Managing documentary collections through digitization

Dr Marilyn Deegan
Digital Resources Director
Refugee Studies Centre
Editor, Literary and Linguistic Computing
University of Oxford
Queen Elizabeth House
21 St Giles
Oxford OX2 3LA
Phone: 01865-270435
Fax: 01865-270721
Email: marilyn.deegan@qeh.ox.ac.uk

Abstract

Digital collection development is part of a broader perspective on collection development, and generally needs to be assessed using the same criteria. However, there is a difference between reviewing collections already held by the institution with digitization in mind, and choosing to acquire digital materials from elsewhere. In-house holdings have been initially acquired or retained because they are perceived to have some value to the community served by the library, and therefore theoretically they might all be candidates for digitization. In reality, there will be complex factors within both the library and the community it serves which will dictate certain priorities for digital capture and delivery. The amount of material available for potential digitization in any library (however small) is likely to be much greater than resources available and so careful assessments of costs and benefits need to be made before embarking on projects.

This paper will review many of the issues attendant upon digitizing documentary collections, and will also look at the complex reasons for embarking on digitization. These will include widening access for students through the use of electronic reserves; facilitating immediate access to high demand and frequently used items; rapid access to materials held remotely; ability to reinstate out-of-print materials; potential to display materials which are in inaccessible formats, for instance, large volumes or maps; 'virtual reunification'—allowing dispersed collections to be brought together, the ability to enhance digital images in terms of size, sharpness, colour contrast, noise reduction, etc; the potential to conserve fragile/precious originals while presenting surrogates in more accessible forms; the potential for integration into teaching materials; enhanced searchability, including full text; integration of different media (images, sounds, video etc); satisfaction of requests for surrogates—such as photocopies, photographic prints, slides etc. and many other factors.

The presentation will draw upon a wide range of examples drawn from libraries all over the world, and will review documents from all periods and in a number of different languages and scripts.

Biography

Dr Marilyn Deegan has a PhD in medieval studies: her specialism is Anglo-Saxon medical texts and herbals and she has published and lectured widely in medieval studies, digital library research, and humanities computing. She is currently Digital Resources Director of the Refugee Studies Centre at Oxford University, and is Editor-in-Chief of *Literary and Linguistic Computing*, the Journal of the Association for Literary and Linguistic Computing, and Director of Publications for the Office for Humanities Communication based at King's College London. Dr Deegan has just completed a book *Digital Futures: Strategies for the Information Age* with Simon Tanner. This will be published in Autumn 2001.

The benefits of digitization

The amount of material available for potential digitization in any library (however small) is likely to be much greater than resources available and so careful assessments of costs and benefits need to be made before embarking on projects. Digital materials have some significant properties that make them fundamentally different from analogue materials, even though they may contain the same content. Some of these properties prove to be disadvantageous to the presentation and survivability of cultural materials—the evanescence and mutability of the digital object, for instance. On the other hand, there are considerable benefits of digital access to library collections. The digitization of resources opens up new modes of use, enables a much wider potential audience and gives a renewed means of viewing our cultural heritage. These advantages may outweigh the difficulties and disadvantages, provided the project is well thought-out. Institutions large and small are therefore embarking upon programmes of digital conversion for a whole range of reasons. The advantages of digital surrogates include:

- Facilitating immediate access to high demand and frequently used items
- Easier access to individual components within items (e.g. articles within journals)
- Rapid access to materials held remotely
- Ability to reinstate out-of-print materials
- Potential to display materials which are in inaccessible formats, for instance, large volumes or maps
- 'Virtual reunification'—allowing dispersed collections to be brought together

- The ability to enhance digital images in terms of size, sharpness, colour contrast, noise reduction, etc.
- The potential to conserve fragile/precious originals while presenting surrogates in more accessible forms
- The potential for integration into teaching materials
- Enhanced searchability, including full text
- Integration of different media (images, sounds, video etc)
- Satisfaction of requests for surrogates—such as photocopies, photographic prints, slides etc.
- Reducing the burden or cost of delivery
- The potential for presenting a critical mass of materials.

