

Submitting Datasets to the Environmental Research Centre

An overview of procedures

Peter Mooney,

IT Systems Development Fellow,
Environmental Research Centre,
Environmental Protection Agency,
Richview, Clonskeagh,
Dublin 14.

Web: <http://coe.epa.ie>

Email: petermooney78@gmail.com

INTRODUCTION

This document outlines a summary of the guidelines to researchers for submission of LS2 datasets to the COE Data Management team (COE-DMGT) at the Environmental Protection Agency (EPA). The guidelines are intended to provide researchers with a non-technical overview of the data submission procedures and outline a list of any changes or revisions that may be necessary to datasets before submission. A principal aim of the submission procedure is to ensure that no overly burdensome procedural changes on the part of the data collector or researcher are enforced. Researchers are encouraged to submit their data in whatever software format they are currently using to manage their research data.

MANAGEMENT OF RESEARCH DATA

Environmental research data are often irreplaceable; they are always unique, if only in the timing of their collection. They can also be extremely expensive to collection. For these reasons COE-EPA attaches great importance to ongoing development of systems that ensure maximum benefits are derived from research data once acquired. These guidelines have been compiled with computer-based datasets of environmental data generated by research projects in mind. The 'research data' defined will exhibit considerable diversity – output from models, data from automated monitoring devices, statistical analysis, database models. This diversity is reflected in the number of possible data formats the COE-Data-Browser must accept. The listing in Table 2 below outlines the types of data files expected from research projects.

Management of this environmental research data is a core work task for the COE-DMGT. A major product of this work the provision of a **data storage** and **data archive** system for research data. This software system is called the COE-Data-Browser and is supported by a hardware system called ICED-FIRE. The ICED-FIRE system is physically located in HEA-NET offices in Dublin. The COE-DMGT, HEA-NET staff, and EPA-IT provide administration of the system.

The COE-Data-Browser is a web-based tool (accessible with your Internet browser at the address <http://coe.epa.ie/COE-Data-Browser>) to provide a **secure, easy-to-use, always-available**, system for researchers to upload data files and associated metadata. Each project coordinator is given a username and password. This will allow the project coordinator to logon to the system and manage their **private data folders** on the COE-Data-Browser. No special software is required as all functionality is

available to a standard web-browser. The COE-Data-Browser also removes the burden of securing the long-term availability of data from the researcher. Datasets can easily be destroyed unintentionally, or effectively lost, by failure to take adequate and continuing precautions to safeguard data. Data may be held on vulnerable computer media without adequate back-up, such media may become obsolete over time, or data formats may be undocumented. The COE-Data-Browser and ICED-FIRE system have a dedicated data backup policy currently in operation for the protection of all data held in the COE-Data-Browser archive. After the data has been successfully uploaded **the data owner can replace or update these data at any time** in the future.

Datasets can be uploaded to the COE-DMGT in four ways (listed in order of preference):

1. Using the COE-Data-Browser (this is the preferred option)
2. Using our FTP (File Transfer Protocol) services. FTP is particularly useful for large numbers of files for one project or very large datasets.
3. Writing data to CD and posting to COE-DMGT
4. Email Submission of data.

CHANGES TO DATASETS BEFORE SUBMISSION TO COE

In any data storage system the uploaded data must conform to some data quality requirements. In some cases this will require the originator (researcher or data collector) to make some changes to the datasets before upload. The list of requirements below is a general outline of issues each data provider should be aware of. In the case of a dataset exhibiting non-compliance we suggest that the necessary steps are taken to remedy any outstanding issues. Any member of the COE-DMGT will be happy to discuss such formatting issues with individual researchers in greater detail.

WHY CHANGES ARE NEEDED

Data is a resource. In our case environmental data is a resource growing more valuable with time. To **enhance the future usability** of your datasets and facilitate long-term storage of the data into the future some software application specific issues may need to be resolved. An example of a software specific issue is the use of macros in Excel spreadsheet. Another important reason for implementing specified changes is to allow our system (only in agreed specific cases) **to convert datasets to other data formats** such as XML, Text Format, or SAS dataset. Conversion to other formats in almost all instances facilitates the dissemination of the dataset to a wider community of users. In many cases no changes will be required at all on behalf of the data provider. All that is required is the **specification of accurate and descriptive metadata**. Metadata must be provided before any dataset is uploaded to the COE-Data-Browser.

