World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

July 4, 2005

Code Number: 154-E
Meeting: 97 Newspapers

10 Billion Words: The British Library British Newspapers 1800-1900 Project
Some guidelines for large-scale newspaper digitisation

Jane Shaw
The British Library
London, UK

Abstract

The British Library (BL) decided that comprehensive chronological coverage across the whole of the nineteenth century was the key to their project, and as the holder of the 'master collection', the real challenge would be to convert a large volume of text into a searchable online resource knowing that "very little is out there in terms of "mass of content". [1]

The lessons learnt half way through the 'British Newspapers 1800 -1900' (BN) project argue that in order to digitise a large volume of historic newspapers to the highest possible quality, it is necessary to take the planning time to know the characteristics of your source material and to adequately resource your team from the outset.

Existing agreed standards for digitisation from microfilm are defined but not fully followed. [2] The BL therefore set about to establish some standards for filming for large-scale newspaper digitisation and guidelines for best practice.

Problems with both the source material and the digitisation process prompted certain decisions. These decisions included:

- *Setting aside very poor condition volumes.*
- *Condition survey/assessment of source material to act as a benchmark.*
- *Refilming as a platform for digitisation, filming one page per frame to ensure a consistent look.*
- *Only digitising from microfilm for speed, consistency and cost.*
- *Human intervention to aid condition checks, page by page collation and simplified article zoning.*
- *Open source software solution that can be repurposed*

Other points that have been considered in depth are the design of an interface for access to newspapers, how much metadata should be incorporated, and what kind of searches should be enabled.

Introduction

This paper derives from a series of reports commissioned by our funder JISC (Joint Information Systems Committee) and from our experience of developing '**British Newspapers 1800 – 1900 Project.' (BN)** The paper attempts to answer some basic questions about the practical and technical processes involved in the creation of a mass of searchable online newspaper content from microfilm images. Will the entire content of each newspaper be digitised, such as adverts, pictures or only selected articles? What navigation tools will be available; (will readers be able to 'turn pages', will there be keyword searches?) What is the impact of using microfilm produced for one purpose i.e. preservation, for another purpose – digitisation? How does pre-sorting, testing and benchmarking the source material, set the foundations for unimpeded online access to previously difficult-to-access material?

The paper reports on how realistically we planned the project. It describes our decision-making and the cost components for filming to high technical quality standards to yield high quality digitial images and improved OCR.

Background

Early in 2004, the British Library secured funding from JISC.[3] Under the Digitisation Programme, funded with a £10 million grant from the Comprehensive Spending Review, JISC enabled a small number of large-scale digitisation projects that would bring significant benefits to UK Further and Higher Education communities, one of which is British Newspapers 1800 – 1900 (BN) project.

The JISC selection criteria for funding under the Digitisation programme were:

- The materials should be of broad disciplinary interest and form a coherent theme or themes.
- A small number of large-scale projects should be funded that would not be possible without an investment of this size.
- The materials would need to be fully compatible with the common information environment.
- The materials would need to meet rigorous quality-assurance standards and be of value to the wider post-16 education community.

The project was funded to deliver the following – the scanning of the entire microfilmed content; article zoning and page extraction; OCR of the page images; and the production of the required metadata. The main objectives are to digitise up to two million pages of British national, regional and local newspapers, the majority from new microfilm and to offer access to that collection via a sophisticated searching and browsing interface on the Web.[4] This will include names and dates, obituaries, advertisements, regional perspectives and local perspectives to national news.

Aims of the project and how they have guided the selection of newspaper titles

Both the overall goal of the project;

- to provide a mass of historic newspaper content on the web for full text searching by academic communities;

And the main aim,

- to digitise up to 2 million pages of out-of-copyright UK printed material, regional and local newspapers, the majority from new microfilm and to offer free access to that collection via a sophisticated searching and browsing interface on the web

Have not changed in the last year. The project plan however does differ from the original business case in the following main areas;

1. The balance of new filming has increased from 50% to 90%, to enable consistency of images.
2. Filming one page per frame for optimum digitisation.[5]
3. Introduction of an in-house Quality Assurance Team to prepare and repair the volumes, collect both issue level and condition level metadata and filter out duplicates, variants; and identify missing pages, issues, and the last timed edition at the start of the project.
4. Placing an academic User Panel at the core of the project to steer selection of newspapers and advise on the website design.[6]
5. Introduction of two Pilots to survey the physical characteristics of the nineteenth century newspapers, to agree on a methodology for the supply of microfilm and to confirm that specifications are yielding the desired end product, including image quality and OCR results.[7]

Selection Constraints

The original business plan included a preliminary list of many titles, at least 160; split into London national dailies and weeklies; English regional dailies and weeklies; Home Countries newspapers (Scottish national, Scottish regional, Welsh, Northern Irish) and 'specialist sub-clusters'. For copyright reasons, and to keep within the scope of the original project brief, only dates between 1800 – 1900 were selected. However, in the early stages of the selection, additional constraints arose. Owners of incorporated[8] titles still publishing could have objected even if pre-1900 issues are clearly out of copyright. Owners of titles still publishing may be digitising or have plans to digitise their back runs (e.g. *Guardian, Daily Telegraph*) and it would not make economic sense to duplicate their efforts.

