The Centre for Environmental Data Archival Format Audit, Appraisal and Strategy Development

Esther Conway [(1)], Sam Pepler [(1)], Graham Parton [(1)] Wendy Garland [(1)] Charles Newey [(2)] Harry Jones [(1)], Robyn Pinder [(3)]

[(1)] ***CEDA***
STFC, Rutherford Appleton Laboratory, Didcot, OX11 0QX, UK
EMail: esther.conway@stfc.ac.uk

[(2)] ***Aberystwyth University***
Penglais, Aberystwyth, Ceredigion, SY23 3BF, UK
EMail: ccn4@aber.ac.uk

[(3)] ***Southampton University***
University Road, Southampton, SO17 1BJ
EMail: rcp1g11@soton.ac.uk

ABSTRACT

This paper describes the Centre for Environmental Data Archival format audit. It discusses the techniques used to conduct a broad and rigorous audit of an archive containing heterogeneous scientific data which has been acquired over a 25 year period. We discuss the successes and limitations of the CEDA format diagnostic toolkit along with results acquired during the first automated sweep of the archive. We present an analysis of these preliminary results and consider the implications for preservation strategy development.

Keywords: Digital Preservation, Risk Management, Representation Information, Environmental Data, Data Value, Strategy Development

INTRODUCTION

The Centre for Environmental Data Archival (CEDA)[1], hosts a range of activities associated with environmental data. These activities include running a number of key environmental data centres for the environmental research community, including the British Atmospheric Data Centre (BADC)[2] and the NERC Earth Observation Data Centre (NEODC)[3]. These data centres have a responsibility to curate data produced by the Natural Environment Research Council (NERC)[4]. In this capacity it has amassed extensive data holdings from diverse sources, with supporting metadata and documentation. As a result we are tasked with the challenge of maintaining the readability of heterogeneous collections of data, and supporting grey literature.

This paper describes the technical and organisational issues which arise from performing a real-world, archive-wide format audit with limited resources. It examines to the scope and nature of meaningfully defining format types. We proceed by examining a range of techniques which can be employed to identify such formats, appraising their respective benefits, associated costs and inherent risks.

---

[1] Centre for Environmental Data Archival www.ceda.ac.uk

[2] British Atmospheric Data Centre www.badc.ac.uk

[3] NERC Earth Observation Data Centre www.neodc.nerc.ac.uk

[4] Natural Environment Research Council www.nerc.ac.uk

We continue by presenting the preliminary format audit results describing the implications for the archive. We then discuss how format risks can be assessed in terms of data significance/value, format complexity, stability through community of use, availability of formalised description, availability of software and support by external organisations/communities.

We conclude by providing examples of how cross-archive resources can be developed and associated with datasets on a format basis, improving both the immediate (re-)usability and long-term stability of archived data.

THE CEDA DIAGNOSTIC TOOLKIT

State of the art

There are several existing tools that can perform file identification, which are used by organisations such as the UK National Archives and the National Library of New Zealand. One of the earlier-developed toolkits DROID[5] has served as a base for some of the other file format identification tools FITS[6], JHOVE[7], C3PO[8], and more. There are also tools that do not make use of the DROID technology such as the UNIX file command, but none of these seemed appropriate for our purposes, for the reasons detailed below.

The need for a bespoke solution

The preliminary sampling tool produced a list of over 19,000 different file extensions – which was partly due to the vast number of file formats in the archive, and also partly due to the many ad-hoc file naming conventions that had been used by the data producers. The majority of the file extensions were also misnomers – a file extension was very often unrelated to the file's format. This presented a problem in that reliably implementing large-scale format identification would need to actually read the files in question – to extract and compare file signatures, endian-ness, and other information. There are many different tools that do this already – for example, DROID. There are even solutions for text-based files, such as Apache Tika[9], which looks at structured text files such as HTML and XML. However, the largest problem that we encountered with DROID, Tika and other "off-the-shelf" file identification tools is that for the most part, they don't support the vast majority of the specialist formats that the BADC and NEODC have to curate as an archive.

This presented a problem: for the purpose of our task, standard file identification tools were of little use. This led us to develop our own solution based upon the Python libmagic library[10] . We were able to develop a toolkit that also recognised the uncommon file formats that we encountered during the course of the project. The toolkit consisted of a Python script that heavily used regular expressions to match file signatures (magic numbers), file headers, and more. As we encountered unknown files formats, these were opened to see if they contained a pattern that could be matched, whether in the filename or inside the header – the regular expressions derived from these patterns would then be added into a configuration file which was compiled at runtime by the application, and compared against a list of files that was provided to the application.

