INTERNET LIBRARY OF EARLY JOURNALS
(ILEJ)

(January 1996 - August 1998)

FINAL REPORT

March 1999

Contents

Appendices:

Abstract

This report describes an eLib project to provide internet access to substantial (21, 20 or 10 year) runs of three 18th-century and three 19th-century journals, with the objective of improving access to the holdings of research libraries. Four of the titles were digitised from bound volumes using a Minolta PS3000 open-book cradle scanner. Two titles were digitised from pre-existing microfilm copies using Mekel equipment with grey scale facilities. Indexes to two titles were created by the use of uncorrected Optical Character Recognition (OCR). Indexes to four titles were provided by keyboarding volume or cumulated indexes published at the same time as the journals. For one title, *Blackwood's*, the electronic tables of contents from Chadwyck-Healey's *Periodicals Contents Index* was used.

An SGML-based file structure was developed for use with metadata. Images and indexes were mounted on servers at Oxford, using the PAT (Opentext) search engine, and at Leeds using EFS which offered fuzzy search capability. User responses to the service were evaluated by means of questionnaire survey. Issues relating to digitisation, OCRing, metadata and the user interface are discussed and a series of recommendation made. Possible strategies for the continued provision of the service are outlined.

A : Introduction

Objectives

1.   The Internet Library of Early Journals (ILEJ) project was funded by the Joint Information Systems Committee (JISC) as part of the first phase of the Electronic Libraries (eLib) programme.   The project was a consortium undertaking of the research libraries of the Universities of Birmingham, Leeds, Manchester and Oxford.   All are members of the Consortium of University Research Libraries (CURL) and the Research Libraries Group (RLG).

2.   The project sought to enhance access to holdings of research libraries by creating electronic copies of print and microfilm holdings which could be accessed via the Internet.   The specific objective was to create and to provide user access to a corpus of digitised images from three 18th-century and three 19th-century journals.   The runs of each journal would be of sufficient size (a minimum of 20 consecutive years) to provide a critical mass of material as perceived by the user.   The project was therefore multi-purpose, seeking to investigate the following issues:

- the use of both bound volumes and microfilm as the source of images;
- the effects of resolution, bi-tonal or grey-scales and compression on image quality;
- alternative indexing strategies including the OCRing of full text, the keyboarding of existing printed indexes, the use of existing electronic indexes, and the use of fuzzy matching software in conjunction with OCR'd text;
- presentation to the user via the World Wide Web and X-Windows servers in Oxford and Leeds respectively;
- data transfer between sites;
- acceptability to the user of the end product;
- critical mass;
- feasibility of scaling up the scanning process into a large-scale production line operation.

The Journals

3.   The material to be imaged was agreed before the start of the project.   The six titles were chosen according to a set of inter-related criteria to create a critical mass of material which could be considered to be broadly representative of pre-1900 journals as a whole and would test a range of technological variables.   These criteria included:

- wide subject range, covering science and technology as well as the arts;
- perceived user demand in the United Kingdom higher education sector;
- conservation priorities;
- diversity of typefaces and print and paper quality;
- short and long article formats;
- one, two or three-column page format and variable page size;
- balance between text and illustrations (both line drawings and half tones);
- availability of copies in the consortium libraries.

4.   The titles, run lengths and sources chosen were:

> *Notes and Queries* (1849-69): from bound volumes at Manchester
> *Blackwood's Edinburgh Magazine* (1843-63): from bound volumes at Birmingham
> *The Builder* (1843-62): from a copy of a microfilm supplied by Manchester Public Libraries
> *Gentleman's Magazine* (1731-1830): from a copy of a microfilm created by Cambridge University Library as part of the Mellon microfilming project.
> *Philosophical Transactions of the Royal Society* (1757-77): from bound volumes at Manchester
> *Annual Register* (1758-78): from bound volumes at Birmingham

A more detailed description of the six titles appears in Appendix I. *Illustrated London News* was flagged as an additional title, if resources allowed.   Its large size pages and prevalence of illustrations would have further extended the variables to be explored in the project.   In the event, no action was taken with this title.

5.   Bound volume scanning made use of volumes held on the site where scanning took place, with the exception that occasional volumes were supplied by one of the other sites or another library to replace a missing or poor-quality volume.   Before scanning, the run of each journal was sized, checked for missing or poor quality volumes or issues, and pagination characteristics were identified.   All microfilm scanning took place at Oxford.

Image Creation

6.   Minolta PS3000 open-book cradle scanners were selected for scanning of bound volumes and located at Manchester and Birmingham. A Mekel MX500XLG was chosen for scanning of microfilms and located at Oxford.   Installation of both types of scanner was delayed by the unavailability of grey scales, which were still not available for the Minolta at the end of the project.

7.   Image processing, involving cropping, deskewing and compression for display purposes, and OCRing (for two titles only) was undertaken at Leeds and Oxford. A metadata structure, which conformed to agreed standards, was established to provide each image with a unique identifier linked to the paper original and to the indexes.   Images were transferred among the four sites by ftp.   The images were stored on the hierarchical file server (HFS) at Oxford.

Access

8.   User access was central to the project. This required the provision of: distributed Internet access; retrieval and browse facilities; and legible display.   Effective indexing for retrieval was a key element and ILEJ explored two methods:

- OCR'd full text, the first time this had been attempted with this type of 18th- or 19th-century material;

- electronic equivalents of indexes or contents pages published contemporaneously with the journals.

9.   The advent of a Web version of the Excalibur EFS software, combined with the increasing dominance and universality of the Web and Web browsers, resulted in abandonment of the use of X-Windows at an early stage. Web servers were mounted at Leeds and Oxford using EFS (with fuzzy matching capability) and PAT respectively.   Access was via the Oxford server for which a user interface was developed with transparent links to Leeds.   The Web site at Oxford was also used to advertise the project.

10.   In the final three months of the project a user evaluation was conducted by questionnaire survey and telephone interview.

Management and Staffing

11.   The project was managed by Joint Project Leaders at Leeds and Oxford and a Project Executive comprising two members from each site who contributed to and supervised the work on that site, plus the project officers appointed at Leeds and Oxford.   Project officers were full-time staff who were shared with other projects and activities at the site with an average of 50% of their time paid by and assigned to ILEJ.   Scanner operators were appointed for varying periods at Manchester, Birmingham and Oxford.   The list of all staff in the project appears in Appendix II.

12.   The main activities at each site were:

Birmingham - bound volume scanning

Leeds   - image processing and OCR, server with EFS, evaluation

Manchester - bound volume scanning, some image processing

Oxford   - microfilm scanning, server with PAT, image processing and OCR, keyboarded indexes, archiving.

Dissemination

13.   Appendix III outlines the methods of dissemination of information about the project through email lists, paper mailings and journal articles, newsletters and conference papers.

Timetable

14.   This was originally a two-year project starting 1st January 1996. Because of delays in installing scanning equipment it was extended for a further 8 months until 31st August 1998 within the same funding.   Key dates were:

Feb.1996	:	Web site established at Oxford for internal project use
June 1996	:	Minolta PS3000 scanners installed in Manchester and Birmingham (para. 15)
Nov. 1996	:	Operational scanning started in Manchester
March 1997	:	First volume of *Notes and Queries* mounted on Web site
April 1997	:	Mekel (microfilm) scanner installed in Oxford
May 1997	:	Operational scanning of *Blackwood's* started in Birmingham (completed November 1997)
Aug. 1997	:	Operational scanning of *Philosophical Transactions* started at Manchester (completed December 1997)
Aug. 1997	:	First images (10 volumes of *Notes and Queries*) made available for user access.
Oct. 1997	:	Operational scanning of *Gentleman's Magazine* microfilm started in Oxford (completed April 1998)
Dec. 1997	:	Operational scanning of *Annual Register* started at Birmingham (completed March 1998)
Feb. 1998	:	Remainder of *Notes and Queries* (20 volumes in all) made available for user access.
April 1998	:	Operational scanning of *The Builder* started at Oxford (10 volumes completed by August 1998)
March-August 1998	:	Other titles mounted on Oxford server with indexes
June 1998	:	User evaluation

B : Image Creation and Processing

Paper Scanning

15.   Scanning was undertaken at Manchester and Birmingham using the Minolta PS3000P open-book cradle scanner.   Minolta equipment had been selected prior to the beginning of the project on the understanding that it could provide 256 grey-scale capability.   The supply of this equipment was delayed and when it was installed it became clear that it provided only dithered bi-tonal, not grey-scales.   There were further delays while the provision of greyscales was discussed with Minolta. It was eventually decided to accept equipment without greyscales and to proceed with bi-tonal scanning. As a result of these delays operational scanning started 7 months later than originally planned.   Minolta undertook to upgrade the scanners to grey-scales (which would require additional hardware) when these

became available but this did not occur during the lifetime of the project and consequently all images are bi-tonal. Operational scanning procedures were first developed at Manchester before starting scanning at Birmingham.

16. The criteria of image quality were "fitness for purpose" with respect to:

- legibility on a monitor screen or printout
- aesthetic appearance especially with respect to page curvature
- the quality of OCR

OCR quality was the most rigorous of these criteria though only applicable to the two titles for which OCR was used. Quality control for these two titles therefore had to take into account a measure (OCR quality) which could only be observed several stages down the production line from the initial scanning and usually at another site.

17. Much effort was devoted to developing scanning procedures in order to maximise image quality as defined above. A variety of problems were encountered in scanning from bound volumes. The curvature of the bound volume, combined with the buckling of paper, caused shadowing from the scanner's remote light source. Pages would distort and skew on the camera cradle especially at the beginning and end of volumes. Focusing had to be adjusted to take account of changes in page height at different points in the volume. Tightness of binding was a major determinant of the ease of scanning and the quality of output. Other factors that had to be accommodated by scanning procedures were variations in type density, see-through and varying margins (a narrow margin would accentuate the effect of tight binding).

18. The development of production procedures required a long learning curve. Frequent changes of parameter settings and scanner geometry were necessary to provide an acceptable product. Experience with one volume or one title could not automatically be applied to the next. Techniques developed included the use of a white card behind the page to be scanned and the scanning of single pages instead of a double page spread. All (say) right-hand pages would be scanned to the end of a volume, then all left-hand pages. A more detailed description of the scanning process appears in Appendix IV.

19. Though an acceptable product was obtained, this was heavily dependent on high levels of operator skill, judgement and patience. A higher quality product would undoubtedly have resulted from the use of a flatbed scanner with pages from dismembered volumes.

20. During the second year of the project the more sophisticated Zeutschel Omniscan 5000 equipment was purchased for use by the Higher Education Digitisation Service (HEDS) at Hatfield. Our assessment of the Zeutschel equipment was that it offered a much higher level of operator assistance including a glass plate against which the bound volume is pressed. Comparative tests with the same volumes were not undertaken but this equipment would be expected to provide higher quality images and a higher throughput because of greater ease of operation.

Resolution

21. Images for all four titles were scanned at 400dpi bi-tonal, the maximum resolution and colour depth available. However, the application of Cornell benchmarking (**1**) methods to the original material showed that this resolution was not high enough to give dependable capture of the small fonts which appear in parts of some titles, e.g. advertisements in *Notes and Queries*.

Throughput

22. The rate of scanning varied with the degree of difficulty experienced with the bound volumes. For example, for reasons already explained (para. 17) pages at the beginning and end of volumes would commonly take longer to scan then those at the centre. The average scanning rate was around 90 page images per hour, more than three times the rate assumed in the original ILEJ proposal, with extremes of 50 and 140. Appendix IV shows throughput data for individual volumes of *Notes and Queries*. These throughput figures do not include the time taken by the scanner operator to enter low-level metadata

(page number runs and issue start points) into an Excel spreadsheet (para. 47). The complete cycle for an 800-page volume including scanning, entering metadata and FTPing required two operator working days and optimum conditions.

