can readily be accessed, and that its integrity can be ensured over time.

ESAA undertaking

The ESAA undertakes to:

- To maximize the benefits of research data information deposited with it and to make this information readily available.
- Ensure that suitable metadata is available for all the information deposited.
- Manage the information provided, in whatever format, in an effective way and use appropriate technologies to ensure its long-term preservation and availability.

Create a system in which all records, in whatever format or media, can be found readily

For all NERC Research Grants reference should be made to NERC Policy documents see:
http://www.nerc.ac.uk/research/sites/data/policy.asp

Contact details for the ESAA

Rod Bowie,
Manager
Earth Science Academic Archive,
British Geological Survey,
Keyworth,
Nottingham,
NG12 5GG
tel: 0115 9363106
fax: 0115 9363276
email: esaa@bgs.ac.uk

Becky White
Data Collection Officer
Earth Science Academic Archive,
British Geological Survey,
Keyworth,
Nottingham,
NG12 5GG
tel: 0115 9363106
fax: 0115 9363276
email: esaa@bgs.ac.uk

Appendices

Appendix 1 - ESAA Good Data Management Guidelines

Appendix 2 - ESAA Data Management Questionnaire

Appendix 3 - ESAA Data Management Plan

Appendix 4 - ESAA Data Transfer Form

Appendix 5 - ESAA Collection Policy Framework

Appendix 6 - Quick Check list

Appendix 7 - Earth Science Data and Information Life Cycle

Appendix 8 - Example metadata