NDP
National Development Plan 2007 - 2013

epa
Environmental Protection Agency
An Ghníomhaireacht um Chaomhnú Comhshaoil

Environmental Research Centre
Report Series No. 13

Managing Environmental Research Data

STRIVE
Environmental Protection Agency Programme
2007-2013

SSTI
Strategy for Science, Technology and Innovation

Comhshaol, Oidhreacht agus Rialtas Áitiúil
Environment, Heritage and Local Government

Environmental Protection Agency

The Environmental Protection Agency (EPA) is a statutory body responsible for protecting the environment in Ireland. We regulate and police activities that might otherwise cause pollution. We ensure there is solid information on environmental trends so that necessary actions are taken. Our priorities are protecting the Irish environment and ensuring that development is sustainable.

The EPA is an independent public body established in July 1993 under the Environmental Protection Agency Act, 1992. Its sponsor in Government is the Department of the Environment, Heritage and Local Government.

OUR RESPONSIBILITIES

LICENSING

We license the following to ensure that their emissions do not endanger human health or harm the environment:

- waste facilities (e.g., landfills, incinerators, waste transfer stations);
- large scale industrial activities (e.g., pharmaceutical manufacturing, cement manufacturing, power plants);
- intensive agriculture;
- the contained use and controlled release of Genetically Modified Organisms (GMOs);
- large petrol storage facilities.

NATIONAL ENVIRONMENTAL ENFORCEMENT

- Conducting over 2,000 audits and inspections of EPA licensed facilities every year.
- Overseeing local authorities' environmental protection responsibilities in the areas of - air, noise, waste, waste-water and water quality.
- Working with local authorities and the Gardaí to stamp out illegal waste activity by co-ordinating a national enforcement network, targeting offenders, conducting investigations and overseeing remediation.
- Prosecuting those who flout environmental law and damage the environment as a result of their actions.

MONITORING, ANALYSING AND REPORTING ON THE ENVIRONMENT

- Monitoring air quality and the quality of rivers, lakes, tidal waters and ground waters; measuring water levels and river flows.
- Independent reporting to inform decision making by national and local government.

REGULATING IRELAND'S GREENHOUSE GAS EMISSIONS

- Quantifying Ireland's emissions of greenhouse gases in the context of our Kyoto commitments.
- Implementing the Emissions Trading Directive, involving over 100 companies who are major generators of carbon dioxide in Ireland.

ENVIRONMENTAL RESEARCH AND DEVELOPMENT

- Co-ordinating research on environmental issues (including air and water quality, climate change, biodiversity, environmental technologies).

STRATEGIC ENVIRONMENTAL ASSESSMENT

- Assessing the impact of plans and programmes on the Irish environment (such as waste management and development plans).

ENVIRONMENTAL PLANNING, EDUCATION AND GUIDANCE

- Providing guidance to the public and to industry on various environmental topics (including licence applications, waste prevention and environmental regulations).
- Generating greater environmental awareness (through environmental television programmes and primary and secondary schools' resource packs).

PROACTIVE WASTE MANAGEMENT

- Promoting waste prevention and minimisation projects through the co-ordination of the National Waste Prevention Programme, including input into the implementation of Producer Responsibility Initiatives.
- Enforcing Regulations such as Waste Electrical and Electronic Equipment (WEEE) and Restriction of Hazardous Substances (RoHS) and substances that deplete the ozone layer.
- Developing a National Hazardous Waste Management Plan to prevent and manage hazardous waste.

MANAGEMENT AND STRUCTURE OF THE EPA

The organisation is managed by a full time Board, consisting of a Director General and four Directors.

The work of the EPA is carried out across four offices:

- Office of Climate, Licensing and Resource Use
- Office of Environmental Enforcement
- Office of Environmental Assessment
- Office of Communications and Corporate Services

The EPA is assisted by an Advisory Committee of twelve members who meet several times a year to discuss issues of concern and offer advice to the Board.

Managing Environmental Research Data

Environmental Research Centre Report

Authors:

Peter Mooney and Adam C. Winstanley

ENVIRONMENTAL PROTECTION AGENCY

An Ghníomhaireacht um Chaomhnú Comhshaoil
PO Box 3000, Johnstown Castle, Co. Wexford, Ireland

Telephone: +353 53 916 0600 Fax: +353 53 916 0699
E-mail: info@epa.ie Website: www.epa.ie

© Environmental Protection Agency 2009

ACKNOWLEDGEMENTS

This report is published as part of the Science, Technology, Research and Innovation for the Environment (STRIVE) Programme 2007–2013. The programme is financed by the Irish Government under the National Development Plan 2007–2013. It is administered on behalf of the Department of the Environment, Heritage and Local Government by the Environmental Protection Agency which has the statutory function of co-ordinating and promoting environmental research.

The authors wish to thank Dr Shane Colgan (EPA and ERC Co-Ordinator) who worked closely on the development of the user interface and data sharing aspects of SAFER-Data, Dr Alice Wemaere (EPA) who collected metadata and data from STRIVE- and ERTDI-funded projects, Conor McDonagh who was co-developer for the very early versions of SAFER-Data. Conor helped in choosing Java and JSP as the implementation language for the software; and the IT Department of the EPA which has been particularily helpful in the set-up, management, and administration of the hardware and software components of the ERC Server.

DISCLAIMER

Although every effort has been made to ensure the accuracy of the material contained in this publication, complete accuracy cannot be guaranteed. Neither the Environmental Protection Agency nor the author(s) accept any responsibility whatsoever for loss or damage occasioned or claimed to have been occasioned, in part or in full, as a consequence of any person acting, or refraining from acting, as a result of a matter contained in this publication. All or part of this publication may be reproduced without further permission, provided the source is acknowledged.

Reports produced through the Environmental Research Centre are intended as contributions to inform policy makers and other stakeholders to the necessary debate on environmental protection.

ENVIRONMENTAL RESEARCH CENTRE PROGRAMME 2007–2013

Published by the Environmental Protection Agency, Ireland

ISBN: 978-1-84095-266-7

Details of Project Partners

Peter Mooney
Environmental Research Centre
Environmental Protection Agency
McCumiskey House
Richview
Clonskeagh Road
Dublin 14
Ireland

Tel.: +353 1 2680181
Fax: +353 1 2680199
E-mail: p.mooney@epa.ie

Adam C. Winstanley
Head of Department
Department of Computer Science
National University of Ireland, Maynooth
Co. Kildare
Ireland

Tel.: +353 1 7083853
Fax: +353 1 7083848
E-mail: adam.winstanley@nuim.ie

Table of Contents

Acknowledgements **ii**

Disclaimer **ii**

Details of Project Partners **iii**

Executive Summary **vii**

1 Introduction **1**

1.1 Importance of Research Data 1

1.2 Review of Similar Approaches 3

1.3 Structure of this Report 4

2 Managing Environmental Research Data **5**

2.1 Organising Scientific Data 5

2.2 What is a Data Resource? 5

2.3 What are Digital Data for Environmental Research? 6

2.4 Documenting Data Resources Using Metadata 7

2.5 Who Provides Metadata? 8

3 The ERC Data Archive Server **9**

3.1 ERC Server Architecture 9

3.2 ERC Server Networking 10

4 The Secure Archive for Environmental Research Data **12**

4.1 Using SAFER-Data during a Funded Project 12

4.2 SAFER-Data Metadata Profile 14

4.3 Data Resource Availability Levels 16

4.4 Uploading Data Sets to the ERC Server Using SAFER-Data 17

4.5 Administration of SAFER-Data 20

4.6 Technical Information for SAFER-Data 20

5 Using SAFER-Data to Access Research Data **22**

5.1 Downloading Data Sets from SAFER-Data 22

5.2 Monitoring Downloads on SAFER-Data 24

5.3 Using Really Simple Syndication Feeds from SAFER-Data 25

5.4 Creating Links to Data Resources on SAFER-Data 26
5.5 Searching for Data Resources on SAFER-Data 27
5.6 Citing Data Resources on SAFER-Data 28
6 Conclusions and Recommendations 30
6.1 Ongoing Development of SAFER-Data 30
6.2 Continued Engagement with Researchers 30
6.3 Introduce a Data Management Plan for Funded Projects 31
6.4 Extend usage of SAFER-Data to Other Agencies 32
6.5 Improve Academic Recognition Strategies for Data Providers 32
6.6 Visitor Download Ticketing System 33
6.7 Final Remarks 34
References 35
Appendix A 37
Appendix B 38

Executive Summary

Environmental science researchers are now using and generating ever-increasing volumes of data and information about our natural world. It is estimated that the Environmental Protection Agency's (EPA's) STRIVE (Science, Technology, Research and Innovation for the Environment) research funding programme will "*involve more than 1,000 researchers and company-based scientists over its seven-year lifetime*"[1]. The EPA's Environmental Research Centre (ERC) expects that large volumes of environmental data and information will be generated by projects funded by STRIVE. One of the key objectives of the STRIVE programme is to make the outcomes and data from this research available "*in a coherent and timely manner which will ensure synergies across the wider research agenda and early availability of these outputs into the formulation of policy*"[2]. Consequently, the STRIVE programme must adopt best international practice in environmental research data management. Management of these 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 irreplaceable; they are always unique particularly 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 that maximum benefits are derived from research data once acquired.

Several European initiatives provide both legislative and policy-based incentives for the management, archiving, and distribution of environmental research data. The European Research Area (ERA) looks to make knowledge sharing more effective amongst universities and research institutions and European citizens. It has become a key reference for research policy in Europe. The ERA is working to overcome the fragmentation of research activities, programmes and policies across Europe. One of the features of this is to provide an "*open and easy access to the public knowledge base*"[3]. The Shared Environmental Information System (SEIS) is "*an approach to modernise and simplify the collection, exchange and use of the data and information required for the design and implementation of environmental policy in Europe*"[4]. While environmental research data are not specifically mentioned in the SEIS, it is well known that "*research results and new scientific findings are critical in the development and assessment of environmental policy and implementation*"[5]. Consequently, environmental research data should be made available to policy makers in a timely manner and without barriers to the access or distribution of these data. The INSPIRE Directive[6] is an initiative that intends to trigger the creation of a European spatial information infrastructure that delivers to the users integrated spatial information services. These services should allow the users to identify and access spatial or geographical information from a wide range of sources, from the local level to the global level, in an interoperable way for a variety of uses. In this way, environmental research data can be part of this infrastructure provided they are distributed in an

1. Ahlstrom, D., 2007. Spreading the seed money. *Science Today Supplement,* p. 17 – *The Irish Times*.
2. EPA, 2007. *Science, Technology, Research and Innovation for the Environment (STRIVE) – an Environmental Protection Agency Programme 2007–2013*. Environmental Protection Agency, Johnstown Castle Estate, Co. Wexford. Ireland.
3. European Commission, 2008. *The European Research Area: New Perspectives.* Technical Report SEC(2007) 412/2 COM(2007)161, Commission of the European Communities, Directorate-General for Research, B-1049 Brussels, Belgium.
4. SEIS, 2008. *Towards a Shared Environmental Information System (SEIS).* Technical Report SEC(2008) 111 COM(2008) 46 final. Commission of the European Communities, Environment Directorate-General, B-1049 Brussels, Belgium.
5. European Parliament, 2002. The sixth community environment action programme (6th EAP). *Official Journal of the European Communities* L242/1–15.
6. European Parliament, 2007. Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Communities* L108/1–13.

interoperable manner and adhere to the various system specifications of INSPIRE. Together these three European initiatives advocate open and barrier-free access to data and information about the environment.

The key step in harnessing the potential of environmental research data is capturing them quickly after they have been collected or generated. There is a risk that the reuse value of the data will be greatly reduced if a significant amount of time passes after the conclusion of the project. Every project funded under the STRIVE programme is obliged to submit all significant data sets and information generated during the project to the EPA at the conclusion of the research project[7]. To support researchers in meeting this requirement, the EPA has developed a large-scale computer system named the ERC Server for the upload, storage, management, dissemination, and long-term preservation of these data resources. This development was assisted by the ERC Postdoctoral Fellowship *Environmental Research Data Management* which ran from 2004 until 2007. One of the key deliverables of this fellowship was the establishment of a software-based data management system for use by the environmental research community. The SAFER-Data system (http://erc.epa.ie/safer) was developed. SAFER-Data is a fully web-based interface to the ERC Server for use by STRIVE-funded researchers and the environmental science community to upload and manage data resources generated during their research. SAFER-Data is also the principal point on the EPA website for the dissemination of environmental research data generated by STRIVE-funded research projects. This report describes the development of the environmental research data archive by the EPA and the development of SAFER-Data to provide web-based access to this system.

SAFER-Data provides a one-stop shop for any stakeholder looking to access data resources generated by EPA-funded research projects. SAFER-Data is more than just another web-based data management system. The report shows that SAFER-Data has very quickly established itself as an important part of the STRIVE programme's support infrastructure while also establishing itself as an important source of data within the environmental science research community. The report also shows that data from SAFER-Data have been downloaded by visitors to the system from all over the world. SAFER-Data is more than just a web-based system. Rather, it is the realisation of the research data management vision of the EPA and the STRIVE programme. SAFER-Data emphasises the philosophy of open and barrier-free access to environmental research data to any stakeholder. The vision of STRIVE to "*disseminate the findings of individual research projects and the overall programme to the widest possible audience in a coherent and timely manner*"[2] is now possible. Finally, this work on SAFER-Data provides a foundation for the EPA and the ERC to continue the development of systems and tools for environmental data management in the EPA which will assist in meeting the obligations of European Directives such as INSPIRE and the SEIS.

While this report is mostly focused on describing the key outcomes of the postdoctoral fellowship, it also serves as a non-technical guide to environmental research data management within the EPA. The report concludes with a number of recommendations to the EPA in regard to the further development of SAFER-Data, the development of a robust data policy for environmental research projects funded by the EPA, and continued proactive engagement with the Irish environmental research community.

7. STRIVE, 2008. Submission of data sets from STRIVE-funded projects. STRIVE Funding Guidelines. Available online at http://www.epa.ie/researchandeducation/research/docs/#datasets.

1 Introduction

The STRIVE (Science, Technology, Research and Innovation for the Environment) programme 2007–2013 follows on from the ERTDI (Environmental Research Technological Development and Innovation) programme 2000–2006. STRIVE has a funding budget of approximately 100 million euro and foresees investment in research and innovation in areas such as environmental technologies and sustainability development. One of the key aims of the STRIVE programme is to build on the success of the ERTDI programme while "*contributing to a better environment by delivering applicable and relevant datasets, information, and knowledge*" (EPA, 2007b). In addition to this aim, STRIVE also aims to "*disseminate the findings of individual research projects and the overall programme to the widest possible audience in a coherent and timely manner*". It is estimated that STRIVE will "*involve more than 1,000 researchers and company-based scientists over its seven-year lifetime*" (Ahlstrom, 2007). To "*make details of research available in a coherent and timely manner to ensure synergies across the wider research agenda and to ensure the earliest possible dissemination into the formulation of policy*" (EPA, 2007b), it is crucial that the STRIVE programme adopts best international practice in environmental research data management. The advent of the Internet has transformed the way that science and scholarly research is communicated. Indicative of this changing landscape has been the steady growth in open access publishing and archiving which facilitate widespread diffusion and free digital access to publications and the latest scientific discoveries (CIHR, 2008). Open access enables researchers to make their research results freely accessible and useable for the international research community thereby enhancing the application of research results.

