The Internet Library of Early Journals: an Electronic Library of Primary Sources on the Internet

Richard Gartner

rg@bodley.ox.ac.uk

Introduction

The Internet Library of Early Journals (ILEJ) owes its existence to the UK Higher Education Libraries Review, chaired by Sir Brian Follett, the report of which appeared in 1993. The "Follett Report", as it is generally called[1] , discussed the implications of information technology in enhancing the work of libraries and proposed a programme of development in key areas of IT within the higher education library sector. These included electronic publishing (including digitisation), on-demand publishing, electronic reserves, resources access and document delivery. Approximately £15 million was made available over 3 years for phases 1 and 2 of the programme, and institutions were invited to bid for funding to finance projects within these areas.

The ILEJ project resulted from a successful bid to the eLib programme by a consortium formed by the Universities of Oxford, Leeds, Manchester and Birmingham. The aim of the project as stated in the bid was to enhance usage of the holdings of research libraries by creating electronic copies of their contents, and to provide user access to a corpus of digitised images from three 18th- and three 19th-century journals. The intention was to provide a critical mass of material, and to investigate the technologies involved in producing such an extensive archive of digitised images.

The six titles chosen for the project were:-
Notes and Queries (1849-69)
Blackwood's Edinburgh Magazine (1843-63)
The Builder (1843-62)
Gentleman's Magazine (1731-50)
Philosophical Transactions of the Royal Society (1757-77)
Annual Register (1758-78)

These titles and date ranges were selected following discussions with members of the academic community, according to a range of inter-related criteria, including the width of the subject range which they covered, perceived user demand, and also the extent to which they could test a range of technological variables (such as resolution, pixel-depth and the possibility of using optical-character recognition to provide full-text retrieval).

From its inception, the ILEJ project has been a widely-distributed operation, each centre taking responsibility for specific areas of operation. The Universities of Birmingham and Manchester were responsible for the scanning of the original bound volumes, whilst all microfilm scanning took place at Oxford. Leeds was responsible for the processing and mounting of images from original volumes, and for maintaining a fuzzy-search engine to allow the full-text searching of some journals. Oxford took responsibility for processing and mounting all

microfilm-scanned images, for the long time archiving of all the project's images and metadata, and for the design and maintenance of its primary WWW site.

Scanning

The material to be digitised was scanned from bound paper originals in the case of four journals, and from microfilm intermediaries in the case of the remaining two: the intention behind this mixed-media approach was specifically to test the differing requirements of these original formats.

A major concern of the scanning procedures was to ensure that they would be non-destructive and would follow sound conservation practice: therefore it was decided to scan from the bound originals rather than use single sheets from dismembered volumes, as is done by other digitisation projects such as JSTOR[2]. To do this we used an open-book cradle scanner, the Minolta PS3000P, which is designed specifically to handle bound volumes, is non-tactile and incorporates de-skewing functions to correct for the curvature of bound spines. All images were scanned at 400dpi (the highest resolution of which the scanner is capable) and at 1 bit-per-pixel.

The microfilm scanning was carried out on a Mekel MX500XL-G scanner, the first production model of its kind to incorporate greyscale facilities. In general, we scanned at 300dpi bi-tonal, although where necessary we would scan individual images (or most images in the case of one title) at 100dpi or 200dpi greyscale. Greyscale scanning was necessary when the original image was blotchy or otherwise disfigured, or when very small fonts were used (as was the case with *The Builder*).

Both scanning technologies presented some problems: in the case of the open-book scanner, greyscale scanning, which was listed as a feature of the machine's specification, proved unachievable without an upgrade which was not installed during the lifetime of the project. The microfilm scanning suffered from deficiencies in the originals scanned: in some cases, the scanner's automatic edge detection malfunctioned because of problems with the images, in other cases variable contrast levels within the originals made the choice of an ideal level when scanning difficult to achieve. Because of these frequent problems, the microfilm scanning proceeded much more slowly than was initially expected.

After scanning, the original TIFF images produced were archived at Oxford in their uncompressed form. For mounting on the WWW they were converted to GIF format (in the case of bi-tonal images) and JPEG (for greyscale originals). Images scanned from microfilm, which consisted of two pages per frame when scanned, were split into separate images for each page: in the majority of cases, where the gutter dividing the pages was centred on the image, this could be done automatically, but otherwise cropping was undertaken manually by the scanner operator.