Processing of Images

23.   The 400dpi tiff produced by the open book scanner required processing for two distinct purposes:

- display
- OCR

24.   For display purposes the tiff images were converted to 120dpi gif images for delivery to a Web browser, thereby removing the need for the use of browser plug-ins. This had the additional advantage of reducing file size. The resolution of the compressed images varied between 120 and 200 dpi for different titles in order to accommodate viewing of different page and font sizes on 800 x 600 pixel screens. ImageMagick was used initially for this process but the conversion process for a 650-page volume could take up to 24 hours (on a SPARC 5), an unacceptable production bottleneck. It was therefore superseded by Image Alchemy, which reduced conversion time to c. 4 hours per volume.

25.   As a prelude to OCRing, Sequoia Scanfix was used to deskew, despeckle and crop the tiff images. Considerable care was necessary in defining Scanfix parameters as a setting that cleanses an unsightly blotch from one page could remove a plate from another. Parameters therefore had to be authenticated for each title using a large, representative sample of pages.

Optical Character Recognition (OCR)

26.   OCR has the potential to be a low-cost method of producing high-value indexes which are of particular value for "newsy" publications such as *Notes and Queries* which are item-based and rich in anecdotal information. OCRing of 19th- and (especially) 18th-century journals is recognised as problematic because of the quality, variability, idiosyncrasies and size of typefaces, and the quality of the paper originals. For display purposes OCR quality in the region of 99.95% or higher is required. For index use, the quality requirement is much lower. Low quality (85-99%) OCR can offer a greater depth of indexing than conventional indexes or contents pages, though the high error rate does introduce inconsistencies in retrieval. As with any full-text search facility, the penalty for high recall is likely to be reduced precision. The low quality of the OCR may be reduced by the use of fuzzy matching software offered by the EFS search engine, and by the redundancies of full-text searching, e.g. if *Nelson* appears once on a page it may appear several times.

Choice of Software

27.   An OCR package integrated with the EFS software is available but, after preliminary tests, OmniPage Pro was preferred. Version 6 with Windows 3.1 was used initially but was subsequently upgraded to Version 8 running under NT. The OmniPage OCR engine, coupled with dictionary look-up, is a powerful tool for English text with modern typefaces. Inevitably it performed less well with images from bound volumes of 18th- and 19th-century type faces which included passages in French and Latin, and portions in extremely small typefaces. The training facilities provided with the software did not produce significant improvement in recognition of archaic characters and typefaces, nor did the use of prescribed page zone specification and limiting character sets. A more detailed description of the OCR process appears in Appendix IV.

Choice of Titles

28.   At an early stage in the project all six titles were tested for suitability for OCR in small-scale pilots using a flat bed scanner offering up to 300dpi and 16 or 256 grey scale. On the basis of these trials it was concluded that the three 19th-century journals could give satisfactory OCR results with bi-tonal

scanning. Use of grey scales could produce significant improvements in OCR quality especially with pages exhibiting show-through, though at the expense of much larger file sizes. The three 18th-century journals with heavy show-through and high variability in typeface quality were unlikely to give satisfactory OCR, particularly with bi-tonal scanning.

29.   OCR quality was subsequently tested for all six titles with:

- images scanned at 400dpi bi-tonal for the four titles for which the Minolta open book scanner was used.
- at various combinations of resolution up to 300dpi and grey scales up to 256 for the two titles for which images were produced from microfilm.

As a result of these tests, we concluded that the OCR output on *Notes and Queries* and *Blackwood's* was of acceptable quality and both were OCR'd in entirety. The other four titles gave results of varying levels of unacceptability. *Philosophical Transactions* was the next most acceptable and might merit further investigation. Failure to OCR *The Builder* was the most disappointing result. In earlier tests with a flat bed scanner it had given the best results of any title. This was attributable to the quality of the microfilm copy and the small print size. *Gentleman's Magazine* gave the least acceptable results.

30.   Apart from the OCR samples from images on a flat bed scanner at the beginning of the project (para. 28), no comparative studies were undertaken of OCR quality from images of the same page obtained by different imaging techniques. However, there are a number of possible strategies that could be used to improve OCR quality though whether these would have rendered the other four titles OCRable remains uncertain. These strategies are outlined in the recommendations in para. 108.

Measurement of OCR accuracy

31.   At the error rates being observed in this project even a casual glance would distinguish a good page from a bad one but quantification is difficult. Macros were written to match words against a Microsoft dictionary but this is not meaningful in relation to proper nouns (frequent in some titles), foreign text and the use of dictionary look-up by OmniPage which can result in the substitution of the wrong word, e.g. *night* instead of *right*. A character-by-character match is the ideal method and, in effect, would be used with high quality OCR intended for display purposes where the text would be read for sense *and* matched against a dictionary to identify any problem word which would then be assessed manually and corrected if necessary. This is a labour intensive process that is impracticable with high error rate "index-quality" OCR. In ILEJ, quantitative measures of accuracy were based on sample pages by manually matching of characters against the paper original. These showed wide variations within a title and between titles. An average *Blackwood's* page exhibited 98.5% accuracy and a poor *Notes and Queries* page could be below 80%. For both titles some very good pages exhibited an accuracy of greater than 99%.

Throughput

32.   With OmniPage Version 6 on a Pentium 90 with 64 Mb RAM a problem-free 650-page volume would take 12.5 hours to process. The software offered limited processing facilities: only 256 files (images) could be batched in one run and poor error recovery resulted in batch failures, which reduced the efficacy of overnight running. The upgrade to Version 8 (and NT) improved OCR quality and, by removing the 256 limit, the overall throughput. With either version of the software processing speed was very processor dependent. A Pentium 233 achieved five times the throughput of the Pentium 90 (both with 64 Mb of RAM).

Microfilm Scanning

33.   A Mekel MX500XL-G scanner, with full greyscale facilities, was delivered in April 1997 and was used for creating images from microfilm copies (positive) of two titles:

- *Gentleman's Magazine* : archival quality microfilm created in the Mellon Project and provided by Cambridge University Library. On the assumption that very high throughput would be achieved with a microfilm scanner, the project plan had envisaged creating images for a 100 volume (1731-

1830) run. In the event throughput was far lower than predicted and only 20 years (1731-1750) were scanned.

- *The Builder* : the microfilm was provided by Manchester Public Libraries.   A ten-year run (1843-52) was scanned rather than the 20 years originally intended.

34. Both the microfilms used for image creation are (positive) copies of originals produced by the scanning of bound volumes.   Both therefore reproduced the characteristics of bound volumes encountered with the open book scanner: page curvature, page wrinkling, see-through and variable density, in addition to some microfilm-specific problems (see below).   As OCR was not being used (cf. para. 29) and aesthetic appearance, especially page curvature, was pre-determined by the microfilm copy, the only quality criterion under our control was legibility. Use of greyscales to improve legibility had therefore to be balanced against greatly increased file size and reduced throughput.

35. There had been some expectation that the combination of fixed geometry and a pre-existing optical image would provide for a smooth workflow with no need to change parameters.   In practice, as had been found with paper scanning, variations in parameter settings were necessary for different frames, different volumes and (especially) between titles.   Procedures had to be developed to deal with variations in the contrast levels within a frame, "false edges" caused in many frames by the binding of the original volume being visible and disrupting the edge detection procedure used by the scanner. The two microfilm products displayed very different scanning characteristics.   A long learning curve and a high level of operator skill were therefore required.   In contrast to paper scanning, in which image processing was divorced from scanning and carried out remotely, Mekel software provided the facility for automatic cropping of images.   As it was scanned, the image was cropped to dimensions defined by halving the width of the original image.   This procedure did not work where the two halves of a frame were of markedly unequal size (again a function of the open-book original).   Operator intervention was then necessary and cropping was undertaken with Adobe Photoshop. In practice the Mekel cropping facility was used for almost all frames in *The Builder* microfilm but none of those in the *Gentleman's Magazine* microfilm.

36. Scanning at 300 dpi bi-tonal gave acceptably legible images for 66% of frames of the *Gentleman's Magazine* microfilm but 34%, clustered in specific volumes, required the use of 100 dpi with 256-level greyscale.

37. Because of the large size of the original pages and the small font, all images of *The Builder* were scanned at 200 dpi with 256 greyscales, producing file sizes of 10 megabytes.   The large file size resulted in extended cropping time and disk space bottlenecks on the Mekel.   The scanning and cropping therefore had to take place in batches of 50 images. This problem was partly alleviated by a change in the scanning procedure.   The microfilm was scanned twice, the first time for the verso sides and the second for the rectos to produce ready-cropped images in each case.

38. The cropped bi-tonal images were converted to gif files and the greyscale images to medium quality jpegs, using Image Alchemy (para. 24).

Throughput

39. The original ILEJ proposal assumed a throughput of 250 pages an hour, ten times that for paper scanning, and the manufacturer's specification gave a throughput of 600 pages per hour at 200 dpi grey scale, and much higher in bi-tonal mode.   Observed throughputs were well below these values though, as explained above (paras. 35, 37) the reasons for decreased output were different for each microfilm. Under optimum conditions, with no requirement for parameter changes, a rate of 200 images (100 frames) per hour could be achieved but average throughputs were 40 images per hour for *The Builder* (including automatic cropping) and 70 per hour for the *Gentleman's Magazine* (excluding cropping).

C : Conversion of Printed Indexes

40.   Two types of index were created for image retrieval:

- OCR'd full-text (cf. para. 26)

- electronic versions of indexes and tables of contents pages published contemporaneously with the journal. These were created either by keyboarding from the printed versions or, for *Blackwood's* only, by using the pre-existing electronic version of tables of contents already available in Chadwyck-Healey's *Periodical Contents Index* (PCI).

The following indexes and tables of contents were made available to users:

Notes and Queries: subject index, taken from the original printed indexes to each volume

Blackwood's Edinburgh Magazine: author and title indexes, compiled from the tables of contents information in Chadwyck-Healey's *Periodical Contents Index.*

Gentleman's Magazine: subject index, taken from the cumulated index for volumes 1-20.

Philosophical Transactions: author, title and subject indexes, compiled from the tables of contents and subject indexes in the original printed volumes.

Annual Register: subject index, taken from the original printed cumulated index.

The Builder: none

41.   After initial experiments with in-house keyboarding, the task was outsourced to a commercial supplier, Offshore Keyboarding Corporation.   Charges varied from £0.73 to £0.85 per 1000 characters.   The procedure followed was:

- a quotation was obtained on the basis of sample pages of the index and an estimate of the total number of characters, based on a character count for a small sample of pages, usually obtained with the aid of OCR software. (More subjective estimates for the first title, *Notes and Queries*, proved inaccurate).

- the index pages from the printed volumes were photocopied, marked-up in a simple SGML-like form and sent to the keyboarding agency.   The following tags were used in the mark-up:

    - <e>…</e> **entry**, delineating the start and end of each index entry

    - <pr>..</pr> **primary term**

    - <s>..</s> **secondary term**

    - <v>..</v> **volume number** (if a cumulative index to several volumes)

    - <p>..</p> **page number**

An example of the marked-up, keyboarded index entries from *Philosophical Transactions* is shown below:

```
<e><pr>Abaris</pr><s>account of him</s><p>341</p></e>
<e><pr>Abdias</pr><s>quotations from his Apostolical history</s><p>82</p></e>
<e><pr>Abraham</pr><s>case of his offering Isaac discussed</s><p>176</p>
<s>His posterity</s><p>33</p>
<s>Paid tythes to Melchizedeck</s><p>269</p></e>
<e><pr>Abrabam</pr><s>rejoiced to see my day, explained</s><p>177</p></e>
<e><pr>Achilles</pr><s>speech of his horse</s><p>56</p></e>
<e><pr>Acts and monuments</pr><s>of Mr Fox, placed in churches</s><p>105</p></e>
<e><pr>Addison</pr><s>passage from his Battle of the Cranes and Pygmies</s><p>51</p>
```

- the indexes were double keyed with an estimated accuracy of 99.995%, and returned in machine-readable form by email;

- macros in Microsoft Word were used to process the keyboarded files, and to load each index entry into its respective place in the SGML files (using the SGML ID system).

