Emulation Options for Digital Preservation: Technology emulation as a method for long-term access and preservation of digital resources

1.0 Project Summary

A team of researchers at the University of Michigan and research staff in the UK from the Cedars project, being run at the Universities of Leeds, Oxford and Cambridge, under the aegis of CURL (Consortium of University Research Libraries), seek funding from NSF and JISC, respectively, for an international digital library intiative to investigate the potential role of emulation in long-term preservation of digital resources. The project will develop a small suite of emulation tools, evaluate the costs and benefits of emulation as a preservation strategy for complex multi-media documents and objects, and develop models for collection management decisions that would assist people in making 'real life' decisions about how much effort and resources to invest in exact replication within preservation activity (as opposed to preserving 'raw' intellectual content). We will develop preliminary guidelines for the use of different strategies (conversion, migration and emulation) for managing and preserving digital collections.

The project objectives are:

- To explore the options for long-term retention of the original functionality and "look and feel" of digital objects including technology emulation;
- To investigate technology emulation as a strategy for long-term preservation and access to digital resources;
- To consider how and where emulation fits into a suite of digital preservation strategies.

The project will build on the research strengths of the University of Michigan in digital libraries, user evaluation, and long-term preservation, and the practical experience of the Cedars project team in the UK <*http://www.leeds.ac.uk/cedars/*>. If the bid is successful, it is expected that the work already underway in digital preservation through Cedars will act as a springboard for this project, thereby enabling new work to begin relatively quickly.

2.0 Research Issues

The rapid obsolescence of computer hardware and software and the accelerating rate of change in systems and software pose the greatest threat to the longevity of digital information.[1] The lack of technologically feasible and affordable methods for preserving digital information is a major impediment in building digital libraries that will endure. Archivists and librarians have implemented preservation strategies that work effectively for certain types of digital information, but these methods are not necessarily a viable option for preserving increasingly complex digital objects. The most widely used methods for preserving digital information involve a combination of adopting standards

that limit the variety of digital formats that a digital repository accepts, converting digital materials to a standard format when they are accessioned into a digital repository, and migrating the digital information from obsolete to current formats so that the information can be accessed using current hardware and software.[2]

These methods have been deployed effectively for certain types of digital information. Some specialized data archives have successfully migrated numeric and statistical data across several generations of hardware and software.[3] SGML-encoding is a widely adopted standard for the management of electronic text files that facilitates long-term maintenance of the digital files. TIFF is a commonly used format for page image files.

There are also significant problems with these strategies. First, the creators and producers of digital information may be unable or unwilling to create documents, databases or other digital objects in formats that conform to the standards acceptable to digital repositories. Second, conversion to a standard acceptable to the repository, when it is technically feasible, often requires detailed analysis and large investments of resources to write conversion programs. If the original format is unusual or unique, the conversion may be a one-off process with no potential for reusing the conversion program. Third, conversion from the original format to a format that is manageable for long-term preservation may entail significant loss of information or functionality, or both. Fourth, some older digital materials in obsolete formats cannot be rescued without emulating their original software environment. Finally, for certain classes of digital objects, no acceptable long-term storage formats exist, and, therefore conversion is not a viable option.

In today's digital environment, electronic resources are increasingly being made available in multi-media formats, with the intellectual content being bound to the structure, form and behaviour of the digital medium in which it has been produced/published. This presents particular challenges for long-term preservation, not only in terms of the technical solutions required to preserve data where the "look and feel" may be an integral part of understanding the intellectual content of the digital object, but also in terms of collection management decisions and cost-benefit implications of preservation decisions.

Some computer scientists have proposed emulation as an alternative strategy for long-term preservation.[4,5] According to this approach, emulation of obsolete systems on future unknown systems would make it possible to retrieve, display and use digital documents with their original software. One purported advantage of emulation as a preservation strategy is that this approach retains not only the intellectual content of digital information, but also the "look and feel" and functionality of the original. In theory, emulation may also be more cost-effective than conversion or migration because once an emulator is written to access any digital document of a particular class, that same emulator can be reused to access all other documents in the same class.

The emulation strategy has met with skepticism for a variety of reasons. Emulation has been proposed as a method that, in theory, could enable retrieval, display and reuse of digital information in its original software environment, but emulation has not been

applied to archival documents or objects. Some computer scientists consider emulation far too complex even in the context of known systems to warrant serious development and testing of this strategy. Another challenge in assessing the feasibility of emulation as a preservation strategy is that there is no conceptual understanding of which aspects of original computing environment need to be retained or replicated in future systems to ensure that digital documents and objects remain understandable and meaningful.

