APSR
AUSTRALIAN PARTNERSHIP FOR
SUSTAINABLE REPOSITORIES

Survey of Data Collections: a research project undertaken for the Australian Partnership for Sustainable Repositories

Kevin Bradley, National Library of Australia and Margaret Henty, Australian National University

December, 2005

Aims

This report draws together the findings of the Survey of Data Collections conducted in the APSR partner Universities in 2005. The information derived from the process is intended to feed into the next stage of APSR work, providing specific data for the design of software tools, development of repositories, assessment of risk and development of risk management approaches, implementation of preservation metadata and the development of supported formats. The survey of the collections was also developed to help characterise the types of data repositories, to determine if different sorts of data sets needed different facilities, technology or management. The process raised a number of issues which were covered in the survey questions and analysis. This was expected and where appropriate will inform the APSR process and identification of issues. The survey process has also helped to identify datasets with whom APSR will be able to undertake further sustainability investigations with other data sets.

The survey was conducted with managers and owners of collections of data generated within or for the university. The interview process was selective, and intended to be representative, depending on the ability of the particular University to identify datasets. The interviewees were chosen for both the types of data they manage, and knowledge of the area, as well as for pragmatic reasons, such as availability and willingness to be interviewed.

Method

A formal set of survey questions were developed which targeted specific areas of knowledge required by the project (see appendix 2). Answers to the questions were coded and entered into Apollo, the online polling system utilised by the ANU. A statistical report based of the numerical and graph-able data is included in Appendix 4.

Audio recordings of the interviews were made, and participants were asked to be as discursive as they wished on topics relating to the management of data sets. These recordings were transcribed. From these discussions a number of common or pertinent issues have been distilled. Interviewees were asked for permission to allow their comments to be used in this r eport. The audio and transcripts are to be loaded onto the ANU’s institutional repository, the D-Space installation, Demetrius. The ANU Ethics committee determined that given the technical nature of the information, EC permission was not required for the project.

Report Structure

The main body of the report considers the issues raised by the survey, and includes where appropriate, recommendations to address the issues. Some of those issues are already being addressed by APSR. The remainder of the report is included in the appendices. The report is structured as follows:

Discussion of Issues

The common issues raised in the survey are considered under the following headings:
Data Duplication, Storage and Back-up
File Formats
Institutional and specialist data repositories
Responsibilities
Designated repositories
User Communities and Community Expertise
Future use and designated life of materials
Funding
Sustainable Paths for Data-intensive Research Communities
Metadata
Rights and Copyright
Digitisation

Appendices

Appendix 1 Annotated list of interviews.
    1.1 ANU
    1.2 USyd
    1.3 UQ
Appendix 2: Survey Questionnaire
Appendix 3: File Formats
Appendix 4: Statistical Report

Contributors

The survey questionnaire was developed by Kevin Bradley and Margaret Henty, both of whom conducted most interviews and compiled the results for the report. Belinda Weaver, University of Queensland, and Su Hanfling, University of Sydney, coordinated the interviews for their universities, and supported the interview process or conducted supplementary interviews according to their particular circumstances. The data on file formats and types derived from the interview process was supplemented by discussion and analysis with the support of Scott Yeadon and Peter Raftos at ANU.

Issues

Data Duplication, Storage and Back-up

Data back up is a primary and fundamental issue. While back up does not constitute sustainability, all long term preservation actions are dependant on the digital byte stream being reliably stored. The respondents had a number of approaches to data back up:

1. Relying on institutional IT services
2. Undertake own back up of data duplication
3. Possessing legacy technology of backed up material
4. Limited or nonexistent back up procedure.

Relying on institutional IT services

Of those who relied on institutional IT services, these had two distinct approaches: Those who deposited with an archival service, such as astronomical data stored with the Australian Partnership for Advanced Computing (APAC), or social science data sets stored with the Australian Social Sciences Data Archives (ASSDA), or indeed data stored on the growing Library based institutional repositories: Those who store the data on local servers and depend on local IT server back up to maintain their data.

The former respondents tended to be well informed about the roles and responsibilities of the data storage facility and their own responsibilities. The relationship between the data managers and the data owners meant that they were well informed about the strategies being taken and were aware of the short and long term limitations.

The latter maintained, on average, a more remote relationship with technology, and were not readily able to describe the technical extent of support they expected or received. There were exceptions with pockets of expertise in individual non technical communities who would often act as the technical broker; examples of this might be the review of the Coombes server at ANU undertaken to determine content and management of the data server. However, the local IT facilities do not count preservation, sustainability and data curation as their responsibility, meaning that there is no identified area with responsibility for long term preservation.

Recommendations:

- Provide sustainability oriented technical support and training to communities relying on local IT servers.
- Encourage the defining of long term preservation responsibilities and issues with regard to data management and IT back-up, including education regarding the issues.
- Provide recommendations regarding appropriate storage.

Undertake own back up or data duplication

This category too, can be divided into two categories. The first category includes those that develop and manage their own digital mass storage system, such as the seismic observational data kept at the Research School of Earth Sciences at the ANU, which has a large storage facility and dedicated staff to maintain it. The second category consists of those smaller collections that duplicate their data locally, either as data back up, or through making copies of individual items, for example where linguists copy audio data to CD-R.

Larger repositories of specialist data exist which have well established data management procedures and specialist expertise to maintain data. These included such repositories as the Sensitive High Resolution Ion Micro Probe (SHRIMP) at the ANU, or the Reef Grid Project housed at the UQ. Typically, the technology used tends to be large scale magnetic devices on tape and disc, and the system is managed by an IT specialist. These repositories are generally well equipped to deal with current access problems. Sustainable funding was highlighted as an issue by managers of such repositories. The tight relationship with the user community and the development of formats means that the format obsolescence aspect of preservation is identified as less of an issue. They tend not to have engaged with issues of long term sustainability and preservation except where it affects immediate access.

The data collections where data owners back up their data locally tend to be small scale collections, either managed by an individual researcher, or a small discipline based group. The task of data management is peripheral to the main task of working on the content and is generally undertaken by the researcher or an assistant. The data is typically stored on formats over which the user community has little control such as image, text or sound files. Technology deployed tends to be a semi-professional use of optical discs such as CD or DVD recordable. The motivation to duplicate data may be simply because local desk top drives become rapidly full, and separate storage seems adequate. There is little evidence of awareness of the limited life of optical carriers, or of any risk management through redundant copies or error testing.

One interesting use of optical media in a professional archive was where the manager of the repository and server that is fed data from the SHRIMP ion probe backs the data up regularly to DVD recordable. The server is managed and backed up by standard IT magnetic tape procedures, but the optical disc is *"essentially a flame copy … once we burn those CDs and DVDs (there's usually two copies) … I'll take one to my home and put it in there. … If the place is going to burn down then we still have the data available."* (Ireland Interview 2'50). This is only possible due to the small file size. Clearly the life of the disc is irrelevant because it is replaced with a new disc every few weeks.

Recommendations:

- Integrate existing large repositories into the digital sustainability and preservation debate, (perhaps through the creation of a University wide groups such as "Digital Preservation for Large Data Sets" or similar).
- Seek for the large data sets to provide input into the debate as many have extensive experience.
- Provide technical information on strategies and approaches to data management (such as recordable CD and DVD guidelines).
- Encourage use of institutional repositories for appropriate data sets.

Possessing legacy technology of backed up material

A number of the researchers interviewed as part of the review process possessed drawers and cabinets full of older data carriers. These ranged from punchcards, old floppy discs (8", 5 ¼", and 3 ½"), to 10" 9 track open reel tapes, DDS DAT, Exabyte and other forms of data tape cassette, as well as large quantities of optical discs. The content ranged from astronomical data sets, and records of crocodilian blood pressure, to data bases of theatre reviews and other text records. Most often the researcher has maintained the data because it is considered important to key publications or significant research issues.

Researchers now face issues retrieving the data from the legacy carriers. As a Queensland astrophysicist commented *"like everyone else we're suffering from the media problem and one of the reasons I'm interested in this project is, as an example, I recently had to go back to some data from only 2001 and a couple of the tapes (*Exabyte*) I had it on proved very difficult to read here because we don't support that kind of tape drive anymore in the department. … (When accessing other sorts of data tapes) there's a few problems that crop up repeatedly, and occasionally we get a tape we fail with, but normally we find that resetting the block size on different computer systems enables them to talk to each other. It takes a while to do that.* (Drinkwater Interview 12'10).