Any library considering digitization of its holdings will need to evaluate potential digitization projects using criteria such as these. They will also need to assess the actual and potential user base, and consider whether this will change when materials are made available in digital form. Fragile originals which are kept under very restricted access conditions may have huge appeal to a wide audience when made available in a form which does not damage the originals. It may also be possible to display originals whose format makes them inaccessible: large maps or older newspapers, for instance. Libraries should be warned, however, that making such facsimiles more widely available might see them faced with more requests to see the original, rather than fewer. This has already been the case in some libraries, and dealing with the requests puts an additional load on library staff.

Formats of materials for digitization

Most libraries hold a wide range of formats of manuscript, printed, visual and audio materials from all periods, in a variety of sizes, on diverse substrates and in different conditions, all of which may be candidates for digitization. Digitization is the process of conversion of *any* physical or analogue item into a digital representation or facsimile. The physical items that may be candidates for digitization may include:

- Paper

- Bound volumes, both print and manuscript
- Photographs—both prints and transparencies
- Microfilm and microfiche
- Video and audio
- Maps, drawings and other large format paper items
- Art works
- Textiles
- Physical 3-dimensional objects.

These may present many problems of handling, and some may need to be assessed for conservation treatment before any digitization can be contemplated. Each format will almost certainly need to be digitized using different methods, and indeed there may be a range of options for any one object, depending on the potential use of the digital collections and the resources available. Basically, anything that is accessible to photography can be digitized. The digitization processes are numerous, but might include the following:

- Image scanning
- Microfilming and then scanning the microfilm
- Photography followed by scanning the photographic surrogates
- Re-recording video and audio onto digital media
- Rekeying of textual content
- Optical Character Recognition of scanned textual content
- Tagging text and other digital content to create a marked up digital resource
- Digital photography – especially for 3-D objects or large format items such as art works.

Any library considering digital conversion may be faced with a large variety of potential originals that may be digitized and many means of digitization and presentation of digitized resources.

What does a digitization project involve?

Whilst there is significant variation in the original materials and the methods of digitization there are some common themes to every digitization project. Firstly, it is essential to assess the original materials to identify the unique characteristics of the collection. These unique characteristics will drive the digitization mechanisms and help define the required access routes to the digital version. Additionally, whether the end product is one data file or thousands they will have to be organized, given file names and placed in some logical structure. Having a clear vision of the information goals to be achieved from the original materials and the means of delivery are essential to a successful digitization project.

It must also be remembered that digital capture is only one of the many processes involved in the highly complex chain of activities which are attendant upon the creation, management, use and preservation of digital objects for the long term. Capture is likely to incur only a relatively small proportion of the total project costs. Any digitization project is likely to involve some or all of the following activities:

- Assessment and selection
- Grant writing and fundraising
- Feasibility testing, costing, and piloting
- Copyright clearance and rights management
- Preparation of materials
- Benchmarking
- Digital capture
- Quality assessment
- Metadata design and creation
- Delivery
- Workflow processes
- Project management

- Long-term preservation.

Without careful planning for *all* these elements, projects are unlikely to succeed. Costs will rise, deadlines slip and acceptable quality may not be achieved.

Some sample projects

Libraries all over the world have been digitizing valuable content over the last ten years, and there are many examples one could show and discuss. For the purposes of this paper, I will look at three very different kinds of materials: Gutenberg bibles, historic newspapers and photographic collections.

Digitizing Gutenberg

Johannes Gutenberg and the Bibles he printed in the mid-fifteenth century have iconic significance in the worlds of both written and digital culture, for Gutenberg's developments in printing ushered in a revolution in the production and dissemination of knowledge. The significant and far-reaching changes that are being brought about through new modes of information production and exchange are often likened to those brought about by the wide-spread adoption of printing technologies. In 2000, the six hundredth anniversary of Gutenberg's birth, he was named 'Man of the Millennium' and to celebrate this the State and University Library of Lower Saxony in Göttingen, Germany, mounted an exhibition of early European printing, the centrepiece of which was their copy of the Gutenberg Bible. Some 48 copies or parts of copies of this survive, out of around 180 that were originally printed. Only four of these are printed on vellum, of which the Saxony copy is one.