Surprisingly, very little information was available about how many pages each title represented and in order to keep to the project schedule a decision was made to

start with a Pilot of a discrete specialist sub-cluster such as the Chartists followed by the first work batch which included obvious titles (e.g. *Examiner, Morning Chronicle, Graphic*). At the same time, an audit into the pagination and condition began of further likely candidates for selection from the preliminary list.

Notwithstanding the above constraints, the User Panel still decided to assess the value of all of the titles from the original list (the long list) from the perspective of potential usage by the HE/FE community.

From the User Panel's prioritised list, a 'wish list' emerged which was sub divided into coherent bundles or work packages.

Four work packages have been selected to date and comprise approx.2 million pages in total. The Work Packages follow a logical mix of UK wide coverage and nationals with regionals. Work Package 1 includes the Pilot work (The Chartist sub-cluster), plus three national titles – a daily, a Sunday and a weekly review. Work Package 2 extends the coverage to include three regional titles, North of England, far South West and central. Work Package 3 continues the national press with one Sunday, one daily, introduces Scotland and Ireland and continues with the English regional press. Work Package 4 continues to extend UK coverage, with Ireland, Scotland and Wales and enhances the English regionals.[9]

Online Consultation and User's Needs

The relevance to actual or potential users needs has been determined not only by an academic panel to inform the selection, but also validated through the exercise of an online questionnaire.

An online consultation with the wider academic community took place during February – March 2005, specifically on the titles to be included within the BN project and also to ascertain what titles should be included if future funding became available either to extend the BN project or pursue new projects.

195 people replied and of these, the majority were from librarians and lecturers working mainly in Universities and FE colleges, with a spread of researchers, students, managers and teachers. Surprisingly, 13% were replies from USA.

The questionnaire asked users to rank in order of priority for digitising (one=strongly disagree 5= strongly agree), the titles from the long list in the business plan. In addition, we asked them to offer comments on any other titles they may want which were not listed.

Overall, it was clear that the replies endorsed the approach for UK wide coverage and the methodology adopted (i.e. a framework of national titles and countrywide coverage with the breadth and depth to form a virtual key to provincial newspapers in any medium). It was also clear that the omission of newspapers from Eire was causing concern. [10] This is being re-evaluated by JISC.

Some Portraits of Newspaper Titles Selected

Morning Chronicle: A London daily. Under the editorship of John Black, the young Charles Dickens was a reporter, and Thackeray worked as an art critic.

Reynolds Newspaper: Achieved sales of more than 350,000 by the early 1870s. In origin a radical newspaper, it remained in control of the Reynolds brothers until 1894.

Poor Man's Guardian: founded by Henry Hetherington in 1831 to further the cause of universal suffrage and the trade union movement. Offices raided in 1835 by courts and their presses seized and destroyed.

Corbett's Weekly Political Register: William Cobbett founded this paper in 1802, to further his parliamentary career. Tory in outlook initially but gradually became more radical.

Birmingham Daily Post: The Birmingham Daily Post was launched in 1857 by Irishman John Frederick Feeney as a Monday to Friday Paper of four pages and priced at one penny. It is still published.
Belfast Newsletter: founded in 1737, for almost two hundred years the Henderson family was closely associated with the newspaper. It is still published.

Copyright

The BL is in continual discussion with publishers, including newspaper publishers, on a range of IPR issues, covering the life cycle management of information – from acquisition to access to preservation.

The Library policy is to proceed with the agreement of rights holders and their representative bodies. In the case of newspapers, recent discussions with newspaper publishers and updated legal advice to the Library means that for this project, the starting point is that no newspaper less than 100years old will be digitised for access by HE and FE.[11]

How the Project was shaped by problem solving

Deliverables

The project will deliver up to 2,000,000 pages, totalling approximately 10,000,000,000 words from British newspapers 1800 – 1900. This equates to around 40 titles.

The digitisation process will deliver an archival master file for each page, in TIFF format, version 6.0. These files will be scanned effectively at a resolution of 300 dpi, 8-bit greyscale.[12]

The service images will be created after the process of article zoning and OCR. The service copies will be delivered as greyscale hybrids, TIFF version 6.0, and as JPEGs

Many of the newspaper titles were filmed on acetate and before National Preservation Standards were adopted systematically (1990). In order to yield highest quality images and to save on costs in the longer term, best practice was to control the quality of the microfilm from the start and to aid elimination of many postproduction queries. Thus reducing QA workload and providing the supplier with a uniform benchmark and a consistent look.[13] The additional work of collating

duplicates and missing pages, condition survey work and collecting issue level metadata was seen to be the more effective and of best value in the longer term for microfilm scanning.

If organisations do not want to refilm and decide to use their existing stock, because it is more time efficient or within policy, there is still likely to be a cost, at the end of the process and more under the control of the suppliers.

Disbinding the newspapers was not an option and when new state of the art Zeutschel microfilm cameras were introduced into the BL Microfilm Unit, the project benefited from new technology by re-filming bound volumes using spine bars as needed, and not under glass. The Library is not trying to correct the printer's errors or areas of missing text. Facsimile images stored as greyscale master files can easily be repurposed and/or disaggregated for many future projects.