1.1 Importance of Research Data

The importance of scientific research data and knowledge cannot be underestimated. Figure 1.1 shows an example of the life cycle of environmental research data. Data are collected in the field, laboratory, or using other data collection/generation means by researchers. These data are then used in data analysis and modelling where research outputs

Figure 1.1. An example of the life cycle of research data.

are produced. These outputs can be used for various types of reporting, e.g. reporting to the EPA, reporting for Local Authorities, or in some cases reporting to support European reporting obligations such as the Water Framework Directive (WFD). Finally, reporting assists in environmental policy development and decision making. In turn, policy development and decision making help form new questions which feed back directly to the environmental research agenda. Any barriers or interruptions to the flow of data from collection through to analysis through to reporting can delay or disrupt the decision-making or policy-making processes. Some authors state that "*alongside human capital and instrumental capital goods, digital research data are developing into a third stream of scientific capital*" (Arzberger *et al*., 2004).

Previous EPA research funding programmes did not have the same emphasis that the STRIVE programme has on data management for the environmental research data and information generated. Many funded researchers did not submit their research project data to the EPA as a result. A relatively small amount of data was actually collected by the EPA at the end of projects. Unfortunately much of the data which were not collected from completed projects are now essentially all but lost. One can speculate that these data resources have simply disappeared into the "*grey dusty archives*" (GISFORM, 2002). These "*grey dusty archives*" take many forms: data sets stored on CDs which are now left unmarked in a filling cabinet; databases left dormant on now decommissioned computers or network drives; or data whose only digital remains are printed tables within PDF reports. The 'grey literature' within these archives is also a very important source of knowledge and information. As Alberani *et al*. (1990) explain this grey literature can include publications that are non-conventional, fugitive, and sometimes ephemeral. They may include, but are not limited to, reports (pre-prints, preliminary progress and advanced reports, technical reports, statistical reports, memoranda, state-of-the art reports, market research reports, etc.), theses, conference proceedings, technical specifications and standards, non-commercial translations, bibliographies, technical and commercial documentation, and official documents not published commercially (primarily government reports and documents). Often these publications assist other researchers in better understanding a research methodology or research data set. If these data resources (both the data sets and the publications) are not included in a dedicated data archive where they can be properly documented, managed, and indexed, then an important source of research knowledge is lost. Search engines such as Google can allow people to find these data resources in Google web searches. However, this is a hit-and-miss approach. In search results for Google or LiveSearch, users are given no scientific metadata corresponding to the data resources and, consequently, it can be very difficult to assess the fitness for use or fitness for purpose of the data resources without downloading them and examining them in closer detail.

The scientific community is now at a critical time point in the Digital Age. The key data resource outputs from scientific research must be captured quickly after the conclusion of the originating project and properly managed within a suitable data management archival system. Researchers are now using and generating ever-increasing volumes of data about our world. Ever-more common and powerful digital instruments and sensors churn out more data in a single session than a human being could deal with in a whole lifetime. By some projections, in 2010, "*there will be more data being generated [annually] than has ever been generated in human history up to 2006*" (Burton, 2007). Today much of the data currently being collected are 'born digital' and lack any analogue or paper counterpart (NSB, 2005). Other data resources are being converted to digital forms and, in the process, are often disconnected from their analogue representations. Digital environmental research data, and the investments made in gathering them, could be lost unless a robust preservation and management plan is created and implemented. Funding agencies have an important role to play in helping to establish a data management infrastructure that fosters closer collaboration between data collectors, data owners, data generators, and data managers. The STRIVE programme must attach critical importance to the issue of data collection from funded environmental research. This is reflected in recommendations from other countries such as Finland where Kuula and Borg (2008) recommend that research funders support open

access to data by recommending or requiring that the data collected with their funding will be made available for the use of the scientific community after the original research has been completed and by supporting this recommendation with guidance of data management, implementation of a binding data policy, and providing assistance to projects on various aspects of data management.

1.2 Review of Similar Approaches

It is useful at this point in this report to give a brief overview of systems for data management implemented by other international research institutes and government agencies. It is not necessary to analyse the types of software or hardware used to develop and run these systems. Such a discussion is beyond the scope of this report. It is more important to highlight how these institutes and agencies provide web-based access to large archives of environmental data and information. There are many web-based systems offering access to various archives of environmental data. The systems discussed below are just a selection of such systems. The URLs for the web locations for these systems are available in the References section of this report.

The European Soil Data Center (ESDAC) provides a web-based listing of all data sets currently available from ESDAC. Some of them are freely available for download while other are restricted due to copyright rules (JRC, 2008). ESDAC does not provide a search interface. Users must work through the categorical listing of the various data sets using the web links provided.

PortalU is the German Environmental Information Portal. PortalU offers centralised access to over 2,000,000 web pages and about 500,000 database entries from public organisations in Germany. PortalU (PortalU, 2008) is a collaboration of the German 'Lnder' and the German Federal Government. More than 200 public agencies currently offer environmental information in PortalU. PortalU does not actually store data sets or make them available from its website. Rather it maintains a very comprehensive database of metadata. The core component of PortalU is a powerful search engine. All of the environmental metadata information can be searched through the PortalU interface using specially designed search interfaces. The search query can be specified in the 'Advanced Search' mode by semantical, spatial and temporal limitations.

EDMERP (EDMERP, 2008) is a European directory of research projects relating to the marine environment. EDMERP describes more than 1,250 research projects, from a wide range of disciplines, from over 300 research institutes from 30 countries across Europe. Research projects are catalogued in a special metadata profile in EDMERP as fact sheets or abstracts with their most relevant aspects. The primary objective of EDMERP is to support users in finding interesting research activities and in connecting them to involved research managers and project results such as data, models, publications, etc., across Europe. Similar to PortalU the EDMERP system does not make data sets available directly. In the special EDMERP metadata, web links and other contact information are provided which allow users to go directly to where the data are available.

The European Register of Marine Species (ERMS) is an authoritative taxonomic list of species occurring in the European marine environment. The register is actively maintained and daily updated in the framework of the MarBEF EU Network of Excellence (MarBEF, 2008) by a board of taxonomic editors, who are world experts on the taxonomy of their relevant taxa. ERMS maintains metadata in a relational database. The database can be downloaded by applying for special access to download. The ERMS does not actively store data on its servers. Rather, it actively updates linkages within the metadata to online data repositories and data sets.

The Economic and Social Data Service (ESDS) in the UK is a national data archiving and dissemination service which came into operation in January 2003 (ESDS, 2008). The service is a jointly funded initiative sponsored by the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC). The ESDS is a distributed service, based on a collaboration between four key centres of expertise: the UK Data Archive (UKDA), the Institute for Social and Economic Research (ISER), Manchester InforMation and Associated Services

(MIMAS), and the Cathie Marsh Centre for Census and Survey Research (CCSR). These centres work collaboratively to provide preservation, dissemination, user support and training for an extensive range of key economic and social data, both quantitative and qualitative, spanning many disciplines and themes. The ESDS provides an integrated service offering enhanced support for the secondary use of data across the research, learning and teaching communities. The ESDS is similar to the Environmental Research Centre (ERC) in that the it stores data sets on its server systems distributed amongst the four locations mentioned above. Users must be authenticated using their university or research institute login permissions. The ESDS also provides data management support for researchers.

1.3 Structure of this Report

The Terms of Reference for this fellowship were directed at the development of a web-based interface to the ERC's research data archive. Chapter 2 discusses some key software concepts of how environmental research data are now managed on the ERC Server systems. The ERC research data archive (ERC Server) is a large storage capacity networked computer system where data collected from EPA-funded projects are stored. Chapter 3 gives a brief overview of the physical computer architecture of the ERC Server. This includes a discussion of the concept of a data resource and the importance of metadata. Chapter 4 introduces the Secure Archive for Environmental Research Data (SAFER-Data) and describes how SAFER-Data is used by researchers and data providers to submit research data to the EPA. Chapter 5 describes the usage of SAFER-Data from the point of view of users who wish to download environmental research data from the ERC Server. Chapter 6 outlines several recommendations for future work in the area of environmental research data management at the ERC. Environmental research data management is now a core activity for the ERC and this report argues that this must remain the case within the STRIVE programme and beyond. The recommendations outlined describe ways in which the ERC and SAFER-Data can develop and grow based on their current success. The report also contains two appendices. Appendix A contains a detailed breakdown of the number of downloads from SAFER-Data in 2008, which was its first full year of operation. Appendix B contains a bibliography of the publications that resulted from the research work in the design, development, and management of SAFER-Data.

2 Managing Environmental Research Data

Before the description of web-based access to the environmental research data on the ERC Server in Chapter 4, it is useful to first provide background to some key concepts in data management for this project. It is clear upon reading the overview of the STRIVE programme that a very wide range of scientific disciplines make up the environmental science community in Ireland. This diverse community includes social science, economics, engineering, physical sciences, computing and ICT, health, and planning. This heterogeneity is coupled with the structure of research funding and academia. In the STRIVE programme several different types of funding exist, including large scale, medium scale, fellowship, PhD scholarship, and desk study. There is a crossover between scientific disciplines, co-operation between research institutions, and in some cases international collaborations. Building a system to manage the data and information generated by such a diverse community is a major undertaking. It should be obvious to the reader that it is not feasible to build a system that imposes a 'one-data-format-fits-all' solution for environmental research data management. Well-established academic and industry standard data formats and data collection regimes are in place in many disciplines. Examples of these include the Recorder format for biodiversity, the GRIB (gridded binary) format for meteorological modelling, ShapeFile and dBase format (DBF) for geographical information sciences. Consequently, it makes sense to encourage the members of the research community to continue using the established standards for data collection and data formatting within their scientific discipline.

2.1 Organising Scientific Data

One of the key problems encountered by researchers when they search for scientific data for analysis or comparison is that it can be very difficult to find details of such data using search engines, library catalogues, etc. In some cases the researcher may know that a certain data set exists or is available but he or she does not have the link to the source or current location of the data. This can lead to frustration and a loss of valuable research time. Very often the data sets do exist and are available for access but the lack of proper data storage and data documentation for these data sets means that they cannot be accessed by search engines, catalogue servers, or metadata catalogue harvesters. As a publicly funded agency, the EPA has a fundamental interest in ensuring that the knowledge and findings that result from the research it funds, including research publications and publication-related data, are available to the widest possible audience, and at the earliest possible opportunity. Advancements and decision making in environmental policy and environmental science are made possible through widespread and barrier-free access to cutting-edge research and knowledge, enabling scientists, policy makers, and the public to use and build upon these data and knowledge. The common denominator amongst every environmental research project is **metadata**. Metadata are 'data about data' and are the first step towards barrier-free access to data. Metadata describe **what** the data and research mean, **why** this research was carried out, **where** the data were collected, **how** the data were collected and **how and when** the research was carried out and any other information that would be useful to help another scientist or potential user of the data understand the data better. This chapter describes the concept of metadata, data resources, and general data management for the ERC Server.

2.2 What is a Data Resource?

As outlined above, a single data format does not exist for all of environmental science. To understand the key concepts behind the management of environmental research data by the ERC it is necessary to explain the term 'data resource' in greater detail. For a given project, a data resource can be thought of as a composite object. It is the combination of metadata and the actual data sets and/or other information generated by that research project. Put simply **metadata** + **digital data** or **digital information** = a **data resource**. This concept is illustrated in Fig. 2.1. The rectangular box to the left of the figure is a **data**

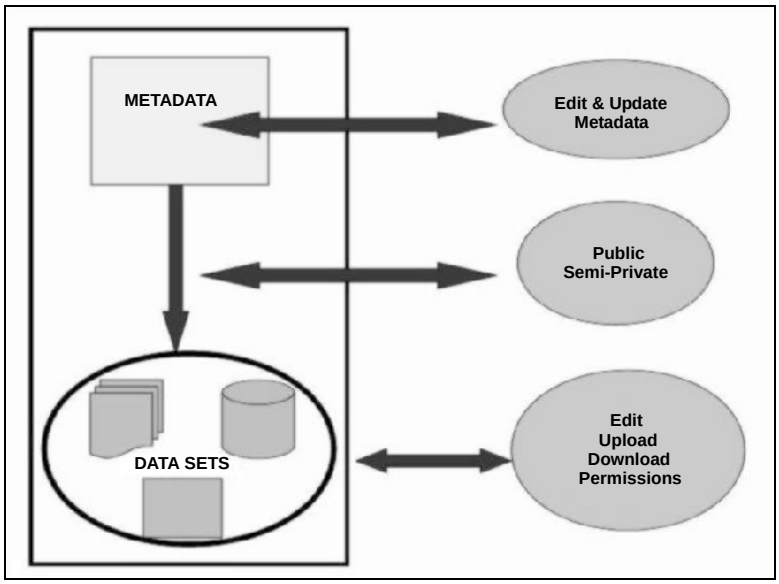

Figure 2.1. An illustration of the concept of a data resource within SAFER-Data.

resource. At its minimal definition, a data resource contains only metadata. Consequently, data sets and other digital information objects may be added to this resource at a later stage. This definition of a data resource allows each project to choose the most appropriate data formats and information management techniques for the data it shall generate and/or collect during the course of the project.

The text bubbles to the right of Fig. 2.1 describe the type of functionality that is available to data resource owners when using SAFER-Data. This shall be described in greater detail in Chapter 4. The concept of a data resource is very important to the philosophy of SAFER-Data. As explained above, a data resource is a composite object. In this way, the data and information are inextricably linked with the metadata. Digital data and digital information submitted from the research projects cannot be stored on the ERC Server using SAFER-Data without being linked to a corresponding metadata description.

2.3 What are Digital Data for Environmental Research?

In the broad field of environmental research every project funded by the EPA is different in some way. Some projects carry out field-based measurements, others carry out simulations on computers, others work in lab-based environments, others are desk-based studies, etc. Every project produces different output. The output from the majority of funded projects is data sets, i.e. actual measurements of some environmental variable(s) captured in the field, lab, or simulations. Other projects may output reports that synthesise information and data from other studies in the area. Questionnaires and surveys are outputs from other projects. Finally, some projects may output cartographical data and maps using a Geographical Information System (GIS). Given that there are many different outputs, it is generally easier to refer to these as digital data sets or digital information than to use their full technical descriptions. The types of digital data or digital information managed by SAFER-Data include:

- **Data sets:** actual measurements of some environmental variable captured in the field, lab, or simulations
- **Reports:** theses, discussion documents, synthesis reports, end of project reports, commentary, literature reviews, user manuals, standard operating procedures, etc.
- **Qualitative data:** output from surveys, derived data products from survey data, written transcripts from interviews, audio, video, photography

- **Software for analysing environmental research data:** software code, software libraries, simulation models expressed in software

- **Imagery:** photographs, simulated imagery, remotely sensed images, aerial photography, etc.

