THE DIGITISATION OF JOURNAL LITERATURE: TOWARDS SUSTAINABLE DEVELOPMENT

Michael Breaks

Paper presented at the UKSG 20th Annual Conference, Heriot-Watt University, April 1997

This paper explains what is meant by digitisation and describes a number of current projects. It also covers the preservation of the material and sustainable development.

Michael Breaks is University Librarian at Heriot-Watt University, Edinburgh EH14 4AS

As we move into the digital age, one of the major problems facing scholars is how they can have seamless access to both digital texts and printed resources. There are a number of initiatives taking place to digitise back runs of academic journals to meet the needs of scholars, but before I look at them in detail, I will address some of the broader issues of the digitisation of printed resources.

Digitisation

Firstly, what do we mean by digitisation? Robinson gives a useful definition "A digital image is a computer representation of an object. This digitised image can be stored, viewed, copied, or otherwise manipulated on the computer."[1] He lists many advantages to the availability of digital images:

- Computer images will not decay in storage, if properly refreshed. Although some types of archival film (e.g. high contrast microfilm) have a quoted storage life of 500 years under controlled conditions, the dyes on many films will decay before this.
- Computer images will not decay in copying. It is estimated that each act of copying of microfilm loses ten per cent of the information in the film.
- Computer images will not decay in use. Microfilms scratch easily, and heavily used microfilms need frequent replacement by fresh copies from the master.
- Any one computer image may be located almost instantly in a digital archive, providing it is properly set up. The scholar can call up the image direct from a terminal without any further help.
- Computer images may be distributed almost instantly from a digital archive over a network. A scholar could, once the image is located, request a copy of this from a computer anywhere in the world.

- Computer images are far more tractable than film images. One may scale the images up or down, use enhancement techniques to make them more readable, paste parts of the images into articles or databases, compare different images side by side on the computer screen and print them out.
- Computer images may be distributed far more inexpensively than their printed equivalents.

There are, however, disadvantages in digitisation and these include:

- Storage of files this size is expensive. Most personal computers have hard disks of less than 100 megabytes, many only 20 megabytes. A single image might use up all this space.
- Manipulation of images of this size is beyond the reach of most personal computers. For efficient work, the whole of an image should go into the computer's memory while it is being viewed or edited.
- For images of such size to be stored at all, they may have to be compressed. Before any program can work with them, they then have to be decompressed. This can make the performance of the computer intolerably slow.
- There are many different graphics files formats and compression schemes, and as yet no firmly agreed single standard for these.

Access to digital images depends, of course, on continuing access to the appropriate computer software and hardware and this could be counted as a further disadvantage.

Mechanisms

Although this paper is not intended to cover the technical aspects of digitisation in any depth, it is worth considering at least some of the technical aspects, and to look at some of the choices that have to be made when digitising printed resources. Fresko, in his FIGIT Periodicals Digitisation Study[2] indicates that the digitisation of printed resources is still an immature science and there are a variety of approaches technical approaches, both to the creation of the digital material and the functionality provided to users. These approaches can be summarised as:

- Scan paper pages, keep the resultant 'bitmaps' (i.e. page image files). 'Lightly' index the images (i.e. index each paper or article).
- Scan paper pages, convert the resultant bitmaps to text using ICR (Intelligent Character Recognition) then:
  - use the resultant text as a full text index to the bitmaps, or
  - use the resultant text instead of the bitmaps, lightly indexed, adding images of illustrations, formulae, etc., from the bitmaps, if necessary, or
  - use the resultant text instead of the bitmaps, adding images of illustrations, formulae, etc. from the bitmaps, if necessary, having added 'deep' SGML tagging (i.e., tagging each heading, subheading, concept, illustration, etc.).
- Obtain the text from original typesetting tapes, then use in one of the three ways suggested above for ICR text.
- Use original files created by author to create PostScript and/or TEX files.
- Low or high colour fidelity: in the case of colour publications, the scanning can be performed with or without careful attention to colour fidelity. Colour fidelity implies a large 'bit depth', such as 24 bits, appropriate compression algorithms, use of an (expensive) calibrated scanner, and attention to other colour calibration issues. In most cases the colour fidelity is unlikely to be significant, but in some domains, and for preservation and some research purposes, it will be a requirement.
- Low or high resolution: pages can be scanned at low resolution (say 200 to 400 dpi) for the purposes of reference or bibliographic access only, or at high resolution (say 600 dpi or above) to allow preservation and research.