---

[5] DROID http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm accessed 14/10/2014

[6] FITS http://wiki.opf-abs.org/display/SPR/File+Format+Identification+and+Metadata+Extraction+using+FITS

[7] JHOVE http://sourceforge.net/projects/jhove/

[8] C3PO http://www.openplanetsfoundation.org/blogs/2012-11-19-c3po-content-profiling-tool-preservation-analysis

[9] Apache Tika http://tika.apache.org/

[10] Python libmagic https://pypi.python.org/pypi/filemagic/

Compromises based on speed and scale

BADC and NEODC holdings are extensive, comprising of around 2PB of data and 137 million individual files. The scale of the archive raises significant challenges to deriving meaningful results with a limited timescale. A simple script to parse the filenames of the entirety of BADC and NEODC holdings was run to determining the range of file extensions that were held inside the archive. This process took approximately 5 days to execute and yielded a list of over 19,000 file extensions. The vast number of different file extensions suggested that a complete scan of the archive would be impractical within the time frame that we had, so we were able to conclude some sort of sampling would be needed to derive meaningful results from the project.

While the sampling was necessary, it also exposed the results to the risk of oversampling. Because of the heterogeneity of the data in the archive, files in some datasets were not given the appropriate extension, and in other datasets, file extensions were given not according to their file type, but in accordance with other, older naming conventions. Often the file extension was replaced with a timestamp, or a scientific instrument name – this variance in file extensions led to oversampling of some datasets because of the arbitrary extensions used.

The extension-mapping script also created a sample list of file paths as it was parsing the filenames inside the archive. The script sampled files in each data collection with a given extension so that the sample contained a number of files proportional to the log of the number of files with that data collection and extension combination.

The HEFTI (Helping Environmental Formats Through Identification) software was developed to read the list of file paths from the sample list and perform a lower-level examination of the files in order to categorise them in terms of format. We were able to successfully accelerate the scanning process using two techniques – parallel processing and I/O using a cluster, and extensive sampling of the archive. The BADC has successfully deployed the JASMIN super data cluster [11]which consists of a large number of high-speed servers connected to massively parallel Panasas storage. We were able to take advantage of this parallel storage in order to allow execution to speed up by a factor of eight.

Pattern matching and introduction of specialist knowledge

Pattern matching was used extensively within the project, to recognise filenames and to recognise patterns for recognised file formats. Pattern matching in filenames was important, because often files would be named in accordance with the accepted naming convention at the time. This meant that the vast majority of files in each dataset, while in an unknown format, were of a uniform naming structure, and this allowed us to infer file formats. After grouping unknown files like this, contact with the CEDA data scientist responsible for that data collection allowed extra were information about the unknown formats to be feed back into the format analysis configuration. This cycle is repeated to move toward a comprehensive format list with all the required file format fingerprints.

Context was also important. Due to the nature of the archive's structure, it was possible to infer which dataset was being tested, based on its file path. This made it very easy to write rules within the HEFTI toolkit's configuration files that only tested certain patterns within certain datasets – which helped overcome a problem that we encountered during the course of the project. The problem was that numerous naming conventions for files were used over time, and so many different datasets were named in many different ways. To have a piece of software that could identify a file based on its context was important.

[11] JASMIN http://arxiv.org/abs/1204.3553

Specialist knowledge also played an important role in the project. Because of the multiple bespoke and specialist formats that accumulated inside the archive over time, there was a considerable number of unknown, difficult-to-decipher file formats; for example, the PP format. PP is a bespoke binary format that was created by the MET Office[12], which is very difficult to identify without prior knowledge. PP does not have a simple file signature, just a set of headers based upon a C structure which are unpacked at runtime – the way to test if a file is PP format is to try and unpack these headers, and if that fails, the file is not in PP format.

Furthermore, specialist knowledge of the data was useful when documenting the change in formats used within a dataset. For example, some of the data from the European Centre for Medium-range Weather Forecasts[13] changed formats after a certain date. We were able to be forewarned of this by a CEDA data scientists, and using the toolkit that was developed during the course of the project, we were able to pin down exactly when this format change occurred.

