

University of London Computer Centre

Digital Asset Assessment Tool (DAAT) Project

Final Report

Written by Ed Pinsent and Kevin Ashley

Version number 3.5

Release date 13 October 2006

Version 1.0 written by Ed Pinsent 25/09/2006

Version 2.0 amended by Angela Mott 29/09/2006

Version 3.0 amended by Kevin Ashley 06/10/2006

Version 3.5 presentation copy by Ed Pinsent 13/10/2006

This report is published as a deliverable under the **Digital Asset Assessment Tool (DAAT)** project. The project was funded by the Joint Information Systems Committee (JISC) under the Supporting Digital Preservation and Asset Management in Institutions 4/04 Programme.

Project website: <http://www.ulcc.ac.uk/daat.html>

Programme website: http://www.jisc.ac.uk/index.cfm?name=programme_404

Project contact: Ed Pinsent, e.pinsent@ulcc.ac.uk, 0207 692 1345

© 2006 University of London Computer Centre
20 Guilford Street
London WC1N 1DZ

JISC



Table of Contents

Table of Contents	2
Acknowledgements	4
1. Executive Summary	5
2. Background	6
2.1 Introduction	6
2.2 The National Preservation Office.....	6
2.3 Centralisation	6
2.4 Needs of the HFE community	7
2.5 The value of doing a Preservation Assessment Survey (PAS)	7
3. Aims and Objectives	8
3.1 Aims.....	8
3.2 Objectives	8
3.3 Things which changed during the life of the project.....	9
4. Methodology.....	10
4.1 Overall approach	10
4.2 Background work.....	10
4.3 Collegiate and collaborative approach	10
4.4 Research	11
4.5 Assessment	11
4.6 Database build	11
4.7 Testing within ULCC.....	11
4.8 Publicity.....	11
5. Implementation	13
5.1 PAS questionnaire grouping exercise	13
5.2 Drafting DAAT survey questionnaires.....	13
5.3 Condition questions	14
5.4 Assets held on servers.....	15
5.5 Further questionnaire construction	15
5.6 Questionnaire validation.....	16
5.7 Database build	16
5.8 Database hierarchy	16
5.9 Condition.....	17
5.10 Rationale for why this approach was an improvement on the NPO model.....	17
5.11 Testing the database	18
5.12 Guidance manual.....	19
5.13 Adding of scorecard elements	19
5.14 File format tools	19
6. Outputs and Results	21
6.1 Overview of outputs.....	21
6.2 The database itself.....	21
6.2.1 Functionality	21
6.2.2 Navigation and user-friendliness	22
6.2.3 Behaviour and bugs.....	22
6.2.4 Scoring capability.....	22
6.2.5 Other possible refinements.....	22

6.3 The questions in the database.....	23
6.3.1 Target collection section.....	23
6.3.2 Hardware Section.....	23
6.3.3 System management section.....	23
6.3.4 Preservation section	23
6.3.5 Software section and File format section.....	24
6.3.6 Media condition section	24
6.3.7 Media system management section	24
6.4 The overall D-PAS approach	24
6.5 The textual deliverables.....	25
6.6 Publicity.....	26
7. Outcomes	27
7.1 Project aims and objectives.....	27
7.2 Other outcomes.....	27
7.3 Lessons for other projects	28
8. Conclusions	29
8.1 A Digital Asset Assessment Tool based exactly on the PAS model will not work	29
8.2 Digital assets have special attributes.....	30
8.3 There is not an exact analogy from paper to digital.....	30
8.4 Value and importance is significant, but difficult to measure through such a tool	31
8.5 The role of a ‘central agency’ could be rethought, leading to rethinking the deployment of the tool itself.....	31
8.6 Increased automation is essential.....	32
8.7 The D-PAS tool needs better scoping.....	32
9. Implications	33
9.1 Survey entire systems, not just selected objects.....	33
9.2 Find more assets by widening the range of the survey	33
9.3 Add-on further technical approaches and solutions.....	34
9.4 Make Organisational improvements.....	34
9.5 Consider improving the archival / records management dimension	34
9.6 Incorporate costing elements.....	35
9.7 Perform sensitivity analysis of weightings	35
Gap analysis chart.....	36
References.....	37
Appendices	40

ACKNOWLEDGEMENTS

The Digital Asset Assessment Tool (DAAT) project was funded by the JISC under the Supporting Digital Preservation and Asset Management in Institutions 4/04 Programme. (http://www.jisc.ac.uk/index.cfm?name=programme_404)

Project partners:

- University of London Computer Centre (ULCC) - lead institution
- Arts and Humanities Data Service (AHDS) - development partner
- The National Archives - advice, evaluation and outreach
- Digital Preservation Coalition (DPC) - dissemination and sustainability
- The British Library - evaluation of tool
- The National Preservation Office – Preservation Assessment Survey (PAS) tool and advice
- University of London School of Advanced Studies (SAS) - pilot evaluation
- Kings College London - pilot evaluation

Thanks:

Colin Love, ULCC
Mina Creathorn, ULCC
Kate Bradford, ULCC
Keith Johnson, Stanford University
Alison Walker, National Preservation Office

1. Executive Summary

DAAT aimed to produce and test a Digital Asset Assessment Tool which would enable institutions which had already identified their digital assets to assess which of those assets were at greatest preservation risk, and make informed decisions, informed by both value and risk, on how to deploy possibly limited resources on digital preservation across the institution. DAAT expected to base this tool on an existing tool (PAS) produced by the National Preservation Office (NPO) which was already used in various forms to assess traditional collections in libraries, museums and archives and to assess photographic collections.

The project involved a number of partners with experience in the management of traditional and digital collections. ULCC's Digital Archives Department and the AHDS, as joint lead partners, brought significant practical and theoretical knowledge in digital collection management. The NPO, as the originators of PAS, had a good understanding of its development and its utility to collections. The British Library and The National Archives could bring in a wider perspective on collection management than would be obtained from the HE sector alone, and The School of Advanced Studies and Kings College London would be practical test beds for HE collections.

Not all of the project's aims were realised, but it has performed significant work in producing a digital survey tool, and in producing reports on what risk factors need to be assessed and what attributes an ideal survey tool would possess. In addition, it carried out work on assessing a range of automated file format assessment tools such as JHOVE and DROID as possible components of a future, more automated, digital asset assessment tool. Whilst we expect that the survey tool would require significant additional work before it was usable in a wide variety of settings, we believe that the reports which the project has made available will be of use not only to those embarking on similar ventures, but also to those looking at the wider issues of automation in digital curation.

The project's primary conclusion was that the data collection model common to all varieties of PAS was not suitable for use in the digital environment, and it is possible that even the underlying conceptual model of where risk lies and how it is measured does not translate well from traditional materials to digital collections. The more automated approach exemplified by tools such as Stanford's Archive Ingest and Handling Tool, and DROID's integration with PRONOM, provide more promising approaches in the short to medium term. This must be tempered with the realisation that some of the assets that are most at risk exist in systems which are not accessible to such automated assessments, so any future approach must allow for manual input. In addition, automated tools are still not capable of assigning a value to our assets, a step which is essential in helping us to make critical decisions about where preservation resources must be spent.

Most progress is likely to be made in the near future by designing specific tools to work in specific repository or content management systems.

2. Background

2.1 INTRODUCTION

DAAT was a response to a specific requirement in the JISC 4/04 call, which called for a project to develop a tool to allow the assessment of digital assets within an institution, based on an existing tool developed by the National Preservation Office (NPO) for traditional materials in archives, libraries and museums. The primary partners – AHDS and ULCC – both independently developed draft proposals on very similar lines and subsequently agreed to pool their expertise and resources to produce a successful bid.

The project sought the involvement of the NPO, their parent institution (the British Library) and The National Archives (TNA), in order to build on their expertise with such tools in traditional settings. Other partners from the Higher Education field agreed to test early versions of the tools to provide evaluative feedback as part of the project.

2.2 THE NATIONAL PRESERVATION OFFICE

As noted above, the National Preservation Office had already devised a tool which had been found to work for assessing the preservation needs of libraries, archives, photographic collections and museum objects.

The Library and Archive module (LA-PAS) developed from a research project carried out in 1998 into methods for assessing preservation needs in libraries. The authors determined that a sample of 400 items could be used as an efficient and accurate way of surveying libraries with holdings of more than 5000 items. In 1999 the NPO undertook a number of pilot studies to test the model in library settings. At the same time work was carried out by the Public Record Office (PRO – later TNA) to develop a companion model for use in archives.

Since 2001, the NPO has offered a library and archive preservation assessment survey (LA-PAS) tool, which has received extensive use in paper-based collections. The increasing emphasis on cross-domain working in stewardship and collection care encouraged the desire to adapt and extend its existing model, to create a practical, sample based method for museums to assess the preservation needs of their collections. Work was also done to adapt the underlying methodology to allow it to be used for photographic collections.

The fact that one methodology had been successfully adapted to a variety of collection types made it conceivable that adaptation of the tool to the digital domain could also be attempted.

