The progress of digitization technology, particularly multimedia, within some British cultural institutions

Lesley Carman-Brown,
Public Programs Coordinator
John Curtin Prime Ministerial Library,
Curtin University of Technology
carmanbl@boris.curtin.edu.au

Abstract:
This paper, based on a research visit to the United Kingdom in 1999, discusses the electronic progress being achieved by some major British cultural institutions, particularly in comparison to digitization progress which has been made by Australia's first prime ministerial library. It examines digitization of difficult collection areas such as multimedia and the desirability of upscaling digital projects.

Introduction

With the support of a VALA Travel Scholarship in 1999 I undertook research into the electronic progress being achieved by some of the United Kingdom's most notable cultural institutions. My aim was to assess new Information Technology developments being used to provide effective ways of digitizing, searching and retrieving the diverse range of materials held in cultural institutions such as libraries, archives and museums. While advancements have been made in Australia in digitizing photographic collections, little progress has been made to digitize such difficult collection areas as manuscripts, three dimensional objects, video and audio material (multimedia) so that it can be fully searched and retrieved by users to ensure easier access to our cultural heritage for a greater range of people, regardless of their physical location.

The objectives of the research into overseas cultural institutions, undertaken for the scholarship, were:

- To investigate approaches being made by such institutions to the digitization of their collections, for example of documents and multimedia.
- To examine developing search and retrieval technology and the parameters of its use.
- To visit beta sites to test technology being developed in the areas of sound and moving picture analysis and retrieval.
- To consult with users of the technology and establish networks with colleagues who are undertaking similar projects.

The idea for this research stems from the experience and technological expertise gained by the John Curtin Prime Ministerial Library (JCPML) staff through its development of an electronic research archive to provide digital access to the JCPML collection and research collections from other institutions which relate to John Curtin. The John Curtin Prime Ministerial Library Electronic Research Archive (ERA) is essentially an electronic repository of digitized documents amounting to approximately 13,000 items. Access to ERA is provided through the Internet to both textual and image documents. The technology used – Excalibur Retrievalware – allows a number of options when digitizing material, such as producing an image, image and text or text only.

Background

The development of ERA came from the success we had with several digital projects we developed at the JCPML between 1995 and 1998 and the lessons we learnt from undertaking these projects. Armed with some ideas of what we wanted our electronic research archive to be capable of delivering to users, the JCPML developed three digital projects to test the current technology's ability to provide the services we envisaged.

Project 1: John Curtin: A prime minister and his people
This was a joint venture with the National Archives of Australia (NAA), initiated to test the feasibility of remote-site scanning. A selection of letters to and from the prime minister's office between 1941-45 were selected from the NAA files. None of the material ever physically left the NAA offices in Canberra. The database is composed of nearly 600

images comprising approximately 500 documents providing high-quality digital facsimiles of the original material. The project was launched on the internet in July 1997.

Project 2: John Curtin Memorial Lectures
We undertook this project to test how well the current technology could deal with optical character recognition of older material since many records in the JCPML's collection are pre-1945 typewritten documents. This project was launched on the internet in February 1998.

Project 3: John Curtin: Australia's wartime prime minister
This interactive CD ROM was developed to bring together a range of media from the JCPML collection and test how we could best enhance access to such items as photographs, oral history recordings, textual documents and video recordings. This project has been available on site at the JCPML exhibition since February 1998.

We demonstrated through these projects that we could successfully digitize, and thereby improve access to, archival materials. Our principal purpose for establishing the JCPML Electronic Research Archive is to enhance access to records and create an electronic gateway giving people access to John Curtin-related material held in collections around the world. One of the critical steps involved in developing ERA was the production of the *Electronic Research Archive Management Framework* which established the principles and best practices to be applied to ERA and covered such areas as:

- cooperation between institutions;
- criteria for selection of material for digitization;
- future migration paths; and
- budget considerations.

The JCPML developed ERA to:

- Build upon the expertise and confidence which was developed during the pilot projects by moving into a phase which is fundamental to our vision. This expertise has been retained in-house and continues to be developed within the JCPML and has been a significant factor in our progress.
- Create a research archive that is easy to access and which is not dependent on a physical presence in the reading room.
- Bring together archival materials that are located not only in the JCPML collection but in other institutions and private hands in Australia and around the world.
- Provide integrity in the context of digitized archival records through a variety of control mechanisms relating to intellectual control, administrative and technical metadata.