Improvements and future developments

Several future optimisations could potentially be implemented, including use of more efficient parallelisation technologies than at present. One of the solutions suggested is full integration with the LOTUS cluster. This would enable more efficient resource use and job allocation, as well as shared memory, so that tasks would be executed more efficiently with less resource duplication and lower overheads. Not only that, but this would allow our file identification toolkit to fully exploit the parallel I/O capability of the Panasas storage that we have access to. Another, different approach would be using a MapReduce framework such as Apache Hadoop[14]. This sort of parallel processing infrastructure could speed up execution times, and therefore allow us to make leaps and bounds in terms of the results that we produce.

Such performance increases would allow us to improve the reliability and quality of the results that we produce about the archive, simply because we'd have greater flexibility in the complexity of computation that we could implement – we could, for example, implement a more effective, more rigorous ASCII data-file recognition algorithm, which would allow us to identify more ASCII data file formats, more reliably. Also in future, it would be a possibility that such a toolkit could be implemented into the ingest stage of the archival process, negating the necessity to identify the entire archive on a regular basis.

IDENTIFYING POSSIBLE ACTIONS FOR NON-STANDARD FORMATS

Identifying formats that are lacking in support allows the data centre to take action, adding documentation or archiving software for accessing the data. The diagnostic toolkit described above will highlight many instances of unknown or non-standard formats. Often these are still readable formats that require a little investigation to create a better format signature, however there are distinct classes of action that would be appropriate based in the support infrastructure for the format. By talking to the data scientists and searching for format related resources it is possible to build up a picture of the format by asking three key questions.

- Where is the organisation responsible for maintaining the format?
- What software do I need to read or write data in this format?
- Is there a open format specification?

How easy it is to gain this information depends on the size and habitual rigour of the community that supports it. The answers to these questions allow us to categorise formats based on community type.

---

[12] The Met office http://www.metoffice.gov.uk/

[13] European Centre for Medium-range Weather Forecasts http://www.ecmwf.int/

[14] Apache Hadoop http://hadoop.apache.org/

Globally supported formats: Microsoft Word has an enormous community made-up of the owning company producing extensive documentation and online supporters cataloguing any other issues and aspects of it, and is so popular that almost every computer has software that is compatible with it. As such .doc is likely it be a long-lived file format that doesn't need an archive holding on to its software or its structural information. As such there should be no action to preserve this format by the data centre, instead we should rely entirely on the external community, perhaps checking every ten years or so to see if it is still well supported.

Community supported formats: NetCDF is an example of a format that has a large science community of devoted users and supporters. It is much less commonly used outside science and links from within the archive to outside sites and sources of information are required for its usage. Checking on these supporting sites to confirm that they are still part of an active community yearly should help to protect the life of the format. If the community seems to be dying out then the site should be immediately archived.

Bespoke formats: PC Cora is an instrument specific format used to record data from weather balloons from a particular manufacturer. Documentation is not readily available on the Internet so the data centre needs to make efforts to keep local copies of the format specification and any software used for reading the data.

These examples are from a spectrum of formats in terms of support effort. Characterising data in this way helps make sensible decisions on preservation actions by qualifying the impact of the action.

DATA VALUE

For any one dataset it must be determined whether or not it is worthwhile to put time and effort (and therefore money) into preserving it. There are a number of factors that determine whether data is actually valuable, for example, current data use, commercial value, and science publications reliant on the data. This has been investigated in initiatives such as the Keeping Research Data Safe Benefits Framework . [15]

Within CEDA we have a number of indicators to establish data value – current use as indicated by downloads metrics (see examples in table 1), known data references in scientific publications, and assessment of future data use. Use patterns change with time so assessment of value is not a one off activity and needs to be revisited periodically. Of these indicators the best is often simply to ask the scientists responsible whether or not the data was scientifically value as they had first-hand knowledge of its previous and potential usage as a scientific contribution. A lot of data isn't valuable until enough has been collected over a long period of time (months or years) to show identifiable trends and patterns in use or citation.