2.3 CENTRALISATION

A valuable part of the NPO's PAS methodology is the central collection of survey data from organisations who use the PAS. This allows the NPO to build up a coherent national picture of preservation needs and challenges. A similar service would be extremely valuable for establishing a national view of digital preservation needs and this view was expressed by both the NPO and the Digital Preservation Coalition (DPC) at the time the project proposal was being drafted. The DPC was willing to coordinate the processing of outputs from

individual surveys to create this national picture, a task well-aligned with its national and cross-sectoral role in digital preservation.

2.4 NEEDS OF THE HFE COMMUNITY

Many reports have noted the lack of awareness, expertise or activity in the field of digital preservation amongst most libraries, archives and data centres. This applies to some extent both to those responsible for mixed analog/digital collections as well as those with purely digital collections. Even with such awareness and expertise, deciding how to prioritise often scarce resources for digital preservation presented significant challenges to institutions in the JISC community. Increasing deployment of institutional repositories, and their use to store materials other than the simply textual, made the need for an assessment tool more critical.

2.5 THE VALUE OF DOING A PRESERVATION ASSESSMENT SURVEY (PAS)

Owners of collections need justification to raise money for preservation purposes, and a tool can provide hard evidence for this purpose. The results of a PAS can help an organisation prioritise their own preservation work and allocation of resources; and justify and support applications to funding bodies for preservation and conservation projects. Anecdotal evidence suggested that this was seen as a significant benefit of using PAS in traditional curatorial settings, and the project partners felt that it would be useful to extend such benefits to digital curation.

3. Aims and Objectives

3.1 AIMS

The project's overall aim was to produce a digital preservation assessment tool, targeted at the needs of the UK HE, FE and research sector, but capable of deployment in other sectors such as national libraries, archives and museums and national and local government. Ideally, the tool would be sufficiently aligned with the NPO's PAS model to allow its outputs to be integrated with those from the other incarnations of PAS to assist NPO in providing an overall picture of preservation needs in the UK.

The tool was intended to address the needs of entire institutions or of groups with identifiable collections within those institutions. The aim was that it should be usable by librarians and archivists as well as research group leaders and IT professionals – all of these groups may have, or believe themselves to have, responsibility for digital assets of some form.

The information derived from carrying out a collection survey should allow institutions to adopt appropriate digital preservation strategies and prioritise activity based on an objective assessment of their needs.

It is also important to note that DAAT did not aim to provide institutions with a tool to identify digital collections of which they were unaware, nor did we aim to produce a tool which could be used in settings where the scope, location or existence of digital collections was unknown or uncertain. Like PAS before it, the DAAT project assumed that an institution had obtained sufficient control over its digital collections to be able to say what they were, or at least where they were. If nothing else, it would be impossible to use a sampling methodology of the type employed by PAS if one did not know the universe of objects from which the sample was to be taken. It is a fundamental principle of collection management that one must begin by knowing what the collection comprises. Only then can one begin to obtain control over its fate.

3.2 OBJECTIVES

The initial objective was to evaluate the suitability of the existing NPO PAS tool as a model for assessment in the digital domain. It is accepted that the tool produced was simple to use, widely applicable and capable of use by those who were not expert conservators.

The NPO PAS tool also enabled the production of statistics, gathered according to a common methodology, which allowed a consistent national picture of preservation and conservation needs and challenges to be developed. This was a strong argument for adopting it for DAAT.

There were, however, questions over how applicable the PAS model is to the digital domain – hence this initial objective.

In parallel, the project would define the ideal attributes of a digital collection assessment tool. This set of attributes would be used either to guide the adaptation of PAS to produce D-PAS, or to create an entirely new toolset should PAS prove to be fundamentally unsuited to the problem.

The project intended to develop a tool and pilot its use in a small number of partner institutions within and outside the HE sector. The tool was to be refined following feedback from these pilot tests.

Wider testing was planned within institutions not part of the initial project, coupled with events to raise awareness of the existence of the tool and its utility.

Once sufficient results have been gathered, the DPC intends to be able to use its results to build up a wider picture of digital preservation needs across the country and across sectors of activity. In this way, the DPC could mirror the role taken by the NPO with its PAS surveys.

3.3 THINGS WHICH CHANGED DURING THE LIFE OF THE PROJECT

The project suffered a severe setback at the outset with the unexpected departure of the key staff member (and co-author of the bid) from AHDS and the promotion, without initial replacement, of the key staff member at TNA. This caused delays and later scheduling problems from which the project never fully recovered, and led to the eventual scaling-down of a number of its original aims. In retrospect, the use of a less collegiate approach at certain key stages might well have allowed us to overcome these problems.

The project didn't involve as many participants as originally planned, and in particular was not able to conduct the two rounds of external testing which had originally been planned. In addition, the School of Advanced Studies (one of the first-round external test sites) did not implement its own institutional repository on the timescale originally envisaged and hence was not in a position to provide a suitable institutional setting for testing.

We may have copied the PAS methodology a bit too closely, in that we followed the same top-down structure for assessing a collection, and assumed that everything we need to know could be expressed as YES/NO or multiple-choice questions. Our original plan might have led us to abandoning PAS as a basis for DAAT, which would have enabled more rapid progress in some areas. However, soundings we had taken from elsewhere in the community indicated that there were strong motivations for attempting to use PAS as a model, and so we have continued on this path.

The project did add some work which arose out of changes in the external landscape during the project's life. We became aware of new tools from which we felt we could learn something of value, and decided that it would be useful to carry out evaluations in a way which would allow us to produce reports which could of value to others as well as to ourselves. These included an evaluation of file format testing tools such as DROID, and the Stanford Archive Ingest and Handling Tool, which did not exist when the original project proposal was made.

4. Methodology

4.1 OVERALL APPROACH

The overall project approach taken involved collaboration with project partners, research, assessment, software development, testing, and publicity.

The proposed project was to be carried out in three stages:

1. Assessing the feasibility of modifying the existing NPO PAS methodology to handle digital objects.
2. Development, and internal testing, of an assessment tool.
3. External piloting of the tool at institutions.

Dissemination activities to promote awareness and use of the tool were carried on from the outset, through the life of the project.

The three phases were a natural outcome of the way theme 2 in the 4/04 call was stated. The clear assumption was that any tool should be based, where possible, on PAS. Since we were aware that there was not a consensus that this was necessarily possible, we had to begin by evaluating whether it was possible and why (or why not.)

4.2 BACKGROUND WORK

The project team met with the NPO and secured copies of their database, along with supporting documentation, questionnaire, and other information. Unfortunately, we only got the 'bare bones' database. We did not get an example of a completed questionnaire, or reports showing how NPO analysed their collections, which in hindsight might have been useful.

4.3 COLLEGIATE AND COLLABORATIVE APPROACH

This approach involved gathering contributions from key staff in TNA, AHDS, and Kings College London.

A wiki was built early on in the project, for partners to communicate and gather ideas. It also served as a repository for project documentation. Experience which the lead partners had gained with other projects involving geographically-distributed partnerships, including European collaborations, had convinced us of the usefulness of wikis as a means of keeping communication moving, and of managing project documentation. The wiki is of particular benefit when different project partners are working at different times, enabling progress to be made without the need for tight scheduling of activities. They don't replace the need for meetings, but enable those meetings that do take place to be more focussed and more productive. Wikis encourage partly-developed documents and ideas to be exposed to others, allowing constructive criticism to produce results more quickly. Even where other partners do not wish to contribute to a particular document, its gradual development in full view of others makes everyone feel more informed. The success of TWiki (the tool we chose to use) on this and other projects in which ULCC has been involved has resulted in its use for many other development and administrative activities within the institution.

This approach meant that the detail of particular work packages was defined and agreed by the project group as a whole, but implementation was then typically delegated to an individual.

4.4 RESEARCH

Desk research and collaborative critical evaluation was used to define a set of criteria which the DAAT tool should satisfy. Much research into published guidance took place. A lot of the issues which DAAT touches on have already been researched and the digital preservation world has already published some suggestions and guidelines for best practice. It made sense to re-use a lot of this groundwork in the context of our project. The work that has been done so far has concentrated on the characterisation of individual entities and the ease (or otherwise) of obtaining the information necessary to perform such characterisations. We also needed to build on this to consider issues of usability and practicality in the type of setting in which DAAT might be deployed.

4.5 ASSESSMENT

The DAAT tool was subjected at every stage to expert assessment by project staff to evaluate the extent to which it satisfies established criteria and the feasibility of adapting it where it does not satisfy the criteria. Andrew Wilson was well-informed with sound thinking on particular issues to do with the detail of preservation and transformation, for instance, and we were also able to consider the experience of NPO with the use of PAS in traditional settings.

The other deliverables, which took the form of written reports, were assessed and validated by staff within the lead institutions who had not had direct involvement with their production.

4.6 DATABASE BUILD

Small-scale software development followed, managed through regular review. A ULCC data specialist (Mina Creathorn) was brought in for the database build.

4.7 TESTING WITHIN ULCC

In the initial and secondary test phases, test partners were asked to address specific issues of usability and suitability as well as undertaking evaluations which are specific to their own environment. Test feedback was collected through written reports and through follow-up conversations where necessary. The results were used to decide on changes to be made to the tool and to feed into output reports on the project.

