A way forward for developments in the digital preservation functions of DSpace : options, issues and recommendations

25th July 2003

Paul Wheatley

1.0 Introduction

This report aims to inform the discussion of the MIT and Cambridge DSpace teams and invited digital preservation experts, meeting on the 30th July, 2003. A key aim of the report will be to present possible options for the future development of the digital preservation aspects of DSpace. The report will also provide questions for discussion on important issues, recommended reading for meeting participants and a suggested way forward for developing DSpace. The report will focus exclusively on the long term digital preservation functions of the DSpace system (termed Functional Preservation).

2.0 Recommended Reading

This section suggests a recommended reading list for review by the meeting participants.

Florida Centre for Library Automation (FCLA) [1]
http://www.fcla.edu/digitalArchive/daInfo.htm

The FCLA are currently developing their own digital repository system DAITSS (Dark Archive in the Sunshine State). It is intended that this will be made freely available as open source in a similar way to DSpace. FCLA are also investing considerable effort in developing action plans and related tools to address the preservation of different file formats. This in particular is of great relevance to developments in DSpace and is discussed elsewhere in this report.

"Digital Preservation Testbed White Paper: Migration - Context and Current Status" Testbed Digitale Bewaring [2]
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf

This Migration white paper provides well written analysis of some of the differing migration strategies. The categorization of the different strategies is far from perfect (and several new ones have appeared in since this report was written) but it does provide a good starting point for review of the available choices for DSpace to examine. The report highlights a crucial conclusion:

"File formats and preservation requirements differ so widely that it will not be possible to develop a 'one size fits all' approach. However, migration will almost certainly form part of a wider and more pragmatic strategy for long term preservation of digital objects and archival records."

New digital preservation strategies

A range of new digital preservation strategies that have been developed to address the inadequacies of a conventional migration approach should be considered. These include:

- **TOM : "The Typed Object Model (TOM)"**, Ockerbloom, J, [3] http://tom.library.upenn.edu/

- **UVC "A Project on Preservation of Digital Data"**, Lorie, R, [4] http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2, or in more detail http://www.kb.nl/kb/ict/dea/ltp/reports/4-uvc.pdf

- **Migration on Request**, CAMiLEON, [5] http://www.si.umich.edu/CAMILEON/reports/mor/index.html

Representation Information Systems

Current developments in systems to manage Representation Information are of relevance to the discussion and include:

- **Global Digital Format Registry (GDFR)**, http://hul.harvard.edu/formatregistry/

- **Representation Networks (CEDARS)**, http://www.leeds.ac.uk/cedars/guideto/cdap/

- **PRONOM**, http://www.pro.gov.uk/about/preservation/digital/pronom/default.htm

Note also that a significant element to the JISC Digital Curation Centre (likely to be established early in 2004) will involve the development and service delivery of a Representation Information system.

3.0 Questions for discussion

1. What timescale should be addressed when long term digital preservation is considered? Should strategies be addressing a truly long term problem?

2. Should the focus of digital preservation development facilitate a viable service delivery to the main depositors with the DSpace system, concentrate more on research and development in possibly more complicated areas, or both? Where does the priority lie and what are the overall aims of developing the digital preservation functions of DSpace?
3. Current thinking suggests external Representation Information (RI) systems (like the GDFR) will fulfil key parts of repositories' digital preservation functions. In the medium to long term will DSpace be a user of one or more of these systems? If so, should DSpace address the recording of RI or simply wait for external systems to be developed to perform this role?
4. Is there an advantage to developing specialist knowledge in some file formats, placing DSpace in a strong position to contribute to and influence RI systems as they are developed in the near future?
5. What current developments would be complementary/conducive to eventual use of an external RI system?
6. How can DSpace back up its commitment to the long term preservation of Supported formats?
7. What issues should influence the categorization of formats within the DSpace Supported/Unsupported/Known categories? Is the legal ownership of proprietary formats a crucial consideration?

4.0 Options for the further development of the digital preservation function in DSpace

4.1 Summary of options

- File Format Rendering
- Cost issues for digital preservation strategies
- Representation Information
- Ingest
- Related developments

4.2 File Format Rendering

A number of File Format Rendering and preservation strategies are considered.

- Migration on ingest
- Migration on request
- Viewers
- UVC
- Emulation
- Associated developments
- Preservation strategies and the view from Cambridge and MIT

4.21 Migration on ingest

Migration on Ingest is used to describe the policy of migrating generally proprietary file formats to open standard formats at the time of ingest to a repository.