[15] KRDS benefits framework http://www.beagrie.com/krds.php

Dataset	*Number of Users*	*Number of files*	*Size (MB)*	*Activity days*
midas	1602	94934	5533132	11034
cru	1530	29353	1530789	3574
hadisst	528	16414	191064	1505
surface	518	428566	32991	1849
era40	335	19549797	8600879	1923
faam	308	108960	1483671	5138
ecmwfop	280	6684158	7255019	2632
radios	247	3805120	199514	1398
hadcm3	229	2746846	6067732	1197
cet	209	3163	160	628
nimrod	205	8401894	4976052	3182
assim	189	198205	4515626	7730
meteosat	167	396145	426417	1073
chilbolton	139	1489671	863614	1002
ecmwf-era-interim	131	613179	25068944	777
hradios	126	817235	82845	478

Table 1. Data downloads

RESULTS OF FORMAT AUDIT

Each collection of data files was summarised by a simple text file containing a list of formats found and the number of files found with that format. The example below shows the results from the MST radar data. This files comprises data in standardise community supported formats (NASA-Ames, NetCDF), data files in dataset specific formats (e.g. nerc-mstrf_gps_iwv), quick-look plots (GIF, JPEG), documentation (HTML) and software (e.g. FORTRAN).

```
ASCII text, 1022
Alpha COFF format core dump (Digital UNIX), from 'edt.x', 1
COFF format alpha executable dynamically linked not stripped, 2
Empty File, 1
FORTRAN program, 31
GIF Image, 53
GZIP Archive, 174
HTML document, 63
Ignored BADC Metadata, 53
JPEG Image, 180
NASA_AMES, 56
NETCDF_3_32, 4482
PNG Image, 57
Unknown binary, 2
[nerc-mstrf_gps_iwv], 1
[nerc-mstrf_mst_ctm], 29
[nerc-mstrf_mst_iq_ks], 83
[nerc-mstrf_mst_lis_files], 46
[nerc-mstrf_mst_spectral_ks], 1413
[nerc-mstrf_mst_v0_balloon], 2
[nerc-mstrf_mst_v0_power_binary], 48
[nerc-mstrf_mst_v0_radial_averaged_binary], 4
[nerc-mstrf_mst_v0_wind_averaged_binary], 2
[nerc-mstrf_ral-fmcw-cloud-radar], 33
perl script, 2
```

Figure 1. Example output from format audit scan of MST radar data. Each line includes a short description of the format and the number of files in the sample list that were found of that format.

These file numbers above indicate the number of sample file of that type. By assuming that the files within a dataset with the same extension are of the same format it is possible to scale the sample format findings to produce a summary for file