One recent proposal advocating emulation as a digital preservation strategy proposes a set of "core digital attributes" that must be retained. These include the ability to copy digital documents perfectly, to permit access without geographic constraint, and to disseminate documents at virtually no incremental cost. According to this proposal, archived digital documents must remain "machine-readable" so that they can be accessed, searched, and processed by automated mechanisms that can modify them, reformat them, and perform computations on their contents.[5] Emulation may be unnecessarily complex and unnecessarily costly if we assume that perfect and complete replication of all of the functionality, look, and feel of all digital objects is critical in all cases for future retrieval, display, and reuse.

We lack an understanding of which features are critical for understanding and reusing which types of digital objects by which user communities. Converting digital objects from their original or native form to current formats so that they can be accessed, displayed and manipulated using available hardware and software may sacrifice features of the original document that are necessary for interpretation and reuse. Features such as audio fidelity, color replication, and interactivity are difficult to replicate when objects are converted from one software environment to another. It is important to note here, however, that reformatting has been used as a preservation strategy for books, paper documents, and some drawing and photographs even though some of the look and feel of the original documents is lost. Microfilm, for example, does not capture textures or color, nor does it permit "paging" through documents, but the trade-offs between some of the original look and feel and the advantages of stability have been considered acceptable. We are proposing research that will help us define similar levels of surrogacy for more complex digital objects and to determine where emulation of the complete native software environment is critical.

We also lack a methodology for evaluating the technical feasibility and cost effectiveness of emulation as a preservation strategy. We propose research that will address this problem by evaluating the utility of currently available emulation tools for preservation, carefully documenting the costs of modifying existing emulation programs or developing new ones, and comparing the costs of emulation to the costs of alternative strategies, such as conversion and migration.

The key research issues we will address are:

1. What functionality and which aspects of the "look and feel" of original digital objects must be preserved to retain the meaning and utility of archived digital information?

2. Is emulation a feasible approach for retaining the necessary functionality, look and feel of archived digital objects? Is it cost effective?

3. Where does emulation fit into a suite of digital preservation strategies?

3.0 The Project

3.1 Emulation

The proposal of this research is to conduct an initial feasibility study of emulation using a small suite of emulation tools and testing their effectiveness for retaining the essential attributes of a few types of digital objects. We propose to take a small number of "intertwingled" digital resources, (i.e. where form and intellectual content are bound together) and for each one adopt a number of different emulation/preservation techniques that will provide a greater or lesser degree of replicating the "look and feel" of the original object.

The practical work will focus on current "intertwingled" resources such as multi-media CD-ROMS and other complex digital objects, as well as older digital material which may already be threatened by obsolescence, for example educational technology developed in the 1980s for specific technical environments. Although current material perhaps presents the most complex challenges over the long-term, early digital objects may help us prove that post-hoc rescue of such material is feasible using emulation techniques.

It is proposed to include one example each of various different types of digital objects. The provisional list of these resources is as follows[*]:

- Canterbury Tales multi-media CD-ROM (published by Cambridge University Press)
- SCRAN (UK project - Resource base of Scottish material culture and human history: http://www.scran.ac.uk)
- an interactive game or software (e.g. Simcity)
- architectural drawings created using CAD/CAM software
- Computer Assisted Learning (CAL) packages
- BBC material created for educational purposes in the 1980s (e.g. the BBC Domesday Project on video disk)
- Apple 2 applications

The project plans to engineer emulation tools. However, the technical team will also consider the use of emulation tools developed elsewhere and available in the public domain and will exploit existing, relevant work where possible. There is a great deal of emulation experience and resource available free over the Internet.[6] One aspect of the technical work will include an investigation of these tools to evaluate their suitability for

[*] We are aware that a proposal may be being submitted by the Museum of Modern Art and the University of Leeds, which may complement the work proposed in this project. We would be happy to collaborate with them in their work, if appropriate.

long-term preservation work. Although many may prove initially unfit for long-term preservation purposes, where possible these tools will be modified and re-engineered by the technical team to enhance their longevity and applicability for digital archiving. Where re-engineering of existing emulation tools proves to be impractical the project will develop new generic tools.