Metadata

In any digitisation project of this kind, the scanned images themselves represent only a part of the completed "digital library". To form any coherent collection, they must be accompanied by suitable metadata. For this project, metadata included basic bibliographic information for a journal volume as a whole and for each constituent image, subject, author and titles indexes for some volumes, and, in the case of two journals, full-text produced by optical-character recognition (OCR) software.

Metadata for the ILEJ project are held in SGML (Standard Generalised Mark-up Language) files. Several reasons prompted this choice in preference to a proprietary database format, including SGML's independence of any given application, its archival robustness, its ability to encode a complex web of metadata into a single structure, and its hierarchical structure, which neatly mirrors the structure of the original materials.

Two *Document Type Definitions* (DTDs) are used in the project: the *Encoded Archival*

Description (EAD), an application designed to encode archival finding aids, is used to describe the structure of the collection as a whole, and the *Text Encoding Initiative* (TEI), perhaps the most widely used DTD for general text mark-up, is used to describe each individual volume within the collection.

The EAD was designed for the encoding of archival finding aids, and is therefore well suited for the description of collections as a whole. The ILEJ project uses a single EAD file to describe the entire virtual library: its header provides basic metadata for the collection as a whole (and information on the ILEJ project itself), and its container description elements are used to describe the titles and individual volumes of which the project is comprised. Each journal title is assigned to a CO1 element, and the constituent volumes for each title are given a C02 element nested within their corresponding C01. A link is made to the TEI file corresponding to each volume listed by means of the EAD's EXTPTR element. This provides a very simple structure that neatly mirrors the intellectual arrangement of the virtual collection. Example 1 shows the basic structure for the description of a journal title in the EAD file.

```
<C01 ID="ILEJ.1">
    <DID>
         <UNITTITLE>Notes and Queries</UNITTITLE>
    </DID>
    <C02 ID="ILEJ.1.1">
         <DID>
              <UNITDATE>1849-1850</UNITDATE>
              <UNITID>Volume 1</UNITID>
              <UNITLOC><EXTPTR ENTITYREF="nq.1849-1850.x.x.1.x"></UNITLOC>
         </DID>
     </C02>
     <C02 ID="ILEJ.1.2">
         <DID>
              <UNITDATE>1850</UNITDATE>
              <UNITID>Volume 2</UNITID>
              <UNITLOC><EXTPTR ENTITYREF="nq.1850.x.x.2.x"></UNITLOC>
         </DID>
     </C02>
</C01>
```

Example 1: Sample from the ILEJ Project's central EAD file

Each physical volume of the journal titles digitised in the project is represented by a corresponding TEI file, in which all the metadata (and in the case of two journals, the full-text) of that given volume are encoded, with, of course, pointers to the images themselves. The TEIHEADER, designed as a container for metadata, contains bibliographic information on the volume as a whole, including imprint information. The TEXT element contains a series of DIV elements, each corresponding to a single image. Within each DIV is a short bibliographic description of that image (often just a page number), and a pointer to the image file itself. The DIV element also includes OCR'd text for the two titles for which this is available, and index entries derived from the original printed indexes.
Example 2 shows the header and two sample DIV elements for a journal volume.

```
<TEI.2 ID=”ar.1767.x.x.10.x”>
            <TEIHEADER TYPE=”volume”>
              <FILEDESC>
                <TITLESTMT>
                  <TITLE>The Annual Register - Volume 10: metadata file</TITLE>
                </TITLESTMT>
                <PUBLICATIONSTMT>
                  <PUBLISHER>Internet Library of Early Journals</PUBLISHER>
                </PUBLICATIONSTMT>
                <SOURCEDESC>
          <BIBL>
                <TITLE>The Annual Register, or a view of the History, Politicks,
and Literature, for the year 1767</TITLE>
                <TITLE TYPE=”short”>The Annual Register</TITLE>
              <IMPRINT>
                <PUBLISHER>J. Dodsley in Pall-Mall</PUBLISHER>
                <PUBPLACE>London</PUBPLACE>
                <DATE>1768</DATE>
                <BIBLSCOPE TYPE=”volume”>10</BIBLSCOPE>
              </IMPRINT>
          </BIBL>
                </SOURCEDESC>
              </FILEDESC>
              <PROFILEDESC>
                <CREATION><DATE>14 July 1998</DATE></CREATION>
              </PROFILEDESC>
            </TEIHEADER>
            <TEXT>
              <BODY>

                <DIV ID=”ar.1767.x.x.10.x.x.u1” TYPE=”page”>
                <MILESTONE N=”Title Page” UNIT=”title”>
                <P><FIGURE ENTITY=”ar.1767.x.x.10.x.x.u1”>
                      <FIGDESC>
                          <BIBL>
                           <BIBLSCOPE TYPE=”page”>Title Page</BIBLSCOPE>
                         </BIBL>
                      </FIGDESC>
                   </FIGURE>
                </P>
               </DIV>
              <DIV ID=”ar.1767.x.x.10.x.x.u2” TYPE=”page”>
               <MILESTONE N=”Preface” UNIT=”title”>
                <P><FIGURE ENTITY=”ar.1767.x.x.10.x.x.u2”>
                      <FIGDESC>
                          <BIBL>
                           <BIBLSCOPE TYPE=”page”>Unnumbered Page</BIBLSCOPE>
                         </BIBL>
                      </FIGDESC>
                   </FIGURE>
                </P>
               </DIV>
```