42.   The keyboarded output was converted into INDEX elements within the TEI.   This conversion had to take into account the special characteristics of some titles.   The printed indexes in *Notes and Queries* delineated primary and secondary terms more by context then by typography and required extensive editing to provide a form suitable for incorporation into the TEI file.   The cumulated index used for *Gentleman's Magazine* required separation of the index entries for each of the 20 volumes.   Where there is an entry for the secondary term only for a given volume, the associated primary term had to be extracted.

43.   The intention had been to provide facilities both to search for terms or term combinations within an index, and to display and browse a list of index terms.   The latter, which was not implemented during the lifetime of the project, required the conversion of index information into separate TEI conformant documents, which would act as cumulated indexes to a title with links from each entry to the associated image.

44.   The total numbers of index entries for each journal title are tabulated below. (An index entry is a separate INDEX element in the TEI file, corresponding to a separate primary or secondary index field in the original file.)

	Subject	Author	Title
Notes and Queries	171,390	-	-
Gentleman's Magazine	19,977	-	-
Philosophical Transactions	8,861	817	917
Blackwood's Edinburgh Magazine	-	96	2,004
Annual Register	18,955	-	-

D : Metadata

45.   **Metadata content**

Several types of metadata have been incorporated into the ILEJ project - these include:

- basic bibliographical information on a journal volume as a whole and on each image (intellectual metadata). (cf. para. 46)

- subject, author and title indexes (paras. 40-44)

- OCR'd full-text (paras. 26-32)

- basic Dublin Core metadata, included in the <META> HTML tags within the ILEJ home page, to facilitate discovery by WWW search engines.

No administrative metadata (file resolutions, compression systems used etc.) were included, though bi-tonal/greyscale data were implicit in the file format for each image (gif = bi-tonal; jpg = greyscale).

Input of Bibliographical Data

46.   In conjunction with image creation, the scanner operators keyed in the basic bibliographic metadata into an Excel spreadsheet: page run and milestone indicators (e.g. the start of a new issue of a journal) and, in the case of the microfilm images, whether scanned in bi-tonal or greyscales.   Accurate input was of critical importance as these files provided the control data for all subsequent automated file processing.

47.   The project had hoped to use the SQL-compliant document management database which came bundled with the PS3000P scanner.   However, investigation showed this database to be a closed system that would not allow independent extraction of metadata.   The only other metadata produced by the scanner was a simple list file, matching image number against file name.   Data from the Minolta scanners was therefore input into an Excel spreadsheet and processed by a simple BASIC programme to produce a text (ASCII) file that matched page-run information with the file produced by the scanner.   A set of Perl scripts was then used to generate the SGML identifiers and TEI files automatically from this file. Data from microfilm scanning was also input into an Excel spreadsheet and exported to the FoxPro database programme from which TEI files were constructed.   In this case the FoxPro script was used to produce the TEI files, which were parsed and edited as necessary to ensure full TEI conformance.
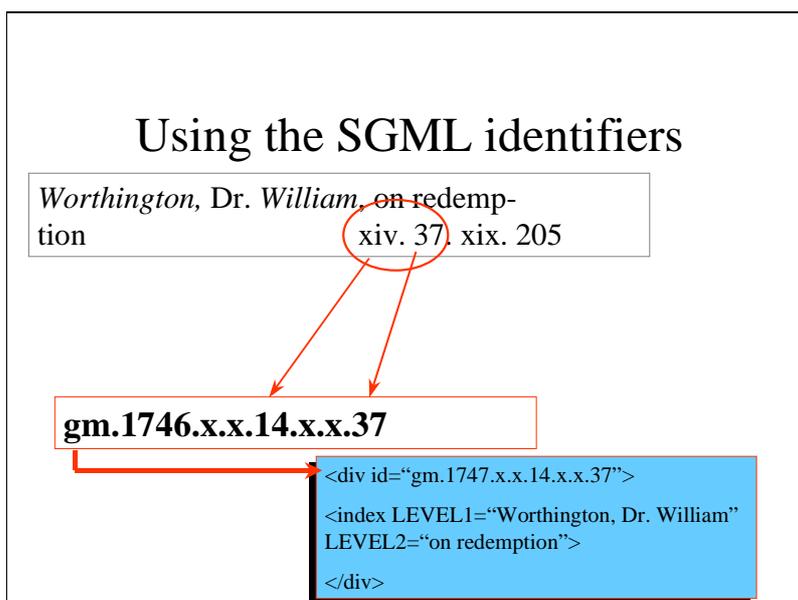
SGML file structure

48.   Metadata for the ILEJ project are held in SGML (Standard Generalised Markup Language) files. Several reasons prompted this choice in preference to a proprietary database format:

- SGML is an ISO-approved format, which is independent of any given application, provides metadata of greater archival robustness and allows easier future migration to new platforms.

- SGML allows a complex web of metadata from the various sources mentioned in para. 45 to be integrated into a single, clearly defined structure.

- The hierarchical structure of an SGML-encoded document neatly mirrors the structure of the original journals, making the creation of their virtual surrogates simpler and more logical.

49.   The ILEJ project uses two Document Type Definitions (DTDs): the Encoded Archival Description (EAD), an application designed to encode collection level descriptions of archives, and the Text Encoding Initiative (TEI), a generic and widely used scheme for encoding texts of many types. The EAD is used to provide a basic skeletal description of the Library as a whole; it lists only the journal titles and the individual volumes of each title.   Each volume is in turn encoded in a single TEI file, which contains bibliographic information for the journal as a whole and for each constituent page, with links to each image file and, where available, to index entries and OCR'd full text.   External links are used in the EAD file to point to each corresponding TEI file.

50.   The approach adopted by the ILEJ project emphasises the physical rather than the intellectual structure of each volume.   Each page of the original is represented in the TEI file by a single <DIV> element, from which a link to its related image file is made.   The intellectual hierarchy, which cuts across its physical counterpart, is represented by empty <MILESTONE> elements, which delimit the boundaries between distinct units such as weekly issues for *The Builder*, articles for *Philosophical Transactions* or monthly issues for *Gentleman's Magazine*.

51.   SGML supports a system of identifiers, which can be applied throughout a document to provide anchor points for cross-referencing.   Every image in the ILEJ database is assigned a unique identifier using the following schema:
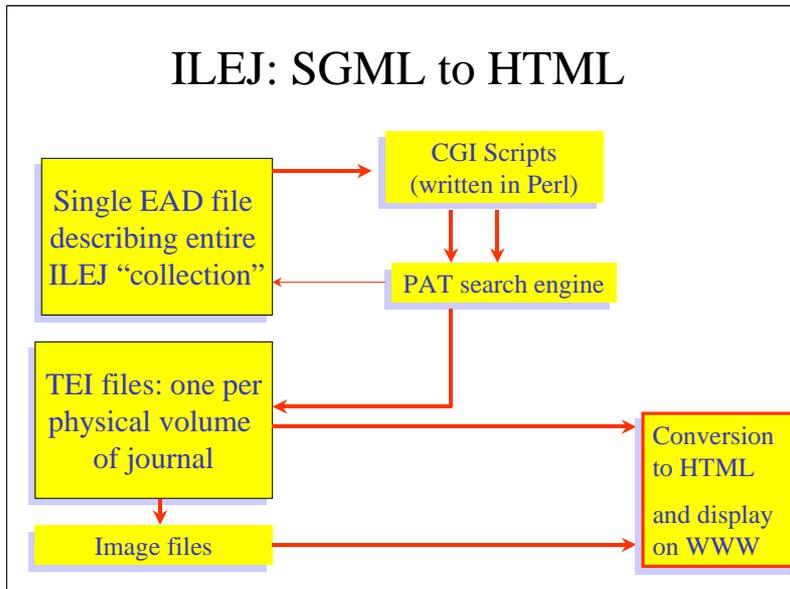
ILEJ: SGML identifiers

gm.1747.1.x.1.x.x.5

Journal

Year

Month

Day

Volume

Part No.

Pagination Sequence

Page No.
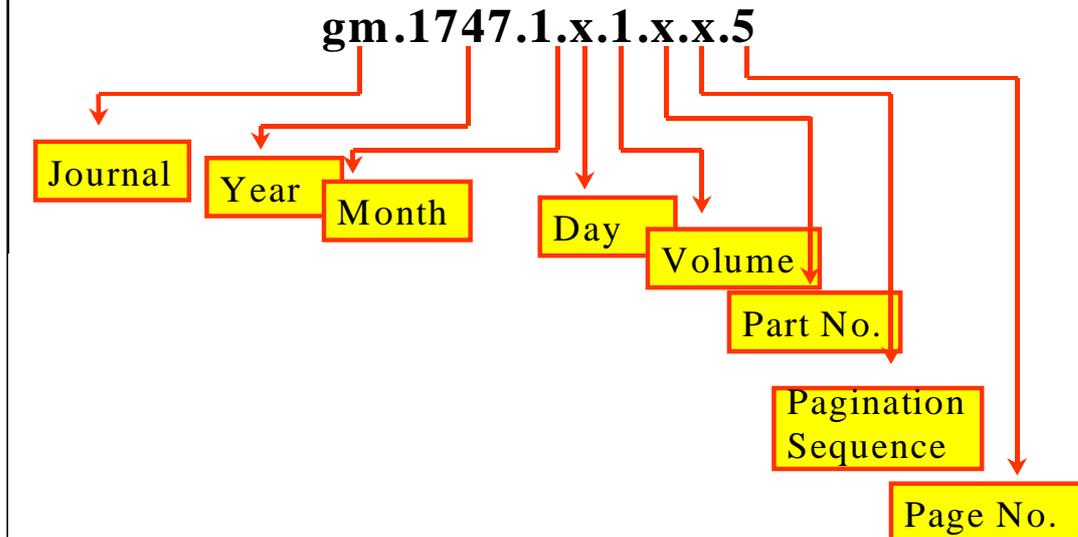
52.   Using this approach, it is possible to construct an ID of this form from any item of data or metadata generated in the course of the Project. This ID can then be used to fit any item into its place in the SGML files.   The following example illustrates how information from a printed index is converted to an SGML ID and the contents of the index item are then allocated to their correct location in the TEI file for its corresponding volume:-

Using the SGML identifiers

Worthington, Dr. *William, on* redemp-
tion   xiv. 37. xix. 205

gm.1746.x.x.14.x.x.37

<div id="gm.1747.x.x.14.x.x.37">

<index LEVEL1="Worthington, Dr. William" LEVEL2="on redemption">

</div>

53.   The overall structure of the ILEJ project in terms of metadata is shown in the following diagram:

ILEJ: SGML to HTML

- Single EAD file describing entire ILEJ "collection"
- CGI Scripts (written in Perl)
- PAT search engine
- TEI files: one per physical volume of journal
- Image files
- Conversion to HTML and display on WWW

54.   A script written in Perl uses the Opentext PAT5 search engine to access the contents of the master EAD file in which the structure of the entire collection is encoded, then the relevant TEI file for the contents of the journal volume and the relevant image files for the item being viewed.   These files and the bibliographic information from the TEI files are then combined and reformatted to HTML for viewing on the WWW.