Researchers with less technical expertise or support fail even more frequently. Many are unsure which of their data carriers are most at risk. Most are desperate to find a facility to manage duplicate copies of their data, and many would like access to facilities that can undertake the extraction of data from legacy carriers.

Some, aware of the sustainability and preservation issues have made attempts to move the data to more sustainable carriers but have been frustrated by having to deal with the changing technical landscape. When asked what would be required of a proposed repository, the curator of the Australian Drama Bibliography Project stated "*I require (the data) to be stored somewhere safe and permanent and I think this is about the third generation of my attempts to do that*" (Kelly Interview 17'46)

Recommendations:

- Develop a web based tool to help prioritise and manage legacy digital formats according to identified risks.
- Support the development of university wide data recovery (digital forensic) facilities.
- Provide and publicise information and guidance regarding risks to data carriers and the benefits of institutional repositories.

Limited or nonexistent back up procedure.

Many researchers are aware that data should be better managed, but are too busy with research issues to undertake the tasks or learn the technology. As one anthropological researcher commented, when asked what would encourage him use an institutional repository "*I need technical help with these things. … I have enough problems and I've got a shortage of time and I don't want to waste my time learning all this technical stuff which becomes outmoded as soon as you learn it. … I've got all my time cut out for me in translating and doing the analysis and that's what I want to do. … So what would help me is someone with that technical expertise who could give me that advice.*" (Gregory Interview 30'27)

Some researchers are still unaware of the fragility of digital records and need advice and information.

Recommendations:

- Develop guidance materials
- Identify a "help desk" person for digital preservation and sustainability issues.
- Publicise

File Formats

One of the motivations of the survey of data collections was to elicit from researches and data managers the types of files held and used in repositories. Such information is useful to repository developers to ensure the tools for ingest and management of data is designed to support the necessary file formats. The table in appendix 3 lists known file formats and will be useful in conjunction with the file format recommendations for long term use.

The files included in the table of file formats covers a wide range of reasonably well known and not unexpected file formats. This clearly highlights the need for ingest standardisation and format migration procedures to minimise the long term requirement to maintain access to data. The files on the table also draw attention to a number of other issues.

Probably the first issue is the level of ignorance by many data owners about the formats that are stored in their repositories. Some of the researchers were not able to name the format they stored their data in, let alone whether it was an access or sustainable format. Once a certain level of functionality was reached many researchers had no further interest in the data formats.

Recommendations:

- Provide support mechanisms and relevant materials to educate data owners and managers in the sustainability issues associated with format choice.
- Encourage the use of sustainable formats from which distribution copies can be derived.

Specialist file formats: Some of the datasets use specialist file formats which are only appropriate to their particular area of investigation. The FITS (Flexible Image Transport System) used to manage astronomical data is one example. There is a need for repositories and associated tools which support these formats within the particular community. However, not all repositories need support these specialist formats. This is discussed below under *Institutional and specialist data repositories*.

Text Files: Files containing text based information may represent many different types of information. A text file might be: spreadsheet (delimited text), table or database, it could contain encoded scientific data, or prose. It might be in word, html, rtf, it may be an ASCII or a Unicode file, it may be encoded in HTML, SGML, or XML. The SHRIMP ion probe at ANU, for example, stores all its data in ASCII files, which, in spite of being text are meaningless without the requisite contextual information.

To understand a text file we need to know explicitly what the data is, what objects are contained within it, what created it and what is needed to read it. We need to know its encoding (which may not always be available), if it is marked up we need to know the mark up language, in XML we need to know the schema, or the DTD if available, any validation rules, any process tools and its end use. If the data is data base information we need headings and explanatory data, and may even need to distinguish SQL from DDL statements. In short, we need detailed contextual information specific to the data type and expected use.

Recommendations:

- Develop recommendations on the documentation of text objects taking into account existing metadata schemas.
- Develop ingest tools conversion tools (eg word to XML).

Open Source and Proprietary Formats: A premise of the survey of data collections was that open source formats were more sustainable in the long term. This was assumed on the grounds that proprietary format owners tended to load their format with non standard features necessitating constant revisions, because the source code of the proprietary software associated with the format is generally not available, and because commercial suppliers tend not to release API documentation, or do so in such a way the accessibility is not guaranteed.

In principle this is still the case, however, there are so many exceptions that the question in the survey regarding open source versus proprietary will not reveal any meaningful statistic. Adobe's PDF, for example, is used very widely in data repositories, the APIs are published and considerable work is being done by archival institutions (notably the UK's National Archives), to ensure the availability of long term technical information that will support sustainable access. In many respects, the requirements of an open source format are met in all but the area of ownership.

Recommendations:

- Use this understanding to inform the recommended file formats process.

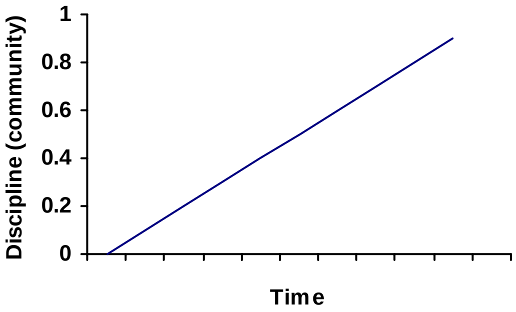
Institutional and specialist data repositories

That there appears to be a distinction between institutional, generalist repositories and specialist data repositories grew out of the interview process. That there are different approaches to data management resulting in different requirements, design and workflow approaches is clear. The actual distinction however, is not so easily determined, as the technologies, issues and processes are all, more or less, shared. Nonetheless, knowing the degree to which a data manager has to be a subject specialist to manage data, or a library and information management specialist is a critical to APSR. It goes to the initial stages of data repository funding and responsibility.

Responsibilities

Specialist technical requirements Critically, a number of scientific data sets claimed that maintaining responsibility for their data was essential to the sustainability process. Many of the scientific data managers could not envisage a role for a general repository for the specialist data. The ASSDA at ANU employs social science graduates to curate their collection, School of Geography, Planning and Architecture employ a Senior Scientific Officer with experience in spatial data, and the data scientist who is doing so much to facilitate the astronomical data at APAC is himself an astronomy doctoral graduate.

The data specialists argue that decisions about the management of the data can alter the meaning of the content, and that only a subject specialist is adequately equipped to make those decision without compromising the content of the repository. This may be the case, but alternately, it may be that the explicit metadata necessary to make a dataset transportable has not yet been developed or made, and that local expertise compensates for the lack of documentation. It can be argued that the need for metadata increases with the distance from either the community, or the creation date. (see graph below).



The need for metadata; the greater the separation in time or discipline, the greater the need. (DCC UK)

As an indicator for APSR deliberations, it might be that if the data cannot be managed away from the designated community, then it will consequently not be possible to sustain the data when there is a large temporal distance between its use and its creation.

Institutional or General Repositories: A design tenet of specialist repositories is that system design should cleave on discipline issues rather than a technical basis, which requires domain expertise and knowledge. However, an institutional repository is predicated on the principle that consolidating technical services will deliver economically sustainable efficiencies, and a concentration of expertise around the issues that inhere in digital preservation and sustainability.

Many of the researchers interviewed identified the need for a repository that would deal with the issues of sustainability and access. Some were concerned to identify a responsible manager of significant data after their academic careers have come to an end. After describing the extended team and process that went into the development of the Australian Drama Bibliography Project, Professor Kelly commented dramatically, *"And I'm the last surviving custodian"* (Kelly Interview).

Professor Grigg stated that identifying a suitable repository was an issue for him and his data sets of zoological data, because impending retirement and consequent loss of university facilities would mean that he could no longer service the requests for his data. He stated that *"this was the problem I was wrestling with .. in the belief that this information might be useful to somebody in the future, rather than locking it away or throwing it out it would be better to put it into some format and location so that other people could be referred to it and/or could have access to it in case it was useful to them for some project that they haven't yet thought of. And maybe they haven't even been born yet, the people who might use it."*

Designated repositories

The dataset owners looking for appropriate specialist, institutional or generalist repositories came from all academic disciplines; technical, scientific and humanities based. The distinction cannot be based on purely content or discipline. As discussed above (see specialist file formats) the use and development of a particular file format can indicate the need for a specialist facility, while the use of general formats might argue for a centralised facility to develop tools and technologies to deal with more general issues.

The following is a table that might be used to distinguish the need for a specialist facility or a general repository. It is intended as a basis for discussion.

