Published on *Reviews in History* (http://www.history.ac.uk/reviews)

ProQuest Historical Newspapers

Review Number:
1096
Publish date:
Wednesday, 1 June, 2011
Date of Publication:
2001
Publisher:
ProQuest
Place of Publication:
Cambridge
Reviewer:
James Mussell

ProQuest Historical Newspapers has been in existence for a decade. The version under review includes runs of 30 newspapers, predominantly from the United States, spanning the years 1764?2005 and totalling some 27 million pages. The newspapers are drawn from three collections: ?US newspapers?, which contains 16 newspapers including the *New York Times* (1851?2007, *Washington Post* (1877?1994), *Wall Street Journal* (1889?1993 and *Chicago Tribune* (1849?1987); ?Black Newspapers, which contains a further 9 American publications including the *Chicago Defender* (1910?75), and *Baltimore Afro-American* (1893?1988); and the much smaller ?International titles?, which contains five publications from Britain, Ireland and India. This rich collection of material is accessed via ProQuest?s new interface, currently being implemented across a number of their products, which permits basic full-text searching across the publications as well as an impressive ?Advanced search? for more sophisticated queries. Given the importance of the newspapers that it contains and the easy way in which this interface provides access, *ProQuest Historical Newspapers* makes a significant contribution to both the study of the press and to modern history more broadly.

ProQuest Historical Newspapers offers unrivalled access to some of the world?s most important modern newspapers. Resources like this, however, do not come cheap, often costing tens of thousands of pounds in the first instance and committing libraries to annual access fees running to the thousands.(1) *ProQuest Historical Newspapers* is competing in a crowded marketplace, where a number of different resources ? some commercial, others funded with public money ? provide access to historical newspapers on a variety of different terms. One of the oldest resources is a commercial product: Heritage Microfilm?s *newspaperARCHIVE* [2]. Initially launched in 1999, this resource now claims to be the largest in existence, containing over 3000 individual publications and running to 115 million pages. A more recent commercial rival is Readex?s (a subsidiary of Newsbank) *World Newspaper Archive*, which currently consists of four major collections ? *African Newspapers, 1800?1922* (43 publications; 20 to come); *Latin American Newspapers, 1805?1922* (36 publications; two to come); *East European Newspapers, 1835?1922* (two currently; 18 to come); *South Asian Newspapers, 1864?1922* (7 currently; two to come) ? with two more

major collections in production. These commercial resources face fierce competition from a number of impressive publicly-funded resources. In the United States the National Digital Newspaper Program (NDNP) recently launched *Chronicling America: Historic American Newspapers* [3] (2007), a resource that currently offers free access to 457 publications and contains around 3.5 million pages. In 2001 the National Library of New Zealand launched *Papers Past* [4], providing free access to 63 publications. In 2008 the National Library of Australia launched *Australian Newspapers* [5], which provides free access to 130 Australian publications and is now cross-searchable with the rest of the library?s holdings via their portal, *Trove*. There is also, of course, *Google News Archive* [6] (2006?), which currently contains around 2500 publications from around the world. The UK press is covered by the British Library and Gale Cengage?s *British Newspapers, 1600?1900*, which consists of two collections, *17th?18th Century Burney Collection Newspapers* and *19th Century British Library Newspapers*.(2) Thanks to Joint Information Systems Committee (JISC) funding it is free to access from within UK Higher Education institutions but outside these is limited to subscribers.

A useful list of digitized newspapers (and their respective conditions of access), organized by country, has been maintained on *Wikipedia* since 2006.(3) This list makes it clear that there is no shortage of this historical material available in digital form, often for free. What *ProQuest Historical Newspapers* provides is access to most of the best-known American newspapers of the past 200 years. Many of these are available elsewhere (for instance, publications still in existence such as the *New York Times* and *Washington Post* provide access on a subscription basis through their own websites), but the ability to cross-search such an important corpus makes this an invaluable resource for anybody interested in the American press or American history more broadly. The publications included in the resource as ?Black newspapers? are a fascinating resource in their own right but, although marketed separately, really need to be consulted alongside the ?US newspapers? with which they share a print culture. The ?International newspapers? collection is much more limited, consisting only of the *Guardian* and *Observer*, *Irish Times* and *Weekly Irish Times*, *Scotsman*, and the *Times of India*. These are all important publications, of course, and presumably ProQuest will enhance this part of the resource, but they sit oddly amongst the strong and diverse selection of American publications that make up its bulk.