Example 2: TEI Header and first two DIV elements for volume 10 of Annual Register

A key element of the SGML approach to metadata adopted by ILEJ is the use of unique identifiers to mark each DIV element: for instance, an ID of the form

ar.1767.x.x.10.x.x.23

is composed of eight units as follows:-

•Journal title	ar	= Annual Register
•Year	1767	= 1767
•Month	x	= not applicable
•Day	x	= not applicable
•Volume number	10	= volume 10
•Part/issue number	x	= not applicable
•Pagination sequence	x	= not applicable
•Page number	23	= page 23

These IDs, which are generated automatically when the SGML files are initially compiled, provides a unique label for every image scanned. They become particularly useful when additional information, compiled from disparate sources, is integrated into the SGML files: they were, for instance, an essential element in the conversion of the original printed indexes for most journals as described below.

The approach adopted in this project emphasises the physical rather than the intellectual structure of each volume: each page of the original is represented by a DIV element within the TEI file, and the cross-cutting intellectual hierarchy is represented by empty MILESTONE tags which delimit the boundaries of conceptually discrete units, such as weekly issues or articles. An alternative approach is to emphasise the intellectual structure of the text, using DIV elements to distinguish conceptually discrete components (such as individual journal issues, articles etc), and embedding phrase-level FIGURE elements within the text to point to image files.

The former approach (adopted by ILEJ) has the virtue of simplicity, and is relatively easy to implement using data from conventional relational databases, but is not ideal if the TEI file is intended to form the basis of a fully-fledged marked up text (such as a critical edition). A later SGML-based project at Oxford is, therefore, experimenting with the latter approach, with some degree of success.

Conversion of Printed Indexes and Full Text

An important value-added component in the ILEJ project is the provision of machine-readable indexes for most of the journals offered, and in the case of two titles, full-text searching. Boolean searching is offered for all indexes and for the full-text when available: in addition, a fuzzy search option is available for full-text searches, utilising Excalibur EFS software based at Leeds University (although this is limited for licensing reasons to five simultaneous users).

A variety of subject indexes and tables-of-contents were keyboarded and incorporated into the project's SGML files. For *Notes and Queries*, for instance, the original subject indexes to each volume were keyboarded, for *Blackwood's Edinburgh Magazine* author and title indexes were converted, and for the *Gentleman's Magazine* the cumulated index for volumes 1-20 was incorporated into the project's metadata.

The keyboarding operation was outsourced to a company specialising in such large-scale conversion projects: in addition to keyboarding, they also marked up the text with SGML-like tagging, specified by ourselves, which delineated primary and secondary terms, and volume and page references. Using Microsoft Word or Emacs macros, these keyboarded terms were

readily converted into TEI-conformant tags and incorporated into the relevant file for each volume.

The TEI allows several ways to include subject indexes within a text: we adopted the approach of using the INDEX element, which is inserted within the DIV element for the portion of text indexed. Incorporating these tags was greatly facilitated by the assignment of a unique identifier to each DIV element within the collection: it proved relatively simple to generate the identifier for each INDEX tag using the volume and page information keyboarded with each entry, and then to slot the entry into the correct DIV element. The only major problem experienced arose from the multiple pagination sequences used by the *Annual Register*: each volume could contain up to three separately paginated sequences, delineated typographically in the original volumes - considerable manual editing was necessary to accommodate these idiosyncrasies.