Opinion is divided over the use of this approach. Those closer to service delivery of preservation in the most part favour migration on ingest to 'standard formats'. Those closer to a less constrained, research approach to preservation are quicker to point out the flaws of migrating on ingest and suggest newer strategies like UVC or Migration on Request. That observation in itself is very telling and suggests perhaps that the newer strategies are yet to be fully realised and tested and hence the more straightforward strategy of migration on ingest is more practical at the current time.

A range of concerns lie with Migration on Ingest, in particular the success of the strategy in the long term. While some are happy to rely on the longevity of current standards, many observers have pointed out that even open and non proprietary standards do not necessarily survive for long. Standards change and go in and out of favour, as history has shown. In the long term, will successive migrations from standard to standard result in sufficiently accurate preservation? Assuming formats can be migrated on ingest successfully, the strategy for preserving the resulting standard formats must also be considered. Can documentation be relied upon to enable the design of rendering tools at a later date? In the emulation field, reliance on documentation for later implementations as originally suggested by Rothenberg has since been widely dismissed as dangerous and impractical. In migration terms the situation is less clear, although there *is* concern about the accuracy and completeness of file format documentation [9].

Migration on Ingest offers a practical way forward to begin to tackle the digital preservation problem now but presents some risks in the long term. Identification of the priorities in addressing the short to medium term dangers of obsolescence versus the real long term dangers will inform the arguments for and against the use of this strategy.

4.22 Migration on Request

Migration on Request [5] aims to address both cost and longevity problems associated with Migration on Ingest strategies. Digital objects are preserved unchanged over time, along with a tool that is itself designed to be preserved or maintained over time that will migrate a digital object to a current usable format at the point of use. If it is impossible to influence the creation of the digital object in order to ease the preservation process, CAMiLEON argues that the focus of preservation should move to the preservation tool, which can of course be designed with longevity in mind.

CAMiLEON also argues that Migration on Request is much more economical in the medium to long term, as the preservation of several similar file formats can be addressed by one Migration on Request tool. CAMiLEON suggests that as long as the Migration on Request tool can be maintained/preserved over time, preservation

accuracy should be greater than in other migration strategies as there is only ever one step between the original and the object viewed by the user.

CAMiLEON and the subsequent Representation and Rendering Project at the University of Leeds have developed working implementations of the Migration on Request strategy but there has been limited testing and no uptake of the strategy elsewhere. Successful Migration on Request preservation relies on the quality of the initial implementation. A flawed initial implementation will make accurate and effective preservation over time very difficult. The strategy aims to reduce overall preservation costs in the medium to long term, but inevitably requires relatively more implementation effort in the initial development phase.

4.23 Viewers

Rather than migrating files as their formats become obsolete the alternative is simply to provide tools to render these different formats. The problems are then moved to ensuring the viewers themselves can be preserved over time, and the technical difficulty of rendering complicated file formats. The UK National Archives are currently trialling a commercial product, Quick View Plus from Stellent [10]. Initial trials are positive but there is no indication of how longevity will be addressed. The risks of depending on the survival of a commercial company and a non open source product are clear. The NA intends to publish the results of its evaluation when it is complete.

There are a great deal of open source viewers and rendering tools, developed primarily by enthusiasts and made available over the Web [9]. These tools could form the basis of developments for preservation rendering tools or act as documentation (in terms of source code) of file format structures.

4.24 UVC

The Universal Virtual Computer [4]] has been designed by IBM as an easily preservable platform on which emulation or migration type tools can be developed. UVC is seen by some as the only strategy which fully takes into account the real *long term* issues in digital preservation. IBM argues that digital preservation can only be addressed by a system expressly designed to be economically and accurately maintained over time. Others conversely argue that the length the UVC approach goes to in order to achieve longevity renders the strategy unwieldy and difficult to implement in practice. Recent work with the Digitale Bewaring Project [2] has yielded results which will hopefully be published soon. It should be noted that the UVC approach has not been sufficiently developed to allow emulation type preservation of interactive resources, and this is recognised by both IBM and external observers as a considerable challenge.

4.25 Emulation

IBM (see above), CAMiLEON, and others, have examined the use Emulation in digital preservation. There are strong arguments to indicate that emulation will play a role in digital preservation both as a primary rendering strategy and as a testing and evaluation tool for other strategies [12]. Key concerns in developing emulation strategies include:

- Achieving sufficient accuracy in the emulation of the targeted computing platform to enable high levels of compatibility and hence a cost effective return of the preservation of many objects or types of objects.
- Longevity of emulation implementations.
- Maintaining sufficient user metadata to facilitate the use of emulated software and systems that have become obsolete and hence unknown to future users.