The indicators proposed are:
File Format: A specialised filed format specific to, or even developed by, a very narrow community would be an indicator of the need of a specialist repository. So, for example, Standard for the Exchange of Earthquake Data (SEED), or the seismic data format ZDF, developed by an ANU academic would be an indicator of the need for a specialist repository; the use of TIFF for images would be an indication of a general or institutional repository.

Metadata Schema: Similar to file format, in that a specialist schema would presuppose a specialist repository, and the use of general standard schemes would suggest an institutional repository. In addition, the paucity of metadata might be an indication of the need of specialist interpretation, and hence suggest a specialised repository.

Impenetrability of data: Machine readable data, or data encoded in such a way that its meaning or purpose is not readily apparent might suggest a specialist repository. Readily accessible data, such as images, would suggest an institutional repository.

Readily Identifiable User Community: A readily identifiable user community of specialists who are easily associated with and relatively able to make decisions about the data, might lead to a specialist repository. A more general community who do not have the structure to provide advice on data sets would suggest an institutional repository. The former might be ASSDA, the latter might be the broad group of historians interested in digitised Australian 19th century texts.

Data Size: Though data size is largely irrelevant, nonetheless a large data set requires more resources to maintain, and may well as a consequence require a specialist repository rather then a centrally funded general repository.

Specialised or general Access Conditions: The degree to which access conditions are specific to the type of material may indicate the need for a specialist repository. The PARIDESIC material, for example requires links with both the community members who were recorded, and the academic who made the recording. Access conditions are not readily encodable, most often requiring interpretation or interaction between owners. The Australasian Pollen and Spore Atlas requires only that signed up members get access, which is relatively simply encoded and would suggest a institutional repository.

Range of potential users: If the users of the data extend beyond the initial community, then a institutional repository may be preferred. The owners of data held in the Reef Grid project eventually expect to provide wide ranging public access. The data created as a result of High Energy Physics experiments is very unlikely to have any value outside of the HEP community and so would gain little value from a generalised institutional repository.

Data Intensive: Data intensive research produces data that is used as a part of the research process, most probably stored in large growing databases containing a range of complex and specific digital items. Standard data usage produces data as a result of research, perhaps preprints or related materials.

	Specialised Repository	**Institutional Repository**
File Format, specific or general	**specific**	**General**
Metadata Schema	**local database**	**standard**
Impenetrability of data	**obscure**	**clear**
Readily Identifiable User Community	**narrow community**	**Broad community**
Data Size	**Large**	**small**
Specialised or general Access Conditions	**special access**	**general access**
Range of potential users	**community specific**	**wide range**
Data Intensive	**research data**	**data as outcome**

Recommendations:

- Review the categories
- Develop appropriate weighting and scoring for the various categories
- Develop it as a guidance tool for decision making.

(this process could be part of the Sustainable Paths for Data-intensive Research Communities program).

User Communities and Community Expertise

As already discussed, researchers look for technical guidance and expertise, and generate and distribute information from and for particular communities. The strength of that relationship depending on the discipline. The responses from the survey questions support the OAIS assumption that communities will play a major role in future preservation actions with regard to the digital collections. Specialised repositories tend to have a strong relationships with particular communities by their nature; institutional repositories do not. The task of building the appropriate relationships necessary with a wide and undefined community is an unresolved complexity, but the degree to which these relationships can be created will impact on future preservation actions and responsibilities. Individuals and researchers who deposited while necessary are not adequate to the full process, as they move on, acquire other interests, or the usefulness of the data may well outlive them.

Recommendations:

- Develop a strategy to manage community relationships for various categories of data in institutional repositories.

Future use and designated life of materials

Some data has a long useful life, and some data is only valuable for a specific period. The ion probe data for example, was considered to have "*a half life of 5 years. So there's always a chance that somebody wants some of the really early data, but it's usually for first order conclusions and after about 5 years old you start questioning the standards and just everything changes.*" (Ireland Interview 30'11). Recordings of lost language, or recording of unrepeatable seismic or astronomical events are believed by the interviewed researchers to be valuable in perpetuity. Linguistic analysis however, was identified by one content creator as being less useful in the long term. It required very specific forms: *"it's just not enough to be able to say – wow, we've got lots of speakers here. They have to be of a particular type and the speech has to be under particular circumstances and things like that, which narrows and narrows and narrows down until you haven't got anything at all. You might just as well go and get your own stuff."* (Rose Interview 34'15)

Recommendations:

- Incorporate a review or disposal option in deposit metadata (which would link to the designated community).

Funding

The majority of those interviewed identified what they saw as a failure in the funding mechanism, in that it was possible to obtain grant to create the data, but that the requirement to archive the data was not required or if it was, not enforced, and it was not possible to obtain sustainable funding to maintain the data.

Recommendations:

- Form a group, perhaps drawing on researchers identified in the process.
- Develop an official position paper on sustainability of data in grants.
- Develop an official paper on sustainable funding for repositories.

Sustainable Paths for Data-intensive Research Communities

The survey process provided a profile of a number of specialist repositories and owners of data intensive research collections. This will allow identification of suitable partners for the Sustainable Paths for Data-intensive Research Communities project. Potential partners have been identified from ANU, Sydney University and University of Queensland.

Metadata

The researchers recorded varying levels of comprehension about the types and quantities of metadata. Some described the data kept on cards about objects, some about technical information, some about exchange standards. Some had very well developed and highly sophisticated understanding of metadata and are involved in developing the standards that govern it.

Most researchers described a need for guidance on metadata that takes into account discipline specific descriptive metadata, technical metadata, preservation metadata, exchange and discovery metadata.

APSR is developing a position on preservation metadata including an implementation and it would be useful to point to other developments appropriate to exchange metadata.

Recommendations:

- Develop some metadata guidance based on current work. (ARROW?)

Rights and Copyright

The survey question regarding rights and access to materials elicited a variety of answers which pointed to a need for guidance on ownership from the legal point of view under copyright laws, as well as clarification from the University as to what rights they intend to enforce, and those they intend to waive.

Rights management

The researchers identified a major distinction in technical rights issues, code-able and uncode-able rights. Code-able rights are where individuals or groups may be identified as belonging to a particular group which has rights to access particular items. The relationship between user and item may be very complex, and a lot of this work is found in MAMS. Uncode-able is where the researcher or data owner wishes to make decisions about access on an individual basis. This is particularly a requirement of researchers still involved developing their work.

Copyright

Many researchers, particularly those with teaching responsibility, are faced with the need of making copyrighted material available on line. They are often unfamiliar with the complexities of copyright, and answers to the survey questions suggest that some would be in breach of copyright if they went ahead with their aims of making the data widely available.

Time Lag Access

Owners and creators of scientific datasets voiced the requirement of making data available after a certain date, generally after a particular publication date, or after the completion of the research. This may also apply to published journal articles.

Recommendations:

- Link to existing projects such as MAMS where these issues are covered.
- Develop guidance for aspects of rights which pertains to research materials
- Ensure repository developers and researchers are both aware of the issues.

Digitisation

Digitisation to create content is not directly described as one of the tasks of APSR Nonetheless it is important to the sustainable use of data to ensure that it done to adequate standards. One of the primary, and simplest, methods of sustaining digitised information is to select a stable and migrate-able format. In the area of sound and video, digitisation is clearly a preservation process as the older analogue technologies fail or become obsolete.

Types of Digitisable material: Not all projects are image based. The types of material which might be digitised include; image, photos, text, magnetic analogue recordings such as audio and video, paper based seismic data, survey questionnaires and microfilm.

Image Digitisation

Digitisation advice was also the most apparent need identified in the APSR survey of data collections. A significant proportion, possibly a majority, of those concerned with digital information were involved in digitisation projects. For image digitisation from two dimensional sources most of those interviewed had used the National Library of Australia’s image digitisation recommendations. The NLA’s standards are a good example of best practice based on the principle of “digitise once, use many times”; they are currently under review to incorporate latest technical changes and understandings.

Also, as has been identified in the SORRT digitisation advice project, the aims of researchers are not always the same as those of archival and cultural institutions, and so advice specific to the university sector is required. However, there is a significant need for “digitisation as a preservation approach” training for researchers whose collections include very important materials which will otherwise be lost. This is something that is held in common with cultural and archival institutions. Likewise, the number of disparate projects with similar technical requirements suggests the need in the universities for some method or structure for sharing technical facilities and expertise.