Example 3 shows some sample subject index entries and their placement within a DIV element.

```
<DIV ID="ar.1767.x.x.10.x.x.221" TYPE="page">
   <P><FIGURE ENTITY="ar.1767.x.x.10.x.x.221">
<INDEX INDEX="subject" LEVEL1="Land tax" LEVEL2="for 1767">
<INDEX INDEX="subject" LEVEL1="Loan" LEVEL2="in 1767">
<INDEX INDEX="subject" LEVEL1="Malt, mum, cyder, and perry"
LEVEL2="the          money raised by this tax in 1767">

<INDEX INDEX="subject" LEVEL1="Militia" LEVEL2="the provision made for
this part of the military service for the year 1767, out of the
national supplies">

<INDEX INDEX="subject" LEVEL1="Pensions, &amp;c" LEVEL2="the duty
assessed in 1767">
                     <FIGDESC>
                         <BIBL>
                          <BIBLSCOPE TYPE="page">Page 221</BIBLSCOPE>
                        </BIBL>
                     </FIGDESC>
                  </FIGURE>
               </P>
              </DIV>
```

Example 3: Index entries for the Annual Register, encoded in the TEi

In the case of two titles, *Blackwood's Edinburgh Magazine* and *Notes and Queries,* the quality of the original type and the scanned images was clear enough to make OCR feasible, so allowing us to provide a full-text searching facility. There were, however, not enough resources to correct the OCR'd text, and so this facility was offered with the stated proviso that the searchable text could well be inaccurate. A Boolean search facility is offered for the full text, although this is less useful than a more fully marked-up text would allow: an AND search, for instance, will retrieve images containing both search terms located anywhere on the same page, which in most cases is of little value.

A further approach adopted to circumvent to some extent the problems of uncorrected OCR is the use of fuzzy searching. A gateway is provided to a server running Excalibur's EFS

software, a proprietary package designed specifically for the retrieval of uncorrected OCR text. Despite the problems of poor precision inherent in fuzzy searching, and the rather slow and confusing interface necessitated by the EFS software (which allows only basic customisation), this facility has proved popular with users.

The Interface

The interface to the ILEJ journals was mostly designed in-house, the exception being some elements of the fuzzy searching interface which were prescribed by the EFS software. A small number of scripts written in Perl, in conjunction with the Opentext 5 search engine, are used to provide browsing and searching facilities, and to convert SGML metadata to HTML for output to the WWW.

The browsing facility operates first of all on the EAD file which describes the structure of the collection as a whole. This allows the user to select a given volume for a title, at which point a link is made to the corresponding TEI file for that particular volume. The software then presents the user with a list of major divisions within the volume (which are defined within the TEI MILESTONE tags): these are usually individual issues, or clearly delineated sections, within a volume. From here, the user can choose the individual page desired for viewing. Most images are formatted to screen width initially, but can then be magnified to their full size if desired.

Figures 1 and 2 show the browsing and image viewing facilities in operation.