4.8 PUBLICITY

The final stages of the project were intended to focus on awareness raising and promotion and were achieved through a variety of means, including conference papers and publicity through partner and project web sites. The project web site (hosted within the ULCC site at <http://www.ulcc.ac.uk/daat.html>) has already been used to disseminate project outputs and posters on the project have been produced for the JISC Joint Programme Meeting in July 2005 and the joint JISC/NDIIPP meeting in Washington DC in May 2006. We intend to

carry out further dissemination activities through outlets such as the DCC's proposed seminar series after the formal end of the project.

5. Implementation

5.1 PAS QUESTIONNAIRE GROUPING EXERCISE

We started with an evaluation of the available survey materials which had been provided by the NPO, in particular the guidance manuals and survey forms. The survey forms were called LA PAS for Libraries and Archives, PHOTO PAS for photographic collections, and MUSEUMS PAS for museum collections. They took the form of questionnaires, they all had areas of overlap, and they broadly asked the same sorts of questions in similar sequences.

The commonality was that all PAS Surveys approached the question of **condition assessment** as a separate line of inquiry. The Museums form in particular was very concerned about this and probed more deeply into the physical nature of different objects. Libraries and archives have only to worry about the condition of paper, leather or parchment. Museums have a variety of materials - metal, glass, stone etc.

To test the suitability of these survey forms for application to digital assets, we did a simple grouping exercise on the forms. For each question component, we asked:

- 1) Does the question continue to apply to digital material? (This was split into two sections, one applying to material held on servers, another for carriers).
- 2) How specific is the question to the survey it was originally designed for?
- 3) Is there a related question that could be added to the question to make it more relevant to digital assets?
- 4) If applicable, could the answer to the question be discovered by automated methods?

The results of this exercise were mapped into a comparative spreadsheet. The immediate result from this was (unsurprisingly) that the condition assessment approach suggested by NPO clearly could not apply to digital materials in their present form.

The value of this simple exercise is that it was more methodical than simply rejecting the PAS approach out of hand. The process of doing the exercise helped us to rethink the questionnaires, and build on the PAS approach.

5.2 DRAFTING DAAT SURVEY QUESTIONNAIRES

We then began to create our own DAAT questionnaires. It was assumed right at the start we would need two separate approaches, and thus two forms:

- 1) questions applicable to assets held on inactive media or carriers
- 2) questions applicable to assets held on servers, also called “virtual objects”.

The distinction may appear to be arbitrary but is, we hope, based on sound reasoning. Some clarification may be of benefit to the reader, however. Material in category one is meant to refer to digital objects which are held on a medium such as a CD, DVD, tape or portable disk drive which is not attached to a computer at all times. Physical handling is thus necessary to carry out any task relating to the objects stored on the carrier, and it is not possible to carry out automated monitoring of the condition of objects on the carrier (or

of the carrier itself.) Problems which may affect the longevity of the carrier are independent of problems which may affect the longevity of the objects stored on it.

The risks which attach to objects on a server (where they will typically be on spinning disk of some sort) are somewhat different. Although all of the object-level risks are the same (since they derive from the way in which a bitstream is used to represent information and the behaviour of that representation) the carrier is now effectively the entire system in which the object is held, not simply the particular disk drive on which it happens to reside at a particular instant. One thus needs to ask different questions in these circumstances.

We're aware that there is not a clear dividing line between these two situations and that a more abstract phrasing of some questions would allow a server to be considered as just another type of carrier. The distinction seemed useful and practical to us and made it easier to frame concrete questions whose answers were (we hoped) clearer for those we expected to be using the tool. We would welcome views on whether the choice we made was a good one.

5.3 CONDITION QUESTIONS

Early on it was not clear if we could assess condition of /damage to digital objects, or if it would be necessary to rethink this part of the survey completely. Clearly media could be damaged, but once damaged a disk often either works or it doesn't; are there 'degrees' of damage which are worth surveying? More specifically, are most institutions capable of carrying out this type of assessment? This highlights both a significant difference in the way that digital and analog carriers tend to behave, and in the understanding of carrier behaviour that tends to exist in traditional and digital curatorial settings.

Digital carriers, such as tapes or CDs, are still fundamentally analog objects, made of materials which are subject to gradual decay as well as to damage through sudden insults such as breakage. But the gradual decay, whether it be the slow loss of magnetisation taking place in domains on a tape or the chemical decomposition of the dye layer in a CD, are not typically visible to us. Digital equipment is expressly designed to work round these phenomena and make them invisible to us until they are no longer correctable. At this point, the media appears to fail completely, suddenly and without warning. Until that point is reached, error correction codes and retries of failed read operations are used to conceal the gradual failure of the underlying medium from us, its users. This is a great contrast to the way in which most traditional materials behave. The fading of ink or the increasing brittleness of paper is apparent even to the untrained eye, and most other phenomena which threaten the survival of an analog object are apparent to the trained eye.

What's more, that trained eye is expected to exist in most libraries, archives and museums. It is a basic assumption of the various forms of PAS that the institution has someone who is capable of making expert assessments of the condition of the objects that it holds. The same is not true of most digital repositories. The authors are aware, for instance, that tools exist to allow one to get detailed information on how much error correction or retries are needed to successfully read various types of optical and magnetic media. We have used one or two of them, but we would not claim to be experts in their use. How, then, can we expect the average digital repository to use such tools to carry out condition assessments of its media? We decided that we could not.

The media handling and storage questions were taken from the published TNA guidance. We also devised a condition survey form for carriers, incorporating questions about damage caused by handling disks, taken from the Fred Byers Council on Library and Information Resources (CLIR) guidance. (See the References section in this report).

5.4 ASSETS HELD ON SERVERS

Using the Online Computer Library Center (OCLC) Digital Archive Preservation Policy and Supporting Documentation as a starting point, we created a questionnaire intended to assess digital objects held on servers. It addressed the question largely in terms of whether the assets have adequate support: from hardware, software, and file formats; and whether these things themselves were adequately supported by the institution, its staff and finance. We thought this could, in digital terms, possibly replace the PAS concept of 'condition / damage' – it seemed more meaningful for our purposes.

The AHDS made detailed suggestions to do with bitstream preservation, checksums and fixity of a digital object. These questions do not address the idea of the static condition of a digital object, but of our ability to detect alterations to it. They are analogous to activities such as stock control or environmental monitoring in traditional collections. Change detection is important in both settings. Although mould, damp and insects threaten traditional materials, theft and vandalism is also a significant threat to the preservation of some collections. In the digital domain, theft is unlikely to be as significant but deliberate alteration of holdings is a real concern. Fixity mechanisms are our protection against this particular threat.

The OCLC document also contributed ideas on organisational risks and preservation / transformation risks. Organisational questions were added to the collection survey form at a high level. It involved adding to the 'Institutional Management' block of questions, which were expanded and sorted into areas of policy, staff, finance and copyright.

Questions about preservation and transformation actions were eventually incorporated at collection level.

We did not consider OCLC's suggestions for risks connected with “Associated Organisations”. This line of thought refers to software vendors and content depositors, and whether they're a monopoly or at risk of going out of business. We felt that these questions will be better addressed in future by acquiring the information from external sources, preferably in an automated fashion. PRONOM is a current, if imperfect, example of such a source.

5.5 FURTHER QUESTIONNAIRE CONSTRUCTION

We added 'Institutional Management' type questions which were inspired by wording in the eSPIDA project on sustainable preservation of digital assets in a university, based at the University of Glasgow.

We added Section 0 on Organisational details. This is just a set of fields for the organisation to identify itself. This was mainly for the benefit of the NPO; it assumes that the results of the surveys will be collated, and that whoever is doing the collation needs to distinguish one organisation from another. We also developed a section for a Target Collection Area.

We now had two fairly coherent survey forms, one for servers and one for carriers; each was in two parts (collection, then condition). We then adjusted the layouts and numbering sequence in order to achieve some sort of parity between their structures. This way all the numbered questions were present in both forms, even if marked 'Not applicable' to carriers. This seemed sensible to prevent the two surveys becoming too divergent.

Drafts of these working spreadsheets are attached as an Appendix. These drafts contain a column with a note of which published source is inspiring each particular question.

5.6 QUESTIONNAIRE VALIDATION

By this point our drafts were already up to over 60 questions, with very little content surviving from the original PAS surveys. A meeting was held to validate the content of the questions; a few further suggestions were made, for example on tricky scenarios involving 'complex objects', and were incorporated.

At this stage there were a few gaps in our understanding of how the questions would be applicable. For example, it was not crystal-clear what the 'unit of sample' would be. Still, we decided to go ahead anyway and put it all into a database. This was a good idea as it resolved a few areas of uncertainty.