METADATA

Metadata is usually loosely defined as “data about data”. A more formal definition would describe metadata as “*a text based description of a dataset indicating information about data generation, collection, research aims and themes, spatial and*

temporal coverage, statistical analysis, potential usage of the data, etc as an aid to interpretation and understanding of the dataset by a 3rd party”. It is self evident that **datasets can only be fully exploited if potential users are aware of their existence**. Metadata greatly simplifies the process of *advertising* datasets and also the understanding of the data. Providing useful metadata will allow third parties to use the dataset without reference to the original collector or generator of the data. Correctly specified metadata also allows Internet-based systems, such as COE-Data-Browser, to **provide data discovery and metadata search facilities** to users. Users can then search the data archive using different queries based on Topic, Theme, Spatial Coverage, Originator, etc. Metadata is provided to the COE-Data-Browser by using a simple online form. This form is only available to authenticated users (those with a username and password supplied by COE-DMGT). Once the form has been filled in correctly and submitted the **data provider can edit and revise this metadata at any time** in the future by using the same online form. The current set of metadata headings are outlined in the table in Figure 2 below. A brief description of each heading is also provided. These headings are a combination of metadata handling techniques from the Dublin Core Metadata specification and the ISO 19115 specification.

The COE-Data-Browser also allows the generic metadata to be extended to incorporate data access constraints providing the owner of the data with **full access control**. This means that the data owner or data provider can easily **specify levels of public access to their datasets**. This functionality is provided with a simple drop-down-list. The options available are tabulated in the list below in Figure 1. As the data provider can edit and revise metadata they can also change the access control levels as they see necessary. The COE-DMGT will never change a users metadata without prior consent of the data owner.

Not Visible to Public	Visible To Public
Raw Data, Results, Samples, and Metadata	
Raw Data, Results, Samples	Metadata Only
Raw Data, Samples	Metadata and Summary Results
Raw Data	Metadata and Sample of Dataset
	Metadata and All Data

Figure 1 Levels of public access to datasets on COE-Data-Browser

Metadata Field Name	Description
Title	Title of the Dataset
Originator	Details of the researcher(s) responsible for the original creation of the dataset
Point of Contact	Full contact details of person(s) available after project lifetime to deal with 3rd party queries
Links to Related Websites	Specify any websites based on this dataset
File Type Descriptions	Software Application File Type (Selected from a list)

Size of Dataset	Size of dataset in bytes (Calculated by COE-Data-Browser automatically)
Start Date	Monitoring Start Date
End Date	Monitoring End Date
Last Updated	Last Date of Update to Dataset
ISO 19115 Theme	Thematic Category of this research (Selected from a list) (See TABLE 1 below)
Metadata Update	Last Date of Update to Metadata
Update Status	Details of when (if applicable) this dataset will be updated again
Dataset Lineage	Lineage information of the dataset
Bounding Coordinates	Spatial Coordinates of the geographical region covered by the research
Public Access Constraints	Specification of level of public access to the data (Selected from a list)
Abstract	Overview abstract about the aims/methodology/results of the research as would appear in a scholarly journal

Figure 2 Description of Extended Metadata Scheme for COE-Data-Browser

WHEN PROVIDING DATA PLEASE:

1. Take time to **decide which datasets are actually needed** by the project and which datasets should be uploaded for archive storage and possible dissemination to the public.
2. Take time to decide or **nominate who will be responsible** for the initial upload to the COE-Data-Browser and for subsequent updates to metadata and datasets.
3. **Provide a sample subset of your datasets** to COE-DMTG. The sample will merely be used to assess any special requirements of the dataset or any changes that we would advise before the full dataset is submitted to the COE-DMTG. This sample will not be stored on COE-Data-Browser nor will it ever be distributed to the public.
4. **State which versions of software you used** to analyse/reproduce the data you are submitting. You may need to contact your system administrator for these details
5. **Supply accurate and detailed metadata** (see Figure 2 above) *in a separate file* or by using the online form provided by the COE-Data-Browser. Your metadata must include the following attributes: *a description of the purposes and aims of the research project, the variables measured and their units, details of the data collectors and analysts, geographical coverage, derived publications, temporal range of the dataset, a description of any flag system or conventions used.*
6. The COE-DMGT welcomes data stored in XML (eXtensible Markup Language) format or GML (Geographic Modelling Language). If this is applicable to your research include the DTD (Data Type Definition), any XHTML, or XSTL code that you have used to support your XML dataset.
7. Consider the *types* of data you have stored in each column of the dataset. For example, ensure that any column with the purpose of storing a date or time variable does not contain text values and time values mixed (with the exception of the column header). This is highlighted in Figure 3 below. The *Date2* column has mixed values of text and date formats. In a similar way numerical data should not

be mixed with text strings in the same column. This is relevant to *Reading1* in Figure 3 below. In this case a convention for numerical values associated with missing values should be used and expressed in the metadata. The COE-DMGT is available to discuss such issues relevant to your data.

