A Case Study: Management and Exploitation of the Nuclear Decommissioning Agency Geoscience Data Archive

Jaana Pinnick[1], Andrew Riddick[1], Robert McLaverty[2] and Garry Baker[1]

1 British Geological Survey, Environmental Science Centre, Keyworth
Nottingham NG12 5GG jpak@bgs.ac.uk
2 Radioactive Waste Management, Curie Avenue, Harwell Oxford
Didcot OX11 0RH

Abstract. The British Geological Survey (BGS) is responsible for managing a major geoscience data archive on behalf of the UK Nuclear Decommissioning Authority (NDA). Much of this geological data was captured during the 1990s and early 2000s using now obsolete software and data formats. This data asset remains an important resource for the NDA and the wider scientific community. The NDA wishes to ensure the data remain accessible and usable for many decades into the future. BGS has been working closely with Radioactive Waste Management (RWM), a wholly owned subsidiary of the NDA, on a programme of data management and digital continuity measures to ensure the long-term usability of the data. This paper describes some of the challenges and outlines the approaches we have taken to address these issues.

Keywords: Data preservation, digital continuity, case study, geoscience

1 Introduction

This paper describes the results of a programme of data management and data preservation activities undertaken by the British Geological Survey (BGS) to actively develop and maintain a geoscience data archive owned by the Nuclear Decommissioning Authority (NDA). Throughout this task BGS has worked closely with Radioactive Waste Management (RWM), a wholly owned subsidiary of the NDA [1], and the organisation tasked with managing the maintenance of the archive. The archive contains a wide range of geoscience data formats collected mainly during geological investigations in the 1990s and early 2000s. Much of the software in which this data was created is now obsolete, and some of the data formats are challenging or impossible to read in modern software. The time which has elapsed since the data was originally captured also means that supporting information about the data, such as details of software tools and methods used, is often limited and generally not present in a consistent form.

The data contained within the archive was captured from analyses and observations, which are difficult or potentially very expensive to repeat but nevertheless remain relevant to RWM's ongoing business operations. There is a key business requirement for this data to remain accessible and usable by RWM and their consultants over many decades into the future. The continued accessibility and usability of this data also remains an important data resource for the geoscience research community in general.

A programme of data management and digital preservation activities has been implemented in order to meet the business objectives and to fulfill an ambition to create an "openly available" archive of intensely geologically studied regions for use by academia, industry, research and members of the public. This work draws on emerging methods and techniques in the field of digital preservation which BGS is increasingly applying to its own data assets and to other geoscience archives which it maintains. Key elements in the digital continuity strategy for the NDA archive include obtaining a comprehensive understanding of the data and file formats, capturing detailed preservation metadata to support ongoing preservation activities and to maintain the understandability of the data, and enhancing the accessibility and usability of the data through the construction of appropriate digital indexes to assist stakeholders in locating and using the data they need.

2 Approach and Methodology

An important element of our approach has been to undertake an initial analysis of the data and file formats present within the archive in order to fully understand the digital preservation issues we need to address. This has been progressed by file profiling of the whole archive using the DROID file format identification tool [2] and the PRONOM file format registry [3] developed by the UK National Archives (TNA). These open source tools have provided a rapid and cost effective means of gathering useful preservation metadata about the file formats which, in turn, allows the identification of digital preservation issues and informs decisions on likely data migration or other data rescue strategies. For example, this analysis has generated data about the true format of the files, which may differ from that suggested by the file extension. This investigation is described in more detail below but has allowed us to classify the data into e.g. those formats which we can identify and often still make use of in their present form, and those which may require some type of transformation or migration actions in order for the data to remain accessible and usable. The preservation metadata gathered in this way has enabled risk assessments to be made for different groups of file formats, allowing resources to be prioritized appropriately for data rescue activities.

The archive includes several database export files in various legacy formats. These contain useful geoscience data sets of long-term value. An important focus of the work programme has been to migrate this data into modern relational database formats (e.g. Oracle ©) to ensure long-term accessibility and usability.

Although this work focusses on digital data management and preservation issues, another work stream within the project has sought to improve the accessibility of analogue data collections, which include various geological samples and thin section microscope slides. It became clear from the initial analysis of the data within the archive that whilst metadata about these analogue collections was available in digital form, it had been assembled by differing consultants at different times and was not gathered in compliance with an easily usable standard. Consequently, it was not always easily to locate samples of interest to a user. This was addressed by the creation of a unified digital index to the microscope slide collection. The usability of the microscope slide collections was further enhanced by imaging of the each of the slides allowing high-resolution digital images to be added to the collection, in formats facilitating long-term preservation.