Given the pivotal cultural significance of Gutenberg's work, it is appropriate that his surviving Bibles should be viewed as candidates for digitization. But there are many other reasons for producing digital facsimiles of this work, not the least being its aesthetic qualities and the possibility of bringing these to a wider audience. In 1996, Keio University in Japan embarked on an ambitious programme of capture of digital facsimiles of Gutenberg Bibles, including its own volume acquired that year. The HUMI Project (Humanities Media Interface, and also a Japanese word for books, writing or learning) plans to digitize 10 copies in all, and has currently completed six of these: the Keio copy, two copies at the Gutenberg Museum in Mainz, two copies at the British Library and one copy in Cambridge University Library.

In 2000, to accompany the anniversary exhibition, Göttingen produced a digital facsimile of their copy of the Bible, together with the fifteenth-century Göttingen Model Book which laid out decorative patterns to be used in the Bible, and also Helmasperger's Notarial Instrument, a document that details a legal dispute between Gutenberg and his investor, Fust. This latter document provides the proof that Gutenberg was indeed the inventor of movable type.

The availability of seven versions of the Gutenberg Bible in digital form that can be compared instantly is of enormous importance to the study of early printing. No two copies of the work are the same, because the illuminations and rubrication were added by hand after printing, and the British Library/Keio team found evidence of changes of plan during the course of the printing process. There is no other way to compare all these fragile and rare bibles side-by-side when they are distributed across the whole world. It is also important to disseminate such vital cultural witnesses to the widest possible audience, which these digital products allow.

Newspapers

Newspapers are particularly difficult both to preserve and to access: they are large in format, prolific in output. Their creators intend them as essentially ephemeral — important today, discarded tomorrow—and so they print them on paper which is produced with cheapness in mind, rather than survival. There is, however, no other medium in our history that records every aspect of human life over the last 300 years—on a daily basis—like newspapers. They are also fearsomely difficult to extract information from unless indexed (a monumental task in itself) or unless the researcher knows the exact dates he or she seeks. There has been grave concern for decades about the survival potential of historic newspapers, given that many of them were printed on acid paper. Major libraries such as the Library of Congress in the USA and the British Library in the UK have been microfilming newspapers for many decades in order to preserve the historical record as well as, or instead of, preserving the objects. But there is also concern about the preservation status of microfilm, which itself deteriorates. Microfilm is also not as accessible a medium as the paper originals. The fate of newspapers has leapt into prominence over the last year with the controversies caused by Nicholson Baker and others about selection and retention policies in the UK and the US.

For users of many kinds, newspapers represent a source of information which is of monumental importance, and they are unparalleled as a primary source medium. However, it takes dedicated researchers to handle broadsheet-sized bound volumes of crumbling paper, or miles of microfilm, especially when most newspapers are minimally indexed. What makes newspapers such a unique resource (their diversity, their multimedia nature) is what also makes them so difficult to manage. Extracting content from the text of newspapers without presenting all the information around it, as well as the layout and typographical arrangement, is an impoverishing exercise, and clippings without context lose much meaning. In historical perspective, too, those aspects of newspapers which are often ignored —such as advertising—become a huge source of social, economic, political and cultural information. But researching these is a mixture of diligence and serendipity.

British Library Online Newspaper Archive

In the first half of 2001, the British Library Newspaper Library, OCLC Preservation Resources, the Malibu Hybrid Library Project at King’s College London, and Olive

Software produced a prototype system for the digitization, indexing and presentation of historic newspapers.