Sound and video digitisation

Many researchers hold large collections of valuable data in audio and video form. The time frame for preservation of this data is critical as the old carriers are rapidly becoming inaccessible through obsolescence of the replay technology and the failure of the carriers. The guidelines for preserving analogue audio are well established. The guidelines for video are developing in various organisations.

Recommendations:

- Continue to develop the SORRT project on digitisation advice.
- Develop or point to guidance on particular technologies and formats.
- Develop guidance or recommendations on digitisation for the purposes of preservation.
- Link with existing digitisation training.
- Suggest in the APSR reporting to DEST the need for a pro-active approach to digitisation.

Appendix 1 Annotated list of interviews.

Australian National University

Allen, Dr Bryant, Senior Fellow, Human Geography, Research School of Pacific and Asian Studies, Australian National University, interviewed by Margaret Henty on 29th April 2005. *The datasets discussed with Dr Allen relate to his work in Papua New Guinea: photographs, slides, prints, field notebook and other papers. These cover the fields of rural development, agriculture, population, land use and date back to the 1950s. They form part of a larger collection derived from PNG researchers, including maps, theses, books and other publications. The primary need of this collection is to get it digitised, which is relatively quick, and given appropriate metadata, which is a time-consuming process as there is no existing database to describe the items. Some of the records have been digitised, notably those of a CSIRO collection of aerial photographs of land use surveys. It is planned that all of these records will be made available via the ANU Demetrius Repository. Dr Allen is very concerned about the lack of funding for digitisation.*

Bowden, Dr John, Research Fellow in Linguistics, Research School of Pacific and Asian Studies, Australian National University, interviewed by Margaret Henty on 25th May, 2005. *Dr Bowden is associated with a number of projects in the field of Linguistics, notably a project funded by the German Volkswagen Foundation to document the Waimaha language of East Timor. He is also looking at the Makasae language as part of an ARC funded project. Dr Bowden has been associated with PARADISEC and has concerns about the need to provide appropriate facilities to document endangered languages in such a way that the research materials are not lost. He has collections of sound files, photographs and moving images, all of which need to be converted into digital formats for preservation purposes.*

Buckhorn, Dr Marcus. Head of Internet Futures, based in APAC Grid and GrangeNet projects, Division of Information, The Australian National University. Interviewed by Kevin Bradley on 17th May 2005. *Dr Buchhorn, though specialising in data management, has extensive expertise in Astronomy. The interview considers large scale data management projects with special emphasis on Astronomical data sets. Also discussed were chemistry, high energy physics, bioinformatics, the management of language data, geosciences and earth sciences. Issues ranged across the interaction of sustainable data and grid interaction, manual data retrieval systems, data retention times and very large data sets (as in high energy physics). Buchhorn set the local data management and storage systems in data intensive disciplines in the national and international data sharing grids and networks. Practical work situations and limitations of human generated metadata were considered, as was the technology needed to support the data sharing and sustainability.*

Cooper, Bob, Photographer in the Photography Digital Archiving Project, Research School of Pacific and Asian Studies at ANU interviewed by Margaret Henty on 26th April 2005. *Mr Cooper has been a Photographer in the University for many years and is now digitising thousands of photographs (colour, colour negative, transparency, black and white, some print images) of events, people and research taken over the last 36 years. His work primarily involves the development of* *standards for digitisation which can be used on his and other collections held by the Research School. The digitised collection is held in the ANU Demetrius repository. The major issues faced are the cost of digitisation and metadata creation, otherwise Cooper believes the work of image digitisation is relatively straightforward.*

Dancey, Kay, Cartographer with the Research School of Pacific and Asian Studies at ANU interviewed by Margaret Henty on 3rd May 2005. *The Chinese Historical Map Collection comprises large, colourful maps of China, produced in the 1960s and with historical interest. The maps have been digitised in sections and metadata provided by the Library. The collection will be made available via the ANU Demetrius Repository with MARC records available through the Library catalogue. The collection has been digitised to improve accessibility and remove the need to access the deteriorating physical copies. From a technical point of view, the biggest hurdle to be overcome in digitising the maps was accommodating the size of the maps and the consequent slowness of the digitisation process.*

Gates Stuart, Eleanor, Acting Director, Centre for New Media Arts, ANU interviewed by Margaret Henty on 20th April 2005. *Ms Gates-Stuart is a practising artist whose interest in digital artworks is twofold: firstly she uses manipulated digital images to incorporate into her work as a print maker, and secondly she wants to provide digitised images of her artwork via the web in order to stimulate discussion with other artists. To this end, she has made available digital images of her exhibition, arcv.pls.txt.scrb.spc.spt.vs.eleanor.gates-stuart via the ANU Demetrius Repository together with some commentary on the work. She hopes to add more in future but is constrained by lack of time and the need for technical support.*

Gregory, Dr Chris, Lecturer School of Archaeology and Anthropology, ANU, interviewed by Kevin Bradley 3rd May 2005. *Gregory's research is focussed on Indian oral epic songs. The project is collaborative and involves expertise from India as well as the performers of the songs. The recordings are digitised, or recorded digitally, and a transcript in the local dialect (Halbi), in Hindi and in English is generated. There is an archival and preservation requirement for the original recordings, as well as a need to maintain the links and annotations between sound recording and text and manage specialist scripts. There are also video recordings and still images in both analogue and digital form with similar storage, access and preservation requirements. Much of the interview discussion focussed on the need for technical support for researchers with technical requirements in such areas of research.*

Greenhalgh, Professor Michael, Sir William Dobell Professor of Art History in the Faculty of Arts at the ANU interviewed by Margaret Henty on 31st March, 2005. *Professor Greenhalgh runs a website, ArtServe, hosted by the ANU and he is moving the contents of this across to Demetrius for sustainability purposes. ArtServe contains images and documents for the study of art and architecture. As at May 5 it contained 485,541 JPEG images, each accompanied by a PNG thumbnail and supported by html files linked to about 20 images per page. Many of these images have been digitised from print, but he is now using an 8 megapixel camera. There are also two books in html form. On ArtServe he offers the capacity to hotspot, zoom and pan, and uses a variety of proprietary application such as Zoomify, PixMaker and DjVu. This is not replicated on Demetrius which offers the JPEG files only.*

Haberle, Dr Simon, Research Fellow in the Department of Archaeology and Natural History, Research for Pacific and Asian Studies, ANU, interviewed by Kevin Bradley on the 27th April 2005. *Haberle is Repository Coordinator of the Australasian Pollen and Spore Atlas, a project recently awarded ARC funding. The project aims to digitise 15,000 samples of pollen types – pollen and spores – derived from Australia, the Pacific and Asia. These pollen, taken from living plants – flowering plants, represent a reference collection for pollen and spore attacks throughout the region, which may be used, amongst other things, to enable identification of fossil material. The project will bring together a range of partner members and will standardise metadata and delivery systems. It will develop a federated discovery system, and is intending to use the local Demetrius D-Space initiative for the ANU part of the repository. The project will implement some specialised micro-image digitisation equipment and procedures.*

Ireland, Dr Trevor Senior Fellow, Earth Chemistry, Research School of Earth Sciences, ANU interviewed by Kevin Bradley June 2005. *Ireland is a researcher, user, and ultimately, area coordinator of the Sensitive High Resolution Ion Micro Probe (SHRIMP) at the ANU. The ion probe is used to analyse and type particular geological samples. The data has commercial value, but only to the data owners who possess the contextual information to describe the sample's location and significance. The raw data itself is useful, according to Ireland, for 5-10 years after which changing technical standards and resolutions are likely to make earlier analysis invalid. Interpreted data has a wider range of users. Data is duplicated and protected against disaster using innovative local techniques, an approach which is only possible due to the small size of each data object.*

Kanellopoulos, Lorena, Electronic Publishing Coordinator in the ANU Division of Information interviewed by Margaret Henty on 22nd March 2005. Brendan McKinley provided further technical information. *The ANU EPress was established in 2004 and publishes peer-reviewed monographs authored or edited by ANU scholars. Items are received in a variety of standard word-processing formats such as Word and converted to XML (DocBook) for storage with images stored as tiff. The books are accessed, either by chapter or as a whole, in pdf format, or for hand-held devices. They can also be ordered in print. The books are archived in the ANU Demetrius repository although access is via a separate website.*