By providing access to (fairly) comprehensive runs of these particular papers in a searchable form, *ProQuest Historical Newspapers* constitutes a major resource of value to anyone interested in modern history. But it does so, like all digital resources based upon this material, by transforming newspapers into something else. Nearly all these resources deliver articles in the form of segments of page images, made searchable via a transcript produced by optical character recognition (OCR) technology. By indexing content and making it available online, this methodology allows digital resources to overcome most of the significant difficulties associated with using the press archive in hard copy. *ProQuest Historical Newspapers* seeks to justify its price by offering itself as the only place where its constituent publications can be consulted at once and in so complete a form. The index, although limited to verbal information and reliant on both the quality of the newsprint and the efficiency of the OCR software, opens up this vast and complex archive, allowing its users to find material within runs and trace it across them. As this content is delivered over the web, access to the archive is available 24 hours a day to anybody with a subscription and an internet connection. The crucial question, then, is not whether *ProQuest Historical Newspapers* is worth its price tag (it is clearly what the market will bear) or whether it will contribute to our understanding of history (there is no doubt that it will, and in significant ways), but on how it transforms the newspapers that it republishes in digital form.

In ProQuest?s marketing materials they make much of the fact that the resource will be useful for both ?casual explorers and serious researchers?.(4) The usual way to accommodate these user groups is through the interface, providing different levels of complexity for different users (or uses). *ProQuest Historical Newspapers* now uses ProQuest?s generic interface, launched in 2010 and currently being implemented across their resources. Its default is a basic search and it is admirably uncluttered, offering a very usable gateway to content. All users are familiar with Google?s web search and so are comfortable when confronted with a simple search box and the minimum of information about what is to be searched. For more detailed

queries, the advanced search allows readers to use the usual range of Boolean operators and take advantage of the metadata encoded within the resource. Some of these are fairly standard (?Author?, ?Document title?, etc.) but there are others such as ?Page?, ?Section? and ?Document type? (?Advertisement?, ?Editorial?, ?Front page?, ?Letter to the editor?, ?News? etc) that respond to the newspapers, allowing users to interrogate them in imaginative ways.

Results are displayed in a list sorted by relevance (although how this is calculated is not detailed anywhere), but the user can opt to sort by date and further refine the list by type of metadata (?Publication title?, ?Document type?, ?Keyword?, ?Database?, ?Date? etc). Results can be exported, cited (via Refworks), printed or saved and there is a nice visualization showing the distribution of hits over time. When the user clicks on a result, the article opens as a pdf document, this familiar format allowing the user to easily grasp the basic functionality. In addition, ProQuest have provided the same options that appear on the results screen (export, cite etc.), as well as a list of the metadata attached to the article, and have included an AddThis [7] box so that users can share links over various social media. They have also built in a crowdsourcing element to the interface, allowing users, when signed-in, to tag articles and share their tags with others. User-generated metadata is a well-established part of many digital resources, but tends to be of most value when users feel they are part of a community. There is much potential here, but it will be interesting to see the extent to which is utilized by the users of these subscription resources.

In offering a generic interface for their products, ProQuest are selling the idea that they are providing access to content, rather than republishing it in a new form. As it is a generic interface, it contains no information about either the resource or its contents. Although users can see a list of the publications within *ProQuest Historical Newspapers*, there is nothing about the newspapers, their historical significance, or even their scope. As the web is only a browser window away, this kind of general information is, perhaps, not so important. But there is also nothing about how the newspapers were selected or digitized. In fact, the entire account of how these articles have got to the screen has been excluded. This includes their derivation (which run? Which edition? From which archive?), their condition (what is there? What is missing?), how they were produced (from hard copy? Microfilm?), how they were edited (cropped? De-skewed?), how they relate to the transcripts (what is the text that is searched? Has it been corrected?), and how the metadata schema was designed and implemented. There are two undocumented histories of production here, one for the printed newspaper and one for digital resource, and both are essential if users are to put what they find in context. Without such information, the user cannot establish how the digital material relates to the print material upon which it is based.

One of the consequences about imposing a generic interface is that search is privileged over browse. This is a missed opportunity: one of the notable aspects of *ProQuest Historical Newspapers* is the amount of metadata that appears to have been encoded (without the documentation, of course, one cannot be sure). If users want to browse the contents of the resource, they can use the ?Publications? link situated alongside the two search options, ?Search? and ?Advanced Search?. This brings up a list of 108 different ?publications?, really different runs of the 30 publications that make up the resource, from which users can select one to explore in more depth. From here, users can browse by decade, year, month and issue, finally opening up a set of results that correspond to the articles, in page order, that make up a particular issue. If a user wants to flick through the pages of an issue rather than browse a list of article titles (and accompanying metadata), he or she must open up an article and then click on ?Page view (clickable)?. Although neither of these methods of browsing are particularly elegant, they are vital if users are to develop a sense of the newspapers as discrete publications, with their own structures and organizational logic.