55.   Since the ILEJ project started, the usefulness of SGML as the basis of metadata for digital imaging projects has become more widely acknowledged.   The Commission on Preservation and Access has recently acknowledged SGML as the most robust format for such metadata (**2**) and the Library of Congress in its American Memory Program has used SGML for encoding finding aids and full-text mark-up (**3**), though metadata used in the SGML files are minimal. References to presentations on the project's approach to metadata are listed in Appendix III.

E : Servers and Access

56.   The original project plan, when formulated in 1995, envisaged two distinct server strategies, reflecting the different expertise of the two sites:

- at Leeds, a server with Excalibur's EFS software and X-Windows offering fuzzy searching which would enhance retrieval capability when searching imperfect OCR.

- at Oxford, a Web server using a standard Web search engine with PAT (Opentext) as the most likely candidate, offering simple search facilities (at the time there was no Web search engine offering a fuzzy matching capability).   A Web server already available at Oxford was enhanced by the purchase of 36 gigabytes of additional storage for the ILEJ project.

57.   The arrival of a Web version of EFS, combined with the increasing dominance and universality of the Web and Web browsers, resulted in a modification of this policy. Web servers were mounted at both Leeds and Oxford using EFS (with fuzzy matching capability) and PAT respectively.   FTP was used extensively to transfer images from the sites where imaging was taking place to the servers.   The hierarchical file server (HFS) based at Oxford was used for all archival purposes.

58.   The server architecture and the transfer flows between sites are shown below:

Schematic ILEJ System Architecture
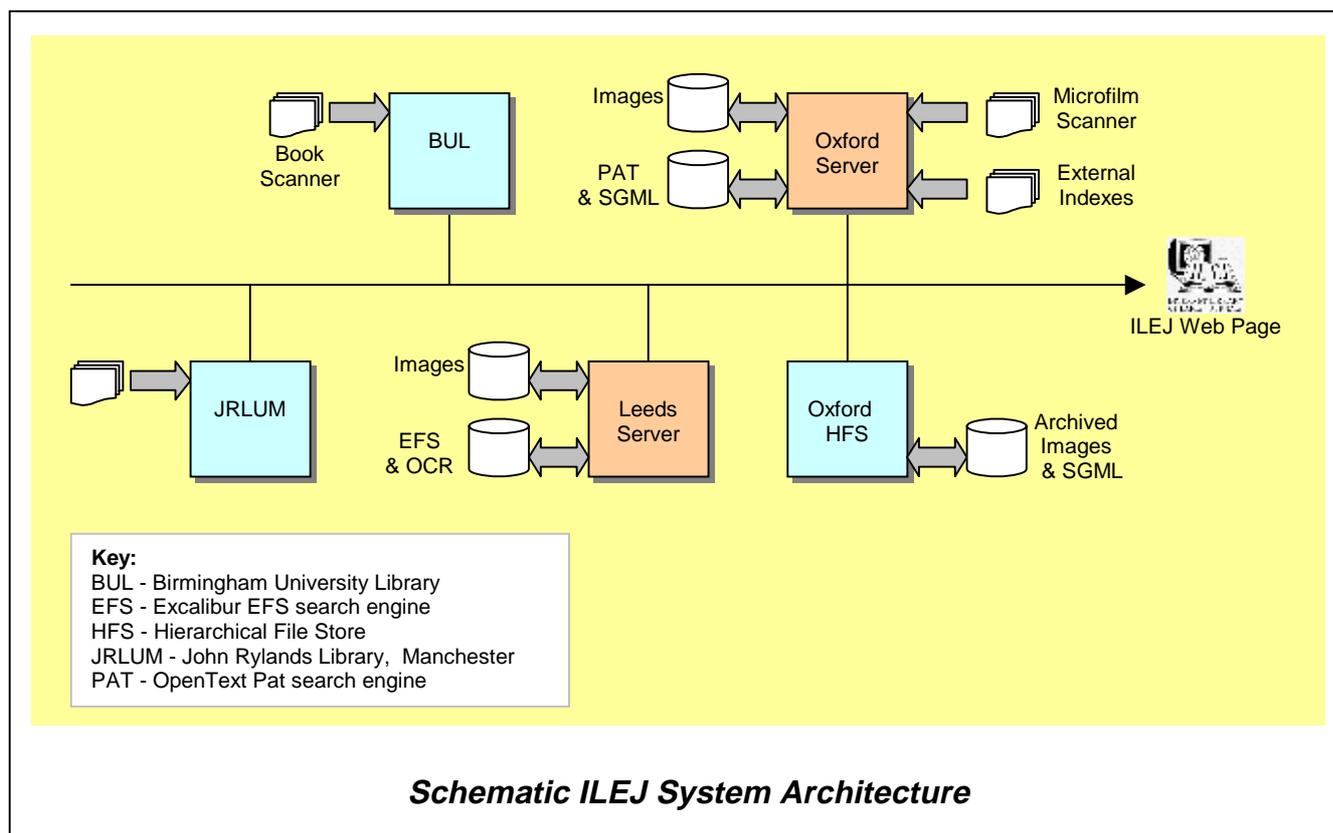
59.   The main elements of the workflow were:

- original (uncompressed) images and associated metadata created in Manchester and Birmingham by scanning bound volumes were sent to Oxford to be stored in the HFS.   Images (and associated metadata) created in Oxford by scanning of microfilm were also stored in the HFS.
- both Leeds and Oxford retrieved images from the HFS for processing to create OCR text and smaller gif images for viewing on the WEB.

- the Web images for four titles are stored at Leeds and for two at Oxford.

- for each journal, an SGML file was created at Oxford and stored in its PAT database. This contains a bibliographic description of each page with its OCR, if any, and keyboarded indexes if available.

- Leeds maintains a copy of the OCR in its EFS database for fuzzy searching.

The Interface

60.   The initial point of entry for all users was the Oxford Web interface which is designed to offer users choices between:

- browsing, searching and (OCR'd text only) fuzzy matching.
- OCR'd text and a variety of keyboarded subject, title and author indexes.

All the journals were browsable but the search options varied: only two offered OCR'd text and   no search options were available for *The Builder* at the end of the project.

61.   The design principles for this interface were:

- to offer the user a synthesis of different types of index linked to image display

16

- transparency
- ease of use.

Screen dumps of the interface appear in Appendix V.

62. When fuzzy searching was required the user was transferred to the EFS server at Leeds. Images were downloaded from Leeds as required. The links from Oxford to Leeds were transparent to the users who were unaware that they had moved to another server, though the EFS interface used for fuzzy searching is significantly different from that at Oxford.

Available Content

63. The images and indexes for each title that were available at the end of the project are listed below. Only the first four titles were available during the evaluation phase of the project.

Journal	Years	No. of images		OCR'd text	Indexes
Notes and Queries	1849-69	26,254	YES	S	
Blackwood's Edinburgh Magazine	1843-63	33,183	YES	P	
Gentleman's Magazine	1731-50	14,181	-		
Philosophical Transactions of the Royal Society	1757-77	18,947		-	S,T,A
Annual Register	1758-78	12,465		-	S
The Builder	1843-52	5,518		-	not available

Key: S = Subject   T = Title   A = Author

Archiving

64. The Hierarchical File Service at the Oxford University Computing Service was used to archive:

- the cropped tiffs for *The Builder*

- the unprocessed tiffs for all the other five titles

- the keyboarded indexes files as supplied by Offshore Keyboarding Corporation

- SGML files

- the OCR'd full-text for *Notes and Queries* and *Blackwood's*

- the gifs (and *Builder* jpegs) created for viewing on the Web.

This archive occupies c. 122 Gb

F : User Recruitment, Usage and Evaluation
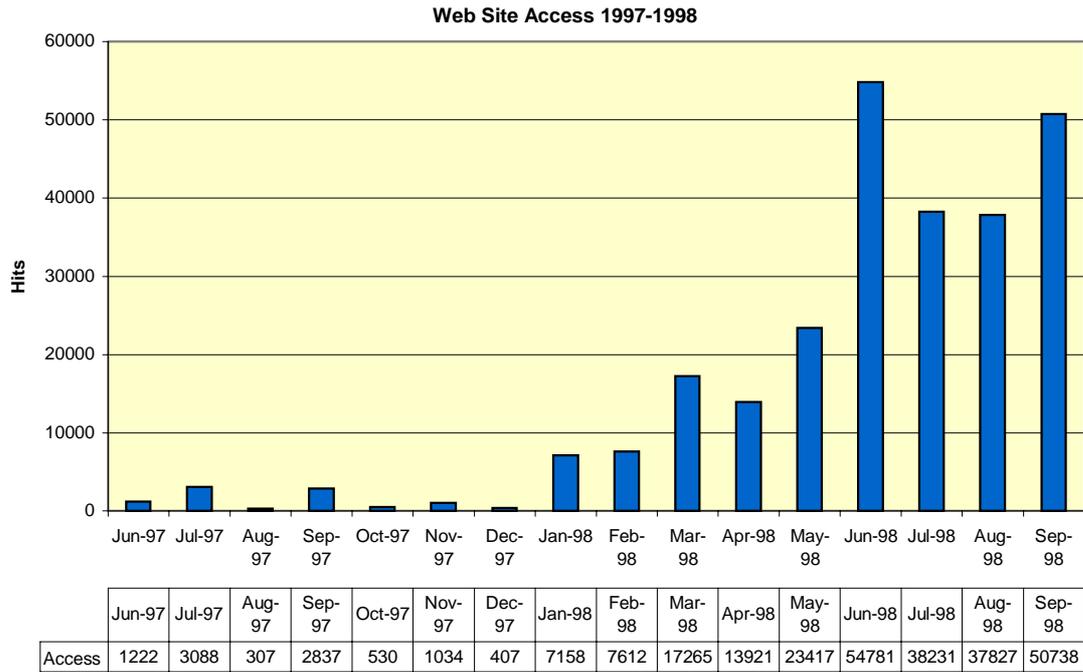
Recruitment