Kercher, Therese, Project Officer, Scholarly Information Services / Library, Division of Information at the ANU interviewed by Margaret Henty on 4th May 2005. *The dataset discussed was the EPrints collection. This is managed by the Division of Information and serves as the official eprints site for the University, containing refereed articles, non-refereed articles, books, book chapters, working papers, theses, conference papers and technical reports. It was originally established using EPrints software, but the contents have more recently been migrated to Demetrius. Articles are received in various formats, most often PDF or Latex, and are converted if necessary into PDF for storage and access. Metadata is supplied primarily by the author but this may be supplemented by the Library. The main issues associated with the data set are not technical, but administrative, including publicity and staffing.*

Maidment, Ewan Manager, Pacific Manuscripts Bureau (PMB), Research School of Pacific and Asian Studies, ANU interviewed by Kevin Bradley and Margaret Henty on 19th April 2005. *The Pacific Manuscript Bureau consists of various library and* *archival materials acquired in, or about, the pacific. The primary thrust of the program has been microfilming. The PMB has undertaken digitisation of audio material as a preservation necessity, some digitisation of photographs, lodgement of digitally acquired databases in ANU's Demetrius, and client driven digitisation of microfilms. The Pacific Manuscript Bureau is a reformatting enterprise, and the form of that reformatting is driven by two concerns, client requirements and technical determined preservation responses.*

Millar, Dr Bruce, Associate Director, Research School of Information Science and Engineering, ANU interviewed by Margaret Henty on 21st April 2005. *The Spoken Language Processing data is made up of recordings of spoken language collected with a view to understanding information processing models. The collection contains data from several sources created or collected since 1980 including the Australian National Database Of Spoken Language (ANDOSL). Data has been created in a variety of audio, and more recently AV media: audio tape, DAT tape, CD-ROM, DVD, EXABYTE tapes and reading some of these has become an issue. Data has been transformed through different digital media and are at various stages of processing. The format is predominantly wav files with software management sensitive header files. Dr Millar is concerned that his data be held for the longer term in such a way that access can be managed properly. For privacy and copyright reasons, access will have to be carefully monitored. It is important to Millar that a suitable repository be found before he retires.*

Osborne, Renata, Manager, Menzies Precinct, Scholarly Information Services/Library at the ANU interviewed by Margaret Henty on 3rd May 2005. *The ANU Library has been digitising parts of its Asian Collections for some years in order to extend availability, and organising the data using ContentDM. The digital collections include out of copyright works in various languages from the rare books collection and political cartoons, converted to pdf with JPEG encoding (in order to accommodate multiple page documents). Text works have not been OCR'd or translated. The Library is moving digital materials across to the ANU Demetrius repository.*

Pallavicini, Rebecca, Manager of the ANU Division of Information Outreach interviewed by Margaret Henty on 18th April with further technical information supplied by Teresa Prowse. *The collection under discussion was the Division of Information Photograph Collection, a collection of 186 photographs designed to be used as a promotional, political and historical tool for the Division. The photographs have been taken with digital cameras, and are stored either as JPEG or TIFF. The TIFF files are converted to JPEG for accessibility purposes, but the original TIFF files are still available if required. The files are all held in the ANU Demetrius repository. JPEG files are held as both thumbnails and branded images, at 1312 x 2000 pixels, 300 dpi, and TIFFS are held at 4016 x 2616 pixels, 300 dpi. Permissions from all subjects are held separately in administrative files and other metadata is supplied using the Dublin Core scheme as used in Demetrius. Technical metadata inserted by the camera resides within the image.*

Pikler, Marianna, Music Librarian in Creative Arts Precinct, Scholarly Information Services / Library at the ANU interviewed by Margaret Henty on 5th May 2005. *The Anthology of Australian Music, previously available on CD, is being converted into digital format with a view to making the collection better known and more accessible.*

Some recordings date back many years and were recorded in analogue forms using a variety of recording media. More recently they have been recorded on DAT tape. All the items (41 CDs containing 415 files) were converted in the first instance by staff of the School of Music onto CD with individual items treated differently according to their origin. The subsequent conversion from CD to bwf has been carried out by the National Library of Australia using their Quadriga Jukebox System. Two issues remain to be resolved: access conditions and permissions. The issue of copyright is not clear as there are performers, composers and technicians to be considered.

Rose, Dr Phil Reader in Linguistics, School of Language Studies, ANU, interviewed by Kevin Bradley April 2005. *The Phil Rose collection comprises a large number of phonetic and phonological recording with predominately speakers of a group of Chinese dialects. The audio material is digitised at a reasonably low bit and sampling rate and analysed using specific linguistic software. The resolution, though much lower than recommended preservation practice, is more than adequate for the current analytical process. The main issues discussed concerned benefits or otherwise of an institutional repository, and the difficulty of making the audio available meaningfully without the involvement of the creator.*

Sambridge, Dr Malcolm Senior Fellow in the Research School of Earth Sciences, ANU interviewed by Kevin Bradley April 2005. *The data set consists of seismic data readings and seismic data modelling. It includes data collected from sensors, and acquired and stored directly in digital form, as well as records converted from earlier analogue instrumentation recorder readings. The Research School maintains its own data storage facility, server and archive. They also employ technical staff to maintain it, and to convert the analogue recordings into digital form. File formats are specialised, but have so far had a long useful life and revisions (1) remain backward compatible.*

Shapley, Maggie, Acting ANU Archivist with the ANU Archives and Noel Butlin Archives Collections interviewed by Margaret Henty on April 22nd 2005. *In 2004 there was a special project to digitise over 300 images from the Canberra Collection and the Noel Butlin Archive of Business and Labour and make them available through Demetrius and through PictureAustralia. The project is regarded as successful, but it would be difficult to continue digitisation in the same way without additional resources. The creation of metadata to the (very high) standard designated for the project was particularly time consuming. The Archives also hold other photographs, digitised as a result of user requests for copies. Ms Shapley was interested in investigating how to add these to Demetrius in order to take pressure off the physical collection.*

Volker, Joye, Manager of the Creative Arts Precinct, Scholarly Information Services (Library) at ANU was interviewed by Margaret Henty on 5th May 2005. *The Art Library Image Database is made up of about 60,000 high quality reproductions of artwork, mainly on 35mm colour slides, acquired for teaching and research purposes over the past 30 years. There is also a small number of digital reproductions of works created by staff of the Art School, notably textiles. The Library has been converting the slides into digital form in order to improve access. Access will have to be limited to ANU staff and students as copyright is not held by the University. Images are scanned as tiff files which will be accessed as jpeg. The images are stored and made available through the ANU Demetrius repository.*

University of Sydney

Cattley, Dr Sonia, Education Officer and Dao Mai, System Administrator of ANGIS: the Australian Genomic Information Service located at the University of Sydney interviewed by Margaret Henty and Su Hanfling on May 30th 2005. *ANGIS is a service located at the University of Sydney and with clients from both the higher education and commercial sectors from all over Australia. Data is brought in from the DNA GenBank, SwissProt and the Protein Data Bank, reformatted and made available to clients. In addition, clients are able to store the results of their own research using the service. There is no particular need to consider longer term accessibility of the purchased data as it remains the responsibility of the suppliers. Data held locally on behalf of researchers is unlikely to be of longer term interest. When a researcher stops using the service, a copy of their data is sent to them.*

Reeves, Professor Peter, School of Molecular and Microbial Biosciences at the University of Sydney interviewed by Margaret Henty and Su Hanfling on May 30th 2005. *The Bacterial Polysaccharide Genes Database of gene names was set up to track gene names as there is no official means of recording these, and to provide a scheme for the naming of all genes of a given function. It contains gene clusters, genes and names. The database was originally created digitally using FileMaker which has been updated as new versions have been introduced. The data is held on a Macintosh in the local laboratory and backed up to a second machine and to CD. The database is seen as being of importance to both academic and commercial interests and Professor Reeves is seeking help from the Library to assist with storage. Longer term sustainability is not at the moment an issue as the database is in constant use and consequently well maintained.*

Short, Professor Andrew, Director of the Marine Studies Centre at the University of Sydney interviewed by Margaret Henty and Su Hanfling on May 30th 2005. *The set of six databases relating to Australian beaches have been developed over a number of years. They contain different kinds of information in different formats. There are three Excel databases which describe 11,535 beaches around Australian and surrounding islands, containing information about their location, zoning, drainage, access, barriers, facilities, etc. A fourth Excel database lists available maps and aerial photographs. There is a database of scanned or digitally created photographs and a database of text descriptions of every beach. There is also a collection of sand. The data collection is now complete and Professor Short is seeking a home for all of it. The data has been collected using funds from the Australian Research Council, the University of Sydney and Surf Life Saving Australia and other water safety agencies. The data is of considerable interest to researchers and to government authorities. Subsets have been sold to local and state governments. There is considerable potential to sell information about individual beaches in book form.*