The approach taken will depend on the type of material that is being digitised and combinations of the above are possible. Robinson has provided a useful categorisation

of approaches to digitising material and proposes that there are three types of digitisation which can be used, depending on the material and its potential usage: a 'minimal' level, an 'average' level and a 'high quality' level.

- The 'minimal' level represents a lowest satisfactory format for display on standard computer monitors. It will not ordinarily print well and some portions of the image are likely to be distorted, blurred or meaningless.
- The 'average' level represents a satisfactory format for display on higher-resolution computer monitors, with or without colour facilities. It will ordinarily print well (though not to publication standard) and obscurity in the image is likely to result from obscurity in the original.
- The 'high quality' level represents the best reproduction now available using reasonably standard hardware and software. It will display well even at several degrees of magnification on a higher-resolution monitor. It will print to publication standard; any obscurity in the image will certainly result from obscurity in the original and use of image enhancement tools on parts of the image may actually make it more readable than the original.

Robinson suggests that projects should begin with the 'average' level and experiment with moving 'down' and 'up', depending on the material being digitised and the use to which it is likely to be put. However, in most cases the purpose of digitising an original is not to replace that original with a digital image, only to make a copy for more effective use.

Project TULIP

There is a considerable number of initiatives taking place in this area and it is only possible to indicate some of them in this paper. The first major digitisation initiative was the TULIP Project funded by Elsevier between 1991 and 1996. The final report[3] provides a useful summary of the work and the lessons learnt. The stimulus behind TULIP was to investigate effective large scale systems for the distribution in electronic form of traditional journal information. TULIP became operational in January 1993 and nine US universities, which had strengths in physical and engineering sciences, decided to participate. The participants set three objectives at the outset:

Technical: to determine the technical feasibility of networked distribution to and across institutions with varying levels of sophistication in their technical infrastructure. *'Networked distribution'* means sending the information both across the national Internet and over campus networks to the desktops of students and faculty.

Organisational and economic: to understand through the implementation of prototypes, alternative costing, pricing, subscription and market models that may be 'viable' in electronic distribution scenarios, comparing such models with existing print-then-distribute models, and understanding the role of campus organizational units under such scenarios.

User behaviour: to study reader usage patterns under different distribution (technical, organizational and economic) situations.

The TULIP project used the paper versions of the Elsevier journals to produce scanned images as the electronic form of the journals. The scanning was done at 300 dpi as no affordable high-volume 600 dpi scanners were available at the time, but the implications of this choice was that colour was not possible and the quality of photographs was sometimes unsatisfactory. After the pages were scanned, they were processed to perform optical character recognition (OCR) to generate a corresponding ASCII file. The page images and OCR were then used in a production editor environment to generate bibliographic records and the SGML file for each article. The intention was then to *'push'* the files via FTP to the participating universities, but difficulties with the Internet led to the decision to send the data on CD-ROM directly to the universities for local storage and access. Elsevier and the participating universities learnt a lot from the TULIP project, and the lessons are fully discussed in the Final Report, but perhaps one of the major issues is that if you only wait for ever, the ideal

technology to solve all the problems will appear!

Chemical Online Retrieval Experiment (CORE)

Another major initiative was the CORE project which intended to create an electronic library of primary journal articles in chemistry, based on about five years of twenty primary journals published by the American Chemical Society (about 425,000 pages). The CORE project, which was based at Cornell University, had five objectives[4]:

- define a suitable architecture for delivery of full text information in a distributed networking environment with heterogeneous workstations;
- convert and mount a critical mass of chemistry journal data in a database format suitable for effective retrieval and display;
- study the elements of full text interface functionality necessary to serve the needs of scholars in a network document delivery environment;
- advance the understanding of suitable document markup for electronic full text databases;
- investigate information retrieval questions germane to the coming era of full text delivery.