- **Cartographical output or GIS output:** GIS map layers, data sets with geographical objects and geographical co-ordinates, electronic maps.

As stated earlier in this chapter it is not feasible to expect a one-data-format-fits-all solution to work for a scientific field as broad as environmental science. Finding a software-based data format which could be specified in such a way that it could be used to store or manipulate any type of data from environmental science is impossible. The ERC employs a simple method for determining the best software format for data collected or generated by the funded research projects. The ERC requires researchers to identify the format or structure for their data and information which is internationally accepted by their peers and in their research community. If a particular project stores all of its data in XML file format and this format is recognised as the international standard or *de-facto* standard for this research area, then the ERC is satisfied to accept data in this particular format. In some cases, if the ERC believes that converting a particular data set or set of data to a different format will assist in making the data more accessible to a wider range of users, then this will be suggested to the researcher or data owner. The ERC will also advise the researchers if the format their data or information is stored in should be migrated to newer technologies. This is an important data curation step and should happen quickly upon receipt of the data. It is unwise to wait for storage media to become obsolete or data formats to become dormant and unsupported. The ERC will encourage researchers to migrate to new technologies and formats as they become available.

2.4 Documenting Data Resources Using Metadata

The concept of metadata was introduced to provide orientation in a space of continuously growing data and information resources (Pillmann *et al.*, 2006). Metadata provide information about the data but do not include the data themselves. In the majority of cases the metadata accompanying a geospatial data resource are represented within a tabular structure in a separate digital file. They may also be represented as rows within a database table (or a set of related tables). The metadata provide an explanation of what the data resources represent, what the map layers available in the web GIS actually represent, etc. The concept of metadata is encountered by people many times during their normal daily routine. The world around us has a vast number of examples of where metadata are used:

- **Digital Photographs:** Metadata for a photograph would typically include the date and time at which the photograph was taken. They would also include details of the camera settings used for that photograph (such as focal length, aperture, exposure). Many digital cameras record these metadata in exchangeable image file format (EXIF) which can be read by most digital photograph software programs.

- **MP3 Audio File:** An MP3 audio file contains many items in its metadata. They can include the album name, the song title, genre, year, composer, contributing artist, track number. They can also include non-text metadata such as a digital image representing the album art.

- **Food package labelling:** The label on any food package usually includes a list of ingredients, the process used in manufacture, the amount of an ingredient that is named or associated with the food, and an appropriate durability indication.

- **Books available in an online store:** The labelling or metadata involved for books is familiar to most people. This would include the author(s) and title, the publisher, and ISBN number. In online shops, these metadata can often be supplemented with customer reviews and comments on the book or product.

All of these everyday examples of metadata share two common characteristics. Firstly, they allow the potential user of the item to make a fitness-for-purpose and fitness-for-use evaluation. By reading the metadata closely users can decide if this book is the exact book they are looking for or is a book they would

like to read, the food about to be bought does not contain ingredients to which the user may be allergic, for example. In the case of food packaging the so-called metadata on the labels allow us, the consumers, to evaluate the food authenticity and whether a food matches its description. Secondly, these metadata allow these products to be organised, classified, sorted, and searched. A collection of digital photographs can be searched by organising the photographs chronologically. MP3 audio files can be sorted and grouped by artist or genre. Without the structure and classification that metadata bring, searching and finding the products you require would almost be impossible.

Waller and Sharpe (2006) remark that the importance of metadata cannot be overestimated for geographic or scientific data. The authors state that metadata add "*a whole new dimension, providing extra richness of contextual or descriptive information at the point of access*". The perception amongst many scientists and creators/maintainers of geospatial data is that the creation and maintenance of geospatial metadata are laborious and unnecessary data management tasks. As the volume of geospatial data about the environment continues to grow so too does the need to properly document these data resources with metadata. An example of metadata in practical usage is in the cards within a library catalogue. These cards give details about the book: the author(s), the publishers, the publishing dates, etc. Without physically accessing the book, readers of the catalogue card can decide if the book is of interest to them or suitable for their research needs. This example extends to online bookstores such as Amazon. Users can browse the online book catalogue and read these same metadata about a given book. The metadata in an online catalogue are often supplemented with additional metadata items such as user comments and testimonials. Again the user does not obtain physical access to the book but is provided with sufficient information on the book to make a decision to purchase or not. This is very similar to the concept of making a decision on whether a data set or data resource is suitable (fit for purpose) for a particular application. If good quality and informative metadata are provided with a data resource then third-party users can make a decision on whether the data resource is suitable for their application or research without actually downloading or accessing the data sets.

2.5 Who Provides Metadata?

Without metadata, environmental research data resources may lie dormant and undiscovered on the Internet. This invisibility may give rise to duplication of effort in creating, accessing, and managing these data resources. The INSPIRE Directive (European Parliament, 2007) states that "*geospatial data must be managed as close to the source as possible*". In relation to this, the representation of key knowledge about these data resources in the metadata must be performed initially as close to the original data creator source or scientific expert group as possible. This vital specialist knowledge provided by the original creators or owners of the data resource can be the difference between the data resource being widely disseminated as opposed to lying in "*dusty archives and grey literature*" (Wilson, 2002) or lost within laboratory or field notebooks. Metadata are also used to provide users with search facilities on collections of metadata. When a user types in a search term or keyword(s) the various fields of the metadata information are searched for appropriate matches for this search query. More details about the metadata profile used by the ERC is given in Section 4.2.1.

3 The ERC Data Archive Server

A computer server system is required in order to store data and information generated by funded research projects. The ERC Research Data Archive (from this point referred to as the ERC Server) is a computer server system with approximately 2 terabytes of storage capacity. The term 'ERC Server' describes the combination of a hardware system and the software applications which together provide the EPA with a data management system for environmental research data. The ERC Server provides storage and hardware infrastructure for data and information generated from research carried out on EPA-funded research programmes. Data sets and other files are organised into folders on the ERC Server. The ERC Server is exclusively used for this purpose. More formally the ERC Server is a Storage Area Network (SAN) system. The SAN architecture is highly fault tolerant which greatly improves the possibility of 24/7 access to the ERC Server. The fault-tolerant nature of the SAN means that if a certain hardware component failed (i.e. a hard disk) then there is a means by which the hardware can continue operating without causing the SAN to crash and ultimately become unavailable. This section gives an overview of the ERC Server from a hardware perspective. There is a web-based front end to the ERC Server called SAFER-Data and it is described in Chapter 4.

3.1 ERC Server Architecture

A full low-level technical description of the ERC Server components is beyond the scope of this report. As mentioned above the ERC Server is a SAN system. The SAN is comprised of two server computers connected using fibre-optic cabling in a fault-tolerant configuration to a very large array of hard disks. Special disk management software running on the ERC Server configures this large array of hard disks to appear as several large-capacity logical drives to other software running on the ERC Server. Figure 3.1 shows the configuration of the individual hard disks which are linked together to form the 2-terabyte storage capacity of the ERC Server. The current hardware configuration

Figure 3.1. Individual hard disks in a disk array connected together.

of the SAN will allow the ERC Server to grow to a storage capacity of 6 terabytes.

Figure 3.2 shows a photograph of a row of server cabinets within the secure facility where the ERC Server is hosted. In addition to the ERC Server there are other EPA server systems managed at this facility. Back-up of data on the ERC server is accomplished using a large capacity tape library connected directly to the SAN. Figure 3.3 shows this large-capacity tape library used for creating back-ups of all data and information on the ERC Server. This tape library contains a tape carousel which allows back-ups to happen in an automated fashion. The ERC and EPA have established a robust back-up policy for the ERC Server. This involves frequent back-ups of all data sets and associated information. Back-up tapes are regularly removed to secure storage off-site in order to

Figure 3.2. A row of server cabinets in the secure facility where the ERC Server is hosted.

Figure 3.3. The ERC Server's high-capacity tape library used for data back-ups.

maximise the security of the back-up procedure. As an added security and management measure, selected IT staff in the ERC and EPA can access and monitor the ERC remotely over a Virtual Private Network (VPN) connection. Physical walk-in access to the hosting facility where the ERC Server is located is strictly controlled by HEA-NET. Only authorised ERC and EPA IT staff can apply for and be permitted walk-in access to the ERC Server.

3.2 ERC Server Networking

The ERC Server is expected to manage a very large volume of environmental research data over the course of the STRIVE programme and beyond. Some funded projects will generate data sets that are expected to be very large while other projects will generate large quantities of relatively small-sized data sets. In both cases, a reliable networking infrastructure

is required to provide the research community with a high-quality upload and download service. The ERC Server is physically located within the hosting facilities of the HEA-NET Managed Network Services and consequently the ERC Server is connected directly to the HEA-NET network backbone. This can be visualised on a map on the HEA-NET website at the location http://www.hea.net. This connection ensures very efficient network performance for the ERC Server, particularly when the ERC Server is accessed from other HEA-NET networks such as Irish university and college campus networks. To provide Internet-based access to data sets and information stored on the ERC Server, a web server (Apache Web Server) and an application server (Tomcat Application Server) are installed and used. The ERC Server is accessible on the URL http://erc.epa.ie and it is also accessible from the Research section of the EPA website at http://www.epa.ie/researchandeducation/. Web-based access to the ERC Server using SAFER-Data is described in more detail in Chapter 4.

4 The Secure Archive for Environmental Research Data

In Chapter 3, the hardware components of the ERC Server were described. This chapter describes the principal software component of the ERC Server: the Secure Archive for Environmental Research Data (SAFER-Data). SAFER-Data is the web-based front end to the ERC Server and is available at the URL http://erc.epa.ie/safer. SAFER-Data is the result of a number of years of research and development into web-based systems for metadata collection, metadata cataloguing, online data set file management, and public access to data. SAFER-Data was formally launched in May 2007. Figure 4.1 shows a screenshot of the homepage or front page of SAFER-Data. The name SAFER-Data or 'SAFER' is also important from the point of view of advertising the system and for public and user perception of the system. The system is easily identifiable and recognised by its name. The name also conveys the emphasis the system places on making the data on the ERC Server safe and secure. It also conveys the meaning that research data are safer when they are managed by a system such as SAFER-Data, where appropriate measures are taken to ensure long-term access to data and long-term preservation of those data.

4.1 Using SAFER-Data during a Funded Project

SAFER-Data is designed to complement the requirements outlined in the EPA's Research Grant Award Documentation Terms and Conditions which specify the requirements for the submission of metadata and data at the completion of a STRIVE-funded project. Figure 4.2 shows a flow chart for the processes involved for projects to submit generated data to the EPA using SAFER-Data. Figure 2.1 illustrates the types of actions that a data owner can perform on a data resource. The actions are displayed within the bubbles on the right of the figure. When a data resource is created both the metadata and the connected data sets can be edited, updated, replaced, and managed at any time in the future. The steps in Fig. 4.2 are now outlined in detail as follows:

1. When a project is awarded funding from STRIVE, one of the first tasks for the project manager should be to allocate responsibility for the management of data generated or collected by

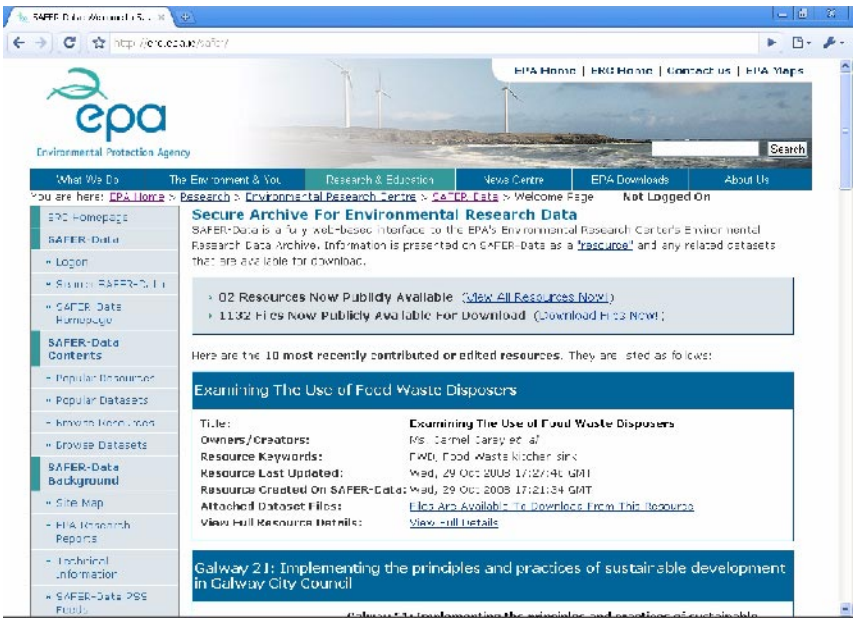

Figure 4.1. Screenshot of SAFER-Data within a web browser window.

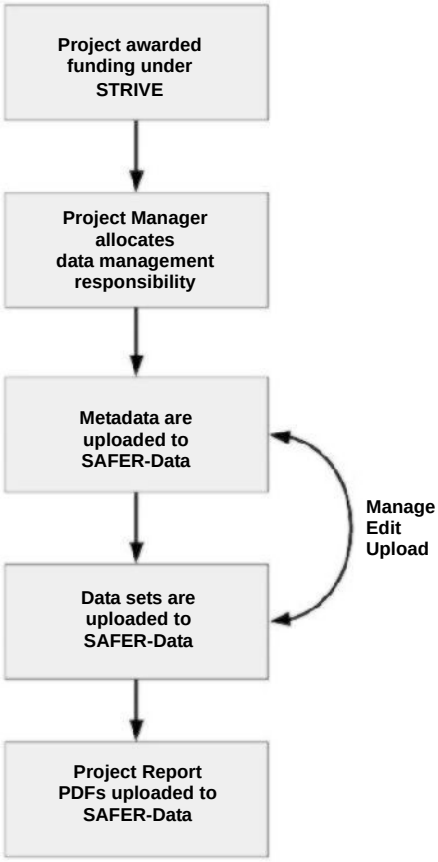

Figure 4.2. The steps involved in submitting data generated in research projects to the EPA.

the project. This individual or group of individuals should agree on the format(s) data will be collected in, what new data will be generated, what data will be submitted to the EPA at the conclusion of the project, and what barriers exist that will prevent public access to the data collected or generated by the project.

2. When the project is coming towards a conclusion, the project must create metadata for the project data on SAFER-Data. The ERC will create a new account on SAFER-Data for the project and supply the project manager with the username and password for this account. Individuals directly involved in the project can use this account. For larger projects, the metadata creation should be carried out collaboratively amongst the relevant project partners. The metadata become fully publicly available immediately and are immediately accessible from SAFER-Data. Figure 4.3 shows a partial screenshot of the metadata entry form on SAFER-Data. The form contains only standard HTML form elements. There are 22 fields, seven of which are selected using drop-down-list selections.