BN project team decided it would be valuable if the information could be viewed in context, as originally published, and with a full-page image. [14]In addition, the three selection criteria of: complete runs of each newspaper, UK wide coverage and spanning the whole of the nineteenth century would provide substantial access to news stories by 'simple' search terms.

The project is not about digitising eighteenth century issues or variant editions, nor are the rich resources of British Colonial newspapers included. Only the latest timed edition of each issue is being filmed and some occasional supplements e.g." *The Graphic* " *Stanley Number,* 30 April 1890*.* BL made these decisions in order to keep duplication of effort to a minimum and to include the maximum number of original pages in our two million total. The User Panel created by the Project also decided that it wished to have as much coverage of different newspapers as possible. This is preferred to including variant editions.

From other digitisation projects in the BL, there is an awareness of issues surrounding both the preservation and digitisation of newspapers, and knowledge of strategies for digitizing them. This paper describes how the BN project team made an informed decision about the appropriate strategy for this particular project. The density of nineteenth century texts makes machine identification of breaks between articles a more difficult task thus requiring a balance between automatic metadata generation and human intervention. Working with a supplier with many years experience, the BL took the view that human intelligence would give the best quality result and therefore shaped the project around computer assisted/human intelligence throughout the whole cycle.

Due to the shape of the project the per page cost for digitisation is estimated at,

100% dupe from existing film = 75p per page
100% new filming = 98 p per page

Moreover, this is in line with the original budget, (£1 per page).

Physical Characteristics of the source material, underlying problems and how these could affect the digitization process

BL holdings of Newspapers

In common with other large collections of historic newspapers, BL's holdings are predominantly bound together in volumes. Generally, this was seen to be the best method to preserve them against the effects of long-term handling. Due to the nature of the bindings, some volumes are very tightly bound and others have started to disintegrate leading to damage around the edges of the pages and consistently within the end papers.

There are two problems due to the way the newspapers are bound, the text being bound into the spine and the curvature of the papers towards the spine. Text can be lost during filming due to gutter shadow, text can be skewed, and this may affect the percentage of the last column that can be OCR-ed. Other examples of problems the OCR will have to overcome include; problems with "set through", (able to see the printing on the reverse of the page through the paper), often due to poor quality paper, heavy inking or a combination of the two, and printers errors due to paper slippage or creased/folded paper causing breaks in the font during printing. The OCR engines could have problems in overcoming these errors and in recognising complete words.

A further complication is, duplicate issues and variant editions are usually bound in together and different titles can be mixed in the same volume. For this project, only using the last timed edition, variants and duplicates had to be weeded out. We decided to hand weed at the initial preparation stage rather than later pre-or-post scanning.

Early nineteenth century newspapers reveal a significant amount of printer's errors, (e.g. a page creased during printing results in text loss when the crease is ironed out). There is also the appearance of the hyphenated word, mainly found at the end of 2 column newspaper texts. [15]Formats and structure change frequently and unexpectedly, from a 4-page issue to a 6-page issue, the order of the content often switched, the appearance of the "editorial essay" invested with the signature of the editor and sudden style changes from a "Two-Penny Trash" to a broadsheet. They represent a rich resource in the history of print and the development of radical argument and opinion is reflected through innovations in typography and layout. They also represent a significant challenge to the OCR, which prefers even text layout, even tone and larger print. We did not accept significant loss of text from gutter shadow or uneven density film from old laminated pages.

BL holdings of Microfilm

It is within BL policy to use existing film and to complete gaps in master negative or dupe runs, to avoid double handling of the collections. However, a high proportion of the BN selected titles are on early acetate. The worse the condition of the film, the longer it takes to copy, thus reducing output and increasing unit costs.[16] In addition, the existing microfilm is too variable for a steady workflow, comprising acetate, pre National Preservation Standard polyester, post standard polyester and post 2004 new camera standard and there may be additional quality problems and post production costs due to this mix of unsuitable film. Overall, the BL decided that

managing complicated workflows due to the different speed work steams to match the mix of variable film, could lead to unacceptable delays and compromised quality. The existing stock of film may include duplicate issues and all variant editions for any bound title, as historically this was haphazard and the current policy is to film the latest timed edition, which is in accordance with the project. A scanning operator would have to learn which parts of the film to scan and which to ignore. To overcome these problems, the BL decided that for this digitisation project, where condition and binding of the material allows, the BL Microfilm Unit on the new cameras would film most newspapers in-house.

Most of the pages have been clipped to keep them flat during filming. This may look odd and appear as black if we are presenting a facsimile image of the page to users. Nearly every reel has a splice and this is because of missing pages identified at the checking stage of production or retakes due to various technical problems. These have been areas of concern.

Data Capture Sources.