5.7 DATABASE BUILD

The database task began in April 2006 and was assigned to ULCC's database expert Mina Creathorn. We used MS Access 2003. This was a simple evolution of approach adopted by PAS rather than an explicit decision that this was the ideal form in which to create the survey tool. We had considered very different mechanisms for collecting information at the project's outset, and considered retro-fitting them to PAS. We had wondered why PAS had chosen a standalone MS Access database rather than some form of networked or web-enabled solution for data collection, particularly as we knew that this caused problems for some institutions that didn't possess the necessary licences for use of MS Access. But it became apparent that online solutions would actually be more work in most traditional curatorial settings. Sampling and answering of the condition assessment questions is typically carried out in library stacks or archive strongrooms. It's possible to bring a laptop to most of these areas (and it usually has to run off its battery) but network connectivity is rarely available. Thus, the answers would either need to be written down and then typed into the online forms later, or a more sophisticated tool developed which could handle offline data collection and then submit the data online when connectivity could be established. We weren't sure that these types of restrictions would necessarily apply to digital collections, but it certainly meant that developing one online mechanism that could be used in all settings was more challenging than we had originally anticipated.

In the same way that we adapted PAS survey forms, we used the MPAS database for our first attempts, with the understanding that it would need reworking and re-design.

The initial version had no score values assigned to any of the answers and simply contained the text of our questions. The same format (YES/NO answers) was used for the most part, although some of the questions were multiple-choice.

5.8 DATABASE HIERARCHY

Our design evolved into a nested hierarchy of tables. This means the survey moves from a macro to micro structure successfully. We automated this workflow aspect with buttons on the forms, which navigate the user in a logical way from one form to the next, thus taking the user down through the five levels.

At Level 1 - profile statements about the entire organisation.

At Level 2 - questions about the policies and strategies of that organisation.

At Level 3 - questions which identify the target collection area(s). For any successful survey, there will probably be several of these.

At Level 4 - questions which cover each of the nominated target collections areas.

At Level 5 - questions about individual assets within each target collection area.

The questions at levels 1 and 2 would only need to be answered once per survey.

At level 3, we realised the user would need to make decisions about what the target collections are, and what's in scope. At this point we found we were moving towards the notion that a target collection could be a discrete 'container' for a set of digital assets. So a target collection could be:

- A server
- A related collection of carriers
- A single tape or disc
- A drive on a single PC
- An external hard drive (including storage drives, mobile phones and digital cameras)

For each nominated target collection, the user would then complete a profile at level four. This profile would survey the following areas of the collection: System management, Asset inventory, Accommodation, and Preservation Actions. Our design came up with tabs in a database form, to accommodate the large number of questions. We assumed that these survey questions only need to be answered once per collection.

Finally, at level five, we survey individual 'items' within each target collection. This part of the database would allow creation of records for as many individual assets as there are in the target collection. We were clear now that an individual asset was probably going to the best unit of sample at this level. In most cases, each asset will be one file; though some complex assets may consist of many files and many formats.

At this 'item' level, the survey would ask questions about individual formats, applications, environment, location, descriptive metadata etc. It would also survey the value and importance, and retention status, of each individual asset.

The various levels would be linked together in the database by common fields such as Organisation ID, Collection ID, and Asset ID. This was designed to allow many-to-one relationships of records belonging to specific target collections.

5.9 CONDITION

The 'Condition' questions were also 'hardwired' into this hierarchical structure design (unlike in the NPO model, where the condition survey is separated).

For virtual objects and servers, we had questions about Hardware, File Formats, and Software, distributed at appropriate levels in the hierarchy. The hardware survey would take place at Level 3, because we assumed the target collection area would all be governed by the same hardware.

File format and software surveys, both applicable to a single asset, would take place at item level, level 5. For media and carriers, the condition questions would all take place at item level.

5.10 RATIONALE FOR WHY THIS APPROACH WAS AN IMPROVEMENT ON THE NPO MODEL

- It made better use of database functionality
- It built a logical tree structure from macro to micro level

- It followed a sequence and guides users through that sequence
- It meant questions don't have to be repeated
- It allowed for greater efficiency in the design
- It was a much more integrated analysis tool, bonding condition questions with the collection questions

Summary of some positive results from the database design stage:

- a) We had a sound hierarchical structure with linking records
- b) We had a clearer notion about what the unit of sample would be
- c) We were clearer about what 'condition' meant in a digital context
- d) We had something approaching an integrated workflow methodology, rather than two separate surveys

5.11 TESTING THE DATABASE

We now had a workable Access database. ULCC began a round of testing within the project first of all, to iron out any obvious defects in its design, wording, and functionality. The first round was to test it with Colin Love, a senior systems administrator at ULCC. His detailed findings are in another report.

Results from Colin Love's input:

- He identified a new type of organisational risk, if external consultants configured a system or bespoke software was utilised. This resulted in additional questions being added to collect information on these potential risks, which result from a potentially lower level of control over system behaviour.
- He suggested that within the Access database we could have dynamic fields that appear depending on the choices made previously. This was considered, but rejected as an over-complication that would add unnecessary work.
- He confirmed our suspicion that no single person could complete a D-PAS survey. "For a large data centre such as ULCC, we can cover all of the questions fairly easily, but for a small organisation with perhaps limited IT skills I think they would struggle."

Further external testing was carried out by the AHDS. The task was undertaken by a member of staff who was proficient with MS Access, and also looked after the operational preservation activities in AHDS. His detailed findings are also compiled in the testing deliverable report.

Results from AHDS input:

- Several suggestions for rewording and applicability of certain questions, particularly in the matter of hardware.
- Identification of certain bugs to do with field behaviour, selection, editing and saving.
- Suggestions for improving navigation.

5.12 GUIDANCE MANUAL

Concurrent with the database design, we started to evolve a guidance manual. Initially this was to be used for the test stage, to help a user navigate their way around. These notes also contained useful discursive information about what the questions mean, taken from the original questionnaire forms, and clarified.

The manual was divided into numbered sections, each relating to a corresponding form or tabbed view within the database.

The guidance manual has been provided as a separate project deliverable.

5.13 ADDING OF SCORECARD ELEMENTS

This involved assigning a numeric value to questions in the survey. This was the methodology:

- Identification of those questions which, if answered YES or NO, would yield results which have some bearing on the life of the digital asset collection.
- Identification of those questions which would not have any bearing on the life of the collection, such as those used for identification or relocation purposes only. These were made into non-scoring questions.
- Assignment of a value of 1 point for every NO answer, and 0 points for every YES answer. This was on the assumption that the higher the score, the worse the preservation problem.
- Looking again at certain areas of the survey where the question was considered to be crucial enough to warrant a higher score than 1 for a NO answer. These questions were mostly found in the organisational part of the survey. Scores were upgraded to 3, 4 or even 5 points, depending on the severity of the consequences. The complete lack of a preservation policy, for example, would yield a score of 5.

The draft scores were mapped into the database. At time of writing, we haven't been able to validate the figures, or create queries to extract them. Appendix 3 contains our thoughts on how this part of the work could be taken forward.

The plan is to arrive at an overall score for the 144+ questions. Other planned refinements include:

- Grading the scores, so that organisations know what they mean and where they stand on a scale
- Banding the scores, in order to bring out preservation problems that are related, or grouped together in one area
- Allowing some form of automated feedback and what-if scenarios, so that users can see more easily what they could do to improve their score

5.14 FILE FORMAT TOOLS

We were always aware that the manual approach which PAS espoused might not be attractive for digital objects, and that the potential existed to collect information about entire digital collections in an automated fashion. Accordingly, we decided to evaluate existing toolsets which carried out this basic task to understand whether DAAT could utilise

them directly, build on their technological approaches, or learn from how they worked. We also wanted to understand what things were not yet amenable to automated assessment.

JHOVE was in existence before we began work on DAAT and was clearly something in which we would take an interest. During the project's lifetime, TNA produced DROID, which had the added advantage of being able to make direct use of PRONOM, one of the information repositories which we saw as a potential supplier of information to an assessment exercise. Late in the project's life, the joint JISC/NDIIPP meeting led to a discussion with one of the key players in the Stanford Archive Ingest and Handling Tool, which further spurred our desire to evaluate automated approaches suitable for bulk handling.

The tools tested were DROID and JHOVE. The data specialist downloaded evaluation copies of each tool, and tried local experiments using a variety of file formats found on drives at ULCC. The detailed results of these tests have been written up as a separate deliverable.

6. Outputs and Results

6.1 OVERVIEW OF OUTPUTS

The project's initial expectations were that its principal tangible result would be the D-PAS database. It is still a tangible result, even if our own findings lead us to question whether this particular incarnation of the tool is the right base from which to continue development. It evolved from being a 'flat' set of questions into a manageable database, and the hierarchical structure of the end product suggested itself through the working process. The value of practical co-operative work shows that ideas and improvements generate from actually doing something, rather than thinking about it too much.

But we also feel that the ancillary written deliverables which informed our thinking are of value to others. The assessment of risk factors report and that on necessary attributes of a digital assessment tool are, we hope, informative documents for anyone either seeking to build on D-PAS or trying to build a different assessment tool from scratch. The evaluation report on the file format assessment tools adds to the body of knowledge and experience which will inform future community use and development of systems such as DROID and JHOVE.

This part of our report thus looks at the outputs from two perspectives. We begin with a detailed description and assessment of the database tool itself, and then examine the documentary deliverables.