3. When metadata have been created, the next step at the conclusion of the project is to upload the most significant data and information resources generated by the project. Initially these data and information resources are hidden from public access and are not available for download. Twelve months after the conclusion of the project these data and information resources automatically become publicly available for download. The uploaded data can be changed and updated at any point in the future by logging onto SAFER-Data using the account details used in this and previous steps. When metadata and data sets are uploaded to SAFER-Data, the owner of this resource can change the content of the metadata and update/replace data sets at any point in the future. This is accomplished using the 'Attachment Management Interface' on SAFER-Data. A screenshot of this interface is shown in Fig. 4.4. The owner of the data resources can choose the availability of each individual file by choosing from the drop-down list.

4. When the final report for the project has been published on the EPA website, other PDF versions of the report are also uploaded to SAFER-Data. They are attached to the metadata in Step 2. These reports are usually uploaded by EPA or ERC staff. When uploaded these alternative versions appear in SAFER/Reports (http://erc.epa.ie/safer/reports). This is the special section of SAFER-Data which displays only research reports. The following versions of the report are usually available:

- **Print-quality version:** For high-quality printing to hardcopy
- **Web-quality version:** This PDF is hyperlinked and usually is smaller in file size to the print-quality version therefore allowing a quicker download.
- **End of Project Report:** The End of Project Report is usually a longer and more technically detailed

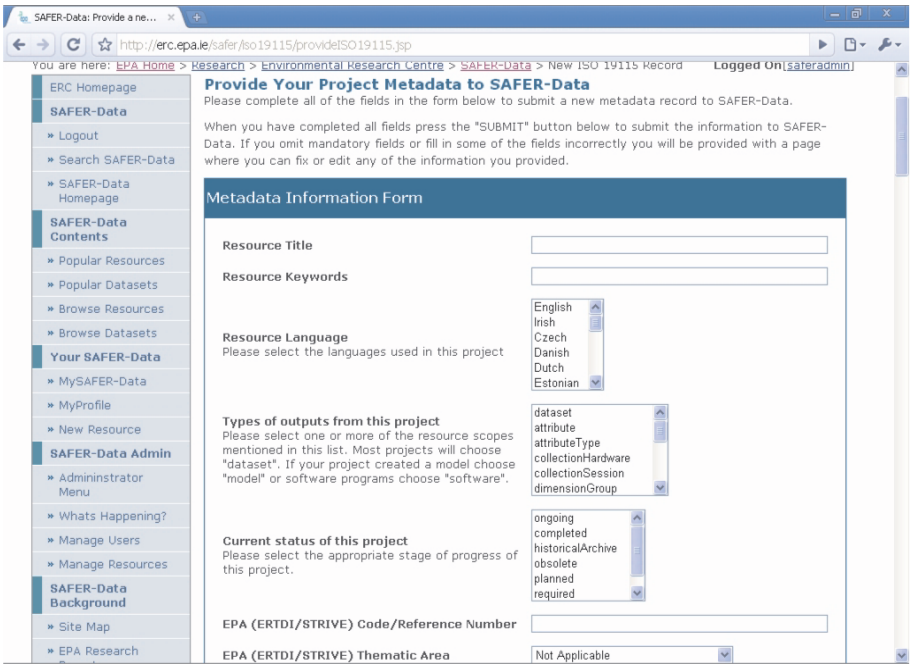

Figure 4.3. Screenshot of the metadata capture form on SAFER-Data.

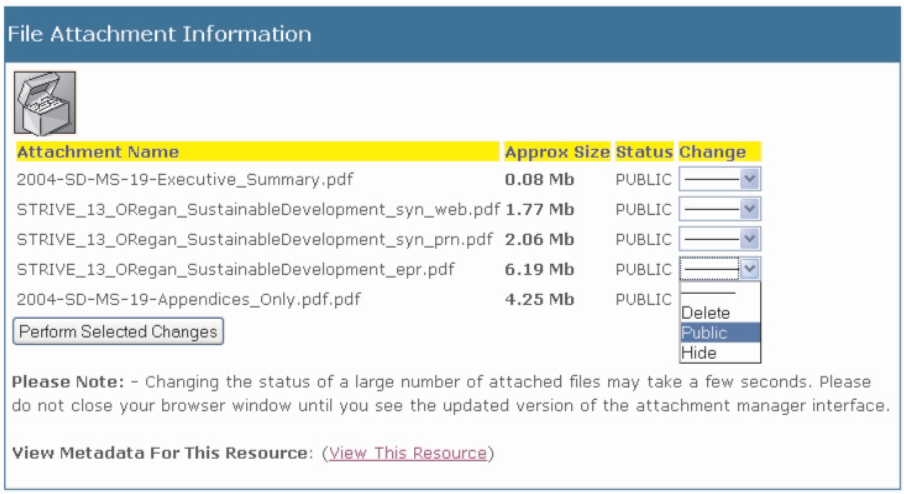

Figure 4.4. Screenshot of how the public availability of data resources on SAFER-Data are changed using the Attachment Management Interface.

report than the final project report available from the EPA website. It usually contains additional appendices and a full bibliography.

4.2 SAFER-Data Metadata Profile

Metadata come in many different 'flavours'. Some metadata standards are more suitable for social sciences while others are designed for library and information science. An example of such a metadata standard is the Dublin Core metadata standard. The ISO 19115 metadata standard is a vast and complex standard defining more than 400 metadata elements and 20 core elements and is suitable for the documentation of spatially referenced data resources and data services. The ERC Research Data metadata profile uses only a subset of the full number of these elements. This subset is outlined in the INSPIRE Metadata Implementation Rules and comprises the

core metadata elements required to properly identify the data set or data resource. The more comprehensive the metadata description provided by the data owners the better third-party users can understand the data resource without ambiguity. The SAFER-Data system collects ISO 19115-compliant metadata. Authorised users (data providers) provide metadata through a series of web-based forms. This removes the burden of having to learn the complexities of the ISO 19115 metadata standard. The following is a brief overview of the key elements in the metadata subset:

- **Who:** Who created the data set? Who owns the data set? Who is responsible for the maintenance and future updates to the data sets?
- **Where:** Where can one obtain access to the data set? Where is the geographical areas or regions represented within the data set resource?
- **When:** When is the data set available for access? What are the date and time ranges (time series) represented within the data set resource?
- **What:** What topic categories does the data set resource exist in? What parameters are measured in the data set? What instrumentation was used? What data quality procedures were applied to the data set?
- **Why:** Why was the data set resource created? Why was a particular geographical study region chosen?

4.2.1 Filling in metadata on SAFER-Data

Metadata technologies and their many formats have been developed in fields such as library science, broadcasting, physical sciences, etc., and are in practical use. Digital content having corresponding metadata are easier to manage and discover (Atarashi *et al.*, 2003) than non-metadata digital content and resources. With the growth of the Internet, metadata have taken on a much more important function in the efficient delivery and dissemination of data resources. Some metadata information can be generated automatically by computers. But the most important aspect of metadata is the capture of information about the digital resource it describes provided by the original owner or creator of that resource. This requires a certain amount of manual input from the original owner or creator whereby they usually are required to manually input information (Shuming *et al.,* 2008). Consequently, metadata has become a term conveniently ignored or avoided by those required to provide or manage data. As a result, large collections of environmental and other geospatial data become 'data tombs' (Han and Kamber, 2006) seldom visited or maintained, key data sets never emerging from "*grey dusty archives or slowly rotting*" (GISFORM, 2002) because users do not know of their existence Metadata creation is seen as 'boring' (Comber *et al.*, 2005) and is usually postponed for as long as possible by most researchers (Mooney and Winstanley, 2007). SAFER-Data attempts to make the task of metadata provision simple and quick for data owners. A partial screenshot of the metadata form in SAFER-Data is shown in Fig. 4.4. There are 10 metadata fields that must be manually filled in by data owners using the metadata provision form in Fig. 4.4. These are outlined as follows:

1. **Resource Title:** The title of the data resource
2. **Resource Keywords:** Several distinct keywords that will help to identify this data resource
3. **Resource Code:** For EPA-funded projects this is an identification code or reference number unique to the project that created this data resource
4. **Resource Temporal Information:** These are the approximate dates (accuracy to 1 month) when the project commenced and the earliest and most recent dates/times recorded referenced by this data resource. If the data set(s) connected to this data resource are not time series based, then the date spans of the earliest and most recent work/analysis on the data set(s) are provided. The temporal information for a resource is selected using a pop-up calendar from which the data owner specifies the required date.
5. **Resource Geographical Description:** This field allows data owners to describe in free text the geographical area(s) studied or analysed in their project and represented within the connected data sets. Data owners are encouraged to be as

verbose as necessary in this field. One can include details of place names, local geographical features, travel and directional information, geographical co-ordinates, etc.

6. **Resource Data Usage Limitations:** The data owner can specify in exact detail the nature of any limitation imposed upon third parties who access and use the data sets connected to this data resource. This should include any special access requirements, licenses, costs, etc., that should be considered before third-party access and usage. One can also specify applications for which these data are not suitable.

7. **Resource Lineage:** Essentially this field specifies 'why' this data resource was created. What were the conditions that caused this project to be funded? What environmental events, gaps in scientific knowledge, national or international environmental legislation or directives, etc., are addressed by the existence of this data resource. This field should contain a general explanation of the data owner's knowledge of the impact of the data resource within the environmental science community.

8. **Resource Abstract:** This field closely resembles the abstract field in a journal paper or report. Data owners are encouraged to be as verbose as necessary. Ideally the abstract should provide an overview of the data resources and the scientific work that created them.

9. **Resource Links:** This field allows data owners to supply a set of hyperlinks to related internet services or websites. These services or websites may offer additional information or data related to this data resource.

10. **Supplementary Resource Information:** This field allows data owners to include any information they feel is necessary for third parties to fully evaluate the data resource. This field offers the opportunity to provide information that cannot be provided using any of the other metadata fields. Examples of information types are as follows: bibliography/references, specification of software requirements, acknowledgements of funding, library information, hardware or instrumentation details, acknowledgements of collaborations, special analysis techniques, etc.

4.3 Data Resource Availability Levels

Section 2.2 discussed the concept of a data resource on SAFER-Data. Traditionally on web-based systems providing digital data or other digital objects for download, there are two means by which a visitor to that web system can download files. The first way is that the files are publicly available and may be downloaded by anyone. The second way is by using authentication whereby visitors must first logon to that web system before they can download any or certain files. The approach to providing digital data or digital information for download using SAFER-Data is similar but exhibits some minor differences. When a researcher or data provider creates a data resource on SAFER-Data he or she firstly creates metadata. During metadata creation the public availability of this data resource must be specified. There are three levels of public availability on SAFER-Data. A researcher or data provider may choose one of the following:

1. **Public:** In this case any digital data or digital information objects attached to this resource can be downloaded by anyone who visits SAFER-Data. This is full and open access. The IP address of the visitor downloading these data is logged in a back-end database.

2. **Semi-Public:** In this case any digital data or digital information objects attached to this resource are hidden from public access. Only metadata are visible to the public. The digital data or digital information objects can only be accessed by the owner of the resource and SAFER-Data administrators. Semi-Public status is only permitted on a resource for a period of 12 months. After this time the resource automatically becomes publicly available.

3. **Private:** In this case neither the metadata nor the digital data or digital information objects (the entire resource) are available for public access and can only be accessed by the owner of the resource and SAFER-Data administrators. The use of Private status for resources on SAFER-

Data for extended periods of time is not permitted. Essentially Private status is only used in cases where resources must be verified by several parties before Semi-Public or Public availability status is assigned to the resource.

4.4 Uploading Data Sets to the ERC Server Using SAFER-Data

SAFER-Data provides data owners and project managers with a special interface through which they can manage their data sets on SAFER-Data. When data owners logs on successfully to SAFER-Data they are automatically redirected to their 'MySAFER-Data' area. The list of data resources for which they are responsible for is listed. Summary information (title, number of data set attachments, date of last update, etc.) about each data resource is provided for easy reference. If the data owners wish to perform some changes to any of their data resources they can do so by accessing the advanced options on that data resource. This is illustrated in the screen capture in Fig. 4.5. From this interface the data owner can perform the following actions:

- **Rename** any of the data set files currently attached to a data resource
- **Upload** new data set files which are automatically attached to the current data resource
- **Edit** the metadata corresponding to any data resource
- **Update** or **Change** the public access availability level for any data set file attached to the current data resource
- **Delete** data set files that are attached to the current data resource
- **Access** simple statistics (number of downloads, number of views, etc.) about the data resources.

SAFER-Data provides upload functionality that is very similar to the functionality that most users would be familiar with from web-based e-mail systems such as Yahoo or Gmail. Using SAFER-Data project, managers and data owners can upload up to four data set files at a time. The reason for this limit is outlined below. When these data set files have been uploaded another four files may then be uploaded. This process is repeated until all of the data set files are uploaded. As outlined in Section 4.1, metadata must have been created for the data resource before data set files can be uploaded. The upload limit of four files at any one time is not an arbitrary value. This value is set to allow multiple files to 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 150 MB and with combinations of four files of total size greater than 100 MB. The following conditions are recommended for data set file upload to the ERC Server:

Figure 4.5. The data resource management interface for data owners within the 'MySAFER-Data' area.

- A fast and reliable broadband Internet connection. SAFER-Data testing has been performed on third-level institution networks and private corporate networks
- A recent version of your Internet browser software
- A 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 data sets. In the case of these projects, alternative delivery methods must be considered. Alternative delivery methods to the ERC must be used if either or both of the following conditions apply to a project:

1. **Large Volume:** The total volume of data generated that will be uploaded by your research project exceeds 300 MB
2. **Large Quantity:** The total number of files that will be uploaded (including ZIP file archives) is over 100 files.

If either of these conditions apply to a project the project manager should contact the ERC in advance to discuss the uploading of the project data sets to the ERC Server. The ERC may recommend that data set files be combined or grouped into logical collections of files.

4.4.1 Uploading data sets containing large volumes of data

This applies to projects where the total data volume for upload exceeds 300 MB. The file upload mechanism in SAFER-Data should not be used if this condition is true and the ERC should be contacted in advance of this upload. The following is a list of alternative means of providing large volumes of data to the ERC:

- Create a compressed archive version of the data set files. In some cases this may reduce the overall file size considerably. If the total data set file size after compression (ZIP, TAR, GZ, etc.) is now less than 300 MB one may use SAFER-Data to upload the data sets to the ERC Server.
- Copy the data sets to a removable media device such as a CD, DVD, USB pen drive, USB hard disk and deliver the chosen media to the ERC using standard post or in person. Appropriate packaging envelopes should be used to protect the media in transit. Upon receipt, within one working week, the ERC will upload your data sets and attach these to the corresponding metadata resource on SAFER-Data. The project manager will be notified by e-mail when this action has been completed.
- Make the data sets available on a HTTP or FTP server. These servers must be accessible to the ERC. The ERC staff will download your data sets from this server. Upon successful download, within one working week, the ERC will upload your data sets and attach these to the corresponding metadata resource on SAFER-Data. The project manager will be notified by e-mail when this action has been completed.