STATE	CONDITION	DECISIONS				
Microfilm Poor Quality	Newspaper Good	Refilm	Dupe	Scan	Enhance	
Microfilm Good Quality	Newspaper Poor	Dupe	Scan	Enhance		
Microfilm Poor Quality	Newspaper Poor	Set Aside	Select another title			
Microfilm Good Quality	Newspaper Good	Dupe	Scan	Reject	Refilm	Rescan
Microfilm Good Quality	Newspaper Good	Dupe	Scan	Enhance		

Reduction and Resolution

When filming the volumes the lowest possible reduction has been used, for example the *Pall Mall Gazette* was filmed at a reduction of 12x. Further to this, it is important to keep the same reduction throughout an entire reel. However, this has not always been possible due to some titles containing fold out illustrations within the pages. The wide range of titles selected, mean that some of the physically larger volumes have been filmed at a higher reduction, up to 20x in some cases. It is worth noting that we have only used full reductions (i.e. no half reductions used).
Despite the concerns over using high reductions, we are achieving some very good resolution figures. In some cases readings of 140 LPM (lines per millimetre) are being regularly achieved, this is further enhanced by tight control of density readings. [17]

Procurement

Due to the likely cost of the digitisation contract being over £153,376 we were subject to OJEU procurement rules which is a lengthy process involving much

learning.[18] The whole process took approximately one year, from request for proposals through to signing of contract.

All of the potential suppliers offered us the suggestion of later high-level conversion. They could all scan and OCR the pages, scan OCR with highlighting of searchable terms, scan, OCR highlighting and viewing the page, but only one could scan, zone, OCR, check and repair zoning and re-key headlines as necessary within a reasonable budget. There was no ability anywhere to present a "completely" clean OCR file.

Learning from the first year of Operation

What we would do again and what we would do differently if we were to start again.

- Preparation is the key: in depth survey and assessment of the physical characteristics of source material to set a benchmark for later QA.
- Page counting and weeding out of duplicate issues and variant editions necessary to produce our work packages and monitor progress against our overall total.
- The format of a long run will often change during a century, so page counting is vital to understand the structure of a long run.
- Place a User Panel, at the core of the project, to act as ambassadors and take ownership.
- Define criteria for selection early on to guide User Panel deliberations.
- Intellectual Property Rights – address issues early, take a robust and consistent approach and maintain an ongoing dialogue with the newspaper industry.
- Conduct an online consultation with user communities during the funding process. A large potential list of titles meant we could not focus.
- User Panel preselected our potential 2 million pages, followed by an online consultation that did not suggest completely new and untried titles apart from Eire ones, so work could start. [19]
- Consider concept of 'Set Asides' – titles too fragile to film as benchmarks for the future. Tolerate and accept gaps in full runs and do not seek to fill these until later on.[20]
- Consider the quality of your microfilm [refilming a proportion of your content will add value in the longer term] in order to adapt many of the risks around heterogeneous originals and variable quality microfilm which should in turn aid image capture. [21]
- The BL's collection of nineteenth century newspapers is in better shape than predicted, less than 2% unfit to film. However, due to the mechanical processes of scanning (too slow and too harsh for vulnerable fragile source material) it was decided to digitise from microfilm.
- Manage future expectations through online endorsement of titles lists by user communities.
- Aspirational production targets do not work. Sustainable targets based on real work done to date should be regularly reviewed and reprofiled as necessary to forecast trends.
- Use "Doubles" or intentional second exposures as a quality assurance technique, after weighing this up in relation to efficient workflows.[22]

Some Digitization Issues

What standards?

We are following some rigid standards e.g. METS, a metadata encoding and transmission standard and BSI standard for microfilming.[23] We believe we will have well formed metadata using people working with original pages and from good quality microfilm, which provides authentication.

The BL requires Dublin Core descriptive metadata records compliant to the British Library Application Profile (BLAP) with the elements encoded as XML and finally an OAI-PMH data provider service as a means to meet our interoperability requirements.

Levels of metadata

There will be four structural levels, title, issue, page and article. Within the title level metadata, there is any changes to the title, publication date, type and sub collection all captured captured by the BL QA team. Issue level metadata comprises, issue number, printed date, normalised date, number of pages and reel identity number. Additonal metadata that will be provided in the XML files includes:

- Date of Issue in standard ISO format.
- Quality rating, (A, B, C) on condition of original material.
- BL copyright statement.
- BL copyright year.
- Conversion credit.
- Placeholder tags for table and illustration credits.
- Placeholder tags for author names.

Greyscale Hybrids versus bitonal

Pages scanned as 8-bit greyscale appear softer, with finer detail and subtler greys, particularly beneficial for text and illustrations and truer-to-source archival images. Speckling interference is reduced, superior deskewing and a higher OCR accuracy with a resolution of 300 dpi.

The project is developing a new product, a hybrid image, to optimise the quality of the images as well as the text. Text areas are converted to bitonals, sharpened using an enhanced version of IZ-Image, and saved to greyscale along with untouched greyscale images of illustrations. These reconstructed hybrids are the service copies that will be accessed via the website. The main disadvantage of this method is a ten-fold increase in image size, which could increase our storage costs and cause problems for 56k modem users. As future users will be mainly broadband and costs are acceptable, the project team accepted the concept of greyscale hybrids.[24]

Of course, there needs to be an exact concordance between the master file and service copy images

Cropping

We felt there was a need to balance legibility and completeness and for the master files to represent as accurately as possible the visual content in the original pages.