8. **Avoid the use of spaces in column names** in Excel or Access. Figure 3 shows all column headers without any spaces in column names.
9. **Detail and explain any conventions implemented for missing values.** Address any anomalies regarding Zero as a missing value and Zero as a valid reading.
10. Specify actions to be taken when readings or measurements are provided “below the level of precision” or “below the level of tolerance”. If mathematical formula are required please specify. This is shown in *Reading2* of the example in Figure 3.
11. *Dates:* It is **preferred that dates are stored in standardised formats** – ie dd/mm/yyyy, mm/dd/yyyy, yyyy/dd/mm, yyyy/mm/dd, dd-mm-yyyy, mm-dd-yyyy, yyyy-dd-mm, yyyy-mm-dd. The use of customised date formats is possible but not advised.
12. *Excel Specific:* If your data is stored in Excel you will be required to specify the range of your data (ie A4:N890) - that is the rows and columns within which you have stored your data. In the case of multiple ranges in sheet(s) where your data is stored these should also be specified[E1]. Ideally only data for storage should be included. Roughwork calculations and cross-check calculations should be removed.
13. **Spell-check variable and filenames.** For example: Thallium and Thalium are not the same to an automated data extraction program but human readers can easily identify the similarity
14. **Datasets containing multiple data files are welcomed** as part of a ZIP file (or TAR file on UNIX systems). Each data file in the ZIP file archive should be checked individually that it conforms to the requirements specified in this document.

<i>Date 1</i>	<i>Reading 1</i>	<i>Date 2</i>	<i>Reading 2</i>
01/04/2001	9.71	05/08/1998	< 0.56
21/04/2001	1.16	26/08/1998	0.76
11/05/2001	1.21	16/09/1998	1.67
31/05/2001	9.87	07/10/1998	1.78
20/06/2001	6.47	28/10/1998	1.85
10/07/2001	4.41	18/11/1998	1.28
30/07/2001	4.14	9th-dec-98	0.68
19/08/2001	N/A	n/a	
08/09/2001	No power	n/a	
28/09/2001	3.21	n/a	
18/10/2001	7.08	n/a	
07/11/2001	0.67	n/a	

Figure 3 Example of dataset containing some data formatting errors

WHEN PROVIDING DATA PLEASE DO NOT:

1. Submit data without first having read the instructions above or contact the COE-DMGT in regard to your first use of the COE-Data-Browser.
2. Use Macros within Excel datasets

3. Password-protect the tables of databases (ie tables in Access) or individual spreadsheets in Excel. All password protection should be removed before upload.
4. Embed links to other files that cause a dialogue box to appear when the dataset is open. This is particular relevant to Excel
5. Use HTML controls in Excel datasets – ie drop down lists, radio buttons
6. Assume that any format can be imported and subsequently stored. Please discuss any data storage format requirements with COE-DMGT before submitting data.
7. Do not provide copyright material (ie OS Maps) without proper authorisation and clearance. Such issues are the responsibility of the data owner.

Table 1 ISO 19115 Environmental Metadata Themes and Categories

ISO 19115 Category
<i>Farming: rearing of animals or cultivation of plants.</i>
<i>Biota: naturally occurring flora and fauna.</i>
<i>Boundaries: legal land descriptions.</i>
<i>Climatology/Meteorology/Atmosphere: atmospheric processes and phenomena.</i>
<i>Economy: economic activities or employment.</i>
<i>Elevation: height above or below sea level.</i>
<i>Environment: environmental resources, protection, and conservation.</i>
<i>Geoscientific Information: earth sciences.</i>
<i>Health: health services, human ecology, and safety.</i>
<i>Imagery/Base Maps/Earth Cover: base maps.</i>
<i>Intelligence/Military: military bases, structures, and activities.</i>
<i>Inland Waters: inland water features, drainage systems, and their characteristics.</i>
<i>Location: positional information and services.</i>
<i>Oceans: features and characteristics of salt water bodies excluding inland waters.</i>
<i>Planning Cadastre: land use.</i>
<i>Society: characteristics of societies and cultures</i>
<i>Structure: man-made construction.</i>
<i>Transportation: means and aids for conveying people and goods.</i>
<i>Utilities/Communications: energy, water and waste systems, comms infrastructure.</i>

Accepted Data Files for COE-Data-Browser
<i>Arc View/Arc Info/Arc GIS Files</i>
<i>Binary Data Files</i>
<i>Comma Separate Values Files</i>
<i>DataSet Dependent Format</i>
<i>DBF Files</i>
<i>GML Files</i>
<i>Images (JPEG, PNG, GIF, PS, EPS, CAD)</i>
<i>Microsoft Access Database</i>
<i>Microsoft Excel Spreadsheets</i>
<i>Microsoft Word Documents</i>
<i>MySql Database</i>
<i>Oracle Databases</i>
<i>PDF Files</i>
<i>SAS Data Set</i>
<i>Shape Files</i>
<i>Tab Separated Values</i>
<i>XML Files (or GML)</i>
<i>Zip Files (Tar Files)</i>

Table 2: List of acceptable data formats for COE-Data-Browser

Page: 5

[E1]Not very clear

|