The Importance of "Yesterday's News": Opportunities & Challenges in Newspaper Digitization

Alison Jones, Tufts University

Introduction

Newspapers have been a part of daily life for centuries. The contemporary desire to stay informed not only about the events in our local communities, but about the world at large is illustrated by the profusion of twenty four hour news stations and the ever increasing presence of online newspapers. Newspapers often seem to be ephemeral these days as content is produced exclusively for online access and frequently disappears without a trace. Thus the long standing questions of how to both preserve and provide improved access to historic newspapers have taken on new urgency in the digital world.

Historic newspapers have always presented researchers with a number of problems regarding their access and effective use. The lack of indexes to newspapers, particularly regional newspapers and those published in the eighteenth and nineteenth centuries present a serious challenge to even the most arduous researcher. The use of historic newspapers usually meant one had to spend days, if not months, in a dusty archive, or more likely, scanning countless reels of microfilm. Furthermore, newspaper collections and holdings are often far flung throughout various archives across the country making their use even more difficult. Despite these difficulties, researchers such as historians, genealogists, and others continue to use older newspapers because of the wealth of information they provide. Newspaper researchers have found an immense amount of information not only in articles and major stories, but in all sections of the newspaper including the advertisements, birth and death notices, property transactions and editorials.

Another issue facing researchers is the frequent paucity of regional and local history provided by major "papers of record." While digitization and access to major newspapers such as the *New York Times* and *The Chicago Tribune* has already been undertaken by major commercial firms, digital access to regional and state papers which may present a lower financial return is a more questionable gambit. Yet it is often the local and smaller newspapers, many of which began to flourish in the nineteenth century, that can prove to be the most important and neglected sources for historians.[1]

In the nineteenth century, the penny press began to largely succeed the earlier party newspapers and mercantile presses. Readership was moving from an elite audience to a broad based readership where street sales became common and the importance of advertising as a source of newspaper revenue became predominant.[2] One scholar has argued that the content of the penny press is "on the one hand reflective of the interests, anxieties, & aspirations of their broad based readership-and on the other hand indicative of the vision and biases of publishers, editors and reporters."[3] Using regional and local newspapers provides an important means of gauging local opinions on historical events, while bearing in mind newspapers could be highly partisan and often represented the views of a particular political party or editors rather than an objective factual representation of events.

---

[1] For two excellent discussions of the potential research uses of newspapers please see, Taft, William H. *Newspapers as Tools for Historians*. Columbia, Missouri: Lucas Brothers Publishers, 1970; Knudson, Jerry W. "Late to the Feast: Newspapers As Historical Sources." *Perspectives*, American Historical Association, October 1993.

[2] Martin, Shannon E. and Kathleen A. Hansen. *Newspapers of Record in a Digital Age: From Hot Type to Hot Link*. Westport, Connecticut: Praeger Publishers, 1998.

[3] Vanden Heuvel, Jon. *Untapped sources: America's Newspaper Archives and Histories*. Prepared for the Newspaper Editors' Newspaper History Task Force by the Gannett Foundation Media Center at Columbia University in the City of New York. Eds. Craig LaMay and Martha FitzSimons. New York: Gannet Foundation Media Center, 1991.

In an effort to preserve these important historical sources, the National Endowment for the Humanities (NEH) and the Library of Congress established the United States Newspaper Program (USNP) in 1985. This program serves as "a cooperative national effort among the states and the federal government to locate, catalog, and preserve on microfilm newspapers published in the United States from the eighteenth century to the present."[4] This focus on the microfilming of newspapers, however, has met with some controversy as critics such as Nicholson Baker and others have argued that microfilm is not an effective preservation method, and that newspapers are only useful as research tools when maintained indefinitely in their original format for browsing purposes.[5]