3 File Format Profiling and Analysis

3.1 Background and Identified Files

An initial data analysis phase showed us that there were 74 different file types in the NDA archive which were recognised by the DROID tool during data scanning, meaning that they have a PRONOM Persistent Unique Identifier (PUID), such as *x-fmt/18* for *.csv* files. Using DROID enabled us to identify different versions of the same file type e.g. Excel and Lotus spreadsheets (Table 1) for prioritizing preservation actions.

Table 1. Different versions of Excel and Lotus spreadsheets

PUID	Version	Total file count
fmt/56	MS Excel 3.0	386
fmt/59	MS Excel 5.0/95	137
fmt/214	MS Excel 2007 onwards	58
fmt/61	MS Excel 97 (8)	15
fmt/172	MS Excel for Macintosh 3	3
fmt/173	MS Excel for Macintosh 4	3
fmt/174	MS Excel for Macintosh 98	3
fmt/175	MS Excel for Macintosh 2001	3
fmt/176	MS Excel for Macintosh 2002	3
fmt/177	MS Excel for Macintosh 2004	3
fmt/178	MS Excel for Macintosh X	3
fmt/60	MS Excel 95 (7)	3
fmt/57	MS Excel 4.0 (4S)	1
x-fmt/117	Lotus 1	2
x-fmt/114	Lotus 2	83
x-fmt/115	Lotus 3	112
x-fmt/116	Lotus 4-5	4

The files identified by DROID included 16,779 *.txt* files, 4,137 *.csv* files, 3,284 *.tif* files and 2214 *.sgy* files totalling ~26,000 out of the ~36,500 files in the archive (corresponding to 72% of the data volume) between these four primary file types alone. Over 13,000 of the *.txt* files were originally created in other formats (e.g. *.in, .edi, .avg*) and subsequently saved in *.txt* format to ensure the preservation of content but compromising on functionality and instant usability. This action was carried out by BGS to ensure we can at least open the file and interrogate the contents while acknowledging the original format. These data include gravity and survey data and geological investigations.

3.2 Analysis of Files not Identified by DROID

A subsequent data analysis task revisited the remaining 28% file types which were unidentified by DROID in the previous phase. This left 280 individual file extensions which were not recognized. In some cases, an extra layer of complexity was added to the identification due to the extension having been manually modified by the creator or user at some point in the past, e.g. *.bgs* or *.bef_nhpi_report*. These local file extensions may have been formed as part of the original data management processes of the source consultancy or contractor.

Identification Based on Folder Directory Structure. In the case of 200 of data formats it was possible to deduce the source software relatively reliably from the location of the data using the embedded logic of folder naming and structure. Some Readme notes were also included in the archive providing detailed information on the creation software, naming conventions and directory structure. These 200 formats consist mainly of Vulcan, Earth Vision and Bentley MicroStation 5.0 data files, and they are currently (July 2017) not represented in PRONOM. The versions of Vulcan and Earth Vision data in the archive are examples of now obsolete geological modelling formats, whereas Bentley MicroStation was used by UK Nirex Ltd as a CAD/GIS software product. All these still exist in a modified and enhanced form. In addition to the unidentified formats, there were 78 files without a file extension but thought to be either Vulcan or Bentley MicroStation files, based on their location in the file directory.

File Formats Unidentified by DROID. The total file count of the remaining 80 unidentified formats (e.g. *.mtx, .sta, .his*) was 4,088 individual files, and these were further analysed for current accessibility. A closer analysis of the files indicated that the unidentified data formats concentrate on one particular folder, the data in which was extracted from one DLT (Digital Linear Tape) tape, making it a localized issue. Other smaller volumes of files, which at a closer inspection turned out to be mostly Vulcan and Bentley MicroStation data, were dispersed elsewhere in the archive.

Partially Readable Files. There are a number of salinity mapping data folders containing 3,440 partially readable files, where the information contained in the file header is accessible using a text editor providing limited metadata of the data creation such as date, time, survey ID or the operator name (Fig. 1).