6.2 THE DATABASE ITSELF

Suggestions for improvement to the tool have arisen from the testing process. Some of these would warrant incorporation in the finished product, if the project had more time to develop it further.

The paragraphs below attempt to evaluate the usefulness of the tool, and take an objective view of the database itself, and the questions within it; and also evaluate aspects of the overall approach.

6.2.1 FUNCTIONALITY

Overall functionality for data entry purposes is good, although it is assumed that the end user will have some familiarity with MS Access. The end product now generates a Unique ID for each organisation that completes it; allows several collections to be added, with Unique IDs; and several assets to be added, also with unique IDs. The end result is that each profiled asset has a unique three-part ID, attaching it to the parent records of collection and organisation. This can all be seen in the relationships diagram for the database.

There was an early bug with first designs, in that Sections 10 and 11 were generating their own Asset IDs; instead they should have 'inherited' this value from Section 8. This has since been resolved and the integrity of the DB remains good.

6.2.2 NAVIGATION AND USER-FRIENDLINESS

D-PAS has a lot of screens to navigate, and sometimes tabs within screens. Early prototypes were not easy to navigate, but the deliverable version has the following user-friendly features:

- Labels on each screen to identify the numbered section
- A basic sequential logic to the navigation, taking the user through the sections in a numbered sequence
- Forward and Back buttons on each screen
- Named target collection appears on each relevant screen, thus allowing users to know what exactly they are profiling
- Symmetrical alignment of check boxes

The usability for data input purposes is pretty good, though could probably be improved by refinements such as:

- Ability to access any section of the survey from any screen
- Additional navigation buttons
- Default to 'Add new record' on every screen
- Conceal the Unique Identifiers from users

6.2.3 BEHAVIOUR AND BUGS

In terms of maintaining the logic of the intended hierarchical structure of the survey, the database works fairly well, although there remain structural loopholes and a few bugs in its behaviour.

One significant structural issue is the behaviour relating to profiles of collections. Ideally, if the user nominates a server collection to be surveyed, the database should accordingly disallow data entry on the media forms (and vice versa).

The bugs spotted so far include:

- Drop-down menu which allows editing (and shouldn't)
- Tick-boxes allowing selection of two opposing attributes
- Size field scrambling numerical data
- Selection boxes which cannot be de-selected
- Selection boxes which unexpectedly fill in automatically, based on previous answers

6.2.4 SCORING CAPABILITY

Not yet available.

6.2.5 OTHER POSSIBLE REFINEMENTS

If further developed, the database could incorporate such refinements as:

- Addition of 'Don't Know' questions
- Dynamic fields, which only appear conditional on data entry in another related field
- Sophisticated reports, queries and what-if scenarios

6.3 THE QUESTIONS IN THE DATABASE

Overall, the D-PAS survey is basically sound from a technical viewpoint, but could be improved. Some specific suggestions arising from the testing process include the following:

6.3.1 TARGET COLLECTION SECTION

The correct identification of a target collection area will be absolutely crucial, if the tool is actually going to work. This needs to be emphasised more in the guidance.

The 'size of target collection', if it remains in the tool, should indicate some form of measurement, for example a number of bytes or objects. If it's a numeric field, there should be some internal logic that forces a target collection to be lower than its parent collection. The usefulness of this question (one of many carried over from the NPO model) is doubtful however, and it may be better to drop it completely.

6.3.2 HARDWARE SECTION

This section has quite a few issues:

- Many of the questions in D-PAS do not appear to apply to hardware.
- The questions imply that hardware is being discussed for storage *and* delivery. This may not always be the case.
- New versions of kit are not backward compatible, the only exception being tape drives which can normally read older formats. The compatibility questions need reworking.
- We ask if the hardware specification is complex, large or ambiguous. This needs to be qualified. It only makes sense in a comparative context; the setup which Colin tested may appear complex compared to a single server.

The hardware questions could be improved, by moving away from a one-to-one relationship between hardware and collection. It may be useful to distinguish between different hardware. An organisation is likely to have several machines that are dedicated to a specific function. A warning could also be raised if the institution is performing preservation action, preservation storage and distributing assets on the same machine.

6.3.3 SYSTEM MANAGEMENT SECTION

Organisations may have audit trails for some tasks, but not all tasks. One single yes-no audit trail answer for the entire collection is too simplistic.

The backup policy may exist, but D-PAS doesn't know if it's adequate. D-PAS asks if there is a backup policy and where it is stored, but it doesn't ask if the backups are actually checked regularly.

6.3.4 PRESERVATION SECTION

This section may lack clarity. Is it probing preservation policy as a whole or preservation action performed on the identified collection?

It may not be possible for a user to indicate if the transformation process is reversible and repeatable for every asset in the collection. Therefore, the YES/NO answer will have only limited value.

Many users will be bewildered by the fixity question, regarding AIP encapsulation or storage as a separate file. The question could be reworded to distinguish between the two options.

6.3.5 SOFTWARE SECTION AND FILE FORMAT SECTION

- D-PAS suffers from confusion over the use of terms 'source code' and 'file format'.
- Some questions are ambiguous.
- Sometimes source code and file format questions appear to be interchangeable.
- Software source code may not need digital rights, encrypted sections, or watermarks.

One way to improve focus in this area would be to include an additional question focusing on the compatibility aspect of the software application, for example "how compatible is the application with older versions of the format". Possible answers could be "compatible with all versions of the format", or "compatible with recent versions of the format". But this is an excellent example of the type of question that we would rather not be asking the end-user at all, but deriving an answer from an external source like PRONOM.

6.3.6 MEDIA CONDITION SECTION

Some of the questions on disk condition may require a N/A option. The questions do not apply to hard disks.

6.3.7 MEDIA SYSTEM MANAGEMENT SECTION

Some distinction may need to be made between media carriers intended for access, and media carriers used for backup only.

6.4 THE OVERALL D-PAS APPROACH

The D-PAS approach as it stands combines the following elements:

1. An analysis of the organisation and its policies regarding its assets
2. An analysis of the value of the assets
3. A technical analysis of the information environment the assets are held in

The project does not yet know if this approach is going to be suitable in general HFE environments. First and foremost, in spite of the external testing that has taken place, the tool has not been tested on an actual collection of assets.

Secondly, based on views from the external evaluator at Kings, the current approach may be weak on the organisational and value aspects (in particular, in Sections 1 and 9); and D-PAS may either need to improve in these areas, or scale-down its approach to something that is more technology-focused.

The Section 1 questions could be too broad; an organisation may have policies, staff, and resources, but how effective are they? Are they being effectively channelled towards the aim of digital preservation?

The Section 9 questions may have omitted or glossed over some important archival / records management issues. For example:

- Ownership of the asset is not really dealt with by D-PAS.
- The 'special value' question conflates a lot of separate and important issues. Legal value is not the same as research value.
- 'Uniqueness' is not an archival value.
- Rights residing in the data are not mentioned in D-PAS. These may affect things like legal access, and the right of the organisation to make preservation copies through refreshment and resaving.
- Academics tend to save and re-use a lot of third-party materials in the context of institutional record-keeping. This too may constitute a risk.
- Questions probing for legal retention issues, FOI and DPA issues are omitted from D-PAS.

These issues are further explored in the Conclusions part of this report.

6.5 THE TEXTUAL DELIVERABLES

Three reports have been produced which encapsulate the thinking and research of the project partners at various stages of the work. We believe that these are of some use independently of the other project outputs and have potentially wider applicability.

Two of the reports deal with matters we had to consider before embarking on the production of the assessment tool. One summarises the risk factors that can affect preservation of digital assets, and could itself be of use to institutions who want to check whether they understand all the risk factors that may be significant for them. This report is relatively brief, and much of the information it presents is not new. Its contents will generally be familiar to anyone who has worked extensively in digital preservation. But it also considers the ease of measuring some of these factors (such as the ability to assess media condition) and hence goes beyond a theoretical assessment of what risk factors exist, into a practical assessment of which ones we can measure and hence have some hope of doing something about. Subject to available resources, ULCC plans to update this document periodically.

The other initial report concentrates on the factors we had consider in the design and implementation of the tool – what were the constraints, what was the ideal, what alternatives would we reject and why? Some of the conclusions are very specific to the project's setting and its initial presumption that PAS was the model of choice, but some are more general and should be of benefit to anyone considering the implementation of such a tool in a general or specific environment. We would be particularly interested in comments on our thinking as presented in this document.

A report which was produced much further in to the project was our assessment of three automated digital object assessment tools, JHOVE, DROID and AIHT. Each are carrying out related, but not identical, tasks. Although JHOVE and DROID can be used to assess bulk collections of objects, only the Stanford tool is expressly designed to deal with bulk collections. Ours is not the only evaluation of these tools, but we believe that at present it is the only evaluation which considers them not just as their authors intended, but as

possible models for the automation of other aspects of digital preservation and curation, such as collection assessment. This report is not a *Which*-style evaluation to tell you which format assessment tool gets our best-buy rating (although it might help you make such a decision for yourself.) But we believe it to be informative about the tools themselves, and about the wider issues of automation in digital preservation.