If appropriate, the ERC will recommend changes to how the data sets are stored and accessed on the ERC Server using SAFER-Data. These changes will be recommended with the dual intention of protecting your data set resources while ensuring that third parties can download and access the files from SAFER-Data without major difficulty.

4.4.2 Uploading data sets containing large quantities of files

In the case of some projects the quantity of data set files that must be uploaded may exceed 100 files. For situations like this, the file upload mechanism in SAFER-Data should not be used. In the first instance project managers should contact the ERC to discuss the proposed quantity of files for upload. To ensure your data sets are distributed and disseminated in the most efficient manner, the ERC recommends grouping these files into logical groups or categories. This can be done in one of two ways:

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

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

- The time period(s) represented within the files
- The geographical area(s) represented by the files
- The working group or thematic area which generated the files
- Any other appropriate project-specific characteristic.

If appropriate, the ERC will recommend changes to how the data sets are stored and accessed on the ERC Server. These changes will be recommended with the dual intention of protecting your data set resources while ensuring that third parties can download and access the files from SAFER-Data without major difficulty.

4.4.3 Preparing files for upload to the ERC

The EPA and the ERC aim to make high-quality environmental research data available using SAFER-Data. There are a very large number of different software tools and software formats used within the field of environmental science. It is not feasible to expect the EPA and the ERC to have expert knowledge of all of the software formats used. However, there are several data management characteristics common to all software formats independent of the thematic area or application. Project managers should ensure that the uploaded data sets exhibit a high data quality standard and appropriate structural formatting. The EPA and the ERC emphasise that data sets uploaded to the ERC Server and made available using SAFER-Data should adhere to the software standards and protocols most suitable to the thematic area they originate from. Some thematic areas have well-defined and widely used software formats. For example in meteorology, the GRIB format is widely used while in the biodiversity domain a text format called Recorder is the *de-facto* standard. In cases such as these, the EPA and the ERC will not convert these data sets to other formats unless there are extenuating reasons for such a conversion. For internationally accepted data formats within specific thematic areas it may not be possible to convert data sets to a corresponding format and retain all characteristics of the original data set. Users downloading these data sets from SAFER-Data for the purposes of further analysis and research are assumed to have access to the appropriate software to use these data sets. In all cases, the metadata corresponding to the data resource should outline in detail the characteristics of the data sets available for download. As a service to the funded research groups, the ERC will in a limited number of cases perform some ‘silent fixing’ of common data management and data quality problems exhibited in uploaded data set files. The following issues should in practice be resolved by the original data creator:

- Renaming of files
- Reformatting of date or date-time series within data sets
- Grouping single files into a compressed file archive
- Uncompressing a file archive and distributing the files within the archive individually
- Adding geographical information to the data set resource either in the metadata or in a separate newly created file
- Version upgrade or downgrade of a file, for example conversion of a document from Office 2007 format to Office 2003 format
- Provision of a README.TXT file to explain any special requirements users will need to meet when downloading the data set files
- Conversion of documents to a platform-neutral format – for example from an MS Word document to PDF or an MS Word document to an open document format.

If any ‘silent fixes’ have been performed the project manager will be informed of the nature and extent of these in an e-mail from the ERC. As far as is possible, the ERC will check the data quality of uploaded files. The EPA and the ERC reserve the right to request that project managers address any data quality issues flagged by the ERC before these data set files are made publicly available from SAFER-Data.

4.5 Administration of SAFER-Data

SAFER-Data offers users 24-7 access to the ERC Server. Data resources are continually changing as the data owners update and edit these resources. Files are downloaded at any time of day or night. As with any computer system there are a small number of users who have administrator access to the system. Specifically authorised users can logon to SAFER-Data with administrative-level access rights. Administrator access to SAFER-Data is an important feature in the maintenance and management of the system itself and of the research data archive. To maintain the security of the system the number of users with administrator-level access is controlled and kept to a strict minimum. SAFER-Data can be fully administered through the web-based interface allowing administrator-level users to access the system at any time from any location. Users with administrator-level access to SAFER-Data can perform the following actions:

- Create new user profiles on the system
- Edit, update, or delete existing user profiles on the system
- Edit, update, or delete any existing data resource metadata
- Change the availability level, with immediate effect, of any existing data resource
- Change the availability level, with immediate effect, of any data set file or digital object currently available for download from the SAFER-Data. This is necessary in the case where a data set file or digital object is uploaded which requires further screening by the EPA before it is made publicly available for download
- View the current visitor access and file download logs
- View the previous day's visitor access and file download logs.

Fine-grained administration of SAFER-Data is performed using various software tools on the ERC Server. It is beyond the scope of this report to outline details of this aspect of the administration of SAFER-Data.

4.6 Technical Information for SAFER-Data

SAFER-Data is a fully Java-driven web application. Apache Tomcat is used as the application server. All static and dynamic content is generated using Java Server Pages (JSP). SAFER-Data stores use a MySQL database as the back-end database engine. This database manages the following data and information:

- Resource metadata (stored in relational tables)
- SAFER-Data user data and logon information
- All hits and visitor patterns to SAFER-Data
- Information on all stored data files and resources.

SAFER-Data was developed to allow customisation of many key pieces of back-end and front-end functionality. These include:

- Cascading Style Sheets (CSS)
- Location of physical data file archives
- Display parameters such as the number of resources displayed on the SAFER-Data homepage.

All software components of SAFER-Data are free and open source software (FOSS): Apache Tomcat, MySQL, various Java software libraries, and OpenLayers JavaScript mapping API. With some additional software development SAFER-Data has the potential to become a software application which could be used by other research funding organisations in Ireland and in other Member States. Throughout the development of SAFER-Data there was a focus on configurability and extendibility. Several aspects of SAFER-Data can be easily configured in order for it to be usable on another system:

- **CSS *Look and Feel*:** Currently the EPA's corporate *look and feel* is displayed on SAFER-Data by linking to the EPA's CSS files. The link to an alternative set of CSS files is easily changed.

- **Data Archive Location:** The physical file location to where all data set files are stored in the ERC Archive can be easily changed by changing the path information to the archive area.
- **Size of Upload Files:** Currently SAFER-Data restricts upload of files to files that are less than 150 MB in size. This can be configured for larger files or for a restriction to smaller-sized files.

5 Using SAFER-Data to Access Research Data

Chapter 4 described the technical development and implementation of SAFER-Data. This chapter describes some of the key features of SAFER-Data for accessing the environmental research data stored on the ERC Server. These include:

- How to provide hyperlinks to specific data resources on SAFER-Data
- How to download data sets from SAFER-Data and how these downloads are monitored on the system
- Using RSS-Feeds to access information from SAFER-Data
- Uploading data to the ERC Server using SAFER-Data.

5.1 Downloading Data Sets from SAFER-Data

The number of data sets downloaded from SAFER-Data is a key performance indicator for the success of SAFER-Data. The ERC clearly recognises that download and usage of data resources from SAFER-Data is a prime motive for its existence and assists the EPA in achieving its target of widest possible dissemination of research output from STRIVE. Figure 5.1 shows the total number of data set files downloaded from SAFER-Data on a monthly basis between October 2007 and December 2008 inclusive. The graph shows an overall growth in the number of data set files downloaded from SAFER-Data each month. The number of files downloaded from SAFER-Data has grown over tenfold from 62 during October 2007 to 640 during October 2008.

Figure 5.2 is a nice example of a power-law graph (purple line) being used to demonstrate the ranking of popularity of downloads on SAFER-Data. The number of times each resource has had a data set file downloaded from it is shown by the vertical bars. The download patterns from SAFER-Data roughly adhere to the Pareto principle (also known as the 80-20 rule, or the Law of the Vital Few). This principle states that

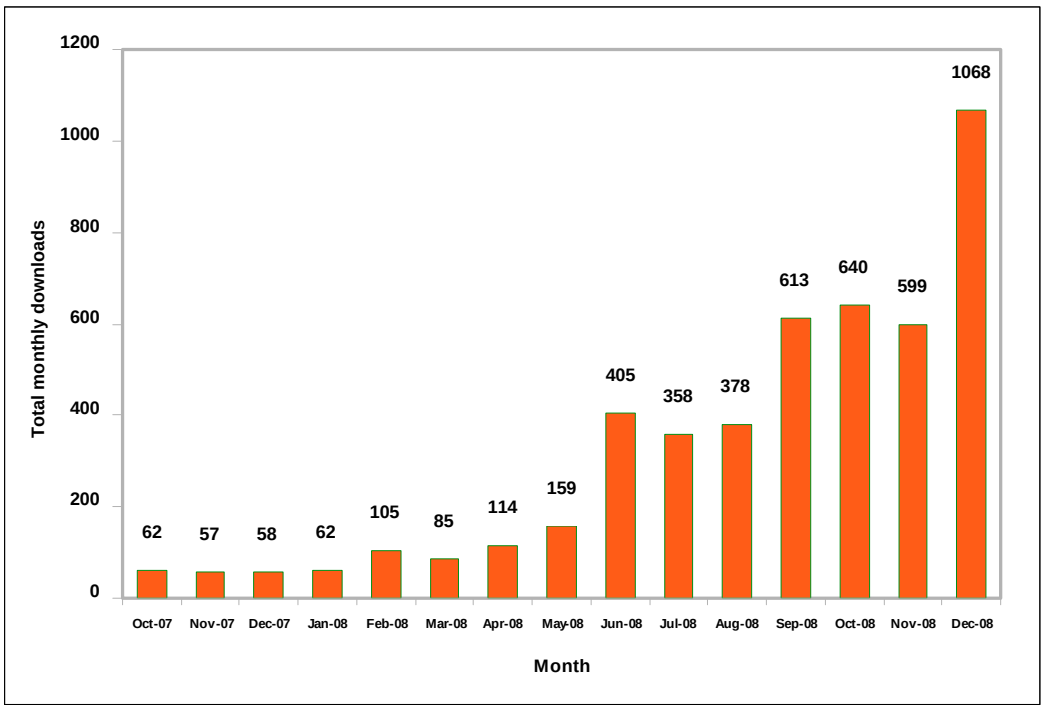

Figure 5.1. Total monthly downloads from SAFER-Data.

20% of the resources on SAFER-Data account for approximately 80% of the downloads.

One of the key messages in Fig. 5.2 for researchers who provide their data for open access through SAFER-Data is that their data will be of value to other researchers and stakeholders who visit SAFER-Data. While several resources will attract a large number of downloads **every** resource uploaded and managed by SAFER-Data will be downloaded. This is shown in Fig. 5.2 for resources that are at the tail of the graph (increasing resource number).

Figure 5.3 shows a map of the geographical distribution of users who downloaded data set files from SAFER-Data between January 2008 and

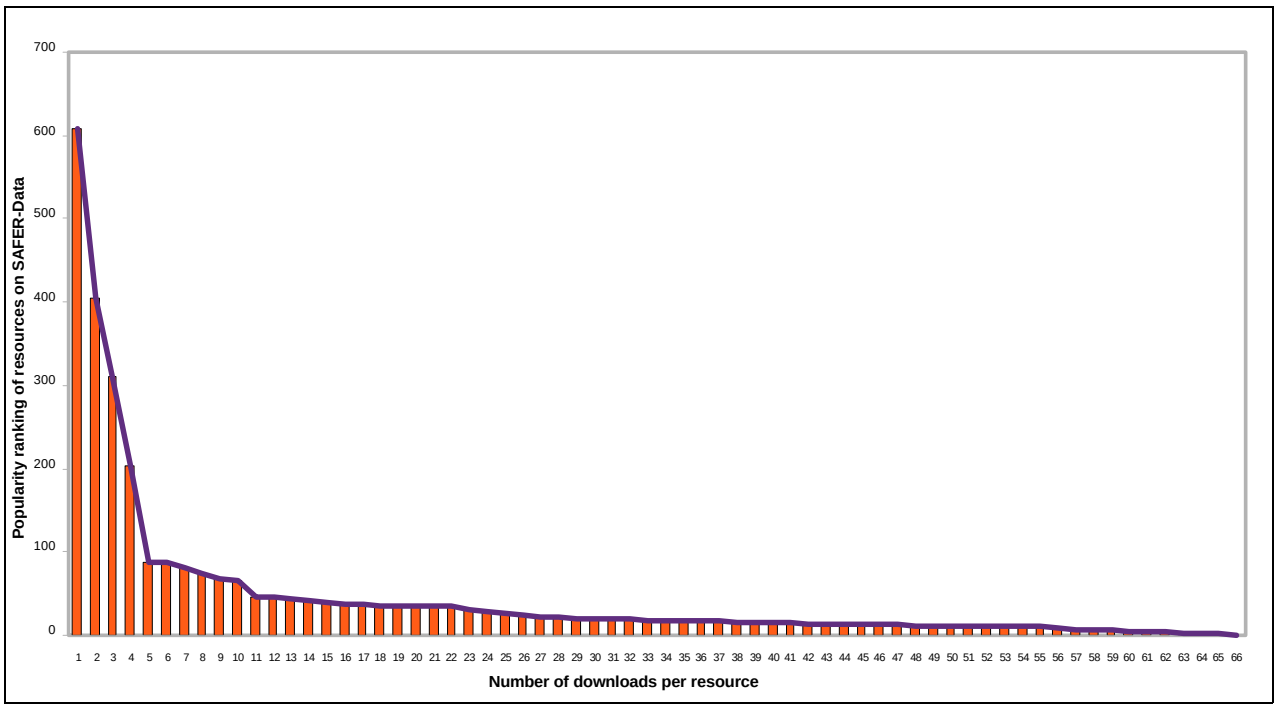

Figure 5.2. A 'long-tail' distribution showing, in sorted order, the number of downloads each individual data resource on SAFER-Data received during 2008.

Figure 5.3. Geographical distribution of downloaders up to December 2008.

December 2008. Dynamic updates to the download traffic information are available on SAFER-Data at http://erc.epa.ie/safer/information/siteTrafficInformation.jsp.

5.2 Monitoring Downloads on SAFER-Data

As discussed previously it is not necessary for a visitor to SAFER-Data to logon to SAFER-Data in order to download publicly available data set files. While this is very convenient for users, it presents a problem in controlling access to publicly available data set files on SAFER-Data. Search engines such as Google, MSN, Yahoo, etc., run programs that 'crawl' over all of the links on a website. If some form of access control is not put in place then search engines and other web crawlers can download all publicly accessible files during a visit to a website. This mass downloading is done automatically.