3.2 User evaluation

A key component of the project will focus on user evaluation of the range of outputs that result from the different technical and emulation solutions that have been adopted, and the impact that this may have on academic research activity. We will design a user evaluation where users are presented with original objects in their native environment, if possible, the original objects running under emulation, and various surrogates for the original objects that have been converted from their original format to formats that are easier to manage from a preservation point of view. We will ask users to evaluate the utility of digital objects in these different formats for specific sets of tasks. This evaluation will allow us to identify attributes of digital documents that are critical to replicate and permit us to define surrogates for original digital objects that are acceptable for long-term preservation.

3.3 Cost Benefit Analysis

We also propose to carry out a cost benefit analysis of the different technical approaches adopted, in order to assist people in making 'real life' decisions over how much effort/resources to invest in exact replication of the original "look and feel" when preserving digital resources (as opposed to just preserving 'raw' intellectual content).

The cost-benefit analysis will consist of several components. First, the technical team will evaluate readily available emulation tools, determine how they must be modified to meet preservation requirements, and re-engineer the tools accordingly. The technical team will also develop new emulation tools as needed. The costs of re-engineering existing emulation programs and developing new ones will be carefully documented. We will then compare the costs of emulation to the costs of alternative strategies, such as conversion and migration, taking into consideration the fact that expenditures for conversion, migration, and emulator development will recur at different intervals of time. We will also take into account long-term maintenance and storage costs of the different approaches. Finally, we will assess whether the benefits that emulation promises, by preserving more of the original attributes and behaviour of digital objects, warrant additional costs if emulation turns out to be a more costly approach.

3.4 Collection Management Issues

The project will address concrete and practical collection management issues relating to the necessary level of emulation and the appropriate techniques to be adopted. This aspect of the project's work, which will be carried out jointly by the University of Michigan and the Cedars team, will include:

- Assessment of user needs in relation to archived digital objects
- Identification of the relative significance of the functionality and "look and feel" of original digital objects to the overall intellectual content
- Guidelines for collection managers to help them to assess relative priorities in order to arrive at appropriate decisions over the use of emulation techniques for preservation
- Models for integrating emulation into programmes for preservation of digital library resources.

This project will evaluate the feasibility of emulation as a preservation strategy. While its primary focus is on the technical feasibility, cost effectiveness, and user acceptance of emulation, we are also concerned with the practical utility of emulation from the perspective of digital collection managers. Therefore, we will assess the technical and staffing requirements that are necessary for collections managers to use emulation for digital preservation. If emulation is demonstrated to be feasible and cost-effective, we will make a variety of emulation tools available to collections managers for use in rescue of digital materials in obsolete formats and as components of comprehensive digital preservation programmes.

4.0 Intended Products

4.1 Preliminary definitions of the attributes of different types of digital objects (e.g. multi-media products, images with textual description, simulations, vector graphics) that must be preserved to satisfy user needs and requirements.

4.2 Cost comparisons of different digital preservation strategies (conversion, migration, emulation)

4.3 Preliminary guidelines for the use of conversion, migration, and emulation in managing and preserving digital collections.

5.0 Proposed Benefits

5.1 The research will produce a set of requirements for digital preservation based on user assessment of different types of digital objects. Librarians, archivists, and digital collections managers will benefit from these requirements because they will be able to make decisions about digital preservation strategies based on assessments of users' needs rather than abstract, *a priori* notions of the essential attributes of archived digital objects[*].

5.2 The research will produce a set of emulation tools that we will make available for use and further testing in digital libraries.

[*] We recognise that current users of digital information are not perfect proxies for future users. Nevertheless, asking current users to test both older materials and contemporary materials is a useful starting point given the absence of research on user requirements for archived digital information.

5.3 If successful, emulation will provide a strategy for preserving types of digital objects that we do not know how to preserve (e.g. simulations, CAD/CAM drawings).

5.4 The project will produce recommendations for more cost-effective approaches to digital preservation.

5.5 The project will train two graduate students at the University of Michigan in user evaluation, assessment of emulation, and digital preservation strategies.

6.0 Project Infrastructure and Management

6.1 Project Management and Division of Responsibility

The project will be led by a team at the University of Michigan who will be responsible for overall project management and support as well as the user evaluation, cost benefit analysis, and exploration of collection management issues. Their work will be based on the tools developed and resources preserved by the technical team at the University of Leeds who will be responsible for the main technical component of the project. The researchers at the University of Michigan will test and evaluate user responses to the emulation tools and various surrogates of the original documents developed at Leeds and evaluate the cost-effectiveness of emulation as an alternative preservation methodology for digital information. The Michigan team will also identify some digital resources for testing. The teams will collaborate on the development of guidance for libraries on emulation as it relates to collection management policies.