65.   Access to ILEJ was freely available on the Internet without password or IP address control.   Users were recruited through advertising on UK mailing lists hosted by mail base, internal mailings (paper and electronic) within the four participating institutions, and through journal articles, newsletters and conference papers.   Promotion continued throughout the project but a major publicity campaign was undertaken immediately prior to the evaluation in June 1998.   Further details of the methods of dissemination of information about access to ILEJ are given in Appendix III.   Throughout the project users were asked to complete a Web registration form, which is reproduced in Appendix V, on first using the service. The form requested information on name, department and institution together with information on research and teaching interests, previous use of the paper versions of the ILEJ titles, access to a Web browser, and the likelihood of using the electronic versions of the titles. 377 Users completed the registration form, though neither compulsion nor reward was offered.
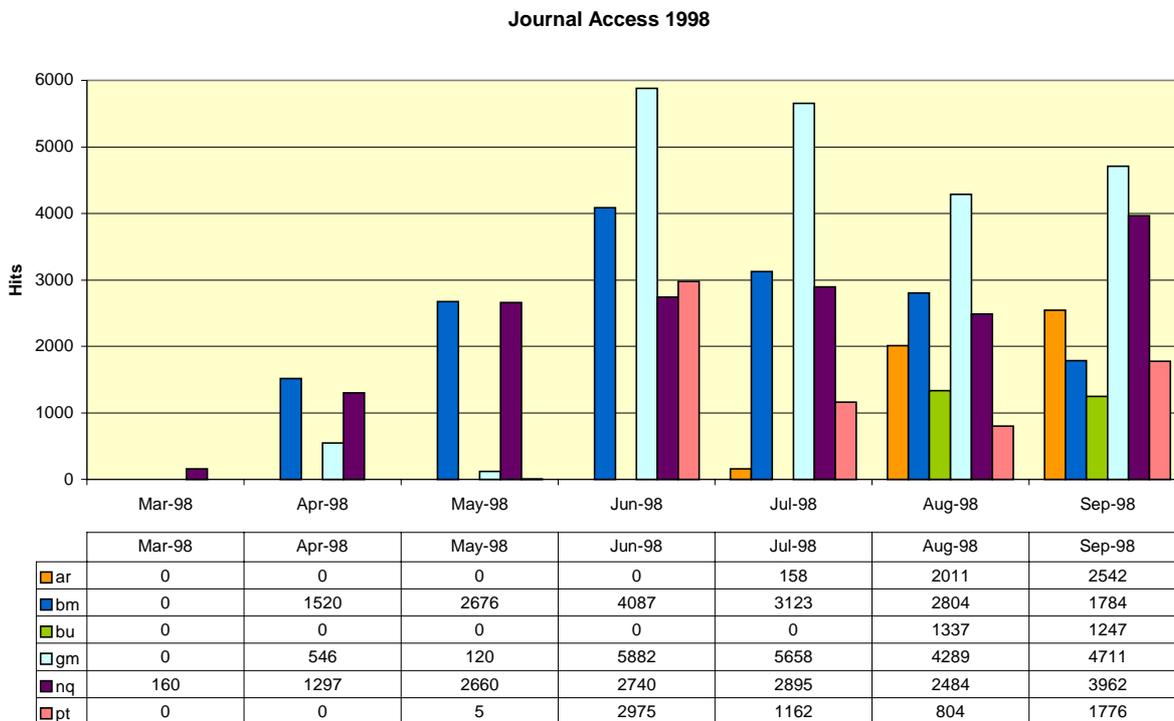
Evaluation Objectives

66.   The project evaluation strategy identified the following objectives for a user evaluation:

- to identify the level of use, and reasons for use, of paper originals, and expectations for an electronic equivalent. Specific data required were:

    - identity of users
    - frequency of use
    - patterns of use
    - reasons for use

- to determine the acceptability of the service to users.   Specific data were required on the acceptability of:

    - images in terms of legibility and appearance
    - the retrieval options
    - presentation
    - content
    - comparison of use of electronic and paper versions

- to identify improvements that could be made in the service.

- to predict, as far as possible, the future viability of such a service.

Usage

67.   The usage data shown below are for all users, not just those who completed the registration form. These data are subject to the normal limitations of Web statistics arising from the use of caching and proxies.   The number of individual pages accessed on the ILEJ Web site in the sixteen months June 1997 to September 1998 are shown in the following bar chart (note that these data include access to pages describing the project, the registration form, the search screen and the lists of volumes, issues and pages in the browse mode, in addition to the page images).

Web Site Access 1997-1998

	Jun-97	Jul-97	Aug-97	Sep-97	Oct-97	Nov-97	Dec-97	Jan-98	Feb-98	Mar-98	Apr-98	May-98	Jun-98	Jul-98	Aug-98	Sep-98
Access	1222	3088	307	2837	530	1034	407	7158	7612	17265	13921	23417	54781	38231	37827	50738

The number of *page images only* accessed for each journal is as follows

Journal Access 1998

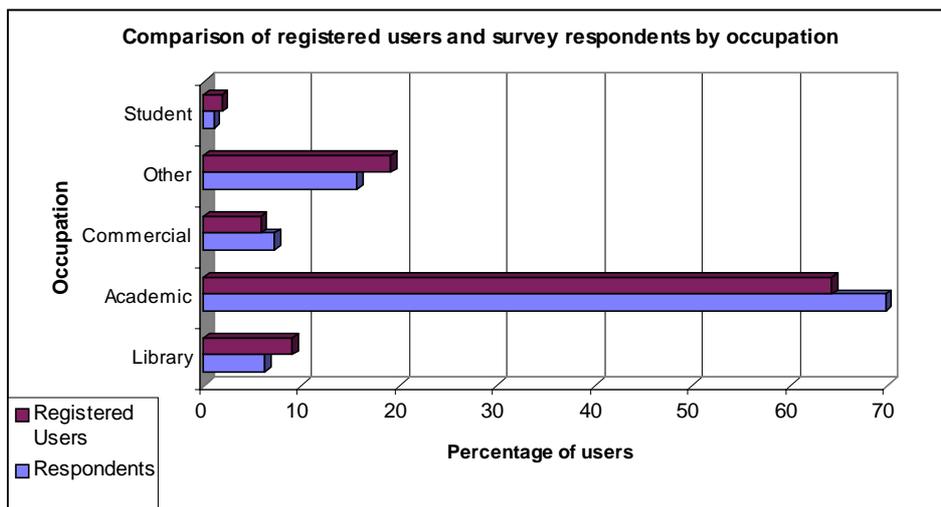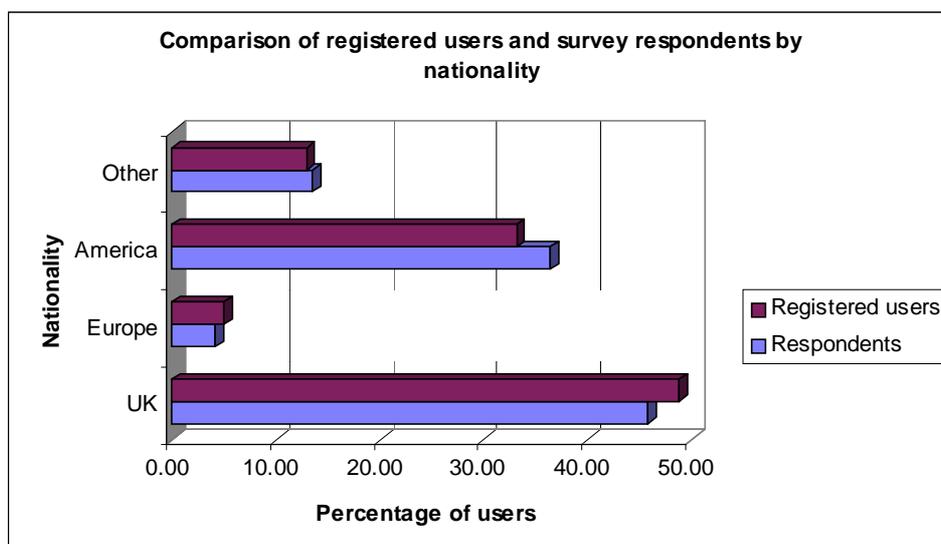|	Mar-98	Apr-98	May-98	Jun-98	Jul-98	Aug-98	Sep-98
ar	0	0	0	0	158	2011	2542
bm	0	1520	2676	4087	3123	2804	1784
bu	0	0	0	0	0	1337	1247
gm	0	546	120	5882	5658	4289	4711
nq	160	1297	2660	2740	2895	2484	3962
pt	0	0	5	2975	1162	804	1776

68.   The peak in June 1998 for both sets of figures corresponds to the publicity associated with the evaluation phase of the project.   In the four months June-September 1998 users accessed 181,577 Web pages in total (1,488 per day) and 58,181 page images (476 per day).   All titles attracted significant usage but *Gentleman's Magazine* (35% of accesses during the four months) and *Notes and Queries* (29%) were the most heavily used.   *Annual Register* and *The Builder* though not available for the

evaluation both attracted significant usage during August and September 1998 (15% and 9% respectively of page images accessed during those two months).

Questionnaire Survey

69.   377 Registered users were emailed a nine-question questionnaire on 2nd July 1998; respondents were invited to return the survey by email, fax or post with the deadline set for 17th July 1998. A reminder was sent out a week after the first emailing. Responses were received from 97 users (a 26% response rate). The 26 UK academics who responded to the survey together with 1 undergraduate student and 1 librarian were asked if they would be willing to take part in a telephone interview. In the event, 6 semi-structured telephone interviews were carried out with 5 academics and 1 librarian. A detailed description and analysis of user feedback appears in Appendix VI, together with the Web registration form and the questionnaire survey. The results are summarised below.

70.   The occupational and geographical distribution of registered users (377) and questionnaire respondents (97) are shown in the following bar charts. Academics were by far the largest occupational group (64% of registered users) and UK users (49%) were the largest geographical group though with a substantial US contribution (33%). The profiles of questionnaire respondents and registered users were broadly similar though respondents showed a slightly higher proportion of academics (70% against 64%) and of US users (36% against 33%) and a lower proportion of UK users (46% against 49%). We concluded that the respondents were representative of the registered user group as a whole, but there is no evidence as to whether the registered users were representative of the total (unknown) user population, other than the journal preference data summarised in para. 73.

Comparison of registered users and survey respondents by nationality

Comparison of registered users and survey respondents by occupation

Reasons for using the Service

71.   The survey data provided an overview of research and teaching interests of respondents using ILEJ. These fall broadly into the following categories:

- 18th and 19th century language and literature
- history of thought, religion and ideas
- history of science, technology, engineering and medicine
- history and biography of specific individuals
- genealogy
- sociology and social relations

72.   ILEJ is being used for a variety of purposes, including:

- supporting teaching on undergraduate and postgraduate courses
- as a tool for primary research
- as a tool for background research
- as a tool for checking known references
- supporting development of future research strands and directions

Journals used

73.   Respondents were asked which titles they had used.   The frequencies of use were similar to those exhibited by the Web usage data for page images (for all use), viz:

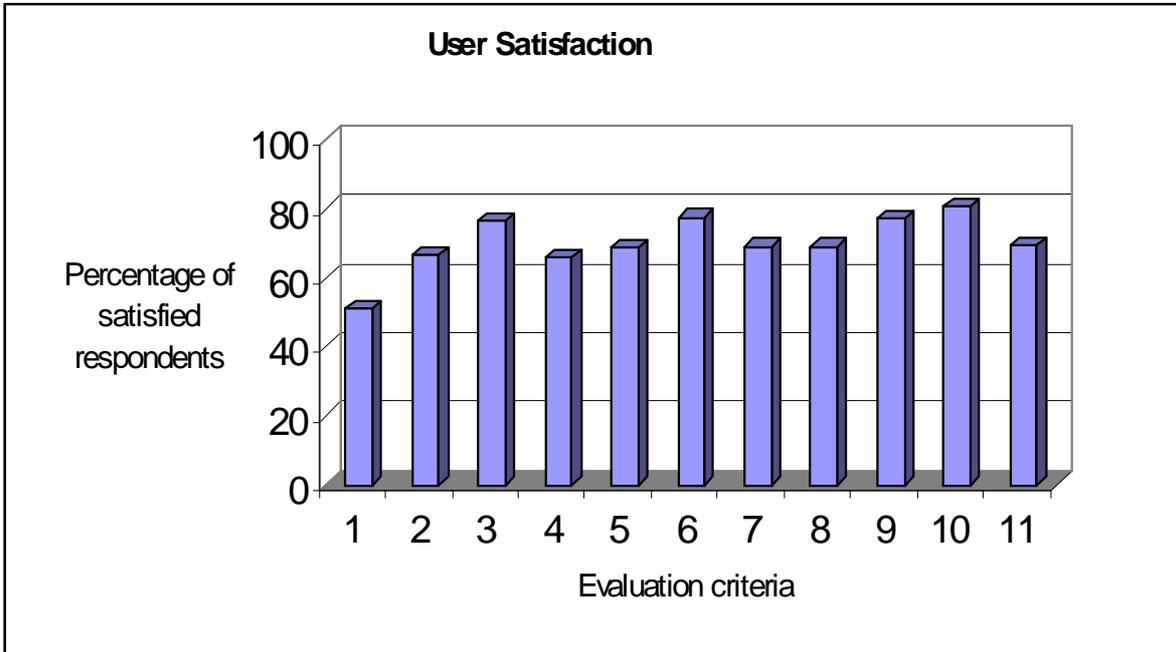|	Respondents	Web usage (June-September 1998)
Gentleman's Magazine	33%	40%
Notes and Queries	26%	23%
Blackwood's Magazine	28%	23%
Philosophical Transactions	13%	13%

66 (68%) Respondents had previously used the paper copies of volumes available in ILEJ. A further 19% had used paper copies of volumes of the same titles which were not covered by ILEJ.

