Providing Large Volumes of Data to the ERC

Environmental Research Centre

Environmental Protection Agency

Introduction To This Document

Management of environmental research data is a core activity for the ERC with particular emphasis on the application of appropriate data management techniques to ensure their long-term availability and accessibility. Environmental research data are often irreplacable; they are always unique, if only in the spatial location and temporal characteristics of their collection. They can also be extremely expensive and difficult to collect or generate. For these reasons the EPA and the ERC attach great importance to the ongoing development of systems that will ensure maximum benefits are derived from research data once acquired.

This document describes data management policy for submitting large volumes of environmental data to the SAFER-Data system. SAFER-Data is an acronym for *Secure Archive For Environmental Research Data*.

The SAFER-Data system is available at http://erc.epa.ie/safer

The term *Dataset* is defined as follows:

- a collection of one or more digital computer files;
- a database which is accessed using database system software;
- part or one of the above as specified by some filtering, querying, or extraction.

Several acronyms are used in the remainder of the document. Unless otherwise stated ERC-DM indicates the ERC Data Manager, ERC-C indicates the ERC Co-ordinator, and EPA-M indicators the EPA Research Manager.

This document is part of a series of ERC Data Management documents. The other documents in this series are "*Metadata Information Requirements*", "*SAFER-Data User Guide"*, and "*SAFER-Data Usage and Usage Policy*". While these documents have been developed as documents for usage by users of data services provided by the ERC the wider distribution of these documents is not restricted. They may be circulated "as is" to other organisations or individuals who may have an interest in data management or whom are developing a data management system or data management policy of their own.

Contact Person: Peter Mooney peter.mooney@nuim.ie

The requirement that all research projects funded by the EPA under ERTDI or STRIVE provide metadata and all generated datasets is outlined in the Terms and Conditions “Guide For Grantees” (Section 8 Final Reporting Guidelines).

http://www.epa.ie/downloads/pubs/other/corporate/oea/research/

Providing Large Volumes of Data to the ERC

Physical Description of ERC Environmental Data Archive

The ERC Environmental Research Data Management System (DMS) is based on the hardware architecture of a highly fault tolerant Storage Area Network (SAN). The current capacity of the DMS is 2 TeraBytes (approximately 2000 GigaBytes). This specification ensures that there is sufficient storage space available for all EPA funded research projects. The network capability of the SAN ensures that 24-7 access to the DMS is guaranteed. SAFER-Data is the web-based interface to this DMS.

Uploading Datasets to The ERC

By default all dataset files uploaded to the ERC DMS are uploaded using SAFER-Data. This upload functionality is very similiar to the functionality most readers would be familiar with from web-based email systems such as Yahoo or Gmail. Using SAFER-Data project managers can upload up to FOUR dataset files at a time. When these dataset files have been uploaded another four files may be uploaded and so on.

This upload limit of four files at any one time is not an arbitrary value. This value is set to allow multiple files be uploaded simultaneously while ensuring that large collections of very large files are not uploaded at once. Web-browser based upload is very computationally intensive. For this reason a limit must be placed on the quantity and size of uploaded data.

The SAFER-Data file upload mechanism has been tested with single files of size 150Mb and with combinations of four files of total size greater than 100Mb.

The following conditions are recommended for dataset file upload to ERC:

1. A fast and reliable broadband Internet connection. SAFER-Data testing has been performed on 3rd level institution networks and private corporate networks.
2. A recent version of your Internet browser software
3. Virus scan is performed of all files that you intend to upload.

Some projects may generate large volumes of data and/or large quantities of datasets. In the case of these projects alternative delivery methods must be considered. Alternative methods of delivery to ERC must be used <u>if either or both of the following conditions apply</u> to your project:

1. ***Large Volume***: The total volume of data generated that will be uploaded by your research project exceeds 150Mb
2. ***Large Quantity****:* The total number of files that will be uploaded (including ZIP File Archives) is over 25 files

If either of these conditions apply to your project the project manager should contact ERC to discuss your project datasets in greater detail and explore the specialist file upload requirements for your datasets. The ERC-DM may recommend that dataset files be combined or grouped into logical collections of files.

Submitting Large Volumes of Data To ERC

This applies to projects where the total data volume for upload exceeds 150Mb.