CORE included both scanned page images and marked-up ASCII text - represented in SGML - from the original machine-readable typography to create index databases. Each page, in print or microfilm, was stored at 300 dpi, but one of the major issues that the project had to address was image storage and presentation - given the importance of pictures to the users. In the selected journals, there is an average of one illustration per page and the illustration is essential to the understanding of the article. This resulted in the creation of a very large database and "even with data from a single publisher, in one format, in one subject area, the job of managing 80Gb of data, building and maintaining the database, and delivering it to a dozen different kinds of workstations should not be underestimated"[5].

Results of the preliminary CORE user studies suggest that users:

- search more effectively with computers than with paper indexes;
- read and absorb contents as effectively with computer displays as on paper;
- prefer paper for close reading of articles;
- wish to organise their own display screen, forcing the interface designers to yield control of the exact placement of interface elements;
- have a wide range of experience and expertise in online information systems;
- particularly like looking at pictures;
- often like to browse, not search the collection; and
- make extensive use of thumbnail-sized versions of the figures in the articles for browsing.

eLib initiatives

One of the areas identified for possible action in the Follett Report[6] was the possibility of digitising back-runs of journals in order to free space in libraries for increased reader space. Soon after FIGIT was established it commissioned a study to investigate the existence of projects to digitise backruns of selected journals[7]. The objective was to identify such projects and ensure that JISC does not fund proposals to digitise backruns which duplicate work done elsewhere. This review led to the funding of two projects in this area, Digitisation in Art and Design (DIAD) and The Internet Library of Early Journals (ILEJ)[8] based at the Radcliffe Science Library. I would like to look briefly at the latter project.

The aim of ILEJ is to offer expanded access over the Internet to digitised page images of substantial runs of already microfilmed 18th- and 19th-century journals, and to evaluate the service in terms of use and acceptability. The core collections for the project will be runs, of at least 20 consecutive years, of:

- three 18th-century journals: *Gentleman's Magazine, The Annual Register,* and *Philosophical Transactions of the Royal Society*

- three 19th-century journals: *Notes and Queries, The Builder,* and *Blackwood's Edinburgh Magazine*

The advantages of scanning from microfilm is that it is no longer necessary to go through the long process of assembling and handling originals which may be fragile, awkward to handle and difficult to position on the camera, as all this was done when the microfilm was produced. However, the potential limitations of using microfilm are obvious: microfilm must already be available; digital image quality is subject to the quality of the microfilm, and a two-stage process is more likely to result in quality loss than a one-stage process. Microform has always been regarded as an efficient, cheap and long-lasting alternative to traditional conservation methods and it is being suggested that originals could be microfilmed specifically in order to digitise. Though some would maintain that "digitisation should be done directly from the original object to the computer image, rather than through intermediate media"[9]. Microfilm is never popular with users because it is cumbersome to use and on occasion difficult to read. Therefore the British Library has been experimenting with a programme of microfilm digitisation and indexing to see what benefits can be gained[10]. The Burney Collection of 17th- and 18th-century newspapers has been chosen for scanning and various methods of indexing are being investigated.

JSTOR

This is a major and continuing project which grew out of *Preserving Digital Information,* a joint report of the Commission on Preservation and Access and The Research Libraries Group[11]. The report suggested that while it is not less expensive for a single library to convert paper to digital formats for the purpose of freeing up shelf space , if the materials to be digitised are held by many libraries, and the costs can be shared, savings can be captured. JSTOR[12] is intended to demonstrate these propositions for scholarly journals and intends to make older material more accessible by converting it to digital formats. William G. Bowen, President of the Carnegie Mellon Foundation, in a paper to the ARL Fall meeting[13], provided an overview of the thinking behind the JSTOR project in the context of the economics of scholarly communication. He explained the background to the decision to focus on the back issues of ten specific journals in economics and history, and the acknowledgment that they were swimming against the proverbial tide and "challenging marketplace solutions". They were able to do this because they had the backing of a major foundation and, unlike commercial entities, the test of success was not the 'bottom line'. Rather, it is how well the project would facilitate teaching and scholarship by improving the mechanisms of scholarly communication. Given these objectives, there were strong pragmatic reasons for focusing on back issues. This literature is

(i) least easily accessible;

(ii) most in need of preservation; and

(iii) most avaricious in its consumption of shelf space (the ten selected journals run to over 750,000 pages).