ERA was initially launched within the JCPML Challenge Bank Reading Room in February 1999; then, following migration to new software, ERA was launched on the internet in July that year. So within four years of the concept of an electronic archive being born in 1995, the idea had become reality. Currently, ERA is being accessed by number of users monthly and this number is expected to grow as more people discover us on the web. Our idea was always to progress beyond the "digital project stage" and offer fully integrated electronic access to our collection. It is due to the foresight of our director, Ms Vicki Williamson, that

our vision for an electronic archive was founded at the same time that our physical collection was being established; and it is due to the professional diligence of our Archivist, Ms Kandy-Jane Henderson, that we have been able to turn our vision into reality so effectively and so quickly, making the retroactive digitization of our collection relatively painless compared with other institutions which have not yet come to terms with what they want to achieve from digitization.

Knowing the progress that we had already made, I wanted to examine the situation overseas and determine what advancements were being promoted that might help us progress to further stages. I visited a number of institutions and in this paper would like to examine four major organisations and two multimedia projects that I learnt about from my discussions.

The British Library

In 1993 the British Library established a digitization programme called Initiatives for Access (IfA) which aimed to develop software applications based on emerging digital technologies. The British Library has been involved in a number of digital projects and at the time of my visit in May 1999 staff were finalising their digital guidelines.

The British Library appears to be approaching the issues of digitization from the perspective, according to Dr Andrew Prescott from the Department of Manuscripts, that digital libraries should not simply be seen to replicate their 19th century counterparts by amassing "large amounts of information in a series of digital stores", but instead should be "thinking of a series of discrete small-scale projects which embody different approaches to information storage and manipulation, but which are linked together to form a wider resource." (Carpenter, Shaw, Prescott 1998 p.21) This appears to be the decision the library has reached after conducting a range of digital projects under its Initiatives for Access umbrella.

Perhaps it is not surprising they have reached this decision given the library's historical perspective. The British Library is a collection of services which were once separate and were only brought together in the early 70s. Historically, some sections offered free services while others tried to be commercially viable. This conflicting philosophy has not been resolved so there are difficult decisions still to be made. As well, the British Library doesn't see its main function as serving the general public. Building and maintaining its collections and servicing other institutions are its main priorities.

Given that historical background, it is easier to understand how Dr Prescott and his colleagues can conclude that "the digital library cannot be reduced to a single technical infrastructure." The cost involved in any kind of retroconversion of documents is also a major stumbling block. The strategy used by the British Library in its approach to the problem of funding has been to develop partnerships and participate in collaborative ventures. In particular it is hoped to involve the private sector more to help provide IT solutions and technical expertise.

One of its collaborative ventures is the Beowulf CD ROM which has been developed in conjunction with the University of Kentucky, a grant from the Andrew W Mellon Foundation and funding from the National Endowment for the Humanities. The CD does not aim to simply make an electronic facsimile of the Beowulf manuscript available, but to promote understanding of the context by providing scans of the rest of the manuscript in which Beowulf is contained, known as the Nowell Codex.

What is particularly interesting about this project is that the technology used included special lighting techniques of optic fibres and ultra violet light to help reveal letters hidden by nineteenth century attempts to preserve the manuscripts – letters which had previously been blotted out – and also enhance some of the poorer quality print. This is a good example of how the technology can be more valuable to the collection than just making a digital facsimile.

But they also see the CD option as offering the best way to recoup some of the costs involved in producing digital material.

The British Library currently faces a crossroads in deciding whether or not to make its collections fully available online. At the same time, however, staff feel that the number of digital projects undertaken has brought together a significant store of digital material. A major problem with this approach is that if the British Library is going to undertake projects it still needs to develop a consistent selection policy as to which items users are given digital access to and whether or not that access will be online or inhouse.