An object image or archival master file will be created during the initial scan from the microfilm by cutting the black borders away from the entire imaged object on a frame-by-frame basis. The page image, to be used for the full-page service copy, is created by cropping the object image to a uniform border around the content of each page, thus removing any unsightly clips or targets. Any impression of scale or changes to page sizes could be lost due to this uniformity, but on the other hand, this method should produce images optimised for web delivery.

Hit term Highlighting

The project does not want to have inaccurate word highlighting and hence a poorer user experience. This will be displayed on the page and is under development now.

Articles without titles

An instance where this is prevalent is in categories such as advertisements, notices, obituaries etc. It was decided that where a category cannot serve as a title, a title would be constructed from the first couple of lines of text.

Deskewing

The human eye will simply address this. Images are de-skewed until they appear straight to the inspector and as very image is inspected prior to OCR this is acceptable. In addition, greyscale images can be corrected up to 10 degrees of skew, whereas bitonal images with as little as 1 degree can affect the quality of the OCR levels.

Article vs. page level

Page images will consist of both articles with search terms highlighted, once completed and full pages. Users will be able to view the page images associated with any of their search matches.

Recommendations for Subject Categories

With such a mass of words to search on and an unprecedented degree of cross-referencing the project is also providing users with straightforward article categorisation, allowing a way into the material to enrich searching by key words.

Our recommendations for subject categories are,

1. **News (Domestic)**; items relating to the UK
2. **News (Foreign)**; items relating to the rest of the world
3. **Advertisements & Notices**; ranging from items for sale to situations vacant to Public notices. Differentiated from news items by format.
4. **Arts & Popular Culture**; Reviews of books, music & the theatre. Poems and serialised fiction. Items relating to travelling shows & fairs.

5. **Births, Deaths & Marriages**; announcements made regarding each.
6. **Obituaries**; specific items relating to the death of someone with a biography of that person.
7. **Court & Society**; relating to the Royal family and the aristocracy.
8. **Crime and Punishment**; Reports from the various courts, sessions and assizes.
9. **Commerce**; business and shipping news, market & Stock Exchange Tables.
10. **Letters**; usually written to the Editor and printed in part or in full.
11. **Sports**; items covering a range of sporting topics.
12. **Editorial/Comment**; items usually, but not exclusively written by the editor on a particular topic or topics.
13. **Miscellaneous**; items that do not fit the above.
14. **None**; Default category.
15. **Illustrations**: without captions.

Steps in the Production Process

The simple steps in the project that the BL is following and hopes will enhance access to historical newspapers are**,**

1. Condition survey: of holdings of source material to understand the structure and to identify underlying problems. Tight bindings causing gutter shadow, how to handle foldouts and supplements, duplicate issues included in films, titles split across different reels and mixed in with other titles.
2. Collation: The results of the survey were collated and used to aid the selection process and the procurement of a supplier. These results included the number of pages per issue, the differences between nationals and weeklies, changes to structure across full runs, page layout, fonts across the century and the general condition of the microfilm stock.
3. Selection: A list of titles was selected using simple straightforward criteria endorsed by the user communities.
4. Microfilming: Tightly bound volumes, poor condition volumes, and sub standard microfilm were treated on a case-by-case basis as to whether to 'set aside' the sub standard volume or microfilm or to work with them.
5. Microfilming: A quality assurance system was built using the results of the surveys at the start of the project and tolerances set for 'set asides', clarity and legibility.
6. Microfilming: A decision was taken on how much to refilm if at all following microfilm quality levels which had been agreed.
7. Batch Definitions: Assumptions were tested in two pilots, one for each work stream and that each were large enough to deliver valid results.
8. Procurement: In order to grow a critical mass of historic newspaper content in the future, both nationally and internationally we knew we had to procure an open source technical solution.
9. Digitisation: scanning from microfilms; generation of greyscale images, QC, perform manual cleanup or not, QC, archive master files.
10. Digitisation: run IZ-Image character sharpening, load complete issue, zone and categorise articles, QC.
11. Digitisation: link multi-page articles, perform OCR and clean up fielded metadata, generate deliverables.

Costs

How much will it cost an institution or an individual to use? The current proposal under consideration is that the project is looking at 'free access' to everyone on the Web after Spring 2007, so usage may be free.

Future and Legacy

Will the online 'stock' be expanded in the future? Due to copyright and IPR issues, it is likely that future projects will select older materials that are out of copyright. There is an urgent need for some form of register or information network of what has been selected for digitisation and what has been worked on that can be checked against before other libraries start to digitise their collections.

With the data created from this project, the BL will be in a good position in future projects to predict what the cost per page will be as all the costs in our project have been analysed.

Summary

The BN project team has spent the last year assessing each of the selected newspaper titles and their associated microfilms for the purpose of digitisation; selecting a final list of up to 40 titles, which were endorsed via an online consultation with academic communities, and appointing a digitisation supplier. The in-house teams have prepared and filmed 650,000 pages since October 2004 and are on course to have completed 1 million frames by September this year. Both the Quality Assurance (QA) and Microfilm Unit (MU) teams work to daily and weekly targets. [25]

The coming year (until September 2006), will see a change of focus, with the start up of both digitisation and website development work streams in parallel with the in-house preparation and filming work.