Preferred method of finding information

74.   The preferred methods of finding information were simple search (47%), browsing (37%) and fuzzy searching (16%).

User Satisfaction with ILEJ

75.   Respondents were asked to indicate their satisfaction with various features of the service on a 1 (= poor) to 5 (= excellent) scale.   The responses for 11 characteristics of ILEJ are summarised below. Overall users expressed satisfaction with the service. Speed of access, which was given the least favourable satisfaction ranking, was given a score of 4 or 5 by 51% of respondents, and of 3, 4 or 5 by 78%. In all other cases at least 60% gave a 4 or 5 ranking and 88% a 3, 4 or 5. Clarity and legibility of images, ease of searching (simple search) and speed of search (fuzzy searching) were given high satisfaction ratings.   However, for each factor there were some users who expressed reservations, experienced serious problems or made critical comments which are of value in considering the future development of the service.

User Satisfaction

Percentage of satisfied respondents

(bar chart, vertical axis 0–100, horizontal axis "Evaluation criteria" labelled 1 to 11)

Evaluation criteria

Key:	
Speed of access to ILEJ	1
Ease of navigation	2
Clarity and legibility of the images	3
Ease of reading on screen	4
Image size	5
Ease of searching (simple search)	6
Speed of searching (simple search)	7
Usefulness of results (simple search)	8
Ease of searching (fuzzy search)	9
Speed of searching (fuzzy search)	10
Usefulness of results (fuzzy search)	11

76.   Comments included:

- some overseas users found access speeds to be unsatisfactory.

- the hierarchical organisation of the ILEJ required too many steps to reach a page image.

- there was a need for a consistent link back to the ILEJ Home Page from all ILEJ screens.

- favourable comments on the retention of the overall appearance of the original page (as opposed to a keyboarded text) and the inclusion of all content, including page numbers and advertisements.

- the legibility of some pages was poor and there was a need for better quality control.

- higher quality OCR could and should be provided.

- full text (OCR'd) searching should be available for all titles.

- mechanisms for refining a search, e.g. limiting by date, would be appreciated.

- fuzzy searching facilities were less heavily used than browse or simple search, but rated highly in terms of ease of searching, speed of searching and usefulness of results. Some users experienced difficulty in achieving satisfactory precision when using fuzzy searching.

A more detailed survey of user comments appears in Appendix VI.

Printing

76. Users were asked about their experiences of printing hard copy from ILEJ was also surveyed. Somewhat unexpectedly, 49% of respondents did not feel the need to print. Many are using the service to track down brief references rather than full articles, so the need for printing is lessened. A further 29% were able to print what they needed. A significant minority (22%) did not get satisfactory results when printing. Problems encountered included difficulties in image size when printed (images getting split over two pages rather than one), cropping of images so that part of the image is lost when printed and problems with slowness of printing.

Comparison of Paper and Electronic Versions

78. Overall, when asked about their preferences for paper or electronic versions of the titles, the majority of respondents felt that they were happy using a combination of paper and electronic, with a preference for electronic where available, though there was a continuing need for an occasional reference to print material. The value of electronic access was strongly recognised, with respondents citing additional flexibility, enhanced searching capabilities, ease of access and time saving in the research process as factors influencing their preference for electronic materials.

Conclusions

79. The feedback indicated general satisfaction with, and enthusiasm for, the service but identified limitations which need to be addressed in future services.

G : Cost Models

80. The identifiable expenditure on the project (Appendix VII) was £338,000 from eLib funding and an estimated £120,000 in institutional contributions of staff time (including overheads). The total of £458,000 represents an expenditure of £4.21 per indexed page image accessible on the Internet. This estimate of expenditure does not take into account the costs of the contribution of the IT and library infrastructures of the four Institutions.

Bound Volume Scanning Costs

81. The production line unit costs for creating the images were, however, much lower. The following spreadsheet shows the estimated costs for creating ready-to-mount, indexed and processed images for each of the four titles scanned from bound volumes. These costs represent production mode experience when development and experimental work had been largely completed and standard procedures established (with a qualification that the learning curve lasted throughout the project with procedures continuing to be refined). These costs do *not* include:

- the general management costs of the project
- initial experimentation and development costs to establish procedures for scanning and processing
- library and IT infrastructure costs for each institution
- costs associated with the purchase and maintenance of servers and mounting images on servers
- archive costs (para. 87)

82. The three main cost categories identified in the spreadsheet are:

- image creation including the preparation of the material prior to scanning, the identification of basic metadata, and the staff and equipment costs of scanning
- index creation costs either by OCR or keyboarding
- image conversion and the subsequent processing of both images and indexes.

Further details of the method of calculating costs appear in Appendix VII

83. The *absolute* values given in the table must be treated with caution.   They are dependent both on the accuracy with which staff costs are estimated, the costs that are included or excluded, and the extent to which procedures are still under development.   With this qualification, the data shows that the cost of creating an unprocessed image is in the range £0.20 - £0.21 for all four titles.   If we assume an error of ± 25% in the data, the cost range would be £0.15 - £0.27.   Similarly, the cost of creating a "ready-to-mount" indexed and processed image is in the range £0.39 - £0.74 (£0.29 to £0.93 assuming ± 25% error).   This broader range is largely the result of differences in index creation costs.

84. Of more significance are the *relative* costs between titles and between different processes which have been estimated on the same basis.   Specific comparisons to be noted, which identify factors which can influence costs, are:

- the three major cost elements are image creation, index creation and processing of both images and indexes.   Index creation costs exhibit the greatest variability of the three.

- the cost of image creation (£0.20 to £0.21) for four different titles and two operators in different locations varied very little. It represented between 26.9% and 51% of total costs.

- the cost of keyboarding indexes is directly dependent on index size (number of characters) and shows the widest variation from £0.26 per image for *Notes and Queries* to £0.02 for *Philosophical Transactions*.   The latter low value reflects the Table of Contents and simple subject indexes of an article-based journal, which differs substantially in character from the other titles in the project.   The cost for *Gentleman's Magazine* indexes (not shown in the table) was £0.08 per image.

- uncorrected OCR offers a greater depth of indexing than keyboarded indexes, usually at a much lower cost (£0.04-£0.05) per image, though with a significant error rate.

- the very low processing costs for *Blackwoods* (£0.134) in comparison with other titles (£0.239 to £0.253) arises because of uncomplicated pagination sequences and the use of PCI indexes (with OCR). As a result, html conversion and file indexing for Open Text were relatively simple.   In contrast the high figure for both these processes for the *Annual Register* reflected multiple pagination sequences and complex indexes.

- the foldout diagrams in *Philosophical Transactions* were the main reason for the high quality control cost (£0.03) compared with the other three titles (£0.01).

- the cost of image conversion varied from £0.04 (*Notes and Queries)* to £0.02 per page (*Annual Register* and *Blackwood's).*

- the cost for file structuring, naming and handling also showed wide variations from £0.07 for *Notes and Queries* to £0.03 per page for *Blackwood's.*

	Notes and Queries		Annual Register		Blackwoods		Philosophical Trans.	

No of images scanned	26254		12465		33183		18947	
Average number of pages scanned per hour	97		98		85		98	

Operation	Cost (£) per image	*% of total cost*	Cost (£) per image	% of total cost	Cost (£) per image	*% of total cost*	Cost (£) per image	*% of total cost*
1 Prescanning/raw metadata	0.02	*2*	0.02	3	0.02	*2*	0.03	*4*
2 scanning (staff)	0.12	*16*	0.13	21	0.14	*34*	0.12	*26*
3 Scanning (equipment)	0.06	*8*	0.06	10	0.06	*15*	0.06	*12*
Total image creation (operations 1 to 3)	**0.20**	***27***	**0.21**	**34**	**0.21**	***51***	**0.21**	***42***
4 OCRing	0.04	*5*	n/a		0.05	*15*		
5 Keyboarding of indexes	0.26	*34*	0.16	27	n/a*	----	0.02	*4*
Total index creation (operations 4 and 5)	**0.29**	***39***	**0.16**	**27**	**0.05**	***15***	**0.02**	***4***
6 Ftping and related activity	0.01	*1*	0.01	2	0.01	*2*	0.01	*2*
7 Image conversion	0.04	*5*	0.02	3	0.02	*5*	0.03	*6*
8 File structuring, naming and handling	0.07	*10*	0.04	7	0.03	*7*	0.05	*10*
9 Quality review	0.01	*1*	0.01	2	0.01	*2*	0.03	*6*
10 SGML conversion	0.07	*9*	0.10	16	0.04	*10*	0.10	*21*
11 File indexing (for Open Text)	0.05	*6*	0.05	8	0.02	*5*	0.03	*6*
12 Equipment and software for 6 -11	0.01	*1*	0.01	2	0.01	*2*	0.01	*2*
Total for image conversion and processing (operations 6 -12)	**0.25**	***34***	**0.24**	**39**	**0.13**	***33***	**0.25**	***53***
Total all operations (£)	**0.74**		**0.61**		**0.39**		**0.47**	

*Blackwood's used periodical Contents Index (PCI) records provided without charge by Chadwyck-Healey
These were not costed but the processing costs include the processing of these and other indexes.

Microfilm scanning costs

85.   The processes for image creation from microfilm were not costed at the same level of detail as those for bound volumes, largely because the experimental and development stages overlapped extensively with production line operation.   However, three factors increased the cost of raw image creation from microfilm, in comparison with that from bound volumes:

- the need to create a copy of the archival microfilm (£0.10 per page for *Gentleman's Magazine*).

- the slower throughput (para. 39) which increased scanning staff costs.

- the higher costs of equipment (£65k compared to £14k per unit) and associated maintenance, partially offset by an assumed (but not demonstrated) greater lifetime capacity for the microfilm scanner.

The overall effect of these three factors is to roughly double the cost from £0.20 to £0.40 per image.

Access costs

86.   The estimated cost of continuing to provide ILEJ on the present or an equivalent server, taking into account hardware and software capital and maintenance and staff costs, is c. £3,300 per annum, £0.03 per page per annum.   This excludes costs of the second EFS server at Leeds which provides the fuzzy matching facility.

Archiving costs

87.   Archiving costs have been excluded from the above discussion. The cost for storing data indefinitely, i.e. including capital replacement, running costs, support etc., is estimated at £20 per Gb per annum which gives a total cost for the ILEJ project (122 Gb) of c. £2,400 per annum, £0.02 per page per annum.

H : Exit Strategy

88.   ILEJ offers distributed access to UK material and is a contribution to the Integrated Information Environment envisaged in the CEI Content Working Group discussion paper **(4).**   It is a potential resource for two hybrid library projects in which ILEJ institutions are involved: BUILDER (Birmingham, Oxford) and MALIBU (Oxford). The ILEJ corpus also complements material of the same era provided by JSTOR (though the pre-20th century content of JSTOR is limited) and by Chadwyck Healey's LION project.   However, the small number of titles and relatively short runs limits the value of ILEJ as a resource.