University of Queensland

Drinkwater, Dr Michael, acting Head of the Department of Physics, University Queensland, (substantive) senior lecturer in astrophysics, interviewed by Kevin

Bradley on the 19th September 2005. *The datasets discussed with Drinkwater were relatively small collections of Astophysical data which were stored on data tapes in an office environment. The data was critical to a number of papers and to research work done by Drinkwater, but because it was small scale individual work, not suited to the large scale infrastructure of the shared repositories. Identifying a likely method of archiving and preserving the data and its meaning was a prime concern. A secondary issue identified and discussed was the growing difficulties in retrieving data from old data tapes which are currently used to hold the information.*

Grigg, Professor Gordon, Head, Department of Zoology, The University of Queensland interviewed by Kevin Bradley and Belinda Weaver on the 20th June 2005. *Grigg holds multiple personal and institutional datasets generated in his career as a zoologist, these include sound recordings, images, and collected data regarding frogs, including many of now extinct or extremely rare examples, aerial surveys of South Australian kangaroo numbers taken over 25 years, and blood pressure readings of crocodiles which were used to resolve the way the crocodile heart works "the most elegant and sophisticated of all of the vertebrate hearts." Griggs is approaching retirement age and is concerned to ensure that this data, much of which is seminal in its field, is available and sustainable. Currently a grant is underway to convert the Kangaroo data to a sustainable digital form.*

Kelly, Professor Veronica, Custodian of The Australian Drama Bibliography Project interviewed by Kevin Bradley at the University of Queensland Brisbane on the 20th June 2005. *The Australian Drama Bibliography Project began in the early 1980s funded by an ARC grant. It soon outgrew its bibliographic limitations and developed into an extensive annotated database of Australian drama listing scripts, all known performances and productions of the plays, including information about the plays such as breakdown of cast, subject matter, theme words and a brief description of the subject. It was created using Inmagic, a commercial library oriented text and data management software and initially stored on the University mainframe. It was soon moved to a single PC, and has been migrated through a number of computers and operating systems. All the creators and managers of the data, excepting Kelly, have retired, left or passed away. Kelly has recently negotiated with the University of Queensland Library to take responsibility for the data. It has been transferred as a text file, though significant amounts of contextual information still exists in paper form.*

Lynam, Col senior observer, QUAKES Centre, The University of Queensland Interviewed by Kevin Bradley talking on the 20th June 2005. *Lynam described a large range of seismic data collected over a number of decades. Data on seismic events is available in ink and paper form, on photosensitive paper, or as digital files depending on the time of the event. Internal funding priorities means that maintaining this seismic data is no longer possible, and the records and continued reading are managed on a voluntary basis. The two critical issues are the storage of the data, and the digitisation of the old paper based records.*

Overheid, Jurgen Senior Scientific Officer in the School of Geography, Planning and Architecture, The University of Queensland interviewed by Kevin Bradley on the 19th September 2005. *The School of Geography, Planning and Architecture holds an extensive collection of spatial data sets. These include satellite images and aerial photography as well as vector data which may represent contours, cadastra, road centre lines, or network systems. The data is used by students, researchers and lecturers for teaching and is made available on the Schools own server and storage* *facility which is managed by Overheid.*

Pailthorpe, Professor Bernard, Chair of Computational Science, School of Physical Sciences, The University of Queensland interviewed by Belinda Weaver on the 12th September 2005. *The data set discussed is "Reef Grid", a 10 year old project collecting data on The Great Barrier Reef by AIMS, the Australian Institute of Marine Science. UQ is supporting a project to move the measuring and capturing approaches from manual capture and recovery to an automated and continual update project. The legacy data is being converted from Microsoft Access to Unix PostgreSQL open source database and all contemporary data is being included in the data base. There are significant developments envisaged for the project in open source development to support Scalability, portability, interoperability across legacy systems, multiple hardware systems, and multiple sites. The data also includes image and moving image recordings in a variety of lossy formats.*

Stumm, Deb manager of the Fryer Library and university archive and Anne Horn, Executive Manager, Social Sciences and Humanities Library Service, The University of Queensland, interviewed by Kevin Bradley on the 20th June 2005. *The University of Queensland Libraries and Archives hold a wide range of digital data sets. Apart from catalogues, databases, and collection guides, there are also websites and contextual information, sound recordings and digital and digitised images, off air recordings and video. The issues included sustainable systems and digitisation standards.*

Ulm, Dr Sean, Director of the Mill Point Archaeological Project, and Karen Murphy, sub-project manager, form the department of Aboriginal and Torres Strait Islander Studies, University of Queensland, interviewed by Kevin Bradley with Belinda Weaver (UQ) on the 19th September 2005. *The Mill Point Archaeological Project is a a multi site project involving the University of Queensland, the Environmental Protection Agency and Queensland Parks and Wildlife Service. The dataset is a complex and comprehensive database developed in Microsoft Access that stores images, detailed descriptive and provenance metadata about the excavated objects, GPS (Global Positioning System), differential GPS and EDM (Electronic Distance Measurement) data which can place an excavated object in three dimensions to within 10mm referred to total station values. The dataset will continue to grow as information is added. Current plan include converting the data to an XML database.*

Western, Emeritus Professor John, School of Social Science and Karen Hargreave Admin Assistant and part time IT Support interviewed by Kevin Bradley on June 20th 2005. *The interview concerned survey data accumulated over thirty years. The data comes from major Australian population surveys and another smaller one. While the coded data derived from the surveys is held by the Australian Social Science Data Archive, the original paper questionnaires, of which there are about 6,000 may contain uncoded data which could be useful to later researchers. Professor Western is therefore seeking a means of storing the questionnaires as he is due to retire. One of the imperatives to sustaining the questionnaires is the increasing capacity for data mining, as researchers make use of older survey and statistical data derived from a variety of sources.*

Appendix 2: Survey Questionnaire

Q1 Date of interview:

Q2 Name

Q3 Phone:

Q4 Position:

Q5 What is the name of the data/project?

Q6 Association:

Q7 What is your relationship with this?
- □ Primary creator
- □ Other creator
- □ Repository provider
- □ Administrator
- □ Other

[Q8, Q9]

Q10 Who has prime responsibility for it?
- □ Primary creator
- □ Other creator
- □ Institutional repository
- □ Administrator
- □ Other

Q11 General description of the dataset

Q12 How many objects/digital objects are in the collection?
- □

Q13 What do you define as a digital object?

Q14 Is it ongoing, or closed, collection?
- □ Ongoing
- □ Closed
- □ Open but not yet closed
  When to be closed?

Q15 What is the source of the data?
- □ Created by researcher/s
- □ Published research output
- □ Re-use of purchased data
- □ Re-use of other data

Q16 Who owns the copyright to this data?
- □ Creator
- □ Individual contributors
- □ Public domain
- □ Other
  Who?

Q17 What formats was it created in?
- □ Analogue
- □ Commercial software
  Which?
- □ Open standard
  Which?
- □ Locally developed

Q18 Does it need to be, or has it been, converted into another format?
- □ Yes – some
- □ Yes - all
- □ No

Q19 If yes, who will be/was responsible for doing this?
- □ Creator
- □ Local system administrator
- □ Local IT support
- □ External entity
- □ Other

Q20 Is/was the process of conversion documented?
- □ Yes
- □ No

How?

Q21 Is/was anything lost in the conversion process?
- □ Yes
- □ No

What?

Q22 Did it matter?

- ❑ Yes
- ❑ No

Q23 What kind of quality control is/was used when converting the data?

- ❑ Visual scan
- ❑ Automatic checking
- ❑ Other

Q24 What file formats are/will be used to access the data?

- ❑ Commercial
  Which?
- ❑ Open standard
  Which?
- ❑ Locally developed

Q25 What file formats are/will be used to store the data?

- ❑ Commercial
  Which?
- ❑ Open standard
  Which?
- ❑ Locally developed

Q26 What software is required to access the data?

- ❑ Commercial
  Which?
- ❑ Open standard
  Which?
- ❑ Locally developed

Q27 What are the main features of the software that you depend on?

