CENTRE FOR HISTORY AND ECONOMICS

From page to screen: a case study in the digitization of historical resources

Many of the materials and interviews on this website will deal with ramifications of digitization and with larger digitization projects. This brief essay is intended to explain the ‘nuts and bolts’ process which takes an excerpt from a book to a fully-searchable .pdf file accessible online.

The text:

While researching the history of copyright, ER discovered that the volume of the Parliamentary History of England, published in 1813, which included the 1774 ‘Proceedings in the Lords on the Question of Literary Property,’ i.e. the groundbreaking debates on copyright establishing legal precedents which to this day influence the ownership and reproduction of print information in much of the world, was incomplete in the version scanned and made accessible online through the MOME (The Making of the Modern World) data base. The copyright debates are now themselves out of copyright – but had there been digital images of them produced by MOME, those images would still be under copyright.

We here reproduce the process of scanning and performing OCR (Optical Character Recognition) on the pages from the Parliamentary History, providing a step-by-step explanation of the technical issues which we hope will in turn inform our understanding of the scholarly and social consequences of the drive to digitization.

Steps:

There are several different ‘layers’ which go into the texts we can see in online book projects: one is the digital image of the page/document in question, the second is the ‘text layer’. In order to be able to use a search engine to index a scanned image, the image needs to be bundled with plain text which can be ‘read’ and interpreted by computer programs.

Scanning:

We used ordinary software and hardware for this project, itself an example of how technology is beginning to allow democratized access to digitization. We used a HP OfficeJet 5110 All-in-one printer/scanner to scan the document. While large companies like Google and Microsoft, and institutional libraries, tend to use purpose-built, high-end hardware and software to perform high-volume scanning projects, the technology to perform scans of a limited number of texts is generally available in most of the University of Cambridge’s computing resource centres, and indeed on many home computers.

First, we photocopied the excerpt from an edition of The Parliamentary History of England, vol. XVII (London, 1813) at the Cambridge University Library. Because books published before 1850 are generally non-circulating, we could not perform a scan on the original – but we will explain the differences between the two techniques!).

One of the major issues with high-volume scanning projects aiming to harvest as much raw text as possible from publishers’ archives and library collections is that they are not terribly concerned with documenting the details of the ‘containers’ in which this text is stored: the books themselves. Little effort is made to represent different editions; and some editions may be duplicated while others are neglected entirely. The lacunae which emerge in the digital archive thus created can be substantial – our excerpt is just such an item. (In the volume of the Parliamentary History for 1774-1777 which was micro-filmed as part of the Goldsmiths’-Kress Library, which is the basis of the MOME database, a number of pages from an earlier volume had been included in error; the missing pages happened to include the debates over copyright.)

For this exercise, we employed a scanned version of the document captured using the HP scan software shipped with the scanner. We captured at four different settings: 200 DPI black-and-white, 200DPI greyscale, 200DPI colour, 300DPI black-and-white (standard printing resolution for black-and-white printers).

OCR:

We found that the OCR software was the most temperamental of all of the various technical components to this project. While Roman script OCR is now in the public domain and is readily available via Adobe Acrobat (which we used for this text) or through many scanning software packages, OCR is by no means an exact science. The difficulties presented by older typefaces, on deteriorating or ‘foxed’ books and papers, on documents containing non-standard spellings, images or handwritten notes are significant. Thus even leaving aside the issue of non-Roman character recognition, there are a number of serious obstacles to scanning an English-language, out-of-copyright English book printed after 1750!

For the OCR, we used the Adobe Acrobat Professional 7.0 software running on Mac OS X. This software offers the following language specializations, which may be selected when beginning a new scan session: Brazilian Portuguese, Catalan, Danish, Dutch, English (UK), English (US), Finnish, French, German, Italian, Japanese, Norwegian, Nyonorsk, Portuguese, Spanish, Swedish, Swiss German. Adobe Acrobat offers three output options for PDF format: Searchable Image (exact), Searchable Image (compact), Formatted text & graphics. We selected English (UK) and Searchable Image (exact). Adobe Acrobat also offers the option to ‘downsample’ image information; we selected the lowest downsampling level.

Some quality-related considerations: This document will effectively have been scanned twice, once by the photocopier, and once by the scanner on the computer. This leads to a loss of information and hence quality. Ideally, original documents should be scanned for capture and OCR; this can be done using some commercially available photocopiers as well as more general-purpose scanning equipment.

The scanner we used was intended for personal or small-scale use -- higher end solutions offer higher quality and higher volume processing, able to handle large numbers of pages, higher resolution, and with fewer mechanical issues, such as paper jams. The original photocopy was at an angle, in black-and-white, and has book edges and other markings. It will be very interesting to see if the OCR software properly handles hyphenated words, and what impact selection of the language (and hence dictionary) has on its ability to recognize words less frequently found in the current corpus, or specific to that corpus. The Adobe Acrobat Professional software includes software to adjust for these, but the loss of original information due to the first generation scan being black-and-white rather than greyscale cannot be made up for in software. It is not currently clear to us what impact that has had on OCR accuracy --

possibly a higher quality initial scan, losing less information, would result in higher quality OCR.