The need to digitize historic newspapers, either as a means of preserving them or increasing access, is a timely issue for both librarians and researchers. Recently, two major projects to digitize a large number of historic newspapers have been announced by both the British Library and the NEH. The British Library announced in June of 2004 that more than one million pages of nineteenth century U.K. newspapers will be digitized and made available on the Internet. The program will be co-managed by the Joint Information Services Committee and will allow full text searching of various historic newspapers.[6] The NEH who will again be working in partnership with the Library of Congress has requested proposals for partner institutions to participate in the beginning phases of creating a National Digital Newspaper Program.[7]

In addition, there are also a large number of historic newspaper digitization projects which have already been completed as well as a number which are still in production phases. This paper will provide an overview of the main newspaper projects and offer a comparison of the types of content they provide, their searching capabilities, and how they provide access to historic newspapers. It shall also provide a brief discussion of the major software options and technology solutions that are available. Finally, it shall provide an overview of some of the best practices, cost issues and remaining problems that need to be addressed in the future.

Problems with Digitizing Historic Newspapers

The digitization of historic newspapers, particularly those from the nineteenth century and earlier, present a number of challenges. To begin with, the newspapers size, structure and layout are in themselves an impediment to large scale digitization project. Nineteenth century newspapers usually presented a page of text that began simply with an item in the upper left hand corner and read down the column until the item's conclusion, which would be marked by a single line. The next item would usually begin immediately afterwards. The pages were designed to be read from top to bottom going across the page by each column. Items were rarely more than a single column in length and many items did not have headlines. For those items that did have headlines it was normally printed in boldface with the font only slightly larger than the text itself. [8]

The process of successfully scanning newspapers is also a difficult issue to resolve. Most projects have digitized their newspaper collections by scanning their preservation microfilm and then using optical character recognition technology (OCR) to make the text readable, if not always

---

[4] See http://www.neh.gov/projects/usnp.html (visited 7.18.2004)
[5] For a discussion of this debate please see Cox, Richard J. "The Great Newspaper Caper: Backlash in the Digital Age." *First Monday*, 5 (12), Dec 2000, http://www.firstmonday.dk/issues/issue5_12/cox/>
[6] See Kiss, Jemima. "Who Wants Yesterdays Papers." *Dot Journalism*, 16 2004 June. http://www.journalism.co.uk/news/story950.shtml
[7] For the British Library announcement http://www.jisc.ac.uk/index.cfm?name=press_release_newspaper for the NEH announcement http://www.neh.gov/grants/guidelines/ndnp.html. (Both visited 7.28.2004)
[8] "About The Valley Newspapers." The Valley of the Shadow: Civil War Era Newspapers. http://www.vcdh.virginia.edu/xml_docs/valley_news/html/about/about.html.> Accessed July 30, 2004.

searchable. Occasionally, the original newspapers themselves have had to be used when the microfilm was of poor quality, such as in the Waterford City Library project and for some of the collections used in the Utah Historic Newspaper Project. A paper written by the staff of the British Library Newspaper pilot, in cooperation with Olive Software, offers an excellent overview of the most serious challenges faced when digitizing newspapers.[9] There are a number of "material inherent" problems such as complex page layout caused by ten to hundreds of information objects possibly scattered across different pages. The complex layout makes it difficult to identify text areas, which in turn affects the quality of OCR accuracy. Another issue is that there is often little or no space between lines of text, and OCR engines cannot easily deal with dense blocks of text. In addition, the absence of article titles, particularly with pre-1900 newspapers, makes it more difficult to provide meaningful search results. To resolve these issues, the British Library chose to use the customized software ActivePaper Archive (APA) produced by Olive Software.