Q28 Where does the meaning, or the importance of the data reside? (Is it in the accuracy of the colours or the layout, is the relationships or the functionality or what?).

Q29 Is the data available in different versions?

- ❑ Yes
- ❑ No

Q30 If yes, how do you define the different versions?

Q31 What categories of metadata are/will be used to describe the data?

- ❑ Rights and permissions
- ❑ Provenance (documented history)
- ❑ Technical metadata
- ❑ Administrative/management
- ❑ Bibliographic/descriptive
- ❑ Structural
- ❑ Other

Q32 Does this involve the use of any particular known scheme or standard? If so, which?

- ❑ Yes
- ❑ No

Q33 Do you record metadata about different types of entity?

- ❑ Collection
- ❑ Digital object
- ❑ Non-digital source object
- ❑ File
- ❑ Metadata
- ❑ Other

Q34 Who is responsible for the metadata creation?

- ❑ Primary creator
- ❑ Other project team member
- ❑ Individual contributors
- ❑ Research assistant
- ❑ Librarian
- ❑ Programmer
- ❑ Editor
- ❑ Other

Q35 How is the metadata stored and updated?

- ❑ Relational database
- ❑ Bundled with related content files
- ❑ XML database
- ❑ Proprietary database or format
- ❑ Flat files
- ❑ Object-oriented database
- ❑ Other

Q36 Where is the data held at present?

- ❑ Own/departmental server
- ❑ Institutional repository
- ❑ Remote server
- ❑ Own hard drive

- ❑ Floppies/CD/DVD/other media
- ❑ Analog
- ❑ Other

Q37 Is there any backup procedure in place? If so, what is it? Where is the backup kept?

- ❑ Not backed up
- ❑ Backup held locally
- ❑ Backup located elsewhere

Q38 How often is the data backed up?

- ❑ Once a week or more
- ❑ Once a month or more
- ❑ Other

Q39 Who are the main users of this data?

- ❑ Self/research team
- ❑ Researchers in the same discipline
- ❑ Undergraduate students
- ❑ Postgraduate students
- ❑ Other

Q40 How do the users access the data (directly? What medium?)

- ❑ No access allowed
- ❑ Internet – open access
- ❑ Internet - passworded
- ❑ Closed network
- ❑ Transfer on request via other electronic media

Q41 Who else might be interested in using it?

- ❑ Researchers in same discipline
- ❑ Researchers in related disciplines
- ❑ Undergraduate students
- ❑ Postgraduate students
- ❑ Government
- ❑ General public
- ❑ Other

Q42 Do they currently have access?

- ❑ Yes
- ❑ No

Q43 How might it be in the future?

- ❑ Open access likely
- ❑ Could have with permission
- ❑ No future access

Q44 Who funded the creation of this data?

- ❑ ARC grant
- ❑ Departmental/University funds
- ❑ External grant

Q45 Is there funding to sustain the data?

- ❑ Yes – built into existing funding
- ❑ Yes – separate (where from?)
- ❑ No

Q46 Who do you think should have future responsibility for the long-term sustainability of your data?

- ❑ Institutional repository
- ❑ Local area
- ❑ Self
- ❑ Other

Q47 Do you know of other datasets in your area which we could investigate?

Q8 University:

- ❑ ANU
- ❑ Sydney
- ❑ Queensland

Q9 Broad subject:

- ❑ Humanities
- ❑ Social Science
- ❑ Medical Science
- ❑ Science – other
- ❑ Multidisciplinary

Technical data

Files:

Quantity	Size	File format	Version	Date ranges

Details of software environment

Does the storage system record mime types?

Details of data management

Appendix 3: File Formats

As discussed above, the seemingly simple task of listing file formats becomes complex very quickly. There are many causes of this complexity which together highlight the need for accurate and complete preservation metadata and good representation information registries.

There were two approaches to generating the list of formats below. The first was to ask data managers what formats they were storing, with the added question of what role the formats served (archival, access etc). The answers to these questions were limited, due to ignorance of the one hand and essentialised answers on the second. In the former a surprising number of people were unable to answer the question regarding what format they stored in, In the latter case, expert answers only described the primary file format, ignoring the many dependant and associated files. So, for example, a Microsoft Access data base file is described as mdb, but it is unlikely to mention a Microsft Access Macro mam, Microsoft access report, mar, a Microsoft access form, maf, or the myriad of other files associated with Microsoft Access. The table created, though interesting, is insufficient to the requirement of sustainability.

The second approach was to run an enquiry program over the repository. This was done with DROID (http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm) which was passed over the ANU's Demetrius archive (D-Space). It may well be applied to other repositories. Its main limit is that it only describes formats it recognises, and the list of files it recognises is limited. It didn't recognise 715 instances of files, and was only partially correct in its recognition of the audio file.

Table a as result of questioning.

All universities	Open	Proprietary	Other
Formats used to access data	Jpeg, html, xml, tiff, txt, png, seed, sac, zdf, fits, gif, shp, miif, acoverage, suds, edm, gps	Pdf, bwf, wav, doc, mps, mpeg2, xls, gcg, Inmagic, mdb	Application specific community developed format for physics
Formats used to store data	Tiff, xml, ascii, html, png, seed, sac, zdf, fits, gif, shp, miif, acoverage, jpeg, suds, gps, edm	Pdf, bwf, wav, doc, mps, mpeg2, gcg, xls, inmagic, mdb	Application specific community developed format for physics

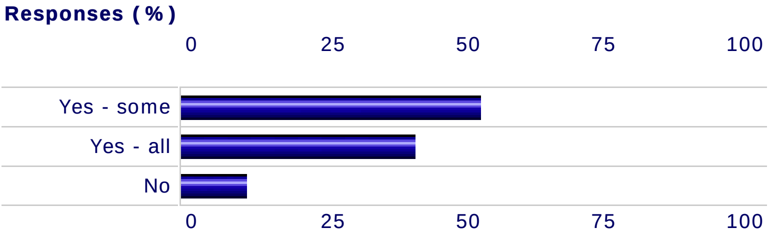
Table Generated by Droid on ANU's Demetrius.

Format	**Version**
Tagged Image File Format	3
Tagged Image File Format	4
Tagged Image File Format	5
Tagged Image File Format	6

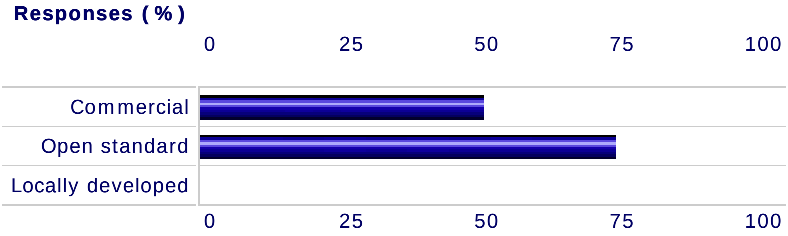
Exchangeable Image File Format (Compressed)	2.1
Exchangeable Image File Format (Compressed)	2.2
Exchangeable Image File Format (Uncompressed)	2.2
Extensible Hypertext Markup Language	1
Extensible Markup Language	1
Graphics Interchange Format	1987a
Graphics Interchange Format	1989a
Hypertext Markup Language	
Hypertext Markup Language	3.2
Hypertext Markup Language	4
Hypertext Markup Language	4.01
JPEG File Interchange Format	1
JPEG File Interchange Format	1.01
JPEG File Interchange Format	1.02
Portable Document Format	1.1
Portable Document Format	1.2
Portable Document Format	1.3
Portable Document Format	1.4
Portable Document Format	1.5
Waveform Audio	

Appendix 4: Statistical results

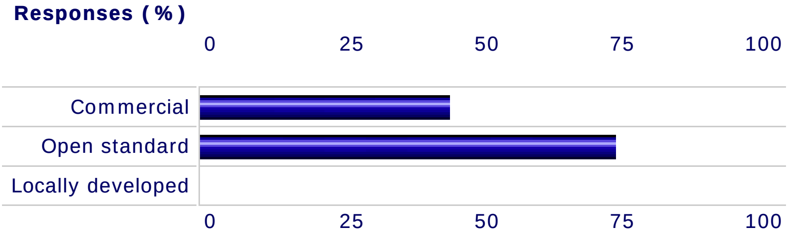
Does the data need to be, or has it been, converted into another format?