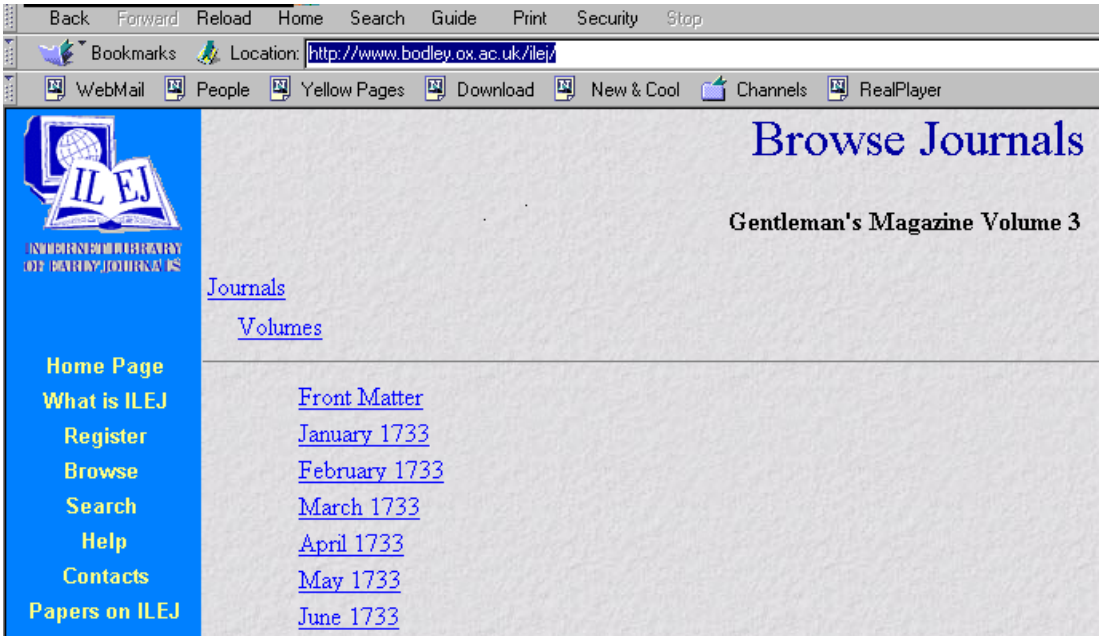
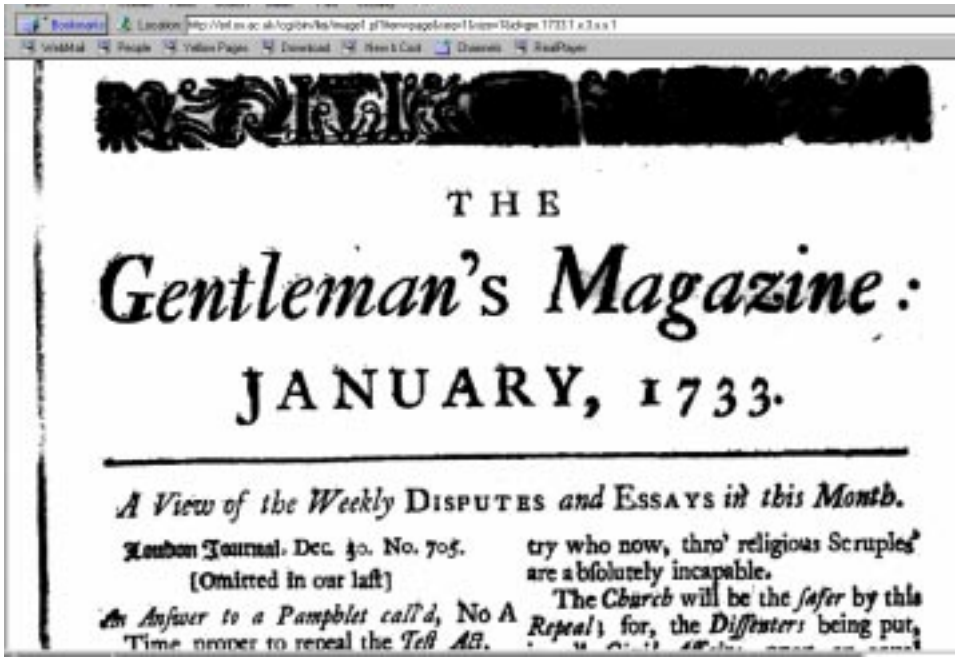


Figure 1: Browsing the journal images: Gentleman's Magazine for June 1733



THE

Gentleman's Magazine:

JANUARY, 1733.

A View of the Weekly DISPUTES *and* ESSAYS *in this Month.*

London Journal. Dec. 30. No. 705.
[Omitted in our laſt]

An Anſwer to a Pamphlet call'd, No A
Time proper to repeal the *Teſt Act.*

try who now, thro' religious Scruples
are abſolutely incapable.

The *Church* will be the *ſafer* by this
Repeal; for, the *Diſſenters* being put,

Figure 2: Viewing a page from the Gentleman's Magazine

The search interface uses a simple HTML form for the input of search terms, which can be linked using Boolean "AND" or "OR" connectors. It is not currently possible to search across all titles as once, owing to the disparity in indexes available for each: therefore, the user initially chooses the journal to be searched, and is then presented with the available search options (subject index, title index, full-text etc.) for that given title.
Figures 3 and 4 show the search screen and the results retrieved for a typical search.