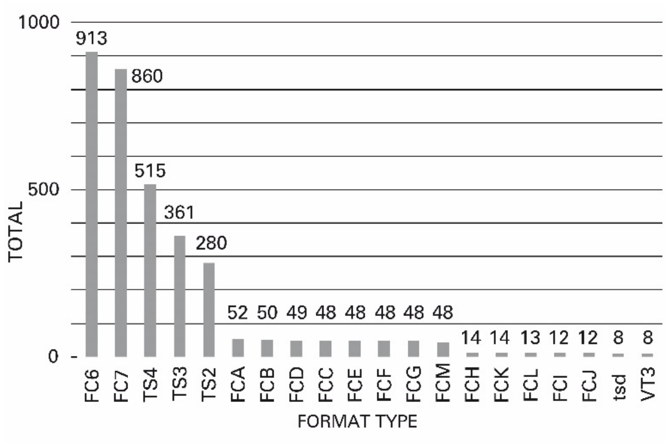


Fig. 1 Top twenty partially readable data files

However, the data following the header is scrambled and presently unreadable. Examples of these file extensions include *.FC6*, *.SOR* and *.M91*. It is not certain whether these data are raw, processed, or modelling data, and this will need to be investigated in conjunction with scientific domain experts. Fortunately, some of the individuals who worked on the data twenty years ago can, in this case, still be contacted to try and obtain more details.

***.dat* files.** The 107 individual *.dat* files vary from containing fully readable and reusable numerical data, to having readable data that lacks metadata and column headers. It was hence not possible to determine what data they contain. However, it has been possible to trace the source software for some of these files from Readme notes added to the archive at the time of data creation or archiving.

Other numerical data files. In a number of cases (59), the entire data file contents are fully readable showing seismic or gravity data with column headings. In a few cases (31), mirroring the behavior of some of the *.dat* files, they consist of gravity or survey data being accessible but lacking metadata e.g. no column headings, making the reusability of the data questionable unless additional metadata can be discovered in the archive.

Unreadable files. The number of completely inaccessible data files, i.e. not readable using a text editor (Fig. 2), is at 79 in 6 file formats, a relatively small subset of the entire archive.

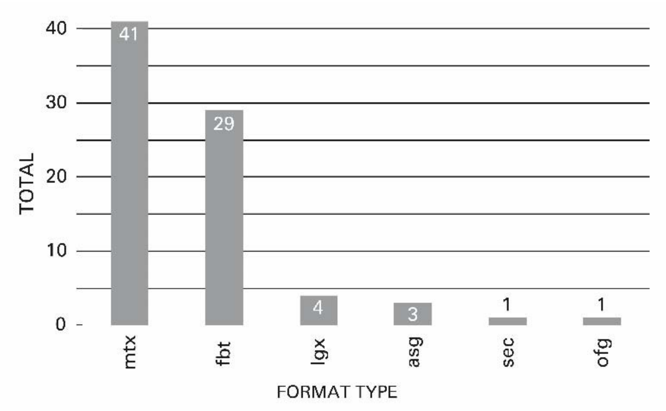


Fig. 2 Unreadable files

3.3 Findings

Although the current data rescue project only started in 2014, many best practice preservation activities have taken place within the archive in the past including transforming a large part of the data into *.csv* and *.txt* formats and creating "Readme" notes within the archive to provide additional contextual and technical metadata. The archive has been stored as three independent copies, one on site, one in a geographically separate location, and remaining copy on tape. This follows the best practice redundancy principle and strengthens the long-term preservation of the data.

However, information on past preservation events has not been systematically captured as fully documented preservation metadata. It is now strongly recommended that a metadata schema be developed to build a comprehensive picture of all migration, transformation and data rescue activities that have taken place so far, and to ensure all future preservation events are added to the metadata record as they take place. Fixity checks, which would show any changes within the data files, were not made initially, but it is now planned to introduce these and to monitor the integrity of the collection. Having multiple copies of the data will enable us to replace any corrupt files if loss is detected.

It is clear from this analysis that the data partially readable using a text editor software is a significant subset of the unidentified data formats, and that decisions are needed on how to deal with these files where only content metadata headers are available. If some of the data is interpreted or modelled data, then it is considered within the context of this project to be a lower priority for data rescue, as the future users would in all likelihood recreate any future models directly from the available raw data.

We have followed with interest the digital signature development work done at the University of York [4] and are considering the possibility of attempting to do this with some of the older and unidentifiable geospatial data formats, resources permitting. However, the long tail of research data within this archive, where the data includes less than ten instances of a particular format, may not be worth the data rescue effort they would incur. It would be difficult to create digital signatures for these formats, as the body of material may not be sufficient to complete the process. This would ideally require us to locate additional examples elsewhere to facilitate the identification of common patterns in the data, as recommended by the National Archives guidance on the topic [5].