The British Museum

The British Museum faces similar problems in regard to the scale of conversion that would be necessary to make its collection available online. The museum has elected to take the same course as the British Library in that it has no immediate plans to make the whole collection available on line. During 1999 the museum developed its own digital access software called COMPASS: Collections Multimedia Public Access System. Testing of the prototype was finalised in 1999 and the system is currently being installed to be available across the internet early in 2000. Staff have gone against the conventional wisdom of selecting a proprietary system and adapting it to their needs and instead chose to develop their own IT system. The reasons they cited for this decision were:

a) the difficulty in developing a system for users who may have no concept of the scope of the museum's collection; and
b) the difficulty in making the system accessible to the 40% of users who are non-English speaking.

COMPASS will provide images of selected objects together with contextual information, at a much greater depth than is portrayed in the museum's galleries. It will use multimedia to provide background and contextual information about objects and link this information, and therefore the objects, thematically to help understand the chronological and cross-cultural relationships between the objects.

For example, many objects in the collection, particularly old artifacts, are damaged or so far beyond our current understanding that users have difficulty knowing exactly what the objects are or how they were used. COMPASS uses multimedia techniques of 3D imaging and computer modelling to recreate damaged objects into their original shapes or show exactly how the objects were used during their lifetime. Hence, what the British Museum is trying to achieve through COMPASS is not simply greater access through the internet but greater understanding of the collection. However, this tack is extremely time-consuming and it is envisaged that COMPASS will provide access to approximately 5000 objects from the museum's computerised inventory of more than 1,000,000 items from the collection.

The criteria developed by the museum for selection of objects include:

- Objects which are famous
- Star objects as determined by their curators
- Objects which are on permanent display within the museum
- Objects for which the use of multimedia is particularly suited for interpretation

COMPASS is central to the museum's Information Technology strategy for the 21st century. Its main public interface will be in the new Reading Room that opens early in 2000 with 50 flat touch screens provided; but simplified access will also be available through the internet to approximately 1,000 objects. The British Museum sees access via the internet as a crucial element because it gets 7,000,000 visitors each year and it is hoped that COMPASS will help visitors design personalised itineraries to plan their visits in advance, as well as possibly relieving the pressure from people wanting to visit the physical premises who may find a COMPASS guided tour an adequate substitute.

Also crucial to the museum is its consultation with other museums and in particular the Museums' Documentation Association (the branch of the Museums & Galleries Commission with responsibility for setting standards). Staff hope that the solutions that emerge, especially in capturing 3D items, will contribute to the creation of national standards.

Public Record Office

The Archives Direct 2001 Programme is an umbrella covering a range of projects to be initiated by the UK Public Record Office which is designed to provide electronic access to their records. Staff face a difficult job. They have 900 years' worth of documents produced by various government agencies to contend with. They have approached this massive job firstly by developing an electronic catalogue – Procat – which will be available through the internet. The retroconversion of 300,000 pages of paper catalogue has been a huge, eighteen-month long enterprise. While they do not anticipate the complete Procat going live on the internet until the year 2001, they are making subsets of it live for users as they are developed. The plan is for users to be able to search digital records – either full descriptions or finding aids – both hierarchically and by subject.

The PRO is also in the midst of a developing a project to digitize 1.5 million documents from their collection of the 1901 census at a cost of 5 million pounds. They have

undertaken smaller digital projects in the past but this is by far their largest undertaking. All the documents are handwritten and at this stage they have no plans to OCR the documents; instead they have chosen to rely on indexing to give users access and navigate the data. The index will link directly to the images of the census returns.

The plan is for the images to be online by 2002. This is an expensive project for which the office has received considerable sponsorship and without such sponsorship the project would not be going ahead. The PRO hopes the 1901 census will prove to be commercially viable and its ambitious plan is to recoup its own proportion of costs within two years of going live. Should this be the case, then the office plans to make other census material available in the same way.

British Broadcasting Corporation

The BBC is heavily involved in making their collection of materials, including sound and video, available online to its staff. There is no intention of making the collection generally available through the world wide web and this allows the BBC more flexibility in its plans because it is not limited by bandwidth constraints of internet usage. In June 1999 it embarked on a pilot project to enable users to browse and search online for video clips. The results were encouraging and in October 1999 it will be going live-enabling its users to search through News 24, the 24 hour news service, for their video needs. The BBC's archivist's intention is to have seamless cross searching available through an integrated front end search engine which will search all multimedia formats from documents and still photographs to video tapes and sound recordings. The BBC archive handles approximately 1 million loans per annum. It has an internal user-pays system for accessing the material neatly exemplified by its photographic access. A green background means the items are free; amber means items may incur a cost and red means there is a charge to use these items.