The project will continue to follow a rolling review methodology[26] within a subset of PRINCE 2, [27]where we are learning about the source material and the technology as we progress. A User Panel having finished their selection task, have changed role to driving and defining potential user's needs, and commented on the design of the website, prior to an early release of a demonstrator for testing early in 2006.

A phased release of the website is likely to begin in September 2006.

BL have also placed human input at the core of the project by introducing a quality assurance team to screen out duplicate issues and variant editions, harvest issue and condition level metadata before filming. In addition, collaborating with an academic User panel to steer which 2 million pages out of a possible 200 plus million to select for digitisation.

The consistency of this approach extends to the choice of supplier, whose content extraction methodology uses computer assisted/human intelligence and has innovated in article categorisation and keyword indexing. What users will have are:

- full text searching of 2 million pages of newspaper texts.
- the ability to search individual newspapers by date.
- like-for-like comparisons of the same subject's treatment by different titles.
- browsing forwards and backwards through a selected issue.

- images of the original pages to read in the usual way.
- it should be possible to save and build searches, to aid in course teaching and collaborative working.
- display of the results of searches at the article level within the context of the original page.
- the ability to search advertisements, obituaries etc.
- the ability to download text versions of the original pages.
- An anticipated 80% accuracy on the OCR conversion.

Value can be added to variable source material in less than pristine condition and even with a poorly printed original, you should still aim for the best image you can get.

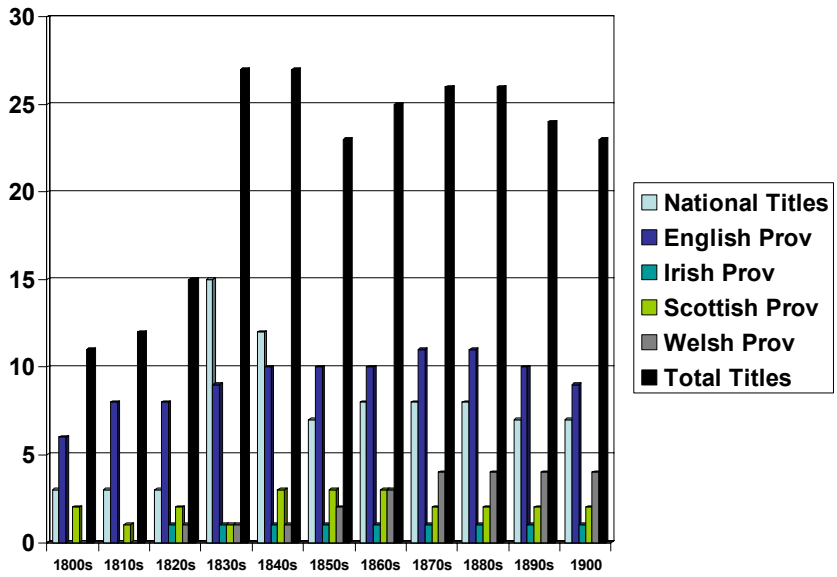
In conclusion, the BN Project is an initiative which has already enhanced the Library's understanding of many issues that digitisation raises. The BL looks forward to its successful conclusion, and, most importantly, that the content will find a wide audience.

Jane Shaw
June 2005

Appendix 1

19th Century British Newspapers

1 Aberdeen Journal
2 Glasgow Herald; Chartist Circular (Glasgow)
3 Caledonian Mercury
4 Newcastle Courant; Northern Liberator
5 Belfast News Letter
6 Northern Echo (Darlington)
7 Preston Chronicle
8 Leeds Mercury; Southern Star (Leeds); Northern Star (Leeds)
9 Hull Packet
10 Liverpool Mercury
11 "Cotton Town" title tba
12 North Wales Chronicle
13 Birmingham Daily Post
14 Western Mail or South Wales Daily News;Baner y Cymru;Goleuad;Genedl
15 Bristol title tba
16 Jackson's Oxford Journal
17 Ipswich Journal
18 Trewman's Exeter Flying Post
19 Hampshire Telegraph (Portsmouth)
20 Brighton Patriot
21 The Chartist; The Champion; Cobbets Weekly political Register; Daily News; The Era; The Examiner; The Graphic: Illustrated Police News; Lloyd's Illustrated Newspaper; London Despatch; Morning Chronicle; Odd Fellow; Operative: Poor Man's Guardian; Pall Mall Gazette; Reynolds's Newspaper