In the context of DSpace it is recognised that there are minimal resources to explore large-scale emulation type strategies. MIT in particular has a requirement to preserve complex interactive objects and so a compromise development may be necessary. Suggestions for this hopefully minimal but useful work include:

- Emulation implementation, but only where (high quality) open source emulators are available as a solid starting point.
- Examination of requirements for emulation, in particular the range of platforms to be addressed and a cost/benefit analysis exploring possibly costly implementation versus strong return from the preservation of many types of objects using few emulators.
- Research into requirements and or development of user metadata (see third bullet above).

4.26 Associated developments

4.261 TOM

The Typed Object Model [3] is a broker system designed to facilitate chained migrations between different formats. TOM is not considered at this time as a complete preservation strategy and the issue of longevity has not been explored, but the TOM technology could possibly be used to enhance developments in migration tools or possibly even integrate migration tools from DSpace and other sources. TOM is currently receiving development funding from the Mellon Foundation and some work is being carried out in collaboration with DSpace.

4.262 FCLA Action Plans

As described above, the FCLA [1] has begun to develop action plans for the preservation of specific file formats. These plans provide useful information and strategies on a limited number of formats which could be exploited by DSpace. They also provide a simple model for addressing and monitoring the long-term issues associated with preservation strategies.

4.3 Cost issues for digital preservation strategies

There are three key cost elements to consider with regard to digital preservation strategies:

- Initial implementation
- Maintenance/continued implementation over time
- "Use", meaning delivery of the strategies following implementation (eg. performing batch migrations).

The distribution of cost between these elements appears to vary considerably between the different preservation strategies.

Identifying the distribution between these elements as well as the factors that influence the overall cost is difficult but possible at the current time. This analysis would be useful in informing the selection of preservation strategies and the planning of implementation. Given that the initial implementation of a preservation strategy is only the first step in a long term preservation process, being aware of the relative level of required spend over time is very useful in ensuring that sustainable strategies are chosen. CAMiLEON has conducted some relevant work in this area.

Identifying the actual costs of implementing preservation strategies with little previous practical work on which to base this analysis is virtually impossible as much of the equation depends on particular file format complexities (which are themselves difficult to identify without in depth investigation). This is arguably a contentious statement but is to some extent backed up by the lack of available cost figures despite a great demand from the preservation community to see them!

Analysis of cost elements may be useful, but should not be considered a priority if it will impact greatly on resources available for development and practical implementation of preservation strategies. Could this work be undertaken, at least in part, by the DSpace business consultants?

Whether or not effort is invested in research and analysis of preservation strategy costing, it is recommended that details of time/effort spent on implementations undertaken by DSpace are recorded and published.

4.4 Representation Information

4.41 Representation Information in context

There is convincing evidence to suggest that a comprehensive digital preservation strategy requires bitstream preservation, practical rendering solutions and sufficient recording of “technical metadata” (or in OAIS terminology Representation Information (RI)) [16].

DSpace must address all three of these concerns in order to fulfil its aim of providing real long term digital preservation. Bitstream preservation is already a staple function of DSpace and much of the rest of this report discusses options for progressing with the development of rendering solutions. The current thinking to address the recording of RI is less clear, but focuses on large scale systems to record, link, manage and describe RI.

4.42 Current and future developments

Section 2.0 lists the main RI initiatives. The GDFR is at this stage still speculative. The DCC is likely to go ahead early in 2004 and is tasked specifically with supporting repositories with the development of RI type systems, including functions for preservation watch, file format recording, etc. Cedars Representation Network technology is not currently being developed but the University of Leeds hopes to take this work forward in the near future. PRONOM is currently at version 2.0. Beginning as a simple file format database, the National Archives have embarked on a process of incremental development which shows considerable promise for the future.

4.43 RI and DSpace

It is likely that one or more RI initiatives will provide for DSpace's RI requirements at some point in the future but it is unclear when a sufficient level of support will be reached. Doing nothing and waiting for appropriate RI systems to appear seems a risky approach but at the same time, addressing this task head on is unrealistic given limited resources.

A proposed middle ground would aim to record some RI and plan preservation strategies, while looking ahead to ensure preparedness for the advent of RI systems as they are developed. This might include:

- Gathering file format documentation for specific file formats.
- Preparing file format "Background Reports" which demonstrate an awareness of long term preservation issues surrounding particular formats as well as "Action Plans" to prepare for preservation action that will need to be taken in the short to medium term [1]
- Integration of this work with developments in rendering tools.