Archivist Adam Lee has a commitment to eventually make the BBC's entire collection available online and this includes:

- 1.5 million items of film and video
- 750,000 audio recordings;
- 3 million photographs,
- 1.2 million commercial recordings,
- 4.5 million items of sheet music,
- 22.5 million newspaper cuttings,
- 550,000 document files and
- 20,000 rolls of microfilm.

The BBC's determination is particularly interesting in light of views expressed by cultural institutions such as the British Museum, British Library and Public Record Office which are still talking in terms of digital projects, and still appear to be committed to only digitizing subsets of their collection.

The BBC uses a number of different databases to provide information to its users. For example it has a number of catalogues including news, books, anniversaries, as well as

ELVIS – Electronic Visual Image Store, Neon – News Information Online; and EDMS – Electronic Document Management System.

In the future the BBC hopes to provide "browse" quality of its audio and video film holdings; management of the digital objects; and continue considering new production techniques which help to reduce costs and improve quality, so that the whole operation will be an end-to-end digital world in which material is digitally stored, accessed, manipulated and delivered.

Two other projects that I would like to mention came out of my discussions with colleagues in the UK. These projects reveal some interesting potential for the use of multimedia on the net. One project is THISL and the other is PATRON.

THISL

THISL – Thematic Indexing of Spoken Language – is a three-year project focusing on the retrieval of multimedia information using an audio interface. The demonstration system has been set up with 100 hours of North American broadcast news and 800 hours of BBC news (currently being added to at a rate of 5 hours per day) to evaluate spoken document retrieval. THISL is concerned with the integration of Large Vocabulary Continuous Speech Recognition (LVCSP), Information Retrieval (IR) and Natural Language Processing (NLP) technology to automatically index and retrieve broadcast television and radio news programmes.

The aim is to have a system which recognizes broadcast speech clearly enough to produce indexing data. It does this by creating approximate transcriptions of the audio documents; these transcripts are then searched by text retrieval technology. Currently, the word error rates in the transcripts is around 30-35%. THISL will investigate the effect of word error rate on retrieval performance. The project is also evaluating a spoken query interface which allows users to interact verbally with the system instead of only by keyboard and mouse pad.

THISL uses the Abbot LVCSR system with a vocabulary of around 65,000 words. The recognizer produces a single best transcription, a word graph and word and phone level confidence measures. The same recognizer is used for both broadcast speech and spoken query interface.

PATRON

PATRON – Performing Arts Teaching Resources Online – is a pilot project to test access to audio, video, music scores and dance notation for students and teachers at the University of Surrey. The PATRON concept centres on the simultaneous delivery of a music score with the corresponding audio segment or a dance video clip with the matching choreographical notation. PATRON is now in its second stage of development which is focusing on integrating this multimedia resource into the university's curriculum. It will embed the use of this resource within particular courses by providing digitized music, dance videos, scores

and notation required as set works for undergraduate learning. Current access is still from workstations within the Library, however, they hope to extend access to the School of Performing Arts when their Asynchronous Transfer Mode (ATM) network is developed on campus which can guarantee the real-time delivery of sound and image data to users.

The work in this project involved digitization of three different types of material:

- scanned pages as images
- audio from CD
- video.

The aim was to produce an electronic resource whose quality exceeds everyday demands but is not necessarily of archival quality.

The facilities provided to users by PATRON include the ability to:

- listen and watch;
- select from the table of contents;
- pan and zoom;
- select and play clips; and
- speed control.

These two projects are exciting indications that what it is possible to currently achieve in terms of access, retrieval and storage for photographs and documents will become the norm also for video and audio.

Conclusion

In conclusion I must say that the trip I undertook raised many more questions for me than it actually answered. One of my major disappointments was in discovering that there was no consistent or clear cut direction being taken by the major institutions I spoke with. Institutions such as the British Museum and the British Library appear to be intimidated by their collections. They are taking small bites-for example the British Library in digitizing some of its special collections such as Beowulf and the Gandharan Buddhist Scrolls, or looking at digitizing its microfilm collections-but seem to be avoiding the larger picture.