The lack of information provided in the interface and the privileging of search make it difficult for the user to fully take advantage of the potential of the resource. As mentioned above, there is a fairly rich set of metadata encoded within the resource, but this has not been fully exploited in its design and, because it is not documented, it is difficult for the user to work out how it might be employed in the course of his or her research. A good example of this is the way that the resource makes images accessible. The press, of course, is a vital repository for a whole range of visual material, from illustrations in advertisements to documentary photographs, and the ability to access this material is incredibly valuable. Although images are marked with

metadata (presumably either ?illustration? or ?Image / Photograph?, judging from the search options), the search actually runs on any verbal information that accompanies the image. As there is no way to browse images, users must try and guess the sorts of words that might appear in accompanying articles or captions. For some users, this works very well: the user who wants images of a particular event, for instance, will be well served and the corpus is large enough to provide some results for even the more esoteric queries; however, those users interested in the images themselves, rather than what they represent, will find it very difficult to locate what they want and, of course, any images that lack accompanying text will be lost within the archive.

Those using the resource to research visual material are likely to be disappointed. Although it is not stated anywhere, the source of the page images appears to be microfilm produced with a high tonal contrast. This means that the tones of the printed image, whether engraving or photomechanical reproduction, have been reduced to large areas of black and white in the digital representation. The quality varies according to the publication, the condition of the page, and the size and type of image. Line drawings, for instance, are reproduced very clearly, but quite a few photographs have become entirely obscured. Given the time and expense involved in producing digital images from hard copy, it is understandable that ProQuest have drawn upon existing collections of microfilm, particularly if they are already in their possession. However, visual material, whether images or typographical features, is an integral component of the newspaper. If this resource is to serve researchers as the principal mode of access to these publications, then it is vital that visual material adequately represented. Or, at the very least, documented so that researchers can understand why it appears as it does.

Taken on its own terms, *ProQuest Historical Newspapers* provides a way to search this material that is so well-executed that it is likely to become the principal mode of access for those lucky enough to have subscriptions. The importance of the newspapers that it contains and the considerable effort expended to make them searchable guarantees that it will have a far-reaching impact on a range of disciplines interested in the 19th and 20th centuries. On these terms, the resource is an undoubted success, opening up the archive in such a way that its value is clear for all to see. Nevertheless, we should not overlook the conditions upon which we are given access to this wealth of content. In privileging search over browse, the article over the page (or, indeed, the issue or run), and the verbal over the visual, *ProQuest Historical Newspapers* conceives the newspaper as a repository of information and encodes this conception within its design. The marketing materials claim that the resource permits users to ?travel digitally back through centuries to become eyewitnesses to history? and it is as indices to their times, rather than cultural agents in their own right, that these newspapers have been digitized.(5) Given the cost of producing a digital resource on this scale, especially one that republishes content that is still in copyright, we can be sure that ProQuest know their market. On the evidence of this resource, ProQuest are targeting casual users who want to read written accounts of past events. It can certainly be used to study the press ? searchable runs of newspapers are themselves a substantial boon for media historians and there is a degree of flexibility in the interface ? but the researcher will necessarily be working against the grain of the resource. ProQuest are competing on breadth of scope, combining content from different newspapers; however, the newspapers themselves have been harvested for articles that are offered to users without the necessary context which would allow them to discriminate between them. As the generic database is rolled out across their different resources, these articles become further removed, taking their place alongside other types of content from other data sets. *ProQuest Historical Newspapers* republishes a glorious set of material from the past but, in order to make it pay, does so in the form of an expensive resource that cannot do full justice to the rich and complex newspapers that constitute its content.

Notes

1. ProQuest asked that I remove the any reference to how much their resource costs as they do not make their prices available in the public domain, preferring instead to provide them on an institution-by-institution basis. Their prices depend on size and type of institution, and they offer discounts to institutions that subscribe to multiple resources. So, if you want to know how many tens of thousands

something like this resource will cost your institution, ask a librarian.Back to (1)

2. Reviewed in *Reviews in History* by Martin Conboy. See ?Review of *The 19th Century British Library Newspapers website*, (review no. 730)? <http://www.history.ac.uk/reviews/review/730 [8]> [accessed 30 April 2011].Back to (2)
3. ?List of online newspaper archives?, *Wikipedia* (2001-) < http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives [9]> [accessed 18 April 2011].Back to (3)
4. ?Centuries of discovery online: *ProQuest Historical Newspapers*? (no date) < http://www.proquest.com/assets/literature/products/databases/HNP.pdf [10]> [accessed 18 April 2011]. Back to (4)
5. ?Centuries of discovery online?.Back to (5)

Other reviews:
[11]

Source URL: http://www.history.ac.uk/reviews/review/1096

Links:
[1] http://www.history.ac.uk/reviews/item/5538
[2] http://www.newspaperarchive.com/
[3] http://chroniclingamerica.loc.gov/
[4] http://paperspast.natlib.govt.nz
[5] http://trove.nla.gov.au/ndp/del/home
[6] http://news.google.com/newspapers
[7] http://www.addthis.com/
[8] http://www.history.ac.uk/reviews/review/730
[9] http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives
[10] http://www.proquest.com/assets/literature/products/databases/HNP.pdf
[11] http://www.history.ac.uk/reviews