EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
 ISBN: 978-3-8322-6397-3

# Improving Environmental Research Data Management

Peter Mooney[1], Adam C. Winstanley[2]

**Abstract**

In this paper we discuss the management of Environmental Data generated by Research programmes in Ireland. We describe a web-based software system to manage metadata and data archival. Research groups have outlined a number of barriers preventing sharing and exchange of environmental data on a national and international scale. By addressing these issues directly a more acceptable approach to research data exchange is possible. The paper also outlines a number of issues for future work in management of environmental research data.

## 1. Introduction

Large volumes of environmental data are generated by scientific research groups. In the majority of cases these groups are based in universities, colleges, and specialized institutes of research. Many of these research groups carry out environmental monitoring and additional analysis of collected data. Other groups focus specifically on environmental modeling and related activities. The vast majority of environmental datasets generated by these groups typically have a longer lifespan than the duration of the research project (usually the funding period) that created them. The specialized nature of this data goes beyond the thematic area within which it was original generated. For example, data collected as part of a Biodiversity study could be used at a later stage as part of a Strategic Environmental Assessment or Transboundary Air Pollution monitoring used in environmental health impact studies.

These datasets have unique temporal characteristics. Another key characteristic of the data is related to detailed local geospatial information. For the majority of research groups significant effort would be required to make this data available for others to use. As a result these valuable datasets often reside on a local server machine, un-accessible outside the local network, and hidden from more national or international access. Generally it is not the responsibility of the research group to develop a fully interoperable database (or data access system) for the data generated by their research (PhD, MSc, Large Scale funded project, etc). Neither is it a general requirement to provide Internet-based access to the data. This results from the simple fact that their primary tasks involve applying their scientific expertise to the analysis and modeling of the generated or collected data.

The research communities to which we (Environmental Protection Agency Ireland) have a data management responsibility agree that they would be more willing to share their environmental datasets if the following conditions were realised:

1. The creation of metadata becomes an easier task. This task should not require the researcher to have detailed knowledge of the metadata standards (ie ISO 19115 or Dublin Core)
2. The data archive system must be completely Internet-based. As a result standard Internet browser software and FTP tools should be sufficient to use the system.
3. The researchers can upload their datasets to a data archive system. This system then makes their data available on the Internet without requiring them to manage any internet server infrastructure.

[1] National Center for Geocomputation, Department of Computer Science, National University of Ireland, Maynooth, Co. Kildare, Ireland. Telephone: +353 (1) 268 0100 (Peter Mooney)
e-mail: {peter.mooney, adam.winstanley}@nuim.ie
[2] Environmental Research Center (ERC), Environmental Protection Agency, Richview, Clonskeagh, Dublin 14. Ireland

4. The researcher can have their datasets stored in a long-term data archive that they do not have to maintain themselves.
5. The researcher can maintain an embargo on the distribution of the actual raw data for a period of up to 12 months after the project end. This will allow the research team to write peer-reviewed journal material and retain exclusive first use of the data.

Pervasive issues such as evolving technology, software requirements, and the multiplicity of data formats continue to increase the complexity of managing geospatial data (Azberger et. al, 2004). The key requirement is to remove the burden of overall management of the data from the researcher. However, we have found a number of reasons for reluctance to share and exchange data:

- Some scientists fear sharing imperfect data because of the perceived negative effects on their scientific reputation or research work;
- Exposure to technology outside the scientists area of expertise – those outside the area of geospatial software development are generally unclear about IS0 19115, Web Services, data warehousing, SDI, etc. and often feel that data warehouses operate in black box fashion.

With these issues in mind the next section gives a high level description of our software system.

## 2. System Description

Several Free and Open Source Software (FOSS) systems are available for management of metadata and the management of geospatial data. One of the best FOSS tools for metadata management is CatMDEdit (http://catmdedit.sourceforge.net/). For management of spatially referenced data resources GeoNetwork (http://geonetwork-opensource.org/), developed by the UN FAO, is a powerful metadata editing, metadata searching, and interactive web map viewer. Both of these tools have many advantages. For example, GeoNetwork has an integrated OGC compliant WMS. However, given the specific needs and requirements of our user community it was decided to develop a software system that would address our specific needs.