Are small-scale digital projects an end in themselves? Or do they serve a valuable purpose in testing the water in preparation for large-scale services? The latter is the view that the John Curtin Prime Ministerial Library subscribes to. By digitizing only selected documents, cultural institutions create a demand from users to access more information. Users don't want to know about constraints that may be limiting institutions from making more information accessible: they simply want information that suits their needs.

According to Brian Cook (1997 p.5), "we are now in an era that will be dominated by digital information, made available from several image formats and a wide range of sources...A critical factor is the relationship that should exist between the impact of

technological innovation that enables electronic information growth to occur in leaps and bounds, and the actual information needs of clients."

I think that these are two critical points when we look at what we are trying to achieve by using Information Technology. People are becoming more and more sophisticated in their use of technology. Today's children are growing up with computers and the internet and it is probably safe to say they will automatically assume that they can access any information they want in this way. It will be their habit of a lifetime. If they can't get what they want from one institution that doesn't provide online access then they are likely to find some institution that does-even if that means settling for lesser information in lieu of easier access.

How long will large institutions such as the British Library or British Museum be able to rest on their reputations if they don't come to the technological party? It is true that technology currently offers no easy solution to the retroconversion of old manuscript or typescript documents. The microfilm project undertaken by the British Library to scan old newspapers proved how difficult it can be to OCR this material so that it is legible even for fuzzy searching retrieval software. However, what may be a mistake on the part of these large institutions is having no consistent policy in place to digitize new acquisitions. Hence the problem of retroconversion continues to grow and institutions' ability to make their whole collections available online decreases.

Obviously the task of making entire collections accessible via the internet involves many factors which I have not discussed, such as copyright issues, standards or the problems of data storage. Undertaking digitization of collections requires a long-term commitment. Institutions need to decide whether that commitment is limited to undertaking projects which make available subsets of their collections, or to providing digital access to their entire collection so that it is as accessible to the user half-way around the world as it is to the user sitting in the reading room.

Finally, it was good to see that little old provincial Perth-provincial even in the eyes of cities such as Sydney and Melbourne, let alone London-can feel proud of the way we have entered the electronic information stakes. Compared with our larger siblings overseas the John Curtin Prime Ministerial Library has proved that having a forward vision and a consistent approach paves the way for successful up-scaling-moving from digital projects to full-scale service.

I would like to thank the Victorian Association of Library Automation for awarding me this travel scholarship and also my director, Vicki Williamson, for her support in giving me time to conduct this research.

REFERENCES

Carpenter, Leona, Shaw, Simon and Prescott, Andrew (eds) (1998*) Towards the Digital Library*, The British Library, London

Cock, Matthew (1999) Author Interview, British Museum

Cook, Brian *The Electronic Library: Critical Issues and Responses in Electronic Library and Visual Information Research* - ELVIRA 4: Proceedings of the 4th UK/International Conference on Electronic Library and Visual Information Research, !997, Aslib, London

JCPML Electronic Research Archive [online] Available: http://john.curtin.edu.au [1999]

Lee, Adam (1999) Author Interview, British Broadcasting Corporation

Lyon, Elizabeth, Maslin, Jon and Baker, Bob *Audio and Video on-demand for the performing arts: Project Patron in Electronic Library and Visual Information Research* - ELVIRA 4: Proceedings of the 4th UK/International Conference on Electronic Library and Visual Information Research, 1997, Aslib, London

PATRON [online] Available: http://www.lib.surrey.ac.uk/PATRON/Patron.htm [1999]

Prescott, Andrew, Smith, Neil, Wright, Nicola (1999) Author Interview, British Library

THISL [online] Available: http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl [1999]

Webster, Alison (1999) Author Interview, Public Record Office

Williamson, Vicki and Henderson, Kandy-Jane (1998) *John Curtin Records Open to the World: How Australia's First Electronic Research Archive was Developed* Paper presented at the Australian Society of Archivists Conference "Place, Interface & Cyberspace: Archives at the Edge" Fremantle, 1998.