6.6 PUBLICITY

The project website was launched March 2006, at <http://www.ulcc.ac.uk/daat>. The final deliverables will be posted to this site.

The project completed two poster / talk sessions. The one at Washington May 2006 did bring a good result, in that it attracted attention from the Stanford AIHT project, and expanded our understanding of the ways in which automation could be used in the context of digital assessment. Further detail is in the file format tool assessment deliverable.

7. Outcomes

7.1 PROJECT AIMS AND OBJECTIVES

Our original aims and objectives included the following:

1. A report on the suitability or otherwise of PAS as a model for DAAT.
2. Working papers on the relative importance of different factors affecting long-term preservation of digital materials.
3. DAAT assessment criteria.
4. DAAT guidance manual.
5. Database and data entry tool for pilot sites based on MS Access.
6. Promotion of the work, including at least two short articles, one conference presentation or poster, four regional seminars, one web page.
7. Report on experience of pilot sites and effectiveness of the methodology.

Of these, 1-5 have been successfully delivered.

For 6, we delivered two poster sessions and developed a web site. We are planning further dissemination activities after the end of project funding from JISC.

7 was delivered, using ULCC and AHDS as pilot sites. We didn't work with external institutions.

7.2 OTHER OUTCOMES

Other planned outcomes included:

An increased awareness of the importance of inventories of digital assets within institutions (since the methodology can only effectively be employed where such an inventory exists) and an improved picture of general digital preservation needs within the HE community. We originally thought it would be possible to group this with the more general picture of conservation needs which NPO produces through use of PAS on traditional materials.

In addition:

- We have delivered a report on testing of file format assessment tools.
- We learned that the NPO approach doesn't really map across, and arrived at this conclusion scientifically.
- We learned a lot more about the dynamic nature of digital objects in a collection, especially for preservation purposes.
- We produced a 'gap analysis' diagram, considering other areas we could explore in the future.

We have not succeeded in producing an assessment tool which we could whole-heartedly recommend to others as a basis for carrying out an assessment within their own institution. The database, and the questions it contains, could be of use to institutions as part of a wider asset assessment strategy and could certainly help in developing such a strategy. We have identified some significant steps that need to be taken on the way to the production

of a more ideal tool, however, and these are outlined in section 9, on the project's implications.

We believe that the report on risk factors will be of use to anyone with responsibility for digital materials who does not yet have a deep understanding of digital preservation matters. It, and the report on the desirable attributes of D-PAS, will be of use to any group looking to construct or evaluate a similar tool.

Finally, the report on file format assessment tools will be of utility to anyone who needs, or thinks that they might need, to make use of such a tool. It will also be informative for systems developers and others who need to consider wider issues of automatic in the context of digital curation.

7.3 LESSONS FOR OTHER PROJECTS

We have already described (see 3.3 above) the problems which arose after key staff members in partner institutions changed roles before the project had really got underway. We allowed this to paralyse development for too long, and probably would not have done so had the project been contained in a single institution and the staff losses happened in that one institution. We had initially been determined that all partners should have the opportunity to participate in all decisions, partly as a reaction to our participation in previous projects which had been over-compartmentalised.

Projects involving many partners where most decisions are taken at the centre, and where one is only informed about decisions which a direct impact on one's own work packages, tend to lead to a sense of disconnection and a lack of collective ownership of the project. This is not conducive to a successful outcome. But an over-reliance on democracy can also be hazardous, as we have found. At times it would have been better to press ahead and reallocate work without waiting for collective consensus on whether this was the right approach. Other projects involving many partners may well already be aware of this potential risk; we, at least, are now better informed.

8. Conclusions

Our conclusions can be summarised by the following statements, which we expand on further in the rest of this section:

1. **A Digital Asset Assessment Tool based exactly on the PAS model will not work**
2. **Digital assets have special attributes**
3. **There is not an exact analogy from paper to digital**
4. **Value and importance is significant, but difficult to measure through such a tool**
5. **The role of a ‘central agency’ could be rethought, leading to rethinking the deployment of the tool itself**
6. **Increased automation is essential**
7. **The DPAS tool needs better scoping**

We think we have built a tool of some value which will be a useful first step. TNA have said “this project is a significant step towards a valuable tool kit for a wide variety of digital collections.” Colin Love of ULCC said “I think this is quite a useful tool if aimed at smaller organisations that maintain a small archive of data, and perhaps larger ones who don’t have much of an idea about what they should be doing.”

8.1 A DIGITAL ASSET ASSESSMENT TOOL BASED EXACTLY ON THE PAS MODEL WILL NOT WORK

Attempting to map the NPO model to a digital collection has illuminated some of the challenges posed by digital collections.

The NPO survey may be a bit too simplistic for the digital realm. It proposes a basic 'Yes/No' approach. If certain things are not being done (i.e. the more 'No' answers you give), the more this increases your score; the higher the score, the greater the preservation risks for your collection. ULCC tried to copy the same Yes/No format. Due to the technical complexity of digital assets, it resulted in a large number of highly specific questions.

This led us to three observations:

(a) No single person could ever answer all the questions, so if this method was used you would need a team of specialists and stakeholders (see the Guidance Manual for suggestions). There remains some ambiguity as to who the questions should be directed at. This will, according to TNA (and in the opinion of the other project partners), remain an issue for some time. Colin Love added “My final concern is the number of questions being asked. For a large data centre such as us we can cover all of the questions fairly easily, but for a small organisation with perhaps limited IT skills I think they would struggle, especially if external consultants configured the system or they bought some bespoke software.”

(b) Selecting a target collection, and appraising individual assets within it, presents difficulties. Some assets may have complicated dependencies which are hard to express in the questionnaire format. There are also questions about the wider information environment, and about whether to look at the fine detail of system and library files. Where

dependencies exist, they have to be captured and understood; and managing dependencies is a key element to the preservation of the affected assets.

(c) This would all add greatly to the time a survey would take to complete. This is important, and easily underrated. People will not fill in a survey that they think is wasting their time. Questions will be poorly answered or skipped all together. The tool really needs the ability to deal with lack of knowledge. We must *assume* that certain questions cannot be answered for some assets, and design a scoring system which can deal with this.

8.2 DIGITAL ASSETS HAVE SPECIAL ATTRIBUTES

This is a truism, and one might counter that the original library-based assessment model was successfully adapted to other domains which would also lay claim to the “special” attribute, such as photographic collections. But we would argue that, for this exercise, digital is more different. These differences relate not just to the assets, but to the understanding of the asset’s properties that their curators have, and in the way they manage their collections. Collections of digital assets are not the same as collections of library books or archives. One single digital asset is not the same as a single library book.

Digital assets are not always "static" objects in the same way as a paper-based collection, particularly not if they are involved in a preservation programme with such things as transformation, migration, and differing source and target formats. They are separated from their carriers, and often independent of them, in a way that analog assets can never be. This allows us to protect them from some risks in a trivial fashion, but exposes them to other risks. But the biggest difference is almost certainly in the knowledge that their curators possess. LA-PAS and its derivatives assume that the collection manager has the knowledge to make detailed risk assessments in relation to a particular item, or can at least easily buy in such knowledge. The strength of LA-PAS then lies in how it exploits that knowledge, leading from a set of observations about particular items to a set of conclusions about the collection as a whole. That assumption that one can make an accurate assessment of a single asset is not a safe one in the digital environment. The tool cannot exploit knowledge which is not present at the outset.

A statistical sample of a digital collection will not necessarily reveal much of value about the whole collection.

8.3 THERE IS NOT AN EXACT ANALOGY FROM PAPER TO DIGITAL

Preservation problems of digital assets are only tangentially analogous to those of paper-based materials. If you preserve a book you also preserve the means of reading that book. A digital asset requires preservation of itself and of a means to represent it. Admittedly, there are similar issues surrounding the maintenance of environmental conditions and security of a computer room and an offsite storage unit, which perhaps can be probed. Likewise, correct handling of storage media is every bit as important as correct handling of an item in a photographic collection.

A paper resource can suffer degrees of damage and still remain partially usable. The same cannot be said of a digital resource, especially not a piece of damaged media, which either functions or doesn’t function.

The issues associated with digital preservation branch out into many technical areas of curation, care, protection and sustainability which do not as yet have (paper) archival equivalents.

Archival and library physical care is a well-established area; it's well known by now what might go wrong and generally agreed what can best be done about it. Digital preservation is not as well-established and there is not the same degree of consensus as to what the solutions are (or, at times, what the problems are). Even where there is such consensus, problems which might at first sight differ only in scale are in fact different in kind. Gradual change which takes place over decades (a typical scenario for some types of risk with paper) is fundamentally different from change which takes place over periods of five years or less, and which is often a step function rather than a gradual change.

8.4 VALUE AND IMPORTANCE IS SIGNIFICANT, BUT DIFFICULT TO MEASURE THROUGH SUCH A TOOL

This could be one of the structural weaknesses in the original PAS survey. One of the concerns raised by the project was whether we placed this line of questioning at the correct level. Some users would like to assign value and importance to individual assets. Others would rather apply it to several related assets within a collection.