Error Correction:

Because the digital ‘books’ produced through this process are made up of two layers, each layer must be edited for viewing and searching. In terms of the image itself, it is a question of balancing considerations of quality with size: images are often large files, making disk space a concern for the organisation hosting the website, and speed of download an issue for the end user. Most image compression formats (.jpg, .gif, .tif, .png etc.) ‘fool’ our eyes by discarding a certain percentage and type of information and relying on the viewer to automatically ‘fill in’ the missing sections. Human eyes and brains are extremely tolerant (unlike OCR!) and can use semantics and context to complete the image.

Editing the text layer is a time-consuming process, consisting of going through the document’s plain-text portion to make sure that it is an accurate text version of the original page. Which brings us to another point: different scanning and OCR methods require vastly different amounts of ‘human hours’ for editing and correction. While for this case study we have spent quite some time making adjustments and considering the impact of different resolutions/types of scan on the end product, many online book projects are interested in quantity over quality and would expect to fix any errors which pop up in the OCR and scanning process in the indefinite future.

Hypotheses to test:

1. That comparing OCR on an original document vs. a photocopy makes (or does not make) a significant difference in OCR and visual quality of the reproduction.
2. That the OCR engine stumbles (or does not stumble) over certain classes of problems specific to this document, such as specific typographical properties (font, etc), the layout of the page (the two columns quite close to each other separated by a thin line).
3. It will be very interesting to see if the OCR software systematically fails to recognize hyphenated words, and what impact selection of the language (and hence dictionary) has on its ability to recognize words less frequently found in the current corpus, or specific to that corpus.
4. Generally, what sort of errors occur frequently, what sorts of corrections are required, etc.

Conclusions:

Photocopying on-site using COTS photocopying equipment does indeed impact our ability to produce a searchable document. The typographical conventions of particular documents, such as this one, also seriously affect our ability to produce a useful OCR of the document.

For demonstration purposes, we have placed a single document online, scanned using an HP desktop scanner in black and white at 300DPI resolution, and prepared using Adobe Acrobat Professional 8.0. The resulting PDF file consists of 29 pages, each representing one leaf of the original photocopied document. Prior to PDF conversion of the document, scanned TIFF files were manually cropped in order to produce individual leaves. Cropped leaves were each have a width of 12cm; leaves were captured either at 18.5cm or 20.5cm height, depending on page layout. Optical Character Recognition (OCR) has been run on the images, providing a rough computer-readable text of the document, but human correction has not been applied. This allows text search of the scanned document, subject to significant accuracy limitations in the OCR process.

OCR proved relatively unsatisfactory for this document, and we believe a significant number of hours of correction would be required to provide full text search with a high level of accuracy from these scans, or to allow accessible read-aloud of the document. The primary factors affecting the quality of the result were low original image quality due to use of a photocopy as the scanning original, and narrow column separator, which the OCR software did not identify, leading to incorrect handling of hyphenated lines or search strings spanning line breaks. Both rescanning in higher quality, and customization of the OCR software for this document type, as well as improved tools for manual review and correction, would make a significant difference in OCR conversion, but are not available from the desktop Adobe Acrobat Professional 8.0 tool.

In this presentation, the document server does not provide facilities to search or directly render stored documents, not does it mechanically maintain meta-data. All rendering and search of the document occurs in the client. For any sizable collection of documents, a more comprehensive approach requiring more extensive server-side facilities for document management, would be required. Original image capture quality could have been significantly improved by using high resolution scanning directly on the original document, rather than on a photocopy, and through greater care in the preparation of the document. For example, most pages were misaligned, in some cases folded, not pressed firmly against the copier glass, or not completely visible in the copy. In one case, a page was copied using the photocopier "zoom" facility, which lead to different scale with respect to other pages when presented in the scanned form.

Notes on Moving Forward with Document Scanning Projects:

Issues to consider in capturing, storing, processing, and presenting digital content:

Capture:

- What physical, photographic, and digital properties does the capture process have, and what format will the resulting data be in?
- To what extent will those properties be maintained in long term storage?
- Will "sacrificed" information be used for meta-data generation?
- What quality assurance/correction can/will be performed on captured images?
- What corrective actions can be taken?
- Will images be OCR'd on capture?
- Will existing meta-data for the document be captured with the document itself?
- What information will be lost for a user who has access only to the captured data?
- How will intellectual property information be captured, both in terms of original copyright/license/capture terms, and for resulting processed data?

Storage and Processing:

- What content is stored?
- What content (meta-data) is generated?
- If OCR was performed at capture time, are allowances made for regenerating OCR data as OCR technologies improve?
- What formats best meet storage, search/index, retrieval requirements?
- What indexing tools will be used?

- What special considerations will need to be taken with respect to scale of the archive?

Presentation:

- Will information be presented in the same format as the back end storage?
- Will end users have access to all content, including OCR, or is that used for indexing/retrieval only?
- If fees are being charged for access, how will that be enforced?
- What license terms will be enforced?
- What, if any, technical means will be used to enforce those terms?
- What access control will generally be required?
- Will the front end need to downgrade information quality to meet transfer or display requirements of end users?
- How will meta-data be displayed to the user ("Next book on shelf", "Next shelf", ...)
- Will the interface provide bookmarking, annotation, tagging facilities? Will the results be stored on the server?

L. Denault & R. Watson, 25 July 2007