This same article also talks about a number of image quality problems that have to be surmounted. Microfilmed pages often contain rotated or curved characters and images, because of the fact that pages were often skewed and still attached to their binding when microfilmed. There is also the problem of "garbage" or "noise" which may have been caused by dirt on the original newspaper page or scanning lens. Other major difficulties can include broken vertical or horizontal lines, broken characters and faded text. All of these issues affect the quality of OCR recognition

Another article written by the staff of the Utah Digital Newspapers Program, further addresses these concerns. In "Microfilm, Paper and OCR: Issues in Newspaper Digitization," Kenning Arlitsch and John Herbert discuss the advantages and disadvantages of scanning from microfilm versus hard copy newspapers. The first newspapers digitized for their collection were scanned from microfilm, but eventual problems with the quality of the microfilm led them to pursue print archives. Among the advantages microfilm presents for digitization projects are inexpensive scanning, low conservation costs, and frequent availability. Digitizing from paper, however, if the paper is in good condition, can have the advantage of providing cleaner digital images, and therefore, more accurate OCR results. The authors ultimately found that "original newspapers provide a ten percentage point improvement in OCR accuracy over microfilm" but suggest that each digitization project shall have to explore for itself whether microfilm or hardy copy better suits their needs.[10]

Perhaps the most significant challenge faced in presenting a digital version of a newspaper is that of how to create one that can be searched, not just browsed. The paper, "Digitizing Historic Newspapers: Progress and Prospects" provides an excellent overview of some of the major problems faced in creating content that is not just readable but actually searchable. The authors argue that "creating searchable content is a much more difficult process, given the complexity of the newspaper page and the mixed media formats….early attempts at Optical Character Recognition (OCR) failed because the quality achieved was too poor for adequate retrieval (and correction too costly) and because the OCR engines operated on linear text, not individual content objects. The structural unit of the page was recognized, not the logical unit of the item."[11] Different software products have been created to address these issues, a topic covered later in this

---

[9] Deegan, Marilyn, et. al. "The British Library Newspaper Pilot." <http://digitalcooperative.oclc.org/digitize/BritishLibraryNewspaper.html> Accessed July 17, 2004.

[10] Arlitsch, Kenning and John Herbert. "Microfilm, "Paper, and OCR: Issues in Newspaper Digitization." *Microform and Imaging Review*. 33 (2), Spring 2004. <http://www.lib.utah.edu/digital/unews/pdf/MicroFilmArticle.pdf> Accessed 7.19.2004

[11] Deegan, Marilyn. et. al. "Digitizing Historic Newspapers: Progress and Prospects." *RLG Diginews*. 6 (4), August 15, 2002. http://www.rlg.org/preserv/diginews/diginews6-4.html#feature2 Accessed 7.27.2004.

paper. While some libraries have chosen to go with customized software packages, others have decided to take a simpler and less expensive in-house approach by scanning in their newspapers directly and using an off-the-shelf OCR package.

Major Digital Newspaper Projects-Basic Overview

There is a great deal of variety in terms of the digital newspaper collections that are currently available. While some projects have been undertaken by major commercial vendors and are targeted for sale to academic institutions such as libraries, there are also a number of freely available and significant digital newspaper collections that have been created and maintained by academic and cultural institutions themselves. There are also more limited projects focused on just one newspaper. In addition, there are also several commercial websites with significant historical newspaper coverage that are targeted directly to end users, like genealogists, over the Internet for annual or monthly subscriptions. This section shall provide a brief overview of these different newspaper collections and later sections will compare the search methods utilized, the display and newspaper viewing options provided, and an overview of the software packages and digitization methods used to create the collections.

Major Commercial Vendors

Several major vendors such as ProQuest Historical Newspapers and Gale have concentrated on scanning the entire run of major papers like *The Times* and *The New York Times*. In the case of Gale, an entire digital archive of *The London Times* from 1785 to 1985 has been created.[12] In the article, "The Thunderer on the Web-The Times Digital Archive 1785-1985" printed in the July 2003 issue of *Library and Information Update*, the authors, who also worked on the database development team, address the project at length. After discussion with groups of scholars and librarians, they decided to make the complete full text and page images of the newspaper available for searching, including display and classified advertising.[13]