The sorts of the questions asked might not deliver completely meaningful answers for actual collections, because they tend to take the asset out of its local context and put it in a national context. This problem is exacerbated in the digital realm.

The future task might be to devise questions that work for collections as collections, such that their value and importance can be assessed in a more contextual way. For example, with collections of particular-instance papers, where the loss of an individual item may not be significant, but the collection's value for analytical / statistical research might be weakened if say 40% of the collection were lost.

See also the records management suggestions, in the 'Implications' part of this report.

8.5 THE ROLE OF A 'CENTRAL AGENCY' COULD BE RETHOUGHT, LEADING TO RETHINKING THE DEPLOYMENT OF THE TOOL ITSELF

The National Preservation Office's approach put the NPO in the centre of an operation where they acted as the agency that would gather results from organisations, analyse those results, and publish the 'big picture' of preservation needs in the United Kingdom. Also arising from this process was that NPO would offer consultative advice and play a crucial part in addressing the preservation problems of individual organisations.

In the context of this project, does the same scenario apply? In the Higher Education field, which D-PAS is targeted at, there is no clear candidate for anyone prepared to act as such an agency (with the possible exception of the DCC), nor is it clear if such an agency is needed. There are several organisations that have a stake in digital preservation in the UK, but outside of the DPC it is not known if any are prepared to act in this advisory role. Who would benefit from a nationwide picture of digital preservation needs in the HFE field? Are Universities and others prepared to share this information?

If the tool were deployed simply as a standalone diagnostic for an HFE Institution, it could proceed on a different footing. The profile information (name, address, phone etc) would not be needed; and the organisation could pre-load the tool with internal policy / procedural information that would affect the scoring outcome. The score could be based on a balance between these internal standards, and external ones based on the behaviour of certain file formats and agreed preservation methodologies.

8.6 INCREASED AUTOMATION IS ESSENTIAL

The DAAT questionnaire as it stands is all assumed to be ‘manual’, ie questions are answered on an individual basis. There was talk early on in the project of ‘automated inventories’, but it’s not clear what information we would want such an inventory to provide.

There is the AIHT tool which has been developed at Stanford, and may become more widely available in the future. It seems to be describing a fundamentally different approach to that of the NPO’s LA-PAS model. Stanford based their model on workflow, not on static collections; they devised software to analyse large numbers of file formats in a hard drive; and created an XML output from the process. The software they devised is called 'Empirical Walker'. They keep stressing in their report that, for best results, metadata extraction had to be automated, not done manually.

AIHT also has some provision for assessing things like context, meaning, value and importance of assets - through questions directed at the content owner.

There may be a way to yoke the D-PAS hierarchical questionnaire with an adapted AIHT tool, to produce something that works on both macro and micro levels. Broadly speaking, the D-PAS tool tends to work top-down, where AIHT tends to work bottom-up.

It may also be possible to develop tools like DROID or JHOVE to analyse assets on a file format level, and again arrive at a sort of hybrid solution. TNA’s Ian Hodges view was that “DROID can be called via an API so I don’t see why not. I would think a hybrid tool would be the best solution since it could secure reliable technical knowledge of the assets being described.” In the short term, such developments are most likely if we consider producing a tool which can work within a more specific environment, such as that presented by a particular institutional repository platform.

8.7 THE D-PAS TOOL NEEDS BETTER SCOPING

The current approach of D-PAS, based to a large extent on the NPO model, is looking at current digital storage in an organisation as if it was the same as preservation. There may be a basic confusion here which could perhaps be addressed by rescoping the tool and its intended audience, and thinking more carefully about the desired outcome.

Should it get bigger by taking on more records management/archival type functions, or should the tool be more focused? Suggestions for improving the former functions are listed in Implications, 9.5 in this report. But perhaps a simpler approach would be to scale it down, and abandon the ‘Value and importance’ and ‘Organisation’ dimensions completely.

This simplified approach might lead to a tool that performs a quick audit of an organisation’s servers, and identifies short to mid-term risks associated with their systems, formats, hardware, software and media storage. The likely target audience for a tool like this would be a systems or repository administrator.

The implication is that any responsible organisation should extract from their current systems the material worthy of permanent preservation, placing it in a trusted digital repository or an OAIS-compliant repository, and that this takes place within the archiving programme. A separate assessment tool, taking an approach based on preservation, migration, and transformation actions, could identify any risks taking place within that sphere of operation.

9. Implications

Our conclusions suggest that the D-PAS tool by itself is not necessarily the optimum way to assess a collection of digital assets. If the ideal DAAT tool could be built in the future, it would also cover several other technical and organisational areas to make it comprehensive and 'holistic'. A few 'blue sky' suggestions along these lines are offered below, together with some more practical and short-term ideas for future development. This is followed by a gap analysis chart.

9.1 SURVEY ENTIRE SYSTEMS, NOT JUST SELECTED OBJECTS

An ideal survey tool would have to be automated as a given. Apart from doing the automated crawl of assets in some way, it needs to include some form of checksum stage. Further, it then needs to analyse all the other IT dependencies as well, starting from the single asset itself and branching outwards and upwards. In short a fully integrated D-PAS tool needs to start with a clearly identified asset, and look at:

- The file itself
- The format of that file
- The metadata of that file
- The properties of that file
- The system files generated by that file
- The system files that the file depends on
- The software needed to create, access, and preserve that file
- The hardware dependencies of that file
- The driver dependencies of that file
- The cross-platform dependencies of that file
- The server dependencies of the network it's stored on
- The management environment in which the asset sits
- The possible use of the asset (in a sense, the designated community for the asset, in the narrow sense of OASIS)

Many of these things have already been suggested and incorporated in the survey tool we have built, but only in a general way, and framed as Yes/No questions.

In the medium term, some of these aims are most realisable if we build a version of the tool that works with a specific repository implementation or content management system, where we can exploit APIs within the system to gain as much information automatically as possible.

9.2 FIND MORE ASSETS BY WIDENING THE RANGE OF THE SURVEY

This would entail performing a survey of 'scattered' collections and 'at risk' formats. Cornell University have tried something like this in their own institution. This way of thinking is intended to address two popular scenarios: the disk or other media being stored in direct sunlight on the windowsill, or the professor who holds the only copy of a valuable thesis in a near-obsolete file format on his PC.

The implication of this suggestion is that not all an institution's assets are actually held in a single "collection". An organisation would need to be more pro-active than the NPO model suggests. However, these tasks are really about enabling institutions to get themselves prepared for an asset survey: identifying where assets exist and making a start on building up an inventory for them. Much of the work which is going on in the domain of institutional repositories is addressing this challenge. However, since these repositories are still focussing primarily on research and teaching collections in the HFE domain, other important assets (such as institutional records) are still likely to be missed.

9.3 ADD-ON FURTHER TECHNICAL APPROACHES AND SOLUTIONS

9.3.1 Probe more deeply into the business of complex relationships of assets; for example complex hybrid files (a Word file with a spreadsheet embedded in it); whether system files and/or library files also need to form part of the risk analysis; and the risks in the wider information environment.

9.3.2 Bring in media testing tools.

9.3.3 Build in some form of technology-watch element, through the use of PRONOM Persistent Unique Identifiers (PUIs) perhaps.

9.3.4 Consider platform and cross-platform dependencies, both inside and outside the organisation. The professor with the only copy of a thesis on his PC could be working at home on Windows 98 (now unsupported) attached to a home network with an Apple Mac. Scenarios such as this argue against the complete automation of the tool; assets which are held in obsolete environments, which are more at risk than most other assets, are also the most resistant to automated attempts to assess them.

9.4 MAKE ORGANISATIONAL IMPROVEMENTS

9.4.1 Use of eSPIDA model (based on the balanced scorecard) to help organisations make a better business case, put the survey on the organisational agenda, raise awareness, etc. Help organisations quantify the consequences of data loss; attach a cost to the risks.

9.4.2 Consider ways that a systems administrator could try and improve their DAAT score through systematic decision-making and actions; for example "if I transformed all the .png files in this drive into another format, what would my score be?"

9.5 CONSIDER IMPROVING THE ARCHIVAL / RECORDS MANAGEMENT DIMENSION

This suggestion comes from the Kings College review. Various archival principles which can affect digital preservation include:

Ownership of the asset: The organisation may hold the asset, but may not be responsible for its long-term storage or preservation.

Value of the asset: A key part of archival methodology is identifying value, and this includes legal and research value, not just national heritage issues.

Retention of the asset: Although incorporated in D-PAS, it may not be possible to identify retention requirements when applied to current or semi-current assets, particularly not at a digital object level. It would be easier to identify related series of assets where their

commonality is that they are governed by established retention rules; but then the D-PAS survey proceeds on a quite different basis.

Rights in the data: This affects things like legal access to the asset, but also affects preservation; an organisation cannot preserve an asset if they do not hold the necessary rights. Refreshing and resaving files constitutes a copying action.