4 Data Rescue Activities

4.1 Early Days (1997-2000)

An earlier project managed by BGS for UK Nirex Ltd (predecessor to the NDA) created the Digital Geoscience Database (NDGD) [6]. The NDGD and its associated Helpdesk was always mindful of the future access and use of the data being captured and held. This was due to the initial scope of the NDGD being documented as providing long-term access (20+ years) to an integrated suite of geoscience data captured by a variety of organisations and consultancies working for UK Nirex Ltd. At the end of the project with this long-term vision in mind, the moth-balling NDGD contract (1997-8) to shut-down the helpdesk and various BGS managed systems including the data held was undertaken with preservation and future access firmly in mind. The data in the NDGD Oracle © RDBMS and GIS systems was gathered and exported in both the best system exports formats (as provided by the software) and critically as standardised data or text files (*.dat*, *.csv*, *.txt*) including appropriate documentation, manuals and scientific dictionaries that had been used to constrain the data held in the database(s). A further step of hard copy print out of the data was also taken for the majority of the data held as a final data rescue option if all else failed and future access was required. For its time this was a very forward thinking suite of preservation activities.

4.2 Preservation Actions Prior to the Current Project

The transfer of the UK Nirex Ltd data archive to the BGS in 2000 included hard copy records, scientific reports and data captured upon a variety of media and systems.

To best provide and maintain the existing archive contents some critical datasets were rebuilt from export files onto the BGS enterprise RDBMS solution (Oracle ©) to ensure easier access and maintenance on currently supported media and with licensed software. The remaining data was, where possible, copied to BGS corporate maintained 'Storage Area Network' (SAN) which included the reading of some historical tapes, drives and digital media types and data migration to the BGS SAN. This work has continued into the current project. The documentation of the formal status of the data contained within the archive, including an 'information asset register' and appropriate documentation in technical reports to facilitate understanding data, formats, volumes, sources and locations was created and stored alongside the archive.

4.3 Work Done Within the Current Project

During the current data analysis it was discovered that a fairly large number of files (~13’000) had been previously transformed into *.txt.* files, with the original file extension having been kept as part of the file name e.g. SA1806.AVG.TXT, BA1108_1.IN.TXT, or MAG4.DAT.TXT. About 160 different file types have been transformed in this way earlier in the data life cycle, presumably as part of a data rescue effort as plain text files are more stable than most proprietary file formats. These events have been recorded within the curation workflow, but information about them will need to be formally and systematically captured as part of the preservation metadata creation and collection activity once the metadata schema has been agreed and finalized.

In addition, two other file types (*.prn* and *.mrg*) have been saved as *.csv* files, which is currently one of the BGS preferred or acceptable formats due to its stability.

Some database export files for critical datasets were migrated back into a modern version of the Oracle relational database management system prior to the current project, as described above. Within the current project, this migration process was extended to several other datasets which had been exported from databases originally created by various consultants prior to the year 2000, to ensure the continued accessibility to this data also. These datasets included the original Nirex Groundwater Database, a time series dataset recording the variations in various chemical entities measured within boreholes, and some chemical analysis data. Migrating these datasets involves re-constructing the original data model using supporting documentation where available and sometimes inferring relationships based on experience.

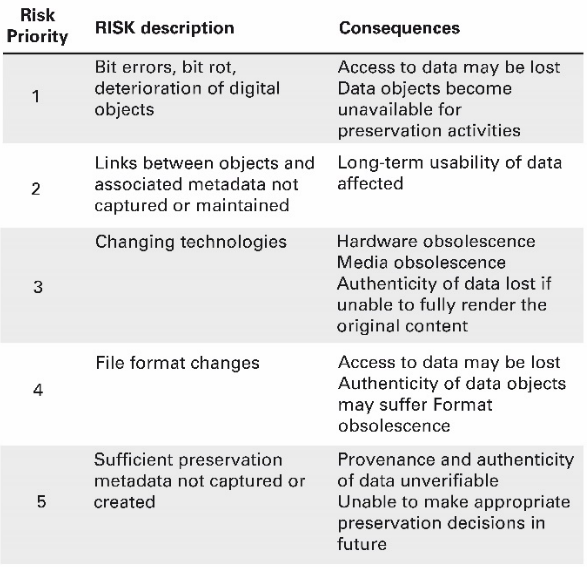
It is likely that the Oracle database platform will remain for a significant time into the future as the basis of the BGS core d ata store. In the unlikely event that this database platform becomes unsupported at some point long into the future, the relational data structure will facilitate ready migration to another chosen corporate product or format.

5 Risk Analysis

Peyrard *et al.* have recommended using the SPOT (Simple Property-Oriented Threat) model as a starting point for requirements analysis for digital preservation metadata [7]. As we had already decided to include a risk assessment in the project, we decided to use the SPOT model [8] which focuses on safeguarding against threats to six essential properties of digital objects: availability, identity, persistence, renderability, understandability and authenticity.