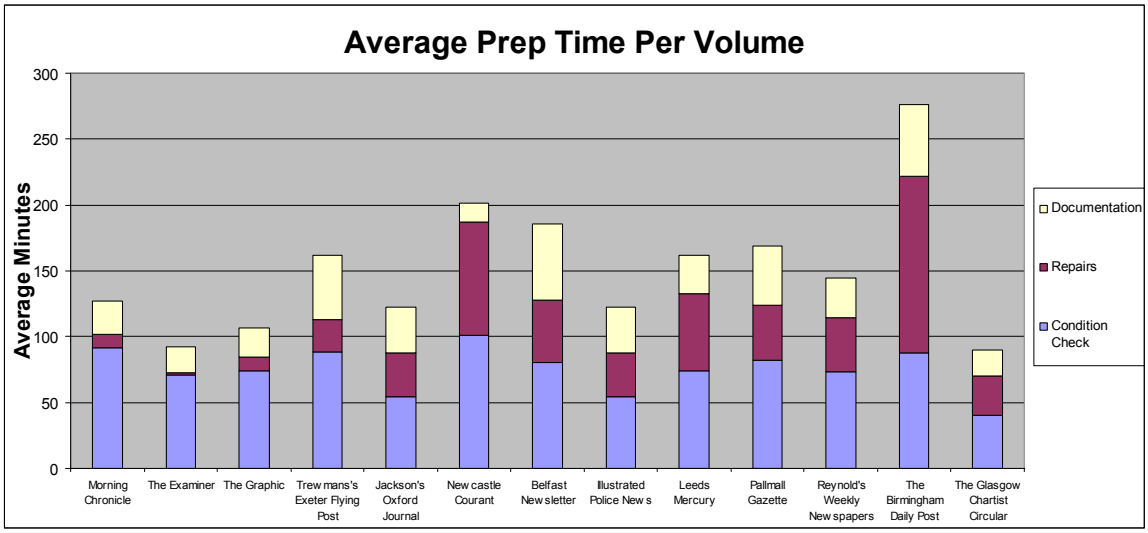
n

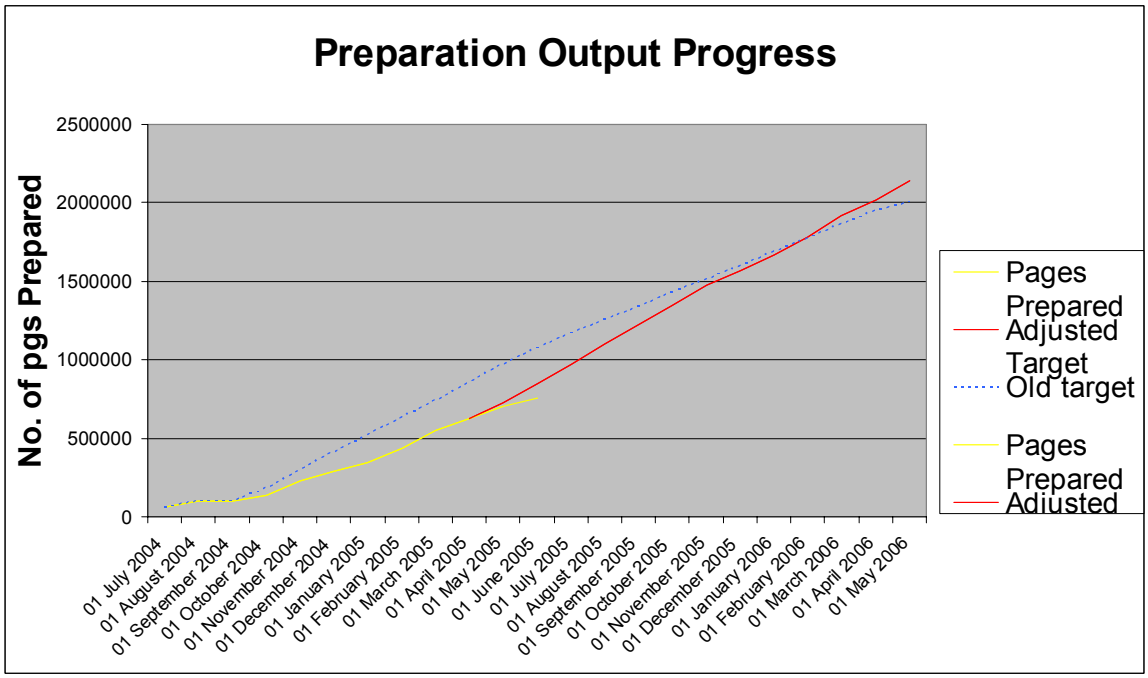
Output as at 17th June 2005

Work Package 1		Work Package 2	
Number of Titles Years	15	Number of Titles	4
Titles Completed	15	Titles Completed	4
Total Pages Prepared	279,836	Total Pages Prepared	95,025
Total Volumes Completed	565	Total Volumes Completed	227
Average Pgs per Volume	495	Average Pgs per Volume	419
Average Vols per Title	38	Average Vols per Title	57
Average Pgs per Title	18,656	Average Pgs per Title	23,756
Work package Complete	100%	Work Package Complete	100%

Work Package 3 (WIP)		Work Package 4 (WIP)	
Number of Titles	6	Number of Titles	20
Titles Completed	2	Titles Completed	0
Total Pages Prepared	357,834	Total Pages Prepared	
Total Volumes Completed	462	Total Volumes Completed	
Average Pgs per Volume	775	Average Pgs per Volume	
Average Vols per Title	77	Average Vols per Title	
Average Pgs per Title	59,639	Average Pgs per Title	
Work Package Complete	33%	Work Package Complete	0%