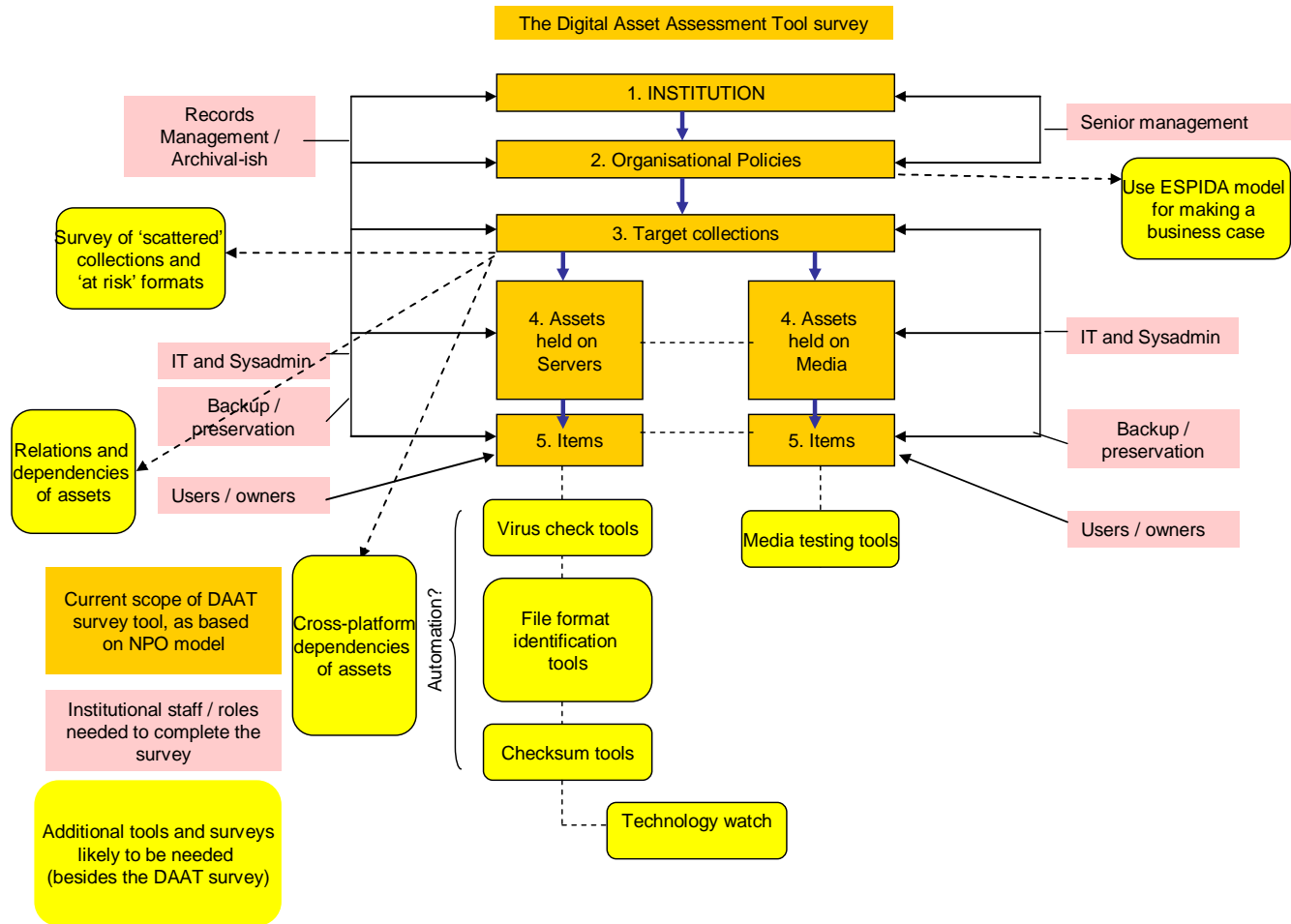
9.6 INCORPORATE COSTING ELEMENTS

Projects such as LIFE have helped us understand the cost implications of preservation actions more clearly, although much remains to be done in this area. Assessing preservation risks is only part of the challenge facing institutions with digital assets; they also need to understand the cost implications of taking action to mitigate those risks. Ideally, DAAT would incorporate knowledge of cost implications of preservation actions and build this knowledge into the recommendations it could make about prioritising preservation action.

9.7 PERFORM SENSITIVITY ANALYSIS OF WEIGHTINGS

The scores assigned in D-PAS, like those assigned in much of the original PAS system, are not as scientifically-informed as they might be. At the current state of knowledge we are unlikely to be able to make these reflect the true significance of each piece of knowledge, but we could improve the tool by understanding how sensitive it is to changes in parts of the system. A sensitivity analysis perturbs the scoring mechanism in different ways and looks for areas where small changes in the inputs result in big changes in the outputs. These suggest either the need to redesign the scoring system (if it can be shown that it is sensitive to changes in measurements that are close to the standard error of those measurements) or to pay particular attention to the accuracy of the weightings in those areas.

Gap analysis chart



References

Richard Anderson et al

The AIHT at Stanford University: Automated Preservation Assessment of Heterogeneous Digital Collections

D-Lib Magazine, Volume 11 Number 12

December 2005

ISSN 1082-9873

<http://www.dlib.org/dlib/december05/johnson/12johnson.html>

Andreas Stanescu

Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology

D-Lib Magazine, Volume 10 Number 11

November 2004

ISSN 1082-9873

<http://www.dlib.org/dlib/november04/stanescu/11stanescu.html>

Fred R. Byers

Care and Handling of CDs and DVDs: A Guide for Librarians and Archivists. (NIST Special Publication 500-252.)

WASHINGTON, COUNCIL ON LIBRARY AND INFORMATION RESOURCES (CLIR) /
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

October 2003

ISBN 1-932326-04-9

<http://www.clir.org/PUBS/reports/pub121/contents.html>

Adrian Brown

Care, Handling and Storage of Removable Media (Digital Preservation Guidance Note 3)
UK, THE NATIONAL ARCHIVES

27 June 2003

http://www.nationalarchives.gov.uk/preservation/advice/pdf/media_care.pdf

Anne R. Kenney and Ellie Buckley

Developing Digital Preservation Programs: the Cornell Survey of Institutional Readiness, 2003-2005

RLG DigiNews, Volume 9 Number 4

15 August 2005

ISSN 1093-5371

http://www.rlg.org/en/page.php?Page_ID=20744

Martin Donnelly

Digital Curation Centre Case Studies and Interviews: JSTOR/Harvard Object Validation Environment (JHOVE), Version 1.0

March 2006

ISSN 1749-8767

<http://www.dcc.ac.uk/resource/case-studies/jhove>

James Currall et al

Espida. Making it Happen by Getting Real.

Sustainable preservation of digital assets in a University.

UNIVERSITY OF GLASGOW

February 2005

And see <http://www.gla.ac.uk/espida>

Frances Halahan and Jennifer Dinsmore

The NPO Museum Preservation Assessment Survey

NPO e-Journal

May 2004

<http://www.bl.uk>

Online Computer Library Center

OCLC Digital Archive Preservation Policy and Supporting Documentation

OHIO, OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

20 January 2005

<http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf>

Alison Walker and Julia Foster

Preservation Assessment Survey for libraries and archives: user's guide

UK, NATIONAL PRESERVATION OFFICE

July 2001

<http://www.bl.uk/services/npo/paslib.html>

National Preservation Office

Preservation Assessment Survey: Museum Module, Guidance Manual

UK, NATIONAL PRESERVATION OFFICE

October 2004

<http://www.bl.uk/services/npo/paslib.html>

Alison Walker

Preservation Assessment Surveys: an Interdisciplinary Approach

THE NETHERLANDS, LIBER QUARTERLY VOLUME 13 No 3/4

2003

ISSN 1435-5205

<http://webdoc.gwdg.de/edoc/aw/liber/lq-3-03/273-280.pdf>

Gregory W. Lawrence, William R. Kehoe, Oya Y. Rieger and Anne R. Kenney

Risk Management of Digital Information

WASHINGTON, COUNCIL ON LIBRARY AND INFORMATION RESOURCES (CLIR)

June 2000

ISBN 1-887334-78-5

<http://www.clir.org/pubs/reports/pub93/contents.html>

Adrian Brown

Selecting Storage Media for Long-Term Preservation (Digital Preservation Guidance Note 2)

UK, THE NATIONAL ARCHIVES

19 June 2003

http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_storage_media.pdf

Appendices

The following appendices contain detailed information which may be of interest to those who wish to gain a greater understanding of the work which led to this report. They should be available from the same source as this report, and will be kept at <http://www.ulcc.ac.uk/daat> for at least 5 years following publication.

APPENDIX 1: Grouping exercise results. Spreadsheets showing the results of grouping the questions on the original PAS surveys, with comments. (MS Excel spreadsheet/PDF, 13pp)

APPENDIX 2: Draft questionnaires / spreadsheets with sources. This spreadsheet shows early pre-database drafts of the D-PAS questions, with a note of the sources that the questions were derived from. (MS Excel spreadsheet/PDF, 12pp)

APPENDIX 3: Scoring table. This is a table of the scores assigned to questions in the D-PAS tool, and descriptions of the ways scores could be used. (MS Word/PDF, 12pp)