4.44 Integration with RI systems

If DSpace is to utilise a service from an external RI systems, some forethought must be given to integration. In most cases this will simply be providing for the recording of unique identifier pointers from objects in the repository to the relevant RI in the external system. In use terms, how the user will be presented the RI is not yet clear. This may be addressed by the external RI systems or repositories may have to provide the functionality to navigate the RI and filter and present relevant information appropriate for the level of user (perhaps more technical detail for repository managers and predominantly rendering information for repository users). The current

DSpace system of categorizing file formats (MIME Types) is inadequate due to the high degree of variance in different versions of a specific format. Choosing a sufficient level of accuracy in identifying file formats will be crucial in enabling integration with RI systems.

4.5 Ingest

The development of facilities to automate the ingest process is particularly important. Little work has been conducted in this area in the field, despite it being recognised as crucial in achieving economical operation of digital repositories. Currently DSpace supports both an individual metadata entry form as well as an XML based "batch loader". Options for development in this area include:

- Automatic extraction of metadata from different file formats. Possibly including extraction/capture of technical metadata.
- Automatic identification of file formats (possibly involving integration with RI systems (see above). Both the UK National Archives and John Mark Ockerbloom (TOM) at the University of Pennsylvania are developing software in this area. In particular Ockerbloom's developments are highly advanced and there is likely to be the potential for collaboration. Categorisation of file formats must also be considered (see above).
- Verification of a digital object's compliance to a relevant file format specification. This could involve ensuring an archival subset is adhered to (eg. PDF/A) or even go as far as identifying external dependencies not included in the submitted object (eg. URLs, externally referenced fonts, etc). This is a key area for development identified by the British Library for their digital deposit system. Collaboration may be possible in this area.
- Development of existing DSpace ingest tools and/or addition of new ingest tools. The Cedars project demonstrated facilities for utilising or repeating entry of metadata based on previous submissions by the same user or example submissions provided by the repository. This will become more important as the range of recorded technical metadata is expanded.

4.6 Related developments

A number of related developments are in progress with which overlap with new work should be avoided. These include:

- SIMILE Project [13], SIMILE is exploring the development of interoperability between different library systems. "Simile will leverage and extend DSpace, enhancing its support for arbitrary schemas and metadata, primarily though the application of RDF and semantic web techniques." (extract from project web page). The project will also develop technology for recording archival provenance and change metadata.

- San Diego Super Computer Centre [14], The Storage Resource Broker project is experimenting with integration with the DSpace system.
- Web archiving – Both Cambridge and MIT are already involved in web archival/preservation work.

5.0 File formats to address

5.1 The problem

The vast number of different file formats and the rapid change, development and emergence of new formats presents a problem of scale for the digital preservation community. Developing preservation strategies for even a small percentage of these formats will require significant effort. Providing reasonable coverage of popular formats submitted to repositories will be difficult to achieve without collaboration and work sharing with the wider digital preservation community.

The resources available to develop preservation strategies on the DSpace project should therefore be allocated wisely. The project must be careful not over stretch its development efforts, but instead concentrate on developing a realistic number of solutions for rendering/preserving file formats.

5.2 File format issues in preservation strategy development

A range of factors affect the difficulty of implementing preservation strategy solutions for different file formats. The chosen strategy will have some impact on this calculation but the following list describes file format implications that impact on the development time of rendering/preservation solutions:

- Size/complexity of format (note graph of size of PDF format documentation in PDF Background Plan, FCLA) [1]
- Structural quality (eg. a well designed format is generally easier to preserve)
- Presence of externally used or referenced objects (eg. fonts)
- Presence of embedded objects of different formats
- Availability and accuracy of documentation [9]

5.3 Complexity of formats to inform research

The experiences of CAMiLEON suggested that the file formats chosen to test and inform development of preservation strategies must be selected very carefully. Formats should be of sufficient complexity to effectively test the theory but not so complex as to require unrealistic levels of development time (allowing a number of formats and strategies to be implemented). For the testing of its Migration on Request work, CAMiLEON chose three vector graphics formats of varying complexity and quality of design which met a middle ground of compromise between these different requirements.

Cliff Lynch suggested in 1999 [15] that canonicalized forms could play a role in digital preservation and presented some very interesting discussion of the possibilities in this area but the simple illustration of bitmap graphics did not test the theory in great detail.