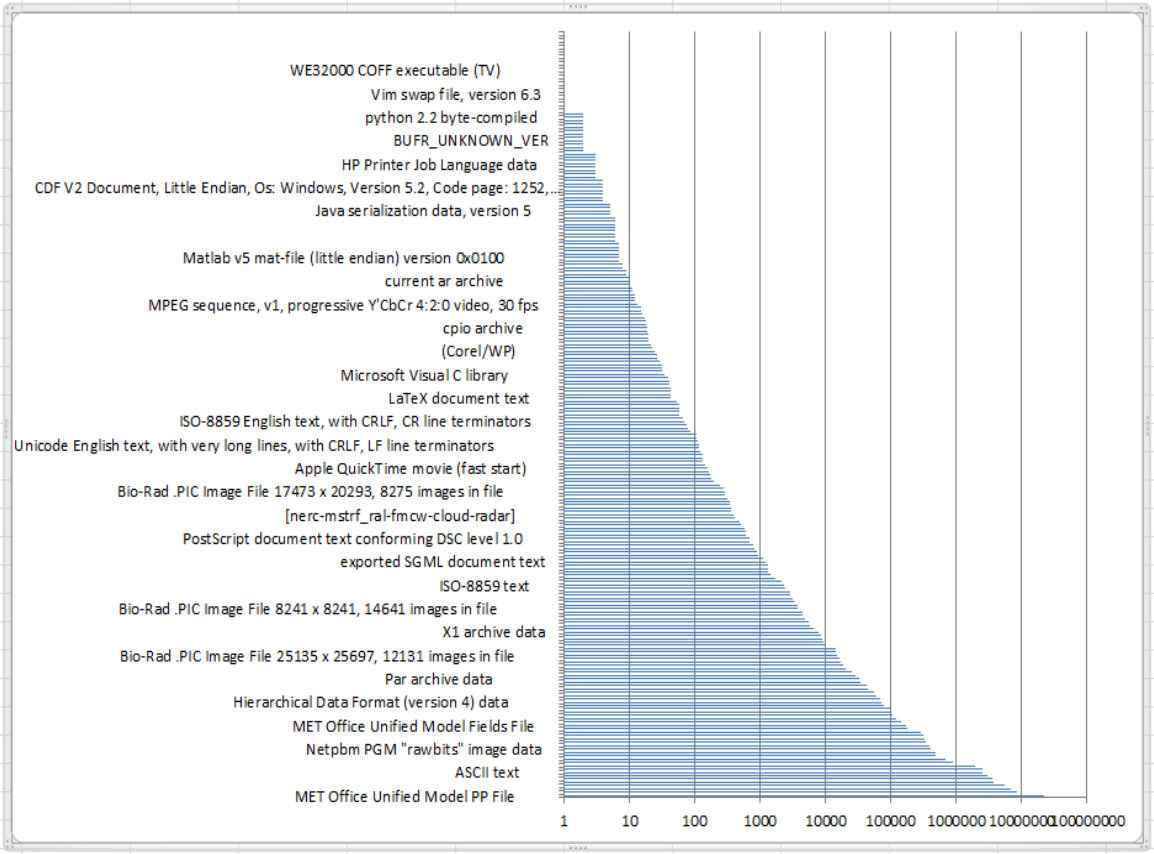

Figure 2. Number of files per format type

It is important to note than a substantial portion of the archive remains unidentified. Formats marked unknown and gzipped need to be explored further. This work will involve manual inspection of file and direct consultation with data scientists and data producers. It is also important to note that file number alone is not a perfect indicator of format significance to the archive. It is entirely possible than high value data exists in a small number of file which have yet to be identified.

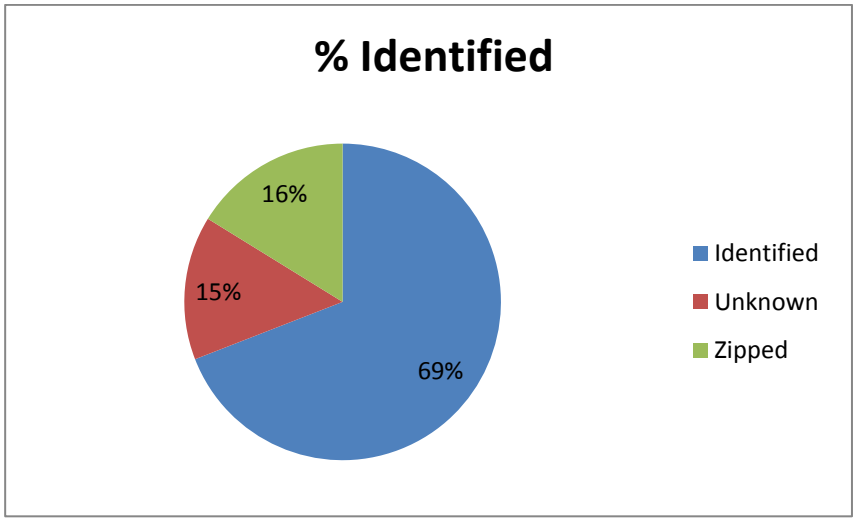

Figure 3. Percentage Archive Identified

There is a long tail of rarer file formats. This may indicate stray files which enable the format audit to be used as a tidying up mechanism, for example core dump files in the archive are like to be there by mistake rather than a deliberate archiving decision. The most prevalent formats identified by the audit were anticipated group. This is reassuring as it confirms we had a realistically predicted the majority of archive content. The area of greatest concern lies with the "Unknown binary" and "ASCII" files. By virtue of the fact that they have not been identified there is a strong possibility they are in high risk preservation state, which must be resolved by manual and consultative means.

Format	*number*	*%*	*format type*
MET Office Unified Model PP File	22395167	33.7%	data
Unknown binary	8283704	12.5%	o
GZIP Archive	6811271	10.2%	z
NETCDF_3_32	5588505	8.4%	data
GRIB_UNKNOWN_VER	3721727	5.6%	data
JPEG Image	3579082	5.4%	i
ZIP Archive	2993459	4.5%	z
ASCII text	2468642	3.7%	doc
PNG Image	2461835	3.7%	i
Symbolic Link	1931604	2.9%	o
GIF Image	899350	1.4%	i
XXXXXXXX	676724	1.0%	o
Hierarchical Data Format (version 5) data	490639	0.7%	data
XML document text	475515	0.7%	data
Netpbm PGM "rawbits" image data	403652	0.6%	i
ENVISAT_PDS	385252	0.6%	data
[nerc-mstrf_mst_spectral_ks]	337452	0.5%	data
NASA_AMES	328175	0.5%	data
raw G3 data	314778	0.5%	data