89.   A proposal was submitted to JISC in November 1997 for an ILEJ-2 project which would use the experience gained in ILEJ to create a corpus of images of 800,000 pages of Victorian journals for the period 1837-1902, in the areas of history and literature broadly interpreted.   The core of this selection would be drawn from the titles listed in *The Wellesley Index to Victorian Periodicals, 1824-1900,* augmented by representative newspaper format journals such as *Punch,* political serials such as *Hansard's,* religious titles for each of the major denominations of the period, and scholarly journals such as *English Historical Review* which began publication in the mid- to late-19th century.   It was proposed that the JISC data centre, MIDAS, which had also been contracted to provide access to JSTOR, should provide access to this corpus and to the ILEJ images.   This proposal was unsuccessful.

90.   In the absence of funding for a follow-up project, our exit strategy requires that the Oxford Web site, together with the full existing browse and search facilities provided by Oxford and Leeds servers, should continue for *at least 12 months* from the end of the project (until August 1999).   During that period we would:

- continue to advertise the existence of the service as widely as possible to the user community, and to encourage links from other Web sites (VADS, Scout Report, etc.)

- publicise the project and the resulting corpus within the library and information community through conference presentations and published papers.

- seek further feedback from users on the level of continuing interest in the collection, the effectiveness of the retrieval and browsing tools and of the user interface, the demand for an expansion in the length of the runs and the number of titles, and the likely impact of charging to recover the cost of providing continuing access.

- maintain usage data.

91.   In the light of this feedback, we would seek to identify a permanent means of providing access, and to encourage the funding of an expanded corpus along the lines outlined in para. 89.   One of the possible options would be for Oxford to maintain a server offering access on a permanent basis beyond August 1999 to both ILEJ and other digitised material resulting from projects in Oxford funded by HEFCE and from other sources.

92.   The Oxford Hierarchical File Server will provide long-term storage for images and indexes created in the project (paras. 64, 87).

93.   The ILEJ corpus is one of the exemplars being used in the eLib3a CEDARS (CURL exemplars in digital archives) project, based at Oxford, Leeds and Cambridge, which is addressing issues relating to the preservation of digital material.

I : Achievements, Failures, Conclusions and Recommendations

94.   The achievements and failures of the project must be measured against the overall objectives of the project stated in para. 2 :

The project sought to enhance access to holdings of research libraries by creating electronic copies of print and microfilm holdings which could be accessed via the Internet.   The specific objective was to create and to provide user access to a corpus of digitised images from three 18th- and three 19th-century journals.

Specific issues to be investigated are outlined in para. 2.

Bound Volume Scanning

95.   The project has successfully defined the scanning parameters and procedures for using the Minolta PS3000 open book scanner to produce images of acceptable quality though with some variation in legibility, aesthetic quality and OCRability.   Specific problems identified and, as far as possible, resolved:

- page curvature linked to the skewing and distortion of volumes in the cradle especially at the beginning and end of volumes;
- differential focusing required at the beginning, middle and end of a large volume;
- page wrinkling;
- see-through from one page to another;
- varying page densities;
- varying type face quality.

96.   A throughput of 80-100 pages an hour, over three times that predicted in the original proposal (25 pages an hour), was achieved. The subsequent processing of images and addition of metadata required more time and resources than image creation itself
(para. 119). A total of 91,000 images from four titles (20 or 21 volumes of each) were created.

97.   Zeutschel Omniscan 5000 scanning equipment, purchased by HEDS during the second year of the ILEJ project (and other equipment now on the market) provides a much higher level of operator support

though at a higher capital cost (x 5). We would expect such equipment to provide a more consistent image quality and a higher throughput.

Microfilm Scanning

98. The project has successfully defined parameters and procedures for the Mekel (MX500XL-G) to produce images of acceptable, though variable quality. The project limited its objectives to the use of two pre-existing microfilms of one 18th-century and one 19th-century title. The problems encountered with scanning from bound volumes (para. 94) are replicated in a microfilm, which is itself created from a bound volume.

99. An effective average throughput of 40 (*The Builder*) and 70 (*Gentleman's Magazine*) images per hour were achieved (2 images = 1). This is less than that achieved with the Minolta and far below the throughput predicted in the original proposal (250 pages per hour) or that claimed in the manufacturer's specification (600 pages an hour at 200 dpi with grey scales). Throughput with the *Gentleman's Magazine* was limited by the need for frequent changes in the parameters and the problems in separating and cropping the two-image frames. Throughput with *The Builder* was limited by large image sizes, which resulted in extended cropping times and disc space bottlenecks.

100. Further insights into the use of microfilm as an image source are provided by three other projects:

- the Yale University Open Book project **(5),** which digitised 2000 books from microfilm using the bi-tonal Mekel equipment and the associated Cornell project (**6**).

- the Australian Co-operative Digitisation project (**7**), which is creating microfilm as an intermediate stage in digitisation.

- UMI's project to digitise 22 million pages from the Early English Books microfilms **(8)**.

101. In addition, publishers are examining a twin-track strategy in which microfilm is created for preservation and digital images for access. This strategy can be implemented either by first creating a microfilm copy then images from the microfilm or vice versa.

102. *We recommend that further investigations should be undertaken of:*

- *the respective effects of the paper original, the microfilming, the microfilm copying, and the imaging process on the quality of the final image.*

- *the use of state-of-the-art microfilming equipment to create microfilm from bound volumes as an intermediate stage in the creation of digital images, the practice being adopted by the Australian Co-operative Digitisation project (**7**). This approach can be justified either because microfilm technology can cope better than that of digital imaging with the bound volume, or by the requirement to create both microfilm and digital forms (para. 101)*

- *the relationship between image quality and variations in the characteristics of microfilm originals from a wide range of sources.*

Impact of Resolution on Image Quality

103. With the Minolta, a resolution of 400dpi bi-tonal gave satisfactory legibility for all titles. Greyscales were not available but, if they had been, might have improved the quality of OCR and made it possible to OCR other titles. With the Mekel, the *Gentleman's Magazine* required either 300dpi bi-tonal or 100dpi with greyscales, depending on the frame quality, to provide legible images. *The Builder* was scanned at a minimum of 200dpi with greyscales (bi-tonal scanning was not effective).

Processing

104. The following processes have been successfully implemented:

- the conversion of tiffs (bi-tonal) to gifs, and of tiffs (greyscale) to jpegs, and re-sizing for display purposes.

- the use of Sequoia ScanFix to deskew, despeckle and crop tiff images prior to OCRing.

Indexing for retrieval

105.   In comparison with the paper indexes, retrieval capability has been enhanced by two types of electronic index:

- OCR'd full-text
- electronic forms of printed indexes or contents pages.

The capability of both was increased by the availability of truncation and Boolean search facilities. Users expressed a relatively high level of satisfaction with search outcomes but no detailed analysis of retrieval capability was undertaken.

Optical Character Recognition

106.   Two 19th-century journals were successfully OCR'd to provide a low cost index (cf. para. 119 below) though at a level of accuracy far below that which would be required for display purposes. The OCR'd texts offer far greater depth of indexing than conventional indexes but with inevitable loss of precision and retrieval failures associated with the high OCR error rate, partly corrected by the use of fuzzy matching software. No quantitative estimates were made of the level of retrieval failures or of the reduction in failures arising from the use of fuzzy matching.

107.   The OCRing process using Omnipage has provided the opportunity to examine isolated problem areas such as:

- small typefaces;
- pages displaying extremes of typeface density;
- pages with complex structure, e.g. mixed fonts.

108.   *We recommend that:*

- *studies should be undertaken to quantify the retrieval failure rate arising from imperfect OCR, the improvements that can result from the use of fuzzy matching and the acceptability to the user of this error rate.*

- *methods of improving OCR quality should be explored. Strategies that might reasonably be expected to result in improvement are:*

    For microfilm products:

    - *use of high quality microfilm obtained by state-of-the-art microfilm technology.*

    - *by scanning from the original paper copies with bound volume or flatbed scanners, i.e. abandoning the microfilm as a source material.*

    For paper products:

    - *use of greyscales on the Minolta*

- *use of the Zeutschel or other higher specification (but higher cost) equipment; either bi-tonal or greyscales.*

- *use of a flatbed sheet feed scanner and dismembered volumes, i.e. abandoning bound volume scanning.*

For either:

- *use of alternative OCR software (not identified in the project). more able to cope with the special characteristics of 18th- and 19th-century typography.*

Conversion of the printed index

109.   The project demonstrated that satisfactory indexes for retrieval could be created by the conversion to electronic form of printed indexes published contemporaneously with the original.   Two conversion methods were used:

- keyboarding by an outside agency from the printed originals
- in one instance (*Blackwoods*) the use of contents pages already in electronic form and provided by Chadwyck-Healey from the Periodical Contents Index (PCI)

Metadata

110.   The complexity of providing appropriate metadata for 18th and 19th century originals was under-estimated. Valuable lessons were learnt, in particular the need to check and verify in advance the pagination of the originals and to allow for the administrative overhead required at the planning stage in order to define a metadata formula for each individual title.   In the journals covered in the project there were examples of pages without numbers, duplicate paginations, inserts, plates and differences in the composition of duplicate volumes in different institutions.

111.   The project created a flexible, extendible metadata structure based on SGML files and on two Document Type Definitions (DTDs):   the Encoded Archival Description (EAD) and the Text Encoding Initiative (TEI).   The hierarchical structure of an SGML-encoded document mirrors the structure of the original journals, making the creation of their virtual surrogates simpler and more logical.   The metadata provides for unique identification of each image and for links to the bibliographic description of the paper copy and to a range of indexes.
It includes Dublin Core compliant metadata in the HTML <META> tags.   No administrative metadata (resolutions, compression systems used, etc.) were included.

112.   *We recommend that future digitisation projects should consider:*

-      *the use of SGML or XML based metadata structures;*
-      *the inclusion of administrative metadata (which ILEJ did not include).*

Data transfer between sites

113.   Procedures were established for routine image transfer between sites by ftp over SuperJanet. Images created at Manchester, Birmingham or Oxford could then be processed at Leeds or Oxford and archived at Oxford.   Because of capacity limits on local stage buffers, image transfers had to be carefully co-ordinated to prevent overflow of disc reservoirs at the draft production stages.   This required careful timetabling and transfer failures could cause major delays because of the need to wait for the next free timetable slot.

Presentation to the User

114. The arrival of a Web version of EFS combined with the increasing domination and universality of Web browsers led to the abandonment of X-Windows as a delivery platform. Web servers were configured at Leeds (using EFS) and Oxford (using the PAT (Opentext) search engine). All user access was to the Oxford *url* with a transparent link to Leeds from which images were downloaded as required. A user interface was developed which offered browse, simple search and fuzzy search capabilities.

115. ***We recommend that a further revision of the user interface be undertaken in the light of feedback from users.***

The Users' View

116. Once all journals were available a substantial level of user activity was generated with 370 users registering and a (probably much) larger number using the site. In June 1998 16,000 page images were accessed with all titles attracting some activity but *Gentleman's Magazine* (29% of accesses) and *Notes and Queries* (25%) the most popular.

117. The evaluation of user responses to the service revealed general satisfaction with respect to content, speed of access, ease of navigation, legibility of image and search facilities. Areas in which the service was criticised or improvements requested included:

   - a more direct (less clicks) route to an individual page of a journal
   - extension of full-text searching to all titles
   - methods for limiting a search, e.g. date
   - expanded content in the form of longer runs of the six journals or more journals, especially for the Victorian period

   Surprisingly, printing was not a major issue with more than half the respondents not feeling a need to print.