The project will build on work being undertaken in the UK through the Cedars project,[7] and on research underway at the University of Michigan on digital archiving and user evaluation. In so doing, it will take advantage of the project infrastructure that is already in place in Cedars and on the considerable research experience in the field of digital preservation that resides at the University of Michigan. The proposed project will exploit the technical expertise that has already developed within the Cedars project team, but will undertake a much more in-depth investigation of the issues surrounding the emulation and preservation of "intertwingled" data than will be possible within the current Cedars project.

The project team will make extensive use of e-mail, video conferencing and other network communications to coordinate their work and share results. Both the University of Michigan and the University of Leeds have well developed video, audio, and telecommunications facilities capable of supporting the distance communication requirements for this project.

6.2 Staffing Requirements

UK Team

The UK team will be led on a day to day basis by Kelly Russell, the Cedars project manager, under the overall direction of Clare Jenkins as Cedars project director. Both have considerable experience of library research projects. Over the past twelve months their involvement with the Cedars project has given them a considerable insight into issues surrounding digital preservation. Additional staff will be recruited to work on the project. A full time technical officer will be hired to investigate existing emulation tools, and to carry out the necessary software development work. This person will be fully integrated into the Cedars technical team at Leeds, and will work closely with Dr. Dave Holdsworth who is an expert in digital archives with a particular interest in emulation technology.

A further half-time appointment will also be made at Leeds to work on the collection management issues that the project will be exploring. This work will be done in close collaboration with the US team.

US Team

The US team will be led by Margaret Hedstrom, project director and an expert in digital archiving. Two graduate student research assistants will be hired to conduct the user evaluation, assist with the technical evaluation of emulation tools, and carry out the cost benefit analysis. Judith S. Olson will provide guidance in the design and execution of the user evaluation. Nathaniel Borenstein and Josph Hardin will assist with the technical assessment of emulation tools.

7.0 Timetable and Preliminary Project Plan

7.1 Timetable

It is expected that this project will run for 3 years and will run through 3 main phases. The following is a preliminary project plan.

Phase 1 - Planning , analysis and preliminary studies. Months 1-6

This phase focuses on project planning; recruitment of project staff; and the establishment of an appropriate infrastructure for project management.

The key tasks for this phase include:

- appoint project staff (both teams)
- produce detailed project plan (both teams)
- agree on a final list of resources to be included in testing (both teams)
- review of related projects (both teams)

- establish project infrastructure (eg. Advisory Board/Management Group)
- establish project co-ordination processes with partner sites
- confirm roles of partner sites (both teams)
- establish dissemination and feedback mechanisms (both teams)

Phase 1: Deliverables
- detailed project plan
- report on related projects
- including review of existing emulation tools available in the public domain
- report on relevant standards and solutions

Phase 2: Emulation work begins on Selected Resources and User Evaluation Underway Months 6-18

- rights negotiation for use of material in the project -some will have been cleared through the Cedars project already (UK team); (Michigan materials are in the public domain)
- user evaluation preparation (US team)
- evaluation of public domain emulators for long-term preservation purposes begins (UK team)
- development of emulation tools (including where possible re-engineering of existing public domain systems) (UK team)
- preserved digital surrogate (plus originals), emulation tools and documentation sent to Michigan for evaluation (UK team)
- cost benefit analysis begins (US team)

Phase 2: Deliverables
- emulation tools (UK team)
- project working papers (both teams)
- annual reports for end year 1 (both teams)

Phase 3: Testing, consolidation and summation, Months 18-36

The focus of this phase will be on further testing of the emulation tools developed; consolidation of achievements to date; and investigation into collections management issues relating to the use of emulation as a digital preservation strategy. Key tasks in this phase include:

- cost benefit analysis of each emulation approach (US team)
- user evaluation analysis (US team)
- guidance for collection managers (both teams)
- international seminar on technology emulation as a strategy for digital preservation (both teams)

Phase 3: Deliverables

- project working papers (both teams)
- cost benefit analysis (US team)
- methodological guidelines for emulation techniques (UK team)
- annual report for year 2 (both teams)
- final report (both teams)

7.2 Evaluation and Review Mechanisms

- Advisory Board

  Its role will be to take a strategic view, to contribute to the formative evaluation process and to provide general guidance and advice. Membership will be broadly based and international in composition, to reflect this role.