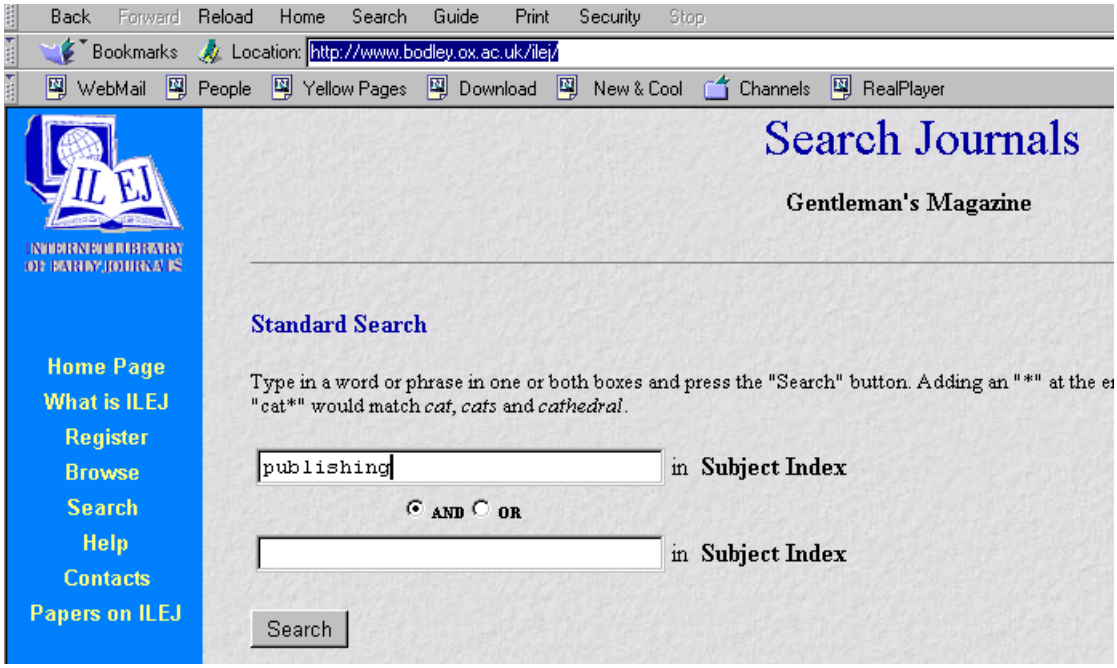


Figure 3: Searching the Gentleman's Magazine subject indexes

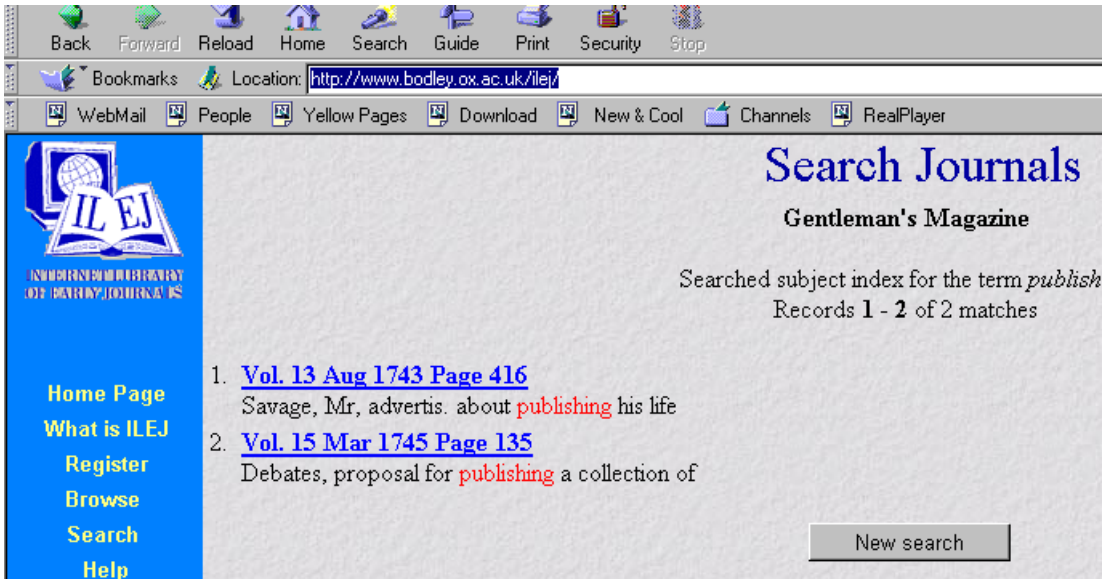


Figure 4: Matches found for the preceding search

User response

An extensive evaluation programme was carried out amongst the 377 users who volunteered to register for the service (which was not compulsory in order to access it). Monthly usage figures averaged around 50,000 hits on the server once the full service became available, of which the *Gentleman's Magazine* and *Notes and Queries* accounted between them for around 60% of accesses. The majority of respondents (approximately 70%) were from the academic sector: their research interests tended to fall within the areas of 18th and 19th century language and literature, the history of thought and science, genealogy and sociology. In addition, a significant number used ILEJ to research the history and biography of specific individuals. 82% of users were located in the UK and the US.