5.4 Format selection

A number of additional factors should influence the choice of formats to address:

- As discussed above, many open source enthusiast developments are ripe for re-use or exploitation by the preservation community. This may well influence format choice in enabling more efficient use of development resources.
- Some developments have been made from within the preservation community although these are often not well publicized. Particular areas to avoid include PDF, email and web related formats. It is suggested that DSpace publicize in advance the formats it decides to investigate, both in an effort to foster interest and possible collaboration but also to avoid duplication with other work.
- Institutional requirements in terms of which formats are most in need of preservation solutions must play a role in format selection for two key reasons:
    - An aim of these developments is likely to be the fundamental support for a preservation service, so the support of at least some popular formats is likely to be required.
    - Creator/depositors of digital objects enthusiastic about the preservation of their work in a DSpace repository will be able to contribute valuable views on the aims of the preservation and the identification of significant properties.
- Legal issues of format ownership and use.

6.0 Suggested developments

The following list is composed of suggested developments for the digital preservation function of DSpace:

- Development of preservation action plans (see description of FCLA activities above) and Migration on ingest tools for around 3 file formats (depending on complexity).
- Development of a single Migration on Request tool to support a number of similar formats.
- Development of ingest processes. In particular suggested development of an automatic file identification module, possibly in collaboration with external developers (see above). In conjunction with this activity a new categorization for file formats, supporting file format versioning, should be designed. This must be undertaken with liaison with RI system development where possible (see above).
- Exploration of DSpace interoperation with TOM.

- Exploitation of externally developed tools.
- Analysis of the distribution of cost elements in preservation strategies.

Assumptions regarding related developments include:

- Developments of the SIMILE Project will extend DSpace support for optional metadata schemas for different subject areas.

7.0 Miscellaneous planning and practical recommendations

Miscellaneous recommendations:

- Prior planning of migration tool development is difficult as development time depends on both the complexity of targeted file formats and the quality of available documentation. It is suggested that formal planning be undertaken after a phase of research and evaluation of the targeted file formats.
- Where possible consideration should be given to the use of software longevity techniques [12] when designing migration tools.
- Migration tools should always generate operation logs, with particular detail given to recording migration losses or omissions. Losses can be acceptable, but only if recorded. This should be undertaken with awareness of the developments of the SIMILE project in developing provenance metadata.
- Preservation using multiple strategies provides redundancy and greater security in ensuring the maintenance of rendering over time. Preservation plans which incorporate multiple rendering strategies (possibly where externally developed solutions are available to complement renderers developed in house) should be pursued where possible.
- Open source viewers, emulators and file conversion utilities could provide the basis for some preservation tool developments and should not be ignored.

8.0 References

[1] Florida Centre for Library Automation (FCLA)
http://www.fcla.edu/digitalArchive/daInfo.htm

[2] Testbed Digitale Bewaring, "Digital Preservation Testbed White Paper: Migration - Context and Current Status" (December 2001),
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf

[3] Ockerbloom, J, "The Typed Object Model (TOM)", http://tom.library.upenn.edu/

[4] Lorie, R, "A Project on Preservation of Digital Data", RLG Digi News, 5,3
http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2

[5] Mellor, P, Wheatley, P, Sergeant, D. “Migration on Request : A practical technique for digital preservation" ECDL 2002, http://www.si.umich.edu/CAMILEON/reports/reports.html

[6] Global Digital Format Registry, http://hul.harvard.edu/formatregistry/

[7] Representation Networks (CEDARS), http://www.leeds.ac.uk/cedars/guideto/cdap/

[8] PRONOM, National Archives (UK), http://www.pro.gov.uk/about/preservation/digital/pronom/default.htm

[9] Wheatley, P, R, “Survey and assessment of sources of information on file formats and software documentation”, http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf

[10] Stellent, http://www.stellent.com/

[11] Gladney H, “Digital Document Quarterly”, http://home.pacbell.net/hgladney/ddq_2_2.htm

[12] Holdsworth, D, Wheatley, P. “Emulation, Preservation and Abstraction" RLG DigiNews, 5, 4

[13] SIMILE Project, http://web.mit.edu/simile/www/,

[14] San Diego Super Computer Centre, The Storage Resource Broker Project, http://www.npaci.edu/DICE/SRB/,

[15] Lynch, C, “Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information”, D-Lib, 5, 9, http://www.dlib.org/dlib/september99/09lynch.html

[16] Holdsworth, D, Sergeant, D, "A blueprint for Representation Information in the OAIS model" http://www.personal.leeds.ac.uk/~ecldh/cedars/nasa2000/nasa2000.html