Figure 3. Top CEDA format types

STRATEGY DEVELOPMENT AND FORMAT SUPPORT

CEDA FORMAT POLICY

CEDA Data centres cannot support a multitude of file formats, because we need to be able to have systems that can parse both the internal metadata and data so as to make them available to services (e.g. catalogues and visualisation/subletting services). We are long past the point where we can operate as a big file store and just rely on a filename convention and a file hierarchy. So there is a very real and major cost with every new format. For this reason it is important clearly identify our supported formats, list the reasons why they have been selected, and give guidance on metadata standards. Our default policy should be that new formats are not acceptable unless:

- it is not possible for the data creators to encode their data satisfactorily in one of the formats we already support, or
- it is not possible for the data users to easily manipulate the data in one of the formats we already support, or
- the data is being provided in a specific format because it conforms to the requirements of a major international programme.

CEDA should take a leadership responsibility to make it easy to create and use existing supported formats (e.g. by development of software like nappy), and to encourage, cajole, and demand that existing supported formats are used wherever possible. In the case of NERC programmes we will insist supported standards are adopted, but even in programmes outside our direct control we should advise and seek to influence the choice of format. CEDA should actively review the formats its supports for ingest, storage and access. The current policy is to ingest, store and access data in the same format. A supported format should satisfy all the requirements below:

1. Producers can make the files on most platforms with available tools.
2. Format should be platform neutral.
3. Required file level metadata can be encoded.
4. Format compliance can be checked.
5. Format should be stable
6. Format should be open (if not one has any software to read the data, user can write there own).
7. Format should be migratable (as improved formats appear the significant properties of the files can be transferred to new formats).
8. Format is in use by the data centre user community

FORMAT ACTION AND RISK ANALYSIS

Action is needed to move unsupported formats onto a firmer footing. This action may be improving a format description, migrating to a more standardised format or archiving software to preserve accessibility. Clearly the actions are carried out to reduce the risk of the format being a barrier to access, however in a large heterogeneous archive which actions should you perform with limited resource? We propose the to prioritise two factors need considering: the cumulative value of the data effective by the action and the reduction in access risk mitigated by the action.

Action benefit = value of data affected * impact of risk reduction action

Given limited resource to spend on the actions it is optimal to carry out actions that are affordable and have the greatest benefit. Risk to format was not something we numerically calculated but rather something we characterised according to

Impact of

- Availability of format specification - complexity of format, skill base of Community, who looks after the specification.
- Availability of Software - Technical fragility, Availability of emulation/virtualisation solutions, is the software actively developing and having bugs fixed.
- Supporting Organisation – Quality, Stability, Levels of co-operation
- Community of Use – Breadth + Stability + Relevance to designated community