The file upload mechanism in SAFER-Data should not be used if this condition is true. The following are a list of alternative means of submitting your data to ERC:

1. Create a compressed archive version of the dataset files. In some cases this may reduce the overall file size considerably. If the total dataset filesize after compression (ZIP, TAR, GZ, etc) is now less than 150Mb you may use SAFER-Data to upload the datasets to the ERC DMS.
2. Copy the datasets to a removable media device – CD, DVD, USB Pen Drive, USB Hard Disk and deliver the chosen media to ERC using standard post. Please ensure to have used the appropriate packaging envelopes to protect the media. Upon receipt, within one working week, the ERC-DM will upload your datasets to your metadata resource on SAFER-Data. You will be notified by email when this action has been completed.
3. Make the datasets available on a HTTP or FTP server. These servers must be accessible to the ERC. The ERC-DM will download your datasets from the server specified. Upon successful download, within one working week, the ERC-DM will upload your datasets to your metadata resource on SAFER-Data. You will be notified by email when this action has been completed.

If appropriate the ERC-DM will recommend changes to how the datasets are stored and accessed on the ERC DMS. These changes will be recommended with the dual intention of protecting your dataset resources while ensuring that 3rd parties can download and access the files from SAFER-Data without major difficulty.

To assist 3rd party users, who download your datasets from SAFER-Data, to understand and use your datasets the ERC-DM encourages project mangers to provide detailed information about the datasets in the “Additional Information” field in the metadata resource. This information should include details on what is contained in the files, how to open them, what statistical analysis or aggregation was performed, etc.

Submitting Large Quantities Of Data To ERC

This applies to projects where quantity of dataset files uploaded exceeds 25 files.

The file upload mechanism in SAFER-Data should not be used if this condition is true. Project managers should contact the ERC-DM to discuss the proposed quantity of files for upload.

To ensure your datasets are distributed and disseminated in the most efficient manner the ERC-DM recommends grouping these files in logical groups or categories. This can be done in one of two ways:

1. Creating a new metadata resource on SAFER-Data corresponding to the subsets of the original dataset files.
2. Create a compressed file archive of all of the dataset files. This compressed file archive is then uploaded to SAFER-Data.

There are several characteristics upon which dataset files can be logically grouped:

1. The time period represented within the files
2. The geographical area represented by the files
3. The working group or thematic area which generated the files.
4. Any other appropriate project specific characteristic.

If appropriate the ERC-DM will recommend changes to how the datasets are stored and accessed on the ERC DMS. These changes will be recommended with the dual intention of protecting your dataset resources while ensuring that 3rd parties can download and access the files from SAFER-Data without major difficulty.

To assist 3rd party users, who download your datasets from SAFER-Data, to understand and use your datasets the ERC-DM encourages project mangers to provide detailed information about the datasets in the “Additional Information” field in the metadata resource. This information should include details on what is contained in the files, how to open them, what statistical analysis or aggregation was performed, etc.

Providing Large Volumes of Data to the ERC

Maintaining The Data Quality of Datasets Uploaded to the ERC

The EPA and ERC aim to make high quality environmental research data available using SAFER-Data. In environmental science there are a very large number of different software tools and software formats used by environmental scientists. It is not feasible to expect the EPA and ERC to have expert knowledge of all of the software formats used. However there are several data management characteristics which are common to all software formats independent of the thematic area or application. Project managers should ensure that the uploaded datasets exhibit a high data quality standard and appropriate structural formatting.

The ERC-DM will in a limited number of cases perform some "silent fixing" of common data management and data quality problems. The following issues should in all cases be resolved by the original data creator. These issues are listed as follows:

1. Renaming of files
2. Reformatting of date or date-time series within datasets
3. Grouping single files into a compressed file archive
4. Uncompressing a file archive
5. Version upgrade or downgrade of a file – example from Office 2007 to Office 2003
6. Provision of a README.TXT file to explain any special requirements users will need to meet when download the dataset files
7. Conversion of documents to a platform neutral format – example from MS Word Document to PDF

If silent fixes have been performed the project manager will be informed of the nature and extend of these in an email from the ERC-DM.

As far as is possible the ERC will check the data quality of uploaded files. The EPA and ERC reserve the right to request project managers address any data quality issues flagged by the ERC-DM before these dataset files are made publicly available.