Selection criteria for the pilot

Eighteen reels of microfilm were selected, containing newspaper issues which centred around British ‘national’ events:

1851 On the 1st May the Crystal Palace was opened, which housed the Great Exhibition of the Industry of all Nations

1856 The Crimean War ended with the Treaty of Paris

1886 The Irish question in British politics; the defeat of Gladstone's Home Rule Bill, and his resignation

1900 A centenary year. Also, the events of the Boer War in South Africa featured prominently bin the national press, particularly the relief of Mafeking on the 17 May 1900

1918 11 November, the Armistice ending the ‘Great War of 1914-1918.

Digitization, searchability and presentation of complex textual objects

The digital images of the newspaper pages were obtained from microfilm images, a relatively cheap and speedy process. Some 20,000 pages were scanned. One of the main problems with digitized newspapers is that it is difficult to obtain good results with OCR from older texts from microfilm, but even more problematic, most OCR works at the level of the page, which can provide acceptable retrieval from simple documents, but not from compound documents. Here, retrieval needs to be at the level of the individual component. In the case of newspapers, retrieval is only truly meaningful from articles, advertisements or other individual objects, rather than from full pages.

With the British Library Newspaper Project, each individual page was ‘zoned’ into its component logical objects (articles, adverts, etc) which are stored as individual images in an XML repository. OCR was then carried out on the components, allowing retrieval of these as separate items rather than whole pages, though items can also be retrieved and viewed in their original context on the page. Boolean searches are possible across the whole database, revolutionizing access to this important resource.

Photographic collections

Many libraries and other cultural institutions hold photographic collections large and small, and these collections are increasingly at risk because of inherent instabilities in the photographic process; even relatively recent materials like colour prints of the last ten or so years are fading and losing definition. Many of the holdings are in urgent need of conservation and reformatting, and some of the materials are actually dangerously unstable such as nitrate film stock, which can ignite or even explode without warning. A survey of photographic materials in Europe carried out by the European Commission on Preservation and Access revealed that the 140 institutions that responded to the survey hold some 120 million photographs, half of which are over 50 years old. This is an astonishing number, and gives some sense of the scale of the world's photographic stock, all of which is at risk. Around four fifths of the survey respondents had already begun digitization of their photographic holdings or were planning to digitize in the future, with protection of vulnerable originals being a crucial reason. Digital surrogates are excellent substitutes for photographic originals and can provide almost all the content information that the original can yield, though of course artefactual information cannot be conveyed other than in the metadata.

Many libraries rely heavily on photographic materials, which are in some demand by users, picture researchers and commercial organizations. The availability of digital surrogates can reduce handling of the originals considerably: the Royal Library of Denmark estimated that to find 3-5 pictures, a user may handle as many as 300 originals. Accordingly, many institutions around the world are digitizing photographic collections. The American Memory project at the Library of Congress is rich in digitized photographic materials; Klijn and de Lusenet (2000) discuss many European libraries which are embarking on digitization of the visual heritage and the European Commission on Preservation and Access (ECPA) has set up the SEPIA project (Safeguarding European Photographic Images for Access) which is funded by the European Union to investigate ways of safeguarding photographic collections, including digitization for access. In the UK, the JISC Image Digitization Initiative (JIDI) digitized photographic collections from libraries in HE institutions for wider access for educational purposes. These collections included the photographs from the personal collection of Gertrude Bell, held at the University of Newcastle and the photographic and slide collections of the Design Council, which are archived in two different institutions, the Design History Research Centre at the University of Brighton and the Department of History of Art and Design, Manchester Metropolitan University. In Australia, the PictureAustralia service provides access to digitized photographs recording all aspects of the country's, with images supplied by a number of cultural institutions.

Conclusion

The bulk of digital materials to be delivered by libraries is likely to be obtained from external sources, both commercial and non-commercial. However, there are huge benefits to institutions in the selective digitization of their own holdings, provided this is embarked on with full knowledge of the issues and costs involved, as well as the potential

markets into which digital materials can be disseminated. Libraries can enhance their prestige, build skills and possibly even create new revenue streams by engaging in these innovative activities.