We then familiarized ourselves with the PREMIS Preservation Metadata Dictionary (version 3.0) [9] with a view of implementing it for the archive next year. This will require customizing the schema to include any specific geoscience and geospatial requirements, and there remains work for us to do on this task.

We started to develop a risk matrix using the properties of the SPOT model as a starting point and wrote a short description for each identified risk in our archive, before aligning them with PREMIS entities to strengthen the use of best practice preservation standards (Fig.4). In some cases the risk description in the matrix linked to more than one essential property in the SPOT model; we attempted to select the most appropriate of these.



Risk Priority	RISK description	Consequences
1	Bit errors, bit rot, deterioration of digital objects	Access to data may be lost Data objects become unavailable for preservation activities
2	Links between objects and associated metadata not captured or maintained	Long-term usability of data affected
3	Changing technologies	Hardware obsolescence Media obsolescence Authenticity of data lost if unable to fully render the original content
4	File format changes	Access to data may be lost Authenticity of data objects may suffer Format obsolescence
5	Sufficient preservation metadata not captured or created	Provenance and authenticity of data unverifiable Unable to make appropriate preservation decisions in future

Fig. 3. An extract from the risk matrix

The next stage involved identifying the consequences of not addressing the risk, as well as the impact, likelihood and rank of the risks, to help prioritise future preservation actions. We suggested some risk management and mitigation methods together with suitable tools, which may aid to minimise the risk in each case. Finally, we identified some risk types within the data. A further prioritisation exercise, undertaken in conjunction with the RWM data managers and taking into account the science and business value of the data, will need to be completed to decide the most appropriate preservation actions. These will be later added to a customised preservation metadata schema.

6 Conclusions

The file profiling analysis undertaken so far has provided a very useful collection of metadata and experience on which to base further work. For example, decisions are needed in conjunction with science and data value judgment by RWM on how to deal with both unreadable files and files where only header metadata is presently accessible. If these are interpreted or modelled data, then further actions are currently thought to be of lower priority than if they were raw data files. In addition, where column headings of numerical data are missing, there is a risk of eventually losing the context and therefore usability of these files, especially when the staff expertise and corporate knowledge of data is lost. The long tail of research data however, where the dataset includes less than 10 instances of a particular file type or extension, may not justify the extra effort on preservation actions (subject to RWM science data value evaluation).

It is recommended to integrate all available or gathered preservation metadata into a schema held upon a relational database built upon the PREMIS metadata model. The PREMIS standard lends itself to implementation in a RDBMS, and the use of unique identifiers, not only for the data objects but also for events, actions and agents. This could include a semantic component to monitor changes in the vocabulary, and hash algorithms to monitor any changes, including bit rot, that may occur in the digital data over time. Other fields to be included in the metadata schema have also been suggested.

An initial risk matrix lists the main twenty areas of risk and relates them to the six SPOT risk types and PREMIS metadata entities. It ranks the risks based on a simple traffic light risk assessment and suggests risk management methods as well as preservation tools that could be used to mitigate against these risks. This may be used to inform the development of the tailored preservation metadata schema.

In addition to supporting the digital continuity of the data archive, this work is also informing the development of data management processes and systems for future geological investigations currently being planned by RWM. This linkage also demonstrates the importance of integrating digital continuity measures within existing processes and of continuing to monitor any changes within and outside the archive, to ensure the preservation activities remain relevant and fit for purpose.

References

1. Nuclear Decommissioning Authority, https://www.gov.uk/government/organisations/nuclear-decommissioning-authority
2. DROID File Format Identification Tool, http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
3. PRONOM File Format Registry, http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
4. Digital Archiving at the University of York, http://digital-archiving.blogspot.co.uk/2016/08/my-first-file-format-signature.html (2016)
5. The National Archives: How to Research and Develop Signatures for File Format Identification (2012)
6. Shaw, R.P., Fortey, N.J., Turner, G., Kemp, S.J., Wheatley, C., Rowley, W.J. and Baker, G.R. Nirex Geological Archive, http://www.bgs.ac.uk/downloads/start.cfm?id=294 (2002)
7. Peyrard, S., Dappert, A. and Guenther, R.S.: How to Develop a Digital Preservation Metadata Profile: Risk and Requirements Analysis. In: Dappert, A., Guenther, R.S., Peyrard, s. (eds.) Digital Preservation Metadata for Practitioners: Implementing PREMIS, pp.11-21. Springer, Cham (2016)
8. Vermaaten. S.: Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. D-Lib Magazine 18 (9/10) (2012)
9. PREMIS Data Dictionary for Preservation Metadata version 3.0, http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf (2015)