Many users see no problem with search engines such as Google, MSN, Yahoo, etc., having access to their publicly available material. The major problem with this approach from the perspective of SAFER-Data is that if a data set file is downloadable directly through a search result link in a search engine the important geospatial metadata corresponding to that data set file are usually missing. The user downloading the data set file in this fashion cannot make a proper decision on the fitness for use or fitness for purpose of the data set file without the corresponding metadata. This type of downloading would also have the effect of severely skewing the statistics on how many data set files were actually downloaded by users from SAFER-Data. As a countermeasure to this a simple commonplace solution is implemented on SAFER-DATA. A completely automated public Turing test to tell computers and humans apart (CAPTCHA) offers a way for web service providers to draw some conclusions about whether a 'user' is human or robot (Kolupaev and Ogijenko, 2008). CAPTCHAs are one of the most popular ways of protecting Internet forms by generating a special picture made up of letters and numbers and then asking the user to type what they see in the picture into a special box. The concept is that it is extremely hard to teach a bot (computer program) how to recognise letters when they are embedded in a graphical display. On SAFER-Data before a user (either logged on or anonymous) can download data set files, he or she must pass a CAPTCHA test. Figure 5.4 shows the standard visual verification CAPTCHA where the user must retype the letters and numbers they see in the graphic. The answer to the CAPTCHA is entered into the nearby textbox. If the user has correctly retyped the CAPTCHA the data set file download begins. If the user has incorrectly retyped the CAPTCHA then a new CAPTCHA is generated and the user must try again.

Figure 5.5 shows the alternative accessibility compliant CAPTCHA. Blind or visually impaired users can launch an audio file within their web-browser application. The audio file, when played by an appropriate media player, reads out the letters within the CAPTCHA. The answer to the CAPTCHA is entered into the nearby textbox. If the user has correctly retyped the CAPTCHA letters that have been read out, the data set file download begins. If the user has incorrectly retyped the CAPTCHA, then a new CAPTCHA is generated and read out to the user. It is necessary to provide an alternative to the graphical-only CAPTCHA (Shirali-Shahreza and Shirali-

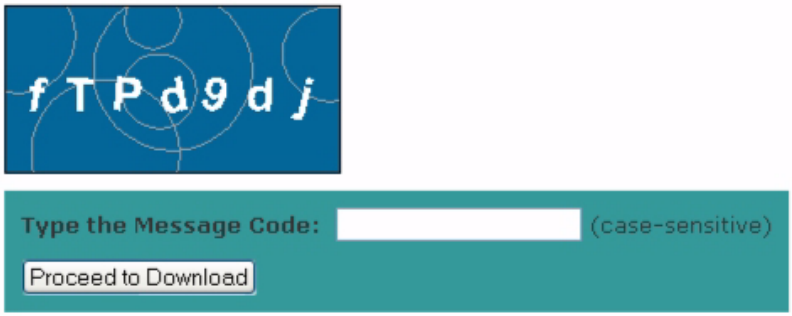

Figure 5.4. The graphical CAPTCHA with the accompanying answer box.

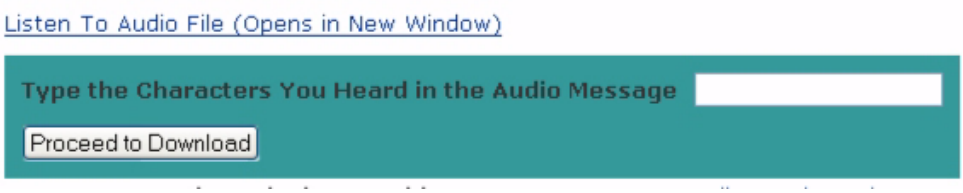

Figure 5.5. An accessibility-compliant CAPTCHA where an audio file is launched which reads out the letters within the CAPTCHA.

Shahreza, 2007). Because answering this problem requires recognising the speech, understanding the characters, and typing the answer to the problem, only a human user can answer this question and current computer programs are unable to solve it. In addition, answering the question is easy for blind people, because the question consists of a small number of natural language characters. Once either CAPTCHA has been correctly solved, users will not be required to retake the test for the duration of their visit session. Currently on SAFER-Data a visit session is 30 min in duration. A visit session is said to expire if the user has not performed any actions on the SAFER-Data website for a continuous period of 30 min. The use of the CAPTCHA allows SAFER-Data to compile very accurate statistics (see Section 5.1) about the quantity of downloads, the rate of downloads per day, and the rate of downloads per user. The graphical CAPTCHA is generated by a back-end software program that creates the CAPTCHA letters randomly. The audio CAPTCHA is chosen from a large pre-generated bank of audio files.

5.3 Using Really Simple Syndication Feeds from SAFER-Data

Really Simple Syndication (RSS) is a format for delivering regularly changing web content. Many news-related sites, web logs and other online publishers syndicate their content as an RSS feed to whoever wishes to access it. Really Simple Syndication solves a problem for people who regularly use the web. It allows them to easily stay informed by retrieving the latest content from the sites they are interested in but without needing to visit each site individually. The user's privacy is protected by not needing to join each site's e-mail newsletter. A variety of RSS readers are available for different platforms. Some popular feed readers include AmphetaDesk (Windows, Linux, Mac), FeedReader (Windows), and NewsGator (Windows – integrates with Outlook). There are also a number of web-based feed readers available. My Yahoo, Bloglines, and Google Reader are popular web-based feed readers. In addition to this, more and more web users are using webtop services such as iGoogle, Pageflakes, etc. These webtops allow users to construct their own personal web space by combining many RSS feeds together.

RSS feeds are suitable for SAFER-Data as they allow advanced web users to access regularly changing content on SAFER-Data in the same way that they would access their favourite news media websites, blogs, etc. RSS is a simple XML format and is very easily generated from the information stored in the SAFER-Data back-end database. Figure 5.6 shows a partial screenshot of the selection of RSS feeds available from SAFER-Data in November 2008. By clicking on the RSS icon users can subscribe to the feed using their web browser or feed reader. Currently, three RSS feeds are available and are summarised as follows:

1. **Most Recently Updated Metadata Resources:** This RSS feed essentially provides the 10 metadata resources that appear on the SAFER-Data homepage.

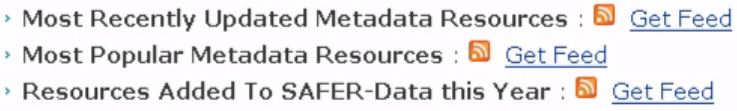

Figure 5.6. The RSS feeds available from SAFER-Data.

2. **Most Popular Metadata Resources:** This RSS feed provides information on the most popular resources on SAFER-Data. The popularity is measured by the number of downloads each resource has received (see Section 5.1).

3. **Resources Added to SAFER-Data this Year:** This RSS feed provides information on all of the metadata resources which were newly created on SAFER-Data this year. The resources are sorted in chronological order.

5.4 Creating Links to Data Resources on SAFER-Data

It is possible to include the direct URL hyperlink to a metadata resource on SAFER-Data. This may be required for inclusion in an e-mail, in a report article, or for reference material in a presentation or similar. There are two means of providing a URL hyperlink to metadata resources on SAFER-Data. Each metadata resource (and its attached data set files) has two unique identification numbers. The metadata identifier is unique to the SAFER-Data system only while the resource identifier is almost universally unique.

5.4.1 Linking using the metadata identifier

The URL that uses the metadata identifier is subject to change. Therefore this method of linking to resources should be confined to situations where the recipient of the URL is likely to use the URL in the immediate future. The metadata identifier currently takes the form: http://erc.epa.ie/safer/iso19115/display?isoID=49.

To obtain this URL the user will need to proceed according to the following steps:

- View the Full Metadata View of the chosen metadata resource
- Scroll to the section Resource Identification and Thematic Information
- Underneath the highlighted ID number is the full URL to the current resource, which can be copied and pasted into e-mails, bibliographies, articles, presentations, etc.

A similar URL is available in the address bar of the Internet browser that the user is currently using to view the data resource metadata. This can be copied and pasted in a similar manner. However, the URL in the address bar may contain extra terms and characters. The metadata identifier always provides the shortest URL representation.

5.4.2 Linking using the resource identifier

The resource identifier is a 36-character string used to uniquely identify any digital object or resource. The character string is generated by the back-end MySQL database. Formally this string is known as a Universally Unique Identifier (UUID). It is almost mathematically impossible that two users or systems will generate UUIDs that are the same (known as a clash). Therefore the UUID is used in SAFER-Data to ensure that regardless of any changes or updates to a metadata resource in the archive the UUID will always remain unique and constant. In SAFER-Data a permalink using the UUID takes the form http://erc.epa.ie/safer/resource?id=fa2287ae-3bf0-102b-950d-28616e04c7da. It is strongly recommended that this form of link including the resource identifier (UUID) is used in preference to metadata identifiers when the hyperlink (or link) to resources on SAFER-Data are being published on web pages, blogs, or message boards. It means that neither the web page, blog post, or message post are affected by changes to the metadata resource (such as change in title, authors, metadata identifier, etc.). It is particularly useful in situations where it will not be possible to inform the owner of the web page, blog, or message board of the change in metadata identifier URL. To find the permalink for any resource on SAFER-Data the user will need to take the following steps:

- View the Full Metadata View of the chosen metadata resource
- Scroll down to the bottom of this page to the Bookmarking section
- The permalink is printed at the bottom of this section
- This permalink may be copied and pasted from here.

5.5 Searching for Data Resources on SAFER-Data

As mentioned earlier in this report, the number of data resources available on SAFER-Data is expected to grow significantly over the lifetime of the STRIVE funding programme. It is very important that the users can easily search for data resources on SAFER-Data. Users will want to perform different types of searches on SAFER-Data. They may wish to search the data resources using specific keywords, names of data owners, project codes, or data set file names. It is very important that the search interfaces on SAFER-Data are predictable and user friendly (Shneiderman, 1997) and care must be taken not to change the type of interface radically from other types of interfaces that users encounter on other systems on the Internet (Hoeber and Yang, 2007). Finally, the search facility should return results in a reasonable amount of time but look towards almost real-time return of results (Huayong and Yiqi, 2004). Figure 5.7 shows a screenshot of one of the search query interfaces on SAFER-Data. The screenshot shows the keyword search interface where users can select which metadata fields they wish to search using the keywords they have supplied. There are four separate search interfaces available for users:

1. **Search by keyword:** Users select the metadata information fields they wish to search and also supply a list of keywords they want to use in their search.

2. **Search by project code:** Users can supply a full or partial project code of an EPA-funded research project and search SAFER-Data to see if this project has supplied any data resources to the ERC Server.

3. **Search by data author:** Users can supply the names of data authors and search SAFER-Data to see if this data owner is the owner of any data resources on the ERC Server.

4. **Search by data set filename:** Users can search for specific filenames of data set files which are available to download from data resources currently publicly available from SAFER-Data.

Search results are presented by showing a summary of the data resource metadata that matches the user's search criteria. Figure 5.8 shows the result of a search

Figure 5.7. The search query interface on SAFER-Data.

Figure 5.8. An example of a successful search on SAFER-Data.

on SAFER-Data for data resources that include the keyword 'water' in the data resource metadata. The search results are presented in a simple text format allowing the user to go directly to the full metadata for the data resources.

5.6 Citing Data Resources on SAFER-Data

Proper documentation of sources is a key component of scholarly research, and the need for documentation is no less important for data sources than for bibliographic sources (Mooney, 2008). It is vital that researchers who use other people's data in their work be clear and explicit as to the sources for their data. Users can download data sets and reports from SAFER-Data quickly and easily. Metadata are provided for all data resources on SAFER-Data to provide third-party users with sufficient information to allow them to make a fitness-for-purpose judgement about the data resource before they begin to download. The contact information of the main point of contact for the data resource is also specified clearly in the metadata. With this open access approach, SAFER-Data facilities the download of data from the ERC Server and potentially the use of these data in research and analysis leading to a publication of a report, thesis, or academic journal paper. SAFER-Data is the conduit for the flow of data between the original data creator or data owner and the third-party user. The third party is not required to contact original data creators or data owners about the download of their data nor is he/she required to provide details of where the data will be used. M.o.Ed. (2005) points out that scientific publishing and distribution has changed and that we are now witnessing "*the 4th cognitive revolution*" where e-communication, open peer-review research results, and direct links from journal texts and books to the original data sources are now commonplace. One of the perceived risks of open access to research data from the point of view of data creators and data owners is the uncontrolled usage of the data by third parties. The fear exists that these third parties will download and use these data in their own research work and subsequently publish results in peer-reviewed journals without any recognition of the original data creators or data owners. Such an action if accomplished would be both unfair and academically unethical. Citing the sources of third-party information and data is one of the cornerstones of academic publishing.

SAFER-Data provides an automatically generated citation for every data resource it manages. A screenshot of this automatically generated citation is show in Fig. 5.9. This citation includes all of the fields third-party users should include when they create a citation for the data resource(s) they downloaded from SAFER-Data and subsequently used in their research. The SAFER-Data citation information is not presented in any particular bibliographical style. The reason for this is because many researchers use bibliographical

Figure 5.9. An automatically created citation for a data resource on SAFER-Data.

management tools when they are writing scientific papers, reports, theses, etc. These tools include well-known bibliography managers such as EndNote and BibTeX. The SAFER-Data citation is bibliography neutral and researchers can easily create a bibliographical entry for the data resource in their chosen bibliography management software using the information fields outlined in the SAFER-Data citation.

6 Conclusions and Recommendations

Publicly funded research data should be openly accessible to the maximum extent possible (Downs and Chen, 2005). This report has described the most prominent reasons why researchers are reluctant to exchange their research data and make it more accessible to the wider research community and general public. To address these reasons, a web-based data management system called SAFER-Data has been developed and deployed for data resources from the environmental research programme. This system greatly simplifies metadata creation, data submission and upload, and long-term data archival for the environmental research community. SAFER-Data offers the environmental research community a unique solution to environmental data access and data management. The systems reviewed in Section 1.2 either (1) stored only metadata containing links to environmental data stored on other systems, or (2) stored data directly on their system but required users to authenticate themselves before download or purchase of the data. SAFER-Data provides free and open access to data generated from EPA-funded research projects. SAFER-Data also provides a secure web-based means for data owners and data providers to manage the data they own which is stored on the ERC Server. There are a number of areas that should form the core of future work on environmental research data management within the ERC. These are summarised in the following sections.

6.1 Ongoing Development of SAFER-Data

The SAFER-Data system is now well established as a dissemination point for environmental research data in Ireland. In order for SAFER-Data to continue this successful contribution to the flow of environmental data and information, the system must continue to be supported, developed, and updated by the ERC and EPA. The direction of the future development of SAFER-Data will be influenced by environmental data management needs and requirements which will change in the future: the INSPIRE Directive will be influential in how national-level data are collected, stored, and managed; the SEIS Directive will emphasise the need for barrier-free access to environmental data for decision makers; and the research community using Internet technologies will expect to see SAFER-Data offer similar technologies and user interfaces, for example online mapping, drill-down interfaces, etc. SAFER-Data should also be given the flexibility to grow organically by responding to specific needs and requirements that become prominent amongst the STRIVE-funded research community and the user base of stakeholders, scientists, and the general public.