Users were asked to rate their satisfaction with the service according to eleven criteria, ranging from speed of access to the clarity of images. Satisfaction levels averaged 75% for 10 of these criteria, with only speed of access falling somewhat lower at around 50% (much of which, was, of course, likely to be accounted for by network speeds outside the control of ILEJ) . About half of the respondents felt it unnecessary to print the images retrieved, finding them usable on screen, but of those who did try to print almost half found it difficult to get satisfactory results: many of these problems were probably browser specific, as most offer only fairly crude facilities for reformatting images to fit neatly onto single sheets. Those requiring more sophisticated viewing and printing facilities had to use helper applications, which are much more flexible.

Amongst the more specific comments received, several users saw the quality of OCR as too poor, and others suggested that full-text should be available for all titles. There were also adverse comments about the legibility of some images: "bleedthrough" and foxing often caused problems of this sort, as did the distortion caused by the tight bindings and narrow central gutters on many of the original materials. But generally, the comments received were favourable, the major complaint being that the date ranges covered were too small. Only 3.2% of respondents stated that they would prefer to use the paper copies if electronic versions were available, and none said that they would use the paper copies exclusively in such circumstances.

Conclusion

The ILEJ project has shown up both the possibilities and difficulties associated with the creation

of a substantial collection of digitised material in electronic form. Substantial advantages in terms of easier access, enhanced searchability through the incorporation of machine readable versions of original subject indexes and full-text, and the tractability of images (admittedly less important for this class of material than, for instance, medieval manuscripts) have already been elaborated at some length in the professional literature[3]. The ready availability of a mass of such material has proved as valuable as we had surmised for those specialising in the periods and subjects covered by the digitised journals.

One major difficulty shown up by this project is the substantial costs involved in compiling a usable digital library collection, costs which are invariably larger than may initially be envisaged. For the ILEJ project, the total costs incurred amounted to 33-42p for each image from a bound original, and 112-116p for each microfilm image (the latter admittedly inflated by technical difficulties experienced which may well be avoided with better quality microfilm). Providing full-text searching of corrected OCR more than doubles the cost involved, making this a very expensive proposition, while the provision of uncorrected OCR, as was done in this project, is likely to prove frustrating to users in the long run.

Inevitably some of the costs involved in the ILEJ project were higher because many of the practices involved in digitising this type of material had not already been well established at the time of the project's inception, and so to some extent much time was spent working out the methodologies to be employed rather than on the production process itself. As time passed, the cost per image for each journal title tended to fall, as its particular idiosyncrasies were identified and assimilated into the production process. The costs of scanning could well be reduced in the case of more suitable microfilms (particularly those with clearer edges and less variations in contrast levels within a frame). In the case of the scans from original materials, the fact that we used bound volumes inevitably made the scanning process slower, and hence more expensive: a operation, such as JSTOR, which uses sheet-fed, dismembered materials will have a much lower unit cost.

It should also be noted that in the case of the ILEJ project, the initial costs of scanning represented on average only a third of the total cost per image. The costs of metadata creation, processing, indexing, OCR when feasible, and image conversion invariably accounted for the majority of the unit cost. It is easy to forget these costs when devising a strategy for such a digital library, but without these processes a digitised collection becomes no more than a list of files, disorganised and meaningless in themselves. It would certainly be virtually unusable to all but its compilers.

The conversion of a sizeable proportion of early material to digital form is, therefore, an expensive proposition, and should be budgeted for accordingly. The costs of metadata provision is easily overlooked, No library, digital or traditional, will work without the input of librarians in organising and describing its contents, and these costs need to be appreciated before undertaking a project of this kind.

mThe ILEJ WWW site can be found at http://www.bodley.ox.ac.uk/ilej/.

[1] Joint Funding Council's Libraries Review. Report (The Follett Report). Bristol: Higher Education Funding Council for England , 1993. Also available at http://www.ukoln.ac.uk/services/papers/follett/report/

[2] http://www.jstor.ac.uk

[3] See, for example, Robinson, Peter, The Digitization of Primary Textual Sources (Oxford Office for Humanities Computing, 1993)