RISK EXEMPLARS

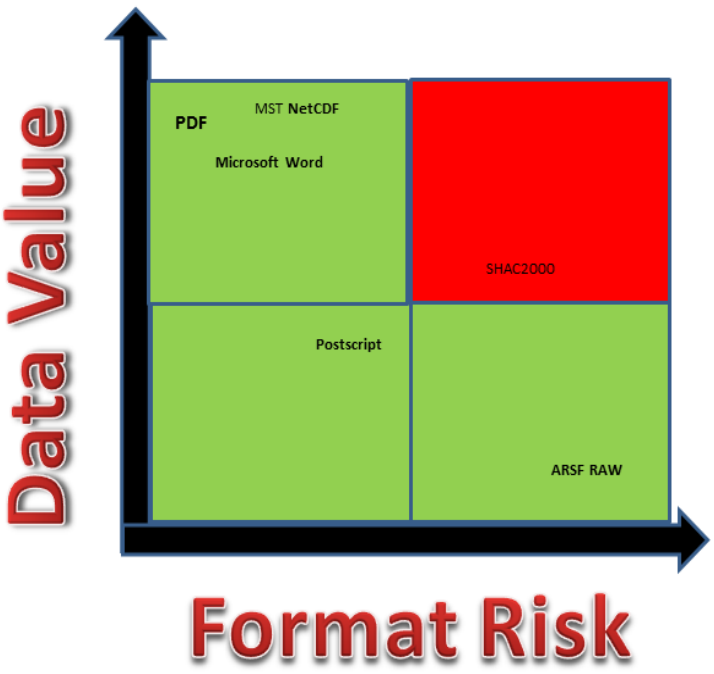

Figure 4. Risk Value Analysis

High Value Stable Data

MST (Mesosphere-Troposphere-Stratosphere) Radar Facility has produced wind profiles to a height of 20km from 1989 to the present day. The data are continually used by researchers and the Met Office and represent a long-term archive at good time and height resolutions. The bulk of the data are in a community supported standard format - NetCDF. The format specification is available; the community are skilled enough to use this format despite it being complex. Additionally, there is extensive community-based software usable on a wide range of operating systems. The supporting organisation (UNIDATA) is well funded and likely to have a long-term future. A broad, well-engaged community uses NetCDF and hence the format is stable. This data is thus of high value and low format risk. The preservation strategy in this is to monitor community use of the NETCDF format.

Low Value Unstable Data

The NERC Airborne Research and Survey Facility (ARSF) flies a plane to take images of the earth's surface for numerous NERC funded research project. The valuable products are the processed images. However raw data are also archived and this illustrates an example of where the data format is high risk. The raw data comprises of information such as emails, raw uncalibrated data, notes etc. However, the data has been deemed low value. The data is still present in the archive for historic reasons purely as a backup for the ARSF team should the data ever need to be reprocessed in the future. However, there is no format specification and no available software to support the format of this data. But because the data value is low, CEDA consider it unnecessary and not justified in terms of return investment of staff time to take any substantial preservation action. While the Plymouth Marine Laboratory possesses knowledge of the format and processing algorithm they have not been mandated with long term support of this lower level data. Therefore no preservation objective is set for this data.

High Value Unstable

SHAC2000 is a campaign that ran sporadically during the year 2000 and covered approximately 7 sites at different times. The campaign consisted of two aircraft: one had the hyperspectral imager instrument and the other had synthetic aperture radar (for generating digital elevation models). It is a dataset which presents a high risk as it contains formats about which little is known. The data scientist has indicated this is of high value because at the time it represented a state of the art combination of both airborne radar data and hyperspectral imagery; it also involved a lot of ground truth by various scientists. This involves using handheld spectrometers at ground level to measure the same variables as the airborne instruments to produce the "ground truth". It is used to quantify the signals lost by atmospheric scattering, absorption etc., and to see how accurate the instruments could be. Although it does not meet today's standards, this data provides a useful comparison reference, for example, to contemporary ARSF data. Many of the SHAC sites were also used as ARSF projects, hence the usefulness of SHAC2000 as a comparison. Therefore preservation action is required, and there is now a preservation strategy for acquisition of format specification and related software.

Documentation Formats

Understanding the risks associated with documentation formats has been a critical aspect of this project. As documents tend to have a much broader user base, the dominant factor in risk appraisal is focused upon the community of use. Some document formats have a high degree of community-based stability, to the point where this becomes the dominant concern. No further preservation action is required beyond preservation watch/community consultation activities. For example, Microsoft Word and PDF documents are distributed through the archive and contain a lot of essential high value information related to the data. The level of stability they currently enjoy leaves a position

where no immediate action is necessary due to the high level of community support. However if a low cost scalable technical preservation solution for conversion of Microsoft word to PDF were to present itself, it would be considered.

On the other hand, formats such as postscript present a different risk profile. This format is not as stable as Microsoft Word or PDF. The supporting organisation Adobe provides a specification and software that would be available for capture. This preservation action acts as a good failsafe but ideally, as scalable conversion technologies mature, the migration of postscript to PDF would occur. There is a greater incentive for postscript to PDF conversion than for Word to PDF conversion, therefore the former would occur at an earlier point than the latter.

CONCLUSION

A rigid framework needs to be implemented with regards to data acceptance saves effort on documenting unknown formats or migrating to new formats. The project itself effectively demonstrated the heterogeneity of such an archive and emphasised the importance that data is only accepted and named in standardised, recognised, risk-assessed formats. This permits greater support, greater longevity, and more widespread usage for the data, and this is absolutely vital for such a resource. Further analysis is also required to determine where preservation action should be employed. Consultation with data scientists and user community is needed to determine where preservation risk should be accepted monitored or mitigated.

Acknowledgements

We would like to thank to the following CEDA scientists for their help and support during this project. Stephen Pascoe, Alan Iwi, David Hooper, Steve Donegan, Anabelle Guillory, Kevin Marsh and Alison Waterfall.