6.2 Continued Engagement with Researchers

During 2007 a series of 14 workshops was delivered in Irish universities and research institutes on the topics of using SAFER-Data and managing environmental research data at the project level. Invitations to other workshops and seminars were accepted and presentations on data management were prepared and delivered. In addition to this, researchers and project managers were spoken to over the phone, communicated with *via* e-mail, and face-to-face meetings were conducted to discuss the specific data management requirements for each project. These forms of proactive engagement must be continued by the ERC. Amongst the many advantages to this type of interaction are enhanced visibility of SAFER-Data amongst the environmental research community in Ireland, increased 'buy in' from data providers who can understand how SAFER-Data works and what the advantages are for their project and data, and less administrative overheads within the ERC in regards to metadata and data collection as researchers begin to use SAFER-Data independently and on a day-to-day basis. The environmental research data management policies and practices of the EPA must be constantly monitored to ensure that they serve the needs of both data providers and data consumers in the Irish environmental science community. The EPA must ensure that the ERC continues to deliver these seminars, workshops, invited talks and

demonstrations, hands-on data management tutorials, etc., to the environmental research community. In addition to this, a number of simple steps are recommended which could be implemented to further establish SAFER-Data within the environmental research community. These include:

- Information about SAFER-Data at the footer segment of outgoing EPA Research programme e-mails
- Hyperlinks to SAFER-Data or resources on SAFER-Data from the EPA homepages
- Promotional items such as post-it notepads, brochures, pens, etc., with the SAFER-Data URL printed on the item which are then distributed at EPA conferences and events.

6.3 Introduce a Data Management Plan for Funded Projects

Over the course of this project it was found that many STRIVE-funded projects postponed planning for data management of their project until the end of their project. Unfortunately this postponement usually means that it becomes very difficult to correct any serious data management problems which the project data may exhibit. This can cause a delay to the upload of the data to the ERC using SAFER-Data and can consume considerable time resources to fix these problems to a satisfactory outcome. One potential solution to this problem is the introduction of the requirement for a data management plan for each STRIVE-funded project. The concept of a data management plan is straightforward. It is illustrated in the flow chart in Fig. 6.1. This is an update of Fig. 4.2 and shows how the data management plan fits into the overall data management workflow for STRIVE-funded

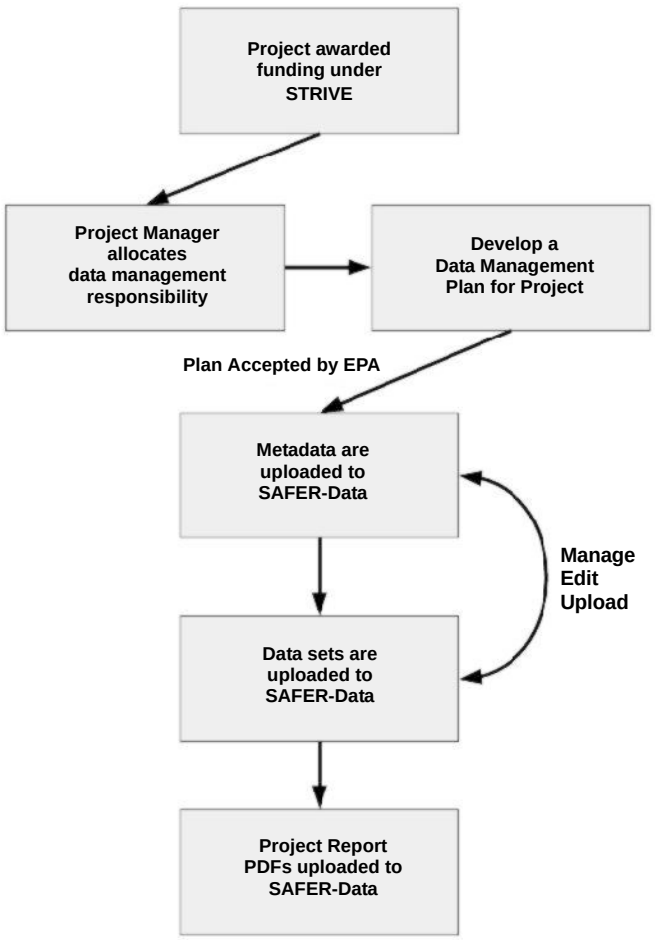

Figure 6.1. The recommended data management schedule for data.

projects. When a project is awarded funding, the first major data management step that it must take is to provide the funding body with a data management plan for the project. This data management plan is submitted to the funding body and reviewed by technical experts (in the case of the EPA the ERC could carry out this task). When the data management plan has been approved, the project is informed of the successful outcome and may commence its data gathering or data generation activities. While this introduces an additional administrative burden on the funding body there are many advantages to its introduction. The most valuable of these advantages is that serious data management problems are identified at the early stages of a project rather than at the very end of the project. These problems and issues can be addressed and solved to the satisfaction of the project partners and the funding body (in this case the EPA). It also allows the EPA to comment on overlooked or under-appreciated opportunities for data access and long-term data preservation. The data management plan would consist of a simple survey that would require the project partners to outline:

- Their plans for methods for the collection of raw data
- Their plans for storage of collected data, including the use of databases
- Current international best practice for data collection and data storage/formatting within their scientific community
- Potential issues that would reduce the opportunities to make the collected or generated data fully publicly available through SAFER-Data. These issues could include survey participant or species privacy and/or protection, intellectual property rights (IPR) issues, and reluctance to share data based on other rationales.

Some research funding organisations in other countries have successfully introduced the concept of a data management plan for funded projects. These include the Natural Environment Research Council (NERC), the UK Social Science Data Archive, and the National Science Foundation (NSF) in the United States.

6.4 Extend usage of SAFER-Data to Other Agencies

It is often the case that the small agencies and research institutions in Ireland hold valuable sources of environmental data. Due to restrictions on technical and financial resources they are not able to make this data publicly available on the Internet in a secure and controlled manner. It is recommended that the EPA should invite other Irish agencies that hold environmental data to use SAFER-Data as a dissemination tool for these data. In particular, the ERC should generate a list of potential data resources that could benefit from management using SAFER-Data. The agency or research institute that owns the data would be able to access and manage its data in the same manner as a STRIVE-funded project. More widespread distribution of these data resources benefits not just the EPA and the agency or research institutes involved but also the wider environmental science community in general. This would also support the EPA in its work towards its 2020 Vision target of "*becoming the key source for environmental data and information in Ireland*" (EPA, 2007a). It would also prevent costly duplication of effort in building other systems independent of SAFER-Data.

6.5 Improve Academic Recognition Strategies for Data Providers

M.o.Ed. (2005) describes the current situation where there is open peer review of scientific research results with the possibility of direct links from journal text to data set source as "*the fourth cognitive*" revolution of mankind. There is often reluctance amongst scientists to upload their data sets to publicly accessible archives and publicly available data repositories for fear that a third party may access their data and gain recognition for scientific work resulting from it. Arzberger *et al.* (2004) state that "*big science or mega science – NASA, NOAA, ESA, etc.*" usually openly share its data and results in public repositories. However, "*small science*", independent, investigator-driven research remains dominant in most scientific fields. Traditionally data from such studies have been extremely heterogeneous and not standardised, with few individuals making their data sets available through public repositories or even openly sharing them. Arzberger *et al.* (2004) point out that most academic

recognition is directed toward peer-reviewed publications. Data sets are not the object of the research (Buckhorn and McNamara, 2006) and, for most scientists, submitting data sets to publicly accessible repositories is a very low priority.

One of the conditions of STRIVE funding is that all significant data sets generated during the course of the research must be provided to the EPA and managed using the SAFER-Data system. The conditions of the contract further state that these data sets will be made fully publicly available 12 months after the final report has been published by the EPA. There appears to be little international agreement on what the appropriate period of time should be until data sets are made publicly available: the European Research Scientific Council (2006) states that this period should be "*ideally, 6 months, and in any case no later than 12 months*", the Australian Research Council "*within two years of the conclusion of any fieldwork relating to the funding*", the US Department of Health (NIH Funding) (National Institutes of Health, 2007) "*no later than the acceptance for publication of the main findings from the final dataset*", and the Biotechnical and Biological Sciences Research Council, UK (BBSRC, 2008) "*within three months of completion of data*".

Consequently, there is a situation of conflicting objectives: the EPA and STRIVE on the one hand who look to disseminate the findings of individual research projects to the widest possible audience and, on the other hand, the individual researchers who (a) gain little or no academic recognition for making their data available, (b) risk unsolicited third-party usage of their data sets as the possibility exists that such third parties could further publish the work, and (c) fear losing ownership of their data sets resources. Specifically this further work should include the following objectives:

- Enhance citation and data set download functionality within SAFER-Data. In this way researchers can better understand the tangible and intangible rewards of improved data access.

- Develop an EPA STRIVE data policy. This data policy will include best international practice in terms of data release policies and open access to research data. It should ensure that both the researchers and the EPA derive maximum added value under this policy.

6.6 Visitor Download Ticketing System

Some data providers are reluctant to allow their data set files be made fully publicly available. The reasons for this were discussed earlier in this report. Given this reluctance, the EPA should not allow these data set files to disappear from public access. Currently SAFER-Data provides three levels of availability for data set files and these are outlined in Section 4.3. Among these availability levels there are two conflicting objectives. On the one hand, the EPA is looking to provide the "*widest possible dissemination of research data*" while, on the other hand, data providers and researchers are worried about the perceived risks of uncontrolled third-party access to their data sets. A potential solution to this situation is the implementation of a visitor download ticket system within SAFER-Data. Broadly speaking this would allow data providers to allow third-party access to their data set files on SAFER-Data for a limited time period. When the specified time period has elapsed the third party can no longer access these files. In more detail, the visitor download ticket system would operate as follows:

1. A visitor to SAFER-Data sees a resource from which he/she would like to download data set files. It is specified in the metadata that the data set files are currently private and not publicly available.

2. The visitor follows a special URL to *Apply for a download ticket for this resource*.

3. The visitor is redirected to a simple web-based form in which contact details must be provided, reasons for requiring access to the data set files, and the number/type of data set files required.

4. The owner of the resource is notified of this request and logs on to his/her account on SAFER-Data.

5. The owner follows a sequence of actions to select the requested data set files for download ticket access and the duration for which this download ticket is valid (i.e. 7 days).

6. A ticket is then automatically generated as a PDF document by SAFER-Data containing an encoded identification number and URL. The data owner e-mails this PDF document to the visitor who requested download ticket access.

7. When the visitor receives this PDF document and accesses the special URL contained within the document he/she is redirected to a special area within SAFER-Data from which these private files can be downloaded.

8. When the specified duration of the download ticket has elapsed the corresponding URL becomes invalid and the private files cannot be downloaded again by this visitor.

This type of solution is fairly common on the Internet today. It is used for situations such as lost or forgotten password retrieval, bidding and bargaining in online auctions, and accessing confidential documents online.

6.7 Final Remarks

In comparison to other similar systems (PortalU in Germany, CESIN in the United States, UK Data Archive, etc.), the ERC Server and SAFER-Data are small systems. All of these systems operate in much larger countries and often with a wider thematic remit than SAFER-Data. However, the ERC Server and SAFER-Data should continue to be developed and managed with the goal of national, European, and international recognition in mind. As mentioned in Chapter 1, the STRIVE programme will significantly enlarge both the environmental science research community in Ireland and the volume of data and information generated by environmental research. Figure 5.3 illustrates the impressive geographical extent to which data sets and resources from SAFER-Data have been downloaded and accessed. In continuing to change the mindsets of researchers towards a more open and coherent data-sharing culture in environmental research in Ireland, the ERC and SAFER-Data will have brought about very significant and long-lasting progress. Advances in web technologies (Web 2.0 and eventually Web 3.0) must be embraced along with the advances in mobile device access to Internet services and data. As the public expectation about data access changes and the public's usage of Internet technologies increases, the ERC and SAFER-Data must keep pace with these changes. Data management services for the research community and data access services for the public and other stakeholders must be provided using the technologies that these user groups are using on a daily basis.

References

Ahlstrom, D., 2007. Spreading the seed money. *Science Today Supplement*, p. 17. *The Irish Times* Newspaper. Thursday 15th November 2007.

Alberani, V., Pietrangeli, P. and Mazza, A., 1990. The use of grey literature in health sciences: a preliminary survey. *Bulletin of the Medical Library Association* **78(3):** 358–363.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Uhlir, P. and Wouters, P., 2004. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal* (*CODATA*) **3(1):** 135–152.

Atarashi, R., Kishigami, J. and Sugimoto, S., 2003. Metadata and new challenges. *Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT-w'03*). IEEE. pp. 395–398.

BBSRC, 2008. *BBSRC Research Fellowships Handbook 2008*. The Biotechnology and Biological Sciences Research Council (BBSRC), Polaris House, North Star Avenue, Swindon, Wiltshire, SN2 1UH, UK. http://www.bbsrc.ac.uk/funding/fellowships/fellowship_handbook.pdf.

Buchorn, M. and McNamara, P., 2006. *Sustainability Issues for Australian Research Data*. Report of the Australian e-Research Sustainability Survey Project. APSR (Australian Partnership for Sustainable Repositories) Publication. PDF version available online from http://hdl.handle.net/1885/44304.

Burton, A., 2007. Avoiding a digital dark age. *ScienceAlert: Australia and New Zealand* **1(1):** 1–2.

CIHR, 2008. *CIHR Policy on Access to Research Outputs*. Online from the Canadian Institutes of Health Research. http://www.cihr-irsc.gc.ca/e/documents/cihr_policy_on_access_to_research_outputs_e.pdf.

Comber, A., Ahlqvist, O., Fisher, P., Harvey, F., Gahegan, M. and Wadsworth, R., 2005. Improving environmental research data management. *Proceedings of the GIS Research UK 13th Annual Conference GISRUK*. pp. 53–57, University of Glasgow, Glasgow, Scotland.

Downs, R. and Chen, R., 2005. Organisational needs for managing and preserving geospatial data and related electronic records. *Data Science Journal* **4(31):** 255–271.

EDMERP, 2008. *EDMERP – European Directory of Marine Environmental Research Projects*. Marine Information Service (MARIS), Koningin Julianalaan 345 A,2273 JJ Voorburg, The Netherlands. System URL: http://www.sea-search.net/mrp/welcome.html.

EPA, 2007a. *2020 Vision – Protecting and Improving Ireland's Environment*. Environmental Protection Agency, PO Box 3000, Johnstown Castle, Co. Wexford. Ireland. PDF version available online from http://www.epa.ie/downloads/pubs/other/corporate/.

EPA, 2007b. *Science, Technology, Research and Innovation for the Environment (STRIVE) – an Environmental Protection Agency Programme 2007–2013*. Environmental Protection Agency, Johnstown Castle Estate, Co. Wexford. Ireland.

European Research Scientific Council, 2006. *European Research Scientific Council Statement on Open Access*. European Commission, Research DG, Brussels. PDF Version available online from http://erc.europa.eu/pdf/open-access.pdf.

ESDS, 2008. ESDS – Economic and Social Data Service. UK Data Archive, University of Essex, Wivenhoe Park, Colchester, UK. System URL: http://www.esds.ac.uk/.

European Commission, 2008. *The European Research Area: New Perspectives*. Technical Report SEC(2007) 412/2 COM(2007)161, Commission of the European Communities, Directorate-General for Research, B-1049 Brussels, Belgium.

European Parliament, 2002. The sixth community environment action programme 6*th* EAP. *Official Journal of the European Communities* L242/1–15.