JSTOR began with a pilot project to provide electronic access to the backfiles of ten journals in two core fields, economics and history. The results of the pilot were considered successful and JSTOR was established as an independent not-for-profit organisation in August 1995, with the broad mission of helping the scholarly community to take advantage of advances in information technologies. Within this mission, the following goals have been articulated:

- to build a reliable and comprehensive archive of important scholarly journal literature;
- to improve dramatically access to these journals;
- to help fill gaps in existing library collections of journal backfiles;
- to help address preservation issues such as mutilated pages and long-term deterioration of paper copy;
- to reduce the long-term capital and operating costs of libraries associated with the storage and care of journal collections;
- to assist scholarly associations and publishers in making the transition to electronic modes of publication;

- to study the impact of providing electronic access on the use of scholarly materials.

JSTOR intends to provide the complete runs of a minimum of 100 important journal titles in 10-15 fields within three years of its launch in January 1997, and will offer access by site licenses, the cost of which will be based on the size of the institution. Individual subscription services are also being developed. The conversion of the backfile archive of the scholarly journal literature in JSTOR involves obtaining the physical copies of the title directly from publishers; checking the contents to provide a publication record for the title; creating appropriate indexing guidelines with input from serials specialists, and shipping the material to Barbados for scanning. Each page is scanned at 600 dpi resolution and to maintain quality. Page images are checked for marks, folds, or skewing, and are rejected if deemed unacceptable. Each page image is then processed by optical character recognition (OCR) software in order to create a digital file of the textual information that it contains. The results are quality checked in order to raise the accuracy of the text to 99.95%, or about 1 error in every 2000 characters. Finally, a table of contents file. This file is keyed for each article in the journal run includes bibliographic citation information and an item type identifier (full length article, book review, advertisement etc.) as well as key words and abstracts if these exist in the original publication. All three digital files - the page image, the OCR text file, and the table of contents file - are then downloaded to CD-ROM. The files are uploaded to the JSTOR file servers, final checks take place and the title is then available for use. Although this project is based in the USA, it is likely that mirror sites will be established in other parts of the world and similar subscription methods established.

Why images?

There is considerable debate on whether databases of digitised material should be image or text-based, and it is worth quoting at length from the JSTOR documents which provide a useful summary of the issues being faced by all projects in this area. "For JSTOR, it is not a question of image OR text, since JSTOR stores the data in both forms. For delivery, JSTOR provides only images to its users, while using a text file (created with OCR software from the scanned images) to facilitate searches. Since JSTOR does perform OCR on the images, some have asked why we don't make the resulting text file available to users. JSTOR's decision to rely on images is not a commitment to images *per se*, but is a result of a careful analysis of the costs and benefits of the alternatives.

If material is primarily character-based, it is generally preferable to deliver it to the user over networks as a text file. Text files are significantly smaller than image files and thus download faster. Text files are directly searchable and can be *marked up* to provide content-sensitive linkages between documents. Lastly, text provides the most flexibility for manipulation by the end-user (e.g. cutting-and-pasting). These advantages, which cannot be duplicated today using image-based files, have led most publishers of electronic editions of current issues to deliver them using some form of text-based file. There are, however, other important considerations. First, there is a substantial amount of non-text material in scholarly journals. Whether it appears as photographs, charts, tables, or special characters and formulae, these components of articles cannot be displayed with 100% accuracy using text-based methods available to standard web browsers. Second, if JSTOR is to serve as a substitute for the journal volumes on the shelves, it must offer an electronic version that is a faithful replication of the original. An image-based approach ensures the integrity of the materials in the database, while also retaining the *appearance* of the journal in its original presentation, which is important to many people.

There is also an economic reason for using images. Although JSTOR creates a text file, the text in it has been corrected to an accuracy level of 99.95%. This level of accuracy is excellent for searching, but it is unacceptable for display, especially from the publisher's point of view. The appearance of *typographical* errors could undermine the perception of quality that the journals have worked long and hard to establish. To pursue higher levels of accuracy