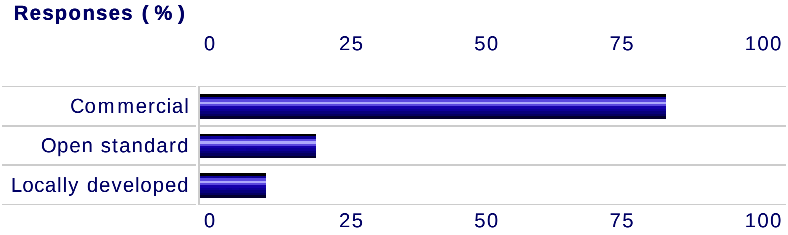
What file formats are/will be used to access the data?



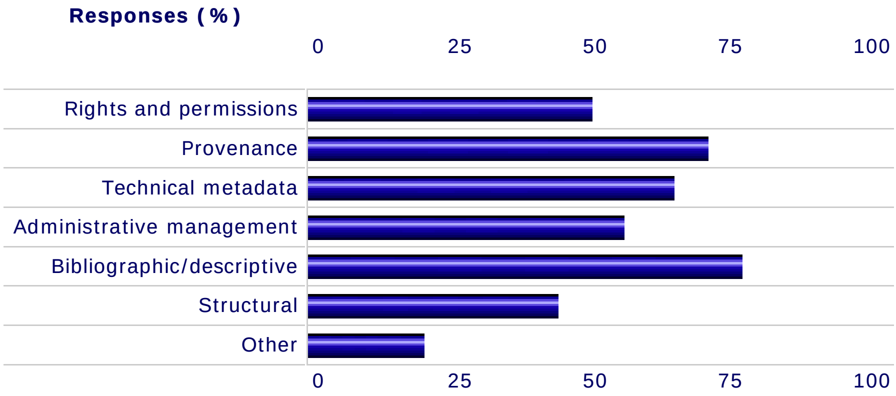
What file formats will be used to store the data?



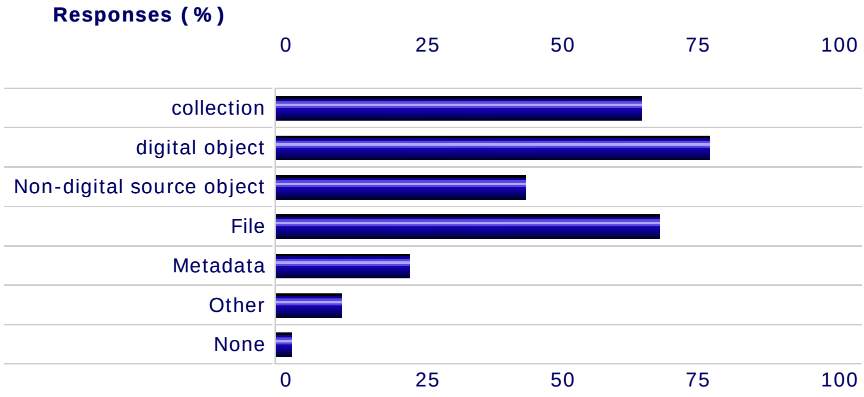
What software is/will be required to access the data?



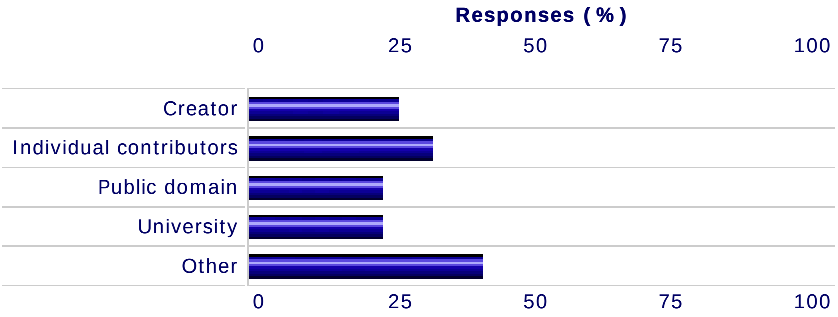
What categories of metadata are/will be used to describe the data?



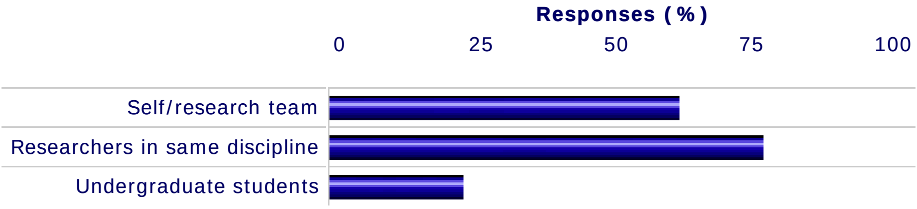
Do you record metadata about different types of entity?

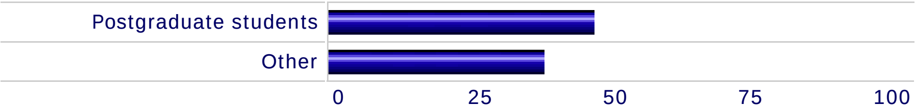


Who owns the copyright in this data?

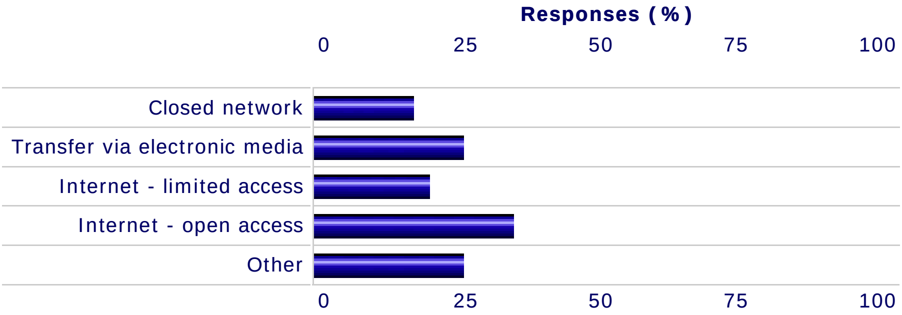


Who are/might be the primary users of this data?

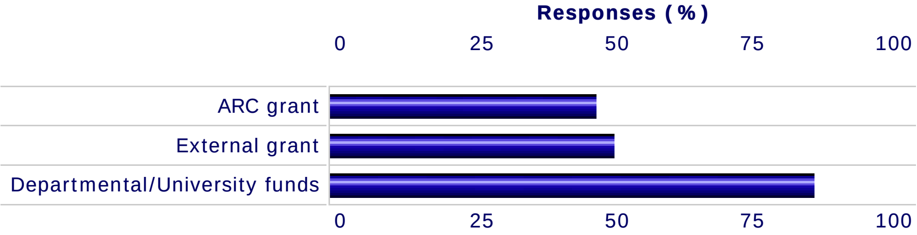




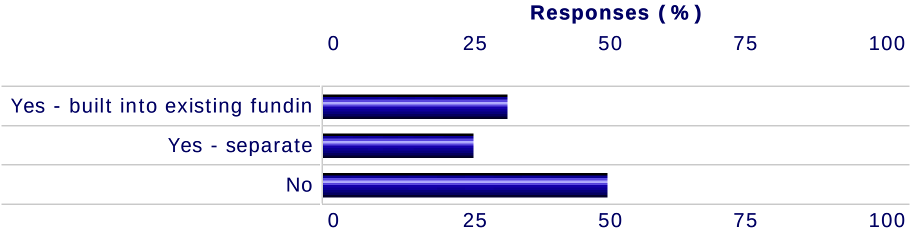
How is the data currently accessed?



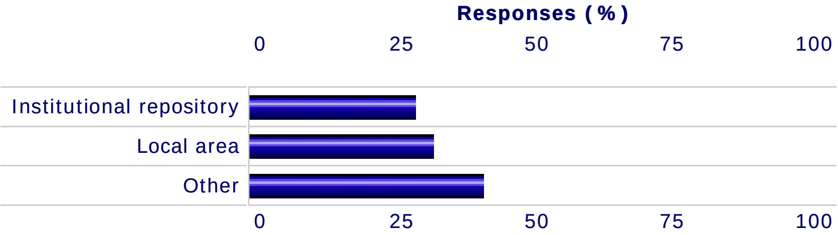
Who funded the creation of the data?



Is there funding to sustain the data?



Who do you think should have future responsibility for the long term sustainability of your data?


Responses (%)
Institutional repository
Local area
Other