European Parliament, 2007. *Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. *Official Journal of the European Communities* L108*I*1–13.

GISFORM, 2002. *Accompanying Measure to Support the Setting Up of a European Territorial Management Information Infrastructure*. European Territorial Management Information Infrastructure, Palazzo Ducale di Colorno, Piazza Garibaldi 23, I-43052 Colorno, Parma, Italy, 6.2.2 edition.

Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems, Chichester, West Sussex, PO19 1UD, England, 2nd edition.

Hoeber, O. and Yang, X.D., 2007. User-oriented evaluation methods for interactive web search interfaces. *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences*. pp. 239–243.

Huayong, W. and Yiqi, D., 2004. An information search interface with soft real-time guarantee. *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*. **2:** 820–824.

JRC, 2008. ESDC – the European Soil Data Center. European Commission – DG Joint Research Centre Institute for Environment and Sustainability. http://eusoils.jrc.it/ data.html.

Kolupaev, A. and Ogijenko, J., 2008. Captchas: Humans vs. bots. *Security and Privacy. IEEE* **6(1):** 68–70.

Kuula, A. and Borg, S., 2008. *Open Access to and Reuse of Research Data – the State of the Art in Finland*. Technical Report, Tampere 2008, Finnish Social Science Data Archive – University of Tampere, Kerlundinkatu 2 A, 5th floor, FIN-33014 University of Tampere, Finland.

MarBEF, 2008. *ERMS – the European Register of Marine Species*. Vlaams Instituut voor de Zee, Wandelaarkaai 7,B-8400 Oostende, Belgium. System URL:

http://www.marbef.org/data/erms.php.

M.o.Ed, 2005. *Recommendations for the Promotion of Open Access in Scientific Publishing in Finland*. Memorandum by the Open Access working group, Ministry of Education, Department for Education and Science Policy, P.O. Box 29, FIN-00023 Government, Finland. Available online as a PDF at http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2005/liitteet/opm_250_tr16.pdf?lang=fi.

Mooney, E., 2008. *Citing your Sources – a Research Guide*. Emory University Libraries, 540 Asbury Circle, Atlanta, Georgia 30322, USA. Published online at: http://web.library.emory.edu/r_guides/how_guides/citationstyles.html.

Mooney, P. and Winstanley, A.C., 2007. Improving environmental research data management. *Proceedings of EnviroInfo 2007*. Warsaw, Poland. pp. 78–81.

National Institutes of Health, 2007. *NIH Revised Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research*. National Institutes of Health (NIH), 9000 Rockville Pike, Bethesda, Maryland 20892, USA. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html.

NSB, 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Board, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA. Available online as a PDF at http://www.nsf.gov/ pubs/2005/nsb0540/.

Pillmann, W., Geiger, W. and Voigt, K., 2006. Survey of environmental informatics in Europe. *Journal of Environmental Modelling and Software* **21(1):** 1519–1527.

PortalU, 2008. *PortalU – the German Environmental Information Archive*. Niederschsisches Umweltministerium, Koordinierungsstelle PortalU, Archivstrasse 2, D-30169 Hannover, Germany. System URL: http://www.portalu.de/ingrid-portal/portal/default-page.psml.

SEIS, 2008. *Towards a Shared Environmental Information System (SEIS).* Technical Report SEC(2008) 111 COM(2008) 46 final. Commission of the European Communities, Environment Directorate-General, B-1049 Brussels, Belgium.

Shirali-Shahreza, M. and Shirali-Shahreza, S., 2007. CAPTCHA for blind people. *Signal Processing and Information Technology.* 2007 IEEE International Symposium. pp. 995–998.

Shneiderman, B., 1997. A framework for search interfaces. *Software IEEE* **14(2):** 18–20.

Shuming, L., Zongkai, Y., Qingtang, L. and Tao, H., 2008. Research of web information retrieval based on metadata and oai. *Granular Computing, 2008. GrC 2008. IEEE International Conference.* pp. 383–386.

STRIVE, 2008. *Submission of Data Sets from STRIVE-Funded Projects.* STRIVE Funding Guidelines. Available online at http://www.epa.ie/researchandeducation/research/docs/#datasets.

Waller, M. and Sharpe, R., 2006. *Mind the Gap: Assessing Digital Preservation Needs in the UK*. Report Prepared for Digital Preservation Coalition. PDF available online at http://www.dpconline.org.

Wilson, J., 2002. *Data and Information Management Strategy and Plan of the Global Ocean Observing System (GOOS)*. Intergovernmental Oceanographic Commission, Paris, France, GOOS Report No. 13 (ioc/inf-1168) edition.

Appendix A Download Statistics from SAFER-Data

Table A.1. Monthly download totals on SAFER-Data during 2008.

Month	Total downloads
January 2008	62
February 2008	105
March 2008	85
April 2008	114
May 2008	159
June 2008	405
July 2008	358
August 2008	378
September 2008	613
October 2008	640
November 2008	599
December 2008	1,068
Total	**4,586**

Appendix B Publications Arising from this Fellowship

Over the course of this fellowship a number of papers were written and in many cases presented at international conference events. Papers that Peter Mooney was lead author on have Mooney, P. as the first author in the author list.

Donnelly, A., Jennings, E., Finnan, J., Mooney, P., Lynn, D., Jones, M., O'Mahony, T., Thrivel, R. and Byrne, G., 2006. Objectives, targets and indicator for use in Strategic Environmental Assessment (SEA). *Journal of Environmental Assessment Policy And Management* **8:** 135–156.

Mooney, P., 2005. Web Accessibility for Dynamically Generated Websites *IASTED International Conference on Human Computer Interaction*. November 2005. Phoenix, Arizona, USA.

Mooney, P. and Winstanley, A.C., 2006. Data Interoperability – Preparing for the Challenges. *GIS Research UK*. April 2006. Nottingham, UK. pp. 10–13.

Mooney, P. and Winstanley, A.C., 2006. Data Management for Environmental Informatics – An Irish Research Perspective. *The 20th International CODATA Conference: Scientific Data and Knowledge within the Information Society*. October 2006. Beijing, China. pp. 79–81.

Mooney, P., McDonagh, C. and Winstanley, A.C., 2006. Processing Large Environmental Data Sets with SAS: Writing Robust and Dynamic SAS Macro Code. *The 31st SAS Users Group International Conference*. March 2006. San Francisco, California, USA.

Mooney, P. and Donnelly, A., 2007. Improving Data Usability for Effective Environmental Indicators in Strategic Environmental Assessment. *The IAIA Conference 07 (Growth, Conservation, and Responsibility)*. June 2007. Seoul, South Korea.

Mooney, P. and Winstanley, A.C., 2007. Proactive Approaches to ISO 19115 Metadata Gathering. *13th GI and GIS Workshop INSPIRE Time: ESDI for the Environment*. July 2007. Porto, Portugal. pp. 198–199.

Mooney, P. and Winstanley, A.C., 2007. Evaluating Interfaces to Publicly Available Environmental Information: *Human Computer Interaction 2007: Human Interface, Part I, HCII 2007, LNCS 4557*. July 2007. Beijing, China.

Mooney, P. and Winstanley, A.C., 2007. Improving Environmental Research Data Management. *Enviro Info 2007*. September 2007. Warsaw, Poland.

Mooney, P. and Winstanley, A.C., 2007. Applying Recommendation Algorithms to Collections of Geospatial Metadata. *The 9th International Conference on GeoComputation 2007*. September 2007. NUI Maynooth, Ireland.

Mooney, P. and Winstanley, A.C., 2008. Making Environmental Research Data Publicly Available – Experiences From Ireland. *Enviro Info 2008*. September 2008. Lunenburg, Germany.

Mooney, P. and Winstanley, A.C., 2009. Representations of Environmental Data in Web-Based GIS *Innovations in GIS 10.* Taylor and Francis, London, UK. In press.

An Ghníomhaireacht um Chaomhnú Comhshaoil

Is í an Gníomhaireacht um Chaomhnú Comhshaoil (EPA) comhlachta reachtúil a chosnaíonn an comhshaol do mhuintir na tíre go léir. Rialaímid agus déanaimid maoirsiú ar ghníomhaíochtaí a d'fhéadfadh truailliú a chruthú murach sin. Cinntímid go bhfuil eolas cruinn ann ar threochtaí comhshaoil ionas go nglactar aon chéim is gá. Is iad na príomh-nithe a bhfuilimid gníomhach leo ná comhshaol na hÉireann a chosaint agus cinntiú go bhfuil forbairt inbhuanaithe.

Is comhlacht poiblí neamhspleách í an Ghníomhaireacht um Chaomhnú Comhshaoil (EPA) a bunaíodh i mí Iúil 1993 faoin Acht fán nGníomhaireacht um Chaomhnú Comhshaoil 1992. Ó thaobh an Rialtais, is í an Roinn Comhshaoil agus Rialtais Áitiúil a dhéanann urraíocht uirthi.

ÁR bhFREAGRACHTAÍ

CEADÚNÚ

Bíonn ceadúnais á n-eisiúint againn i gcomhair na nithe seo a leanas chun a chinntiú nach mbíonn astuithe uathu ag cur sláinte an phobail ná an comhshaol i mbaol:

- áiseanna dramhaíola (m.sh., líonadh talún, loisceoirí, stáisiúin aistrithe dramhaíola);
- gníomhaíochtaí tionsclaíocha ar scála mór (m.sh., déantúsaíocht cógaisíochta, déantúsaíocht stroighne, stáisiúin chumhachta);
- diantalmhaíocht;
- úsáid faoi shrian agus scaoileadh smachtaithe Orgánach Géinathraithe (GMO);
- mór-áiseanna stórais peitreail.

FEIDHMIÚ COMHSHAOIL NÁISIÚNTA

- Stiúradh os cionn 2,000 iniúchadh agus cigireacht de áiseanna a fuair ceadúnas ón nGníomhaireacht gach bliain.
- Maoirsiú freagrachtaí cosanta comhshaoil údarás áitiúla thar sé earnáil - aer, fuaim, dramhaíl, dramhuisce agus caighdeán uisce.
- Obair le húdaráis áitiúla agus leis na Gardaí chun stop a chur le gníomhaíocht mhídhleathach dramhaíola trí comhordú a dhéanamh ar líonra forfheidhmithe náisiúnta, díriú isteach ar chiontóirí, stiúradh fiosrúcháin agus maoirsiú leigheas na bhfadhbanna.
- An dlí a chur orthu siúd a bhriseann dlí comhshaoil agus a dhéanann dochar don chomhshaol mar thoradh ar a ngníomhaíochtaí.

MONATÓIREACHT, ANAILÍS AGUS TUAIRISCIÚ AR AN GCOMHSHAOL

- Monatóireacht ar chaighdeán aeir agus caighdeáin aibhneacha, locha, uiscí taoide agus uiscí talaimh; leibhéil agus sruth aibhneacha a thomhas.
- Tuairisciú neamhspleách chun cabhrú le rialtais náisiúnta agus áitiúla cinntí a dhéanamh.

RIALÚ ASTUITHE GÁIS CEAPTHA TEASA NA HÉIREANN

- Cainníochtú astuithe gáis ceaptha teasa na hÉireann i gcomhthéacs ár dtiomantas Kyoto.
- Cur i bhfeidhm na Treorach um Thrádáil Astuithe, a bhfuil baint aige le hos cionn 100 cuideachta atá ina mór-ghineadóirí dé-ocsaíd charbóin in Éirinn.

TAIGHDE AGUS FORBAIRT COMHSHAOIL

- Taighde ar shaincheisteanna comhshaoil a chomhordú (cosúil le caighdéan aeir agus uisce, athrú aeráide, bithéagsúlacht, teicneolaíochtaí comhshaoil).

MEASÚNÚ STRAITÉISEACH COMHSHAOIL

- Ag déanamh measúnú ar thionchar phleananna agus chláracha ar chomhshaol na hÉireann (cosúil le pleananna bainistíochta dramhaíola agus forbartha).

PLEANÁIL, OIDEACHAS AGUS TREOIR CHOMHSHAOIL

- Treoir a thabhairt don phobal agus do thionscal ar cheisteanna comhshaoil éagsúla (m.sh., iarratais ar cheadúnais, seachaint dramhaíola agus rialacháin chomhshaoil).
- Eolas níos fearr ar an gcomhshaol a scaipeadh (trí cláracha teilifíse comhshaoil agus pacáistí acmhainne do bhunscoileanna agus do mheánscoileanna).

BAINISTÍOCHT DRAMHAÍOLA FHORGHNÍOMHACH

- Cur chun cinn seachaint agus laghdú dramhaíola trí chomhordú An Chláir Náisiúnta um Chosc Dramhaíola, lena n-áirítear cur i bhfeidhm na dTionscnamh Freagrachta Táirgeoirí.
- Cur i bhfeidhm Rialachán ar nós na treoracha maidir le Trealamh Leictreach agus Leictreonach Caite agus le Srianadh Substaintí Guaiseacha agus substaintí a dhéanann ídiú ar an gcrios ózóin.
- Plean Náisiúnta Bainistíochta um Dramhaíl Ghuaiseach a fhorbairt chun dramhaíl ghuaiseach a sheachaint agus a bhainistiú.

STRUCHTÚR NA GNÍOMHAIREACHTA

Bunaíodh an Ghníomhaireacht i 1993 chun comhshaol na hÉireann a chosaint. Tá an eagraíocht á bhainistiú ag Bord lánaimseartha, ar a bhfuil Príomhstiúrthóir agus ceithre Stiúrthóir.

Tá obair na Gníomhaireachta ar siúl trí ceithre Oifig:

- An Oifig Aeráide, Ceadúnaithe agus Úsáide Acmhainní
- An Oifig um Fhorfheidhmiúchán Comhshaoil
- An Oifig um Measúnacht Comhshaoil
- An Oifig Cumarsáide agus Seirbhísí Corparáide

Tá Coiste Comhairleach ag an nGníomhaireacht le cabhrú léi. Tá dáréag ball air agus tagann siad le chéile cúpla uair in aghaidh na bliana le plé a dhéanamh ar cheisteanna ar ábhar imní iad agus le comhairle a thabhairt don Bhord.

NDP
National Development Plan 2007 - 2013

The EPA's Environmental Research Centre (ERC) was established as a centre of excellence under the National Development Plan (NDP) to build capacity in environmental data handling, modelling, assessment and guidance. The objective of the ERC is to allow for a more structured approach to environmental research, through the development of advanced innovative techniques and systems to address priority environmental issues and thereby support environmentally sustainable development.

epa
Environmental Protection Agency
An Ghníomhaireacht um Chaomhnú Comhshaoil

ENVIRONMENTAL PROTECTION AGENCY
PO Box 3000, Johnstown Castle Estate, Co. Wexford, Ireland
t 053 916 0600 f 053 916 0699
LoCall 1890 33 55 99
e info@epa.ie w http://www.epa.ie

SSTI
Strategy for Science, Technology and Innovation

Comhshaol, Oidhreacht agus Rialtas Áitiúil
Environment, Heritage and Local Government