Our system has been developed as follows. All software development has been carried out in Java using Java Server Pages (JSP) for web deployment. Apache Tomcat is used as the application container software. This is integrated with Apache HTTP Server. A MySQL database is used for all backend business logic. The databases the stores all metadata in database tables, manages dataset file information within the archive, user profile management, and system access logging. Several other Java tools assist in providing functionality in the system. The iText (http://sourceforge.net/projects/itext/) Java library allows the production of PDF files from JSP pages and Java Servlets. The Apache Commons File Upload (http://jakarta.apache.org/commons/fileupload/) Java library provides a very robust and scalable set of functionality for data file upload. Finally, the Apache Jakarta POI Java library (http://jakarta.apache.org/poi/) provides a very robust API for access to Microsoft files such as Excel and Word. All datasets in the data archive are stored on a 2 Tb Storage Area Network (SAN). At present the majority of datasets in the data archive are time-series based. There are also several collections of GIS datasets.

Towards the end of the project lifetime (end of PhD or funding period for example) the principal investigator (PI) of each research group is provided with a username and password to the system. The PI delegates a member of the group the responsibility of creating the metadata and uploading the data. When the researcher accesses the system for the first time to upload data they must first create a metadata record for the dataset or dataset collection. Metadata creation is provided through a simple web-based form. Only after the metadata has been successfully created can the researcher upload files associated with this metadata record. After successful metadata creation the user can subsequently edit metadata as required and

upload/replace files in the dataset collection. An important feature of the metadata creation is that the researcher can specify if the datafiles associated with the metadata record are to be made publicly available. This is done through the use of simple drop down list of options where datafiles are either {private, semi-public, fully public}. Semi-public means that other users can see that the metadata resource has a data collection associated with it but they will need to contact the research group directly for permission to access the data. At each logon the authorized user is reminded by the system of any dataset collections which are close to embargo expiration (usually 12 months after initial metadata record creation). After this period datafiles connected to a metadata resource are automatically switched to fully public status.

## 3. Conclusions

Publicly funded research data should be openly accessible to the maximum extent possible (Downs and Chen 2005). In this paper we have outlined several of 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 we have developed an Internet-based data archival system for datasets from environmental research programmes. This system greatly simplifies metadata creation, data submission and upload, and long term data archival for the environmental research community. This system has been developed under as a FOSS system. It is hoped that when the system has reached an appropriate stage of maturity that be made available to other data managers for can deploy on their own systems. It is also intended that the system will expand to other scientific domains outside that of environmental science.

The MySQL database design allows other services on our Internet servers to query the metadata database. This provides simple service chaining and adds value to other services for data discovery and data integration. Detailed logging of system activity is stored in the database allowing us to analyse use-case behavior as well as compiling various statistics about system usage (most popular download, length of user session, etc).

### 3.1 Further Work

The INSPIRE initiative intends to create a European-wide spatial data infrastructure delivering integrated spatial information services to users. The exact benefits of such an infrastructure may not be obvious to those currently working outside the area of SDI. Data managers and database administrators are acutely aware of the difficulties of large-scale geospatial data integration given the characteristics of the current European data landscape. The benefits offered by INSPIRE, such as seamless combination of data from different sources across Europe, are deserving of significant effort from all players. We feel that the environmental science research community must be provided with the proper tools and services to allow them to become part of a European wide SDI. Our system and overall data management approach attempts to remove some of the most commonly cited barriers to data exchange and data -sharing.

Systematic evaluation of the exact re-use of data from this system is difficult. It is extremely difficult to ascertain the exact re-use (if any) of the data that has been made available. The system logs the number of data sets downloaded, queries, and viewed. Other statistics are derived from WWW server log files. Unfortunately this information is of quite limited value. Those downloading the data are requested to cite the original producers of the data and how and where they obtained the downloaded datasets. The embedding of non-removable identify information in uploaded files is one possible way to address this problem.

Another problem we have encountered is that scientists are given little reward for taking time to submitting data to a central data repository. Submitting data often includes auxiliary tasks of metadata creation, data formatting, data documentation, and data QA/QC. Academic recognition is directed toward peer review publication rather than toward the management and sharing of data.

Finally, there is great potential for the digitization of environmental research data that is currently not in digital format. This data exists in multiple representations: photographs, field notebooks, collected specimens, environmental measurement apparatus. While the task to provide access to these resources is difficult the obvious results are tremendous.

## Acknowledgments

This work is being carried out as part of an Environmental Research Center (ERC) Research Fellowship programme (2002-CC-2FS). Financial support for this work is provided by the Irish Environmental Protection Agency's ERTDI programme under the National Development Plan 2000 – 2006.

## Bibliography

Azberger, P., Schroeder, P., Beaulieu, A., Bowker, G. (2004): Promoting Access To Public Research Data for Scientific, Economic, and Social Development. Data Science Journal, 3(29): 135-151

Downs, R. and Chen R.S. (2005): Organisational Needs For Managing And Preserving Geospatial Daa and Related Electronic Records. Data Science Journal, 4(31): 255-271.