118. ***We recommend an extended evaluation of user reactions to content, interface and search facilities in this and in other projects, noting that this is already being done in the SuperJournal project, and that the ILEJ Consortium would hope to undertake some further evaluation during the next 12 months, and would continue to record usage figures.***

Costs

119. An analysis was undertaken of the production-line costs of creating indexed, processed, ready-to-mount images. This analysis excluded the costs of general project management, experimentation and development, providing access, archiving and the contribution of the IT and library infrastructure of the institutions. The results of this analysis are indicative, not definitive, but do demonstrate that:

   - production-line costs are in three broad categories: image creation, index creation and the processing of images and indexes.

   - the total cost of producing indexed and processed images varied from £0.39-£0.74 per image. The major cause of the wide cost range was the variation in indexing costs from £0.02 to £0.29 per image. Processing costs could also vary as a result of differences in the complexities of pagination and indexes.

   - the cost of image creation alone from bound volumes was £0.20-£0.21 per page image and represented between 27% and 51% of total cost. With the materials used in the ILEJ project, the creation of images from microfilms was approximately twice that from bound volumes, and arguably the quality was lower.

   - uncorrected OCR is a genuine low-cost option (£0.04-£0.05 per image) for indexing which can provide greater depth of indexing at lower cost than the keyboarding of printed indexes (with certain exceptions) though with a significant error rate. At the error rate experienced in the

project, the correction of the OCR would undoubtedly have resulted in a many-fold increase in cost to a level similar to that for re-keyboarding of the text (at £0.70-£0.80 per 1,000 characters).

- the ongoing costs of continuing to provide access and of archiving were estimated at £0.03 and £0.02 per image respectively.

Use of bound volumes

120. The project demonstrated that bound volumes (or microfilms of bound volumes) could be used to provide legible images. Apart from initial testing in the early stages of the project, no direct comparisons were made with the use of dismembered volumes on flatbed or sheet feed equipment. However, we are satisfied that the use of dismembered volumes (as in JSTOR) would result in higher throughput, lower cost and higher quality.

121. *We recommend that the issue of using dismembered volumes at least for 18th- and 19th-century material which is still widely available, should be addressed by JISC, CURL and the UK academic community, recognising that the acceptability of this option would decrease with increasing age and rarity of the material.*

Content and Critical Mass

122. The objective of ILEJ was to offer access to a critical mass of journal runs but there is no quantitative definition of what constitutes *critical mass*. Feedback so far indicates that the amount of material was sufficient to be of interest, but there was a demand for expansion in terms of the length of journal runs and the range of titles (para. 117).

123. An unsuccessful bid was made to eLib for an ILEJ-2 project to create 800,000 page images of journals from the Victorian period, based on those listed in the Wellesley Index. There is also (cf. para. 88) considerable commercial interest in offering electronic access to pre-20th century material from publishers such as Chadwyck-Healey and UMI. Hitherto, the electronic forms had been created by keyboarding of full-text, not image capture, and complete journals had not been included. However, the UMI Early English Books Project is creating images from UMI's microfilm.

124. *We recommend that:*

- *the demand in the academic community for access to 18th- and 19th-century journal runs should be further evaluated*

- *in the light of that feedback, the options for developing a larger corpus of journal material should be explored. A collection of journals published in the Victorian period would be one of the options.*

Project Management

125. The work of ILEJ was distributed across four sites with each making a substantial contribution and having technical or clerical staff employed on the project. Project management was vested in the joint Project Leaders, based at Leeds and Oxford, and the other members of the Project Executive (para. 11, Appendix II). This structure was successful in encouraging team working and providing effective co-ordination among the sites. The project benefited substantially from being able to draw on a wide range of expertise on all four sites to facilitate successful problem solving on technical and management issues.

126. As a result of the project, a wide range of expertise, knowledge and experience has been acquired by individuals in four major research libraries. These skills will be put to use in further projects, e.g. hybrid libraries, digital preservation, in which the institutions are engaged, and in the development of in-house and community-wide digitisation strategies. However, skills of scanner operators, employed only during the project, were lost at the end of the project.

127. The project has been publicised through papers and presentations (Appendix III) which have contributed to the development of digitisation practice and strategy.

128. The original proposal assumed that a particular technology – greyscale scanning capability – would be available at an early stage.   The most serious management problems encountered in the project arose because this assumption proved incorrect.   The project was reasonably successful in "working round" the delays in the availability of greyscales (para. 15), but this did cause a major distortion of the project timetable.   As a result, staff resources were spread thinly and there were insufficient time and resources to deal with some of the technical issues which were encountered at a later stage in the project than originally planned.   In particular the evaluation phase was compressed into the final three months of the project.

129. ***We recommend that project plans should take into account the implications of basing projects on*** *promised* **(*however firmly*) *rather than*** *actual* ***technology, even though doing so may be unavoidable.***

130. Other management and resourcing issues were:

- in comparison with a single-site project, the need to co-ordinate across four sites slowed progress and distribution of limited staff resources reduced flexibility.

- project management was vested in senior staff involved in but not funded by the project, with other full-time responsibilities. Combined with the multi-site character of the project this reduced the speed of response to crises and to the need to re-direct resources.

131. ***We recommend that a full-time Project Manager, funded by the Project, should be appointed for multi-site projects of this size.***

132. In addition to the creation of 120,000 images (110,000 actually achieved), the project envisaged index creation by both OCR and keyboarding, the design of metadata, configuration and maintenance of two servers, the provision of Internet access and evaluation.   In retrospective these multiple objectives may have been over-ambitious in relation to resources requested. The thinking behind the original proposal and the resource allocation were dominated by image creation issues and, in particular, equipment costs which represented 56% of total project expenditure.   The three person years of research staff time, even with substantial additional input from members of the executive, was barely adequate to meet the project objectives.

J : Summary of Recommendations from Section H

The recommendations made in the previous section (H) are reproduced below. Paragraph numbers from Section H are given in parentheses.

133. (102) ***We recommend that further investigations should be undertaken of:***

- ***the respective effects of the paper original, the microfilming, the microfilming copying, and the imaging process on the quality of the final image.***

- ***the use of state-of-the-art microfilming equipment to create microfilm from bound volumes as an intermediate stage in the creation of digital images, the practice being adopted by the Australian Co-operative Digitisation project (7). This approach can be justified either because microfilm technology can cope better than that of digital imaging with the bound volume, or by the requirement to create both microfilm and digital forms (para. 101).***

- ***the relationship between image quality and variations in the characteristics of microfilm originals from a wide range of sources.***

134. (108) ***We recommend that:***

- *studies should be undertaken to quantify the retrieval failure rate arising from imperfect OCR, the improvements that can result from the use of fuzzy matching and the acceptability to the use of this error rate.*

- *methods of improving OCR quality should be explored. Strategies that might reasonably be expected to result in improvement are:*

  For the microfilm products:

  - *use of high quality microfilm obtained by state-of-the-art microfilm technology.*

  - *by scanning from the original paper copies with bound volume or flatbed scanners, i.e. abandoning the microfilm as a source material.*

  For paper products:

  - *use of greyscales on the Minolta*

  - *use of the Zeutschel or other higher specification (but higher cost) equipment; either bi-tonal or greyscales*

  - *use of a flatbed sheet feed scanner and dismembered volumes, i.e. abandoning bound volume scanning.*

  For either:

  - *use of alternative OCR software (not identified in the project) more able to cope with the special characteristics of 18th- and 19th-century typography.*

135. (112) *We recommend that future digitisation projects should consider:*

- *the use of SGML or XML based metadata structures;*
- *the inclusion of administrative metadata (which ILEJ did not include).*

136. (115) *We recommend that a further revision of the user interface be undertaken in the light of feedback from users.*

137. (118) *We recommend an extended evaluation of user reactions to content, interface and search facilities in this and in other projects, noting that this is already being done in the SuperJournal project and that members of the ILEJ Consortium would hope to undertake some further evaluation during the next 12 months, and would continue of record usage figures.*

138. (121) *We recommend that the issue of using dismembered volumes at least for 18th- and 19th- century material which is still widely available, should be addressed by JISC, CURL and the UK academic community, recognising that the acceptability of this option would decrease with increasing age and rarity of the material.*

139. (124) *We recommend that:*

- *the demand in the academic community for access to 18th- and 19th- century journal runs should be further evaluated*

- *in the light of that feedback, the options for developing a larger corpus of journal material should be explored. A collection of journals published in the Victorian period would be one of the options.*

140. (129) ***We recommend that project plans should take into account the implications of basing projects on*** *promised* ***(however firmly) rather than*** *actual* ***technology, even though doing so may be unavoidable.***

141. (131) ***We recommend that a full-time Project Manager, funded by the Project, should be appointed for multi-site projects of this size.***

Acknowledgements

A large number of organisations and individuals contributed content, ideas or support to the
project. We would especially like to acknowledge:

Cambridge University Library, and Peter Fox, the Librarian, and Elizabeth Harrison in particular, for providing a copy of the Mellon microfilm of the *Gentleman's Magazine* (1731-1830).

Manchester Public Libraries for providing a copy of *The Builder* microfilm (1843-52).

Birmingham Public Library for providing volumes of the *Annual Register* which were missing from the Birmingham University holdings.

Chadwyck-Healey Ltd. For the supply of the Periodical Contents Indexes (PCI) for the volumes of *Blackwoods* used in the project, and Sir Charles Chadwyck-Healey and Michael Healey for useful discussions on image creation from microfilm.

UMI for the gift of a microfilm copy of *The Builder* (1853-55).

Esteem Computers for assistance in testing of the EFS software.

Michael Alexander (British Library) for advice

Primary Source Media [now a part of the Gale Group], and Mark Holland in particular, for the offer of microfilm volumes of the *Annual Register.*

The organisers and contributors to the Digitisation Workshops at Cornell University Library, which were attended by two project staff.

Ross Coleman and Colin Webb from the Australian Co-operative Digitisation (Ferguson) project.

All users of the service who contributed to the evaluation.

References

1.   Kenney, A.R. and Chapman, S.   *Digital imaging for libraries and archives.* Cornell University
        Library, Department of Preservation and Conservation.   Ithaca, NY 1996.
        ISBN : 85604 207 3

2.   Coleman, J. *SGML as a framework for digital preservation and access.*   Commission on Preservation
        and Access. Washington DC 1997.   ISBN 188 7334 513

3.   *Library of Congress American Memory DTD for historical documents.*
        http://lcweb2.loc.gov/ammem/amdtd.html

4.   *An integrated information environment for higher education : developing the Distributed*,
        National Electronic Resource (DNER).   Committee on Electronic Information (CEI) – Content
        Working Group. December 1997.   http://www.jisc.ac.uk/cei/dner_colpol.html

5.   Conway, P.   *Conversion of microfilm to digital imagery* : *a demonstration project: performance report
        on the production conversion phase of Project Open Book.* Yale
        University Library. New Haven, Conn. 1996.

6.   Kenney, A.R. *Digital to microfilm conversion : a demonstration project, 1994-6. Final report to the
        National Endowment for the Humanities.*   Cornell University Library. Department of Preservation and
        Conservation. Ithaca, NY 1996. http://www.library.cornell.edu/preservation/pub.htm

7.   *Electronic Alchemy : The Australian Co-operative Digitisation Project 1840-45.* National Library of
        Australia.   http://www.nla.gov.au/ferg/jthomp.html

        Also: Webb, C.   *The Ferguson project : a hybrid approach reformatting rare Australiana.*
        http://www.nla.gov.au/nla/staffpaper/cwebb1.html

8.   ProQuest. *Digital Research Collections.*   http://www.umi.com/demo/RCDigital/