- Management Group

  The management group is comprised of the principal researchers of the UK and US teams. The management group will review project plans, monitor progress on the project, and approve periodic and final reports on the project.

7.3 Formal reporting and review procedures

- Annual reports to JISC and NSF (both teams)

Meetings of the Advisory Board will include one meeting a year at which the project team will present an annual report, for formal consideration by the Board. This review process will provide an opportunity for discussion of project progress, the lessons learnt, and any technical and strategic implications for the project arising from these lessons. This annual meeting will contribute to the formative evaluation process, and will encompass both the technical and the organisational aspects of the project.

- Management group meetings

Management group meetings will be held 3 or 4 times a year, alternating between the UK and the US. The purpose of these meetings will be to monitor, at a more detailed level, the project's progress and to advise on specific technical and strategic issues.

- Technical evaluation

It is proposed to exploit the infrastructure and arrangements already in place for the evaluation of the Cedars project, and extend it to the technical work that will be carried out in this project. This will enable us to carry out evaluation at additional marginal cost. The Cedars evaluation is being carried out by an external evaluator. He will be assessing the technical aspects and outcomes of the work at various stages in the project. A full

evaluation strategy will be developed by the UK project team at an early stage in the project.

7.4 Anticipated travel

We estimate the need for three or four trips per year by the principal researchers, alternating between the US and the UK. We plan to make extensive use of telecommunications to minimise the need for travel. We will also attempt to link travel for project management purposes with the need for other meetings, such as the Cedars International Conference on Digital Preservation in 2000

8.0 Dissemination

The project will ensure that the findings of its work are regularly and widely disseminated. The project will, where possible, makes use of existing channels of dissemination such as newsletters and journals. Regular papers will also be given at conferences and seminars. Dissemination for this work will also build on the dissemination strategy employed by the Cedars project and through the research network available to the team at Michigan. Other dissemination activities include:

- synopsis of project progress e-mailed to interest groups
- working papers published on the Web (using the Cedars project Web pages)
- articles in relevant web-based and print publications such as RLG's DigiNews, Dlib Magazine, etc.
- International seminar on emulation as a preservation strategy
- Involvement in the Cedars International Conference on Digital Preservation in 2000
- Articles in scholarly journals
- Presentations at conferences
- Final report

9.0 Budget

We are requesting funding from NSF for support of the US team. This includes stipends, a portion of the tuition support and benefits for two graduate students for three years; one month of summer support for each of the senior personnel; and travel for the senior personnel. The University of Michigan will contribute 76% of the tuition for the graduate students as a cost share.

We are requesting funding from JISC for support of the UK team. The funding requested will cover salary costs of additional staff recruited to work on the project, travel and subsistence for key personnel, and costs of dissemination, evaluation and modest expenditure on desktop equipment. It is estimated that the UK contribution to the project will require 5.5 person years. Of this 4.5 person years will be funded by the project; the remainder will be an institutional contribution from staff already working on the Cedars project.

References

1. Preserving Digital Information: Report of the Task Force on Archiving of Digital Information. Commissioned by the Commission on Preservation and Access and the Research Libraries Group. Washington, D.C. (May 1996).

2. Hedstrom, M. and Montgomery, S., (Dec. 1998) Digital Preservation Needs and Requirements in RLG Member Institutions: a study commissioned by the Research Libraries Group, Mountain View, CA: RLG, Available: www.rlg.org

3. Soete, G., Preserving Digital Information, Washington, D.C.: Association of Research Libraries, 1997: 22-23.

4. Rothenberg, J. "Ensuring the Longevity of Digital Documents," Scientific American 272:1 (Jan. 1995): 42-47.

5. Rothenberg, J. Avoiding Technological Quicksand: Finding a Viable Technical Foundaiton for Digital Preservation. Draft revision 1:981012. Washington, D.C. Council on Library and Information Resources (forthcoming 1999). To be Available: www.clir.org.

6. For example Apple 2 emulators (and various links) can be found at http://www.emulation.net/apple2/index.html and emulators for the BBC material are available at http://emulation.net/bbcmicro/index.html. Another Internet site with numerous emulation tools is the virtual computer museum: http://www.icom.org/vlmp/computing.html#simulators

7. http://www/curl.ac.uk/projects.shtml