Appendix 2


Average Prep Time Per Volume
Average Minutes
300
250
200
150
100
50
0
Morning Chronicle
The Examiner
The Graphic
Trew mans's Exeter Flying Post
Jackson's Oxford Journal
New castle Courant
Belfast Newsletter
Illustrated Police News
Leeds Mercury
Pallmall Gazette
Reynold's Weekly Newspapers
The Birmingham Daily Post
The Glasgow Chartist Circular
Documentation
Repairs
Condition Check


Preparation Output Progress
No. of pgs Prepared
2500000
2000000
1500000
1000000
500000
0
01 July 2004
01 August 2004
01 September 2004
01 October 2004
01 November 2004
01 December 2004
01 January 2005
01 February 2005
01 March 2005
01 April 2005
01 May 2005
01 June 2005
01 July 2005
01 August 2005
01 September 2005
01 October 2005
01 November 2005
01 December 2005
01 January 2006
01 February 2006
01 March 2006
01 April 2006
01 May 2006
Pages Prepared Adjusted
Target Old target
Pages Prepared Adjusted

[1] See OCLC.
[2] For a full description of this issue, see: *RLG Guidelines for Microfilming to Support Digitization*, January 2003. See also, IFLANET, Newspapers Section, *Microfilming for Digitisation and Optical Character Recognition*, December 2002.
[3] The JISC agreed to support the project from April 2004 to September 2006 at a total cost of £2,022,131. See: http://www.jisc.ac.uk
[4] BL Business Plan, 14 October 2003.
[5] There is variable density within images and within exposures on existing two pages per frame films.
[6] Although a complete beginner would have been a useful member to advise as a layperson who just wants to browse.
[7] Reference: Jane Shaw, *JISC Development Programmes, January 2005 Progress Report*
[8] Newspapers published in the nineteenth century can be incorporated in extant newspapers
[9] See Appendix 1 for our list of selected titles.
[10] E.g. *"This is a long overdue project, which has the capacity to transform research on this period. It will make a big difference to Open University history teaching. Ideally, the papers chosen will offer a good geographical, political and chronological spread."*

"Given that newspapers such as the Limerick Chronicle or the Cork Examiner or the Dublin papers of the 19th century were in cities that were part of the United Kingdom, why are none such newspapers included in your project? From an historian's point of view, such an omission is highly illogical and produces an unrepresentative selection. Any chance a few of the major Irish papers might be included in this very interesting and highly promising project?"

"Digitized newspapers are a wonderful resource for History departments. They facilitate independent research and learning, which all of us encourage, via a format, which students enjoy. This is especially welcome given the pressure on library books and journals following increases in student numbers in recent years."

[11] The Chair of the User Panel has suggested that 'Holes in the Selection Process' could be addressed by a longer term licensing agreement such as buying in other licensed sources and by creating relationships with commercial companies.
[12] This is less than the 400dpi that the Library of Congress recommends. BL chose 300 dpi because a higher resolution only gives a significant increase in file size, particularly for greyscale and OCR quality does not improve above 300 dpi, and that on-screen resolution is usually between 72 and 100 dpi.
[13] It also placed the responsibility for the later quality assurance squarely with the supplier.
[14] This decision was revisited and is discussed in detail under Digitisation Issues.
[15] Particularly relevant to Cobbett's and The Examiner where the page layout is two columns per page and words are hyphenated at the end of the line rather than a larger than usual space being left and the full word printed on the next line. What will the OCR software make of both the layout of the page and the way the words are split?
[16] The Burney Newspapers project proved it was possible to scan from older acetate, but with a high production cost.

[17] The resolution used is a minimum of 115 LPM (lines per millimetre) and the reduction is set by the material that is being filmed, although the material is filmed at the lowest possible reduction, depending on the size of the original. We have pages that vary from A4 through to broadsheet. The quality control is done frame by frame for image quality, with each roll tested for resolution and density using a microscope and densitometer. We also use a Quality Control Sheet for each roll of film that documents all the technical information for that roll and any future generations.

[18] The Official Journal of the European Union, an advertisement is placed inviting interested companies to complete and submit a questionnaire. A very thorough RFQ, or request for proposals, produced 26 expressions of interest, from which a short-list of six potential suppliers was made and three companies went through to Best and Final Offer stage (BAFO).

[19] It could be argued that the consultation should have happened earlier, before the Pilot began, but this was prohibited by the possibility of IPR challenges and conversely the best time to consult with the larger communities could be seen to after piloting some of our assumptions.

[20] There have been very few poor condition volumes set aside so far, less than 2%., considerably less than forecast in the business plan. We were able to redeploy conservation money into funding a QA team and refilming.

[21] BL benefited from new technology. Better images from new cameras because lighting is consistent, the gutter of tightly bound volumes is handled via beds that can be raised and the pages are not filmed under glass, just clipped where necessary.

[22] This is a requirement of BL standards, that the page to be doubled is filmed at the normal exposure and then filmed at the new exposure. The digitisation supplier has the job of choosing the best quality image post scanning.

[23] BS ISO 4087 (1991)

[2424] An idea currently under discussion is for users to be able to view the uncorrected OCR text to "repurpose" within their Virtual Learning Experience (VLE)

[25] See Appendix 2.

[26] This is the JISC preferred methodology for their projects.

[27] PRINCE 2 is a product based project management methodology.