ProQuest Historical Newspapers has taken an even more ambitious approach and made the full text and full image of every issue of the *New York Times* available from its beginning in 1851 until 2001. This collection includes digital reproductions of every page from every issue as PDF files. ProQuest has also digitized major parts of the *Wall Street Journal*, *The Washington Post*, *The Christian Science Monitor*, *The Los Angeles Times*, and the *Chicago Tribune*. They have concentrated on making the entire runs of major "papers of record" available and seek to bring "historical research to life".[14]

Another vendor, Accessible Archives, has made the full text only of a number of historical newspapers available online. They offer several different databases including: "The Civil War: A Newspaper Perspective," which contains "the full text of major articles gleaned from over 2,500 issues of *The New York Herald*, *The Charleston Mercury* and the *Richmond Enquirer*, published between November 1, 1860 and April 15, 1865."[15] They also offer access to several nineteenth century African-American newspapers and the *Pennsylvania Gazette* from 1728-1800.

---

[12] See http://www.galegroup.com/Times/ (site visited 7.24.2004)
[13] Readings, Reg and Mark Holland. "'The Thunderer' on the web - The Times Digital Archive 1785-1985." *Library + Information Update*. July 2003. http://www.cilip.org.uk/update/issues/jul03/article2july.html
Accessed 7.23.2004.
[14] See http://www.bellhowell.infolearning.com/ProQuest/features/feature-04/default.shtml (site visited 7.26.2004)
[15] See http://www.accessible.com/about.htm. (site visited 8.01.2004)

There is also one major commercial undertaking that has not yet been released. The company Readex, a subsidiary of Newsbank, is cooperating with the American Antiquarian Society in creating a database called "Early American Newspapers, 1690-1876". The final product will include the full image and text of dozens of newspapers digitized from collections owned by various historical societies and libraries and the first release was scheduled for spring of 2004.[16]

Commercial Sites Aimed At End Users

In addition to the major commercial vendors, a number of smaller commercial operations have also decided to explore the business potential of digitizing historic newspapers and marketing them directly to end users. Three of the major sites are Paper of Record, Ancestry.com and Newspaperarchive.com. All three offer access to historic newspaper archives through either a monthly or annual subscription.

Paper of Record is a commercial service run by Cold North Wind, Inc.[17] This company has digitized over eight million pages of both current and historic newspapers from reels of microfilm by contracting with major cultural institutions to use their microfilm and then scanning it in with their own proprietary OCR technology. The company is unique in that "the project is built on partnerships with organizations that own valuable collections of historical newspapers on microfilm. These partnerships are designed to reap the benefits of a united approach to the digitization, marketing and distribution of this remarkable view of the past."[18] They also provide contract scanning and digitization services. While users can search the archives on their website for free, to view articles you must be a subscriber. This project offers extensive Canadian coverage as well limited coverage of other nations. The dates available for each newspaper range greatly although they do have many newspapers available from the nineteenth century. Cold North Wind has also been involved in the digitization of the major Canadian newspaper, *The Toronto Star*. They digitized the entire content of the newspaper from 1892 to 2001 and made it both searchable and browsable.[19] This database is also available by subscription only.

The website Ancestry.com provides a number of subscription services to genealogists and they recently added a historic newspaper collection to their offerings. According to their website they offer "6 million pages from over 400 different newspapers across the US, U.K. and Canada dating back to the 1700's."[20] In reality, they offer mostly U.S. newspaper coverage with six titles from Canada and nine from the U.K. Newspaperarchive.com offers a similar range of services. The website is a commercial offshoot of the company Heritage Microfilms and advertises that they have historic newspapers from the U.S., Canada, U.K, Ireland, Denmark and Jamaica though the non U.S. content is actually very limited.[21] As with Paper of Record, the date coverage available for the different newspapers varies greatly, although each site offers a fair amount of nineteenth century newspaper coverage.

Major Freely Available U.S. Digital Newspaper Projects

[16] See http://www.readex.com/scholarl/earlamnp.html (Site visited 8.13.2004)
[17] See http://www.paperofrecord.com/ (Site visited 7.27.2004)
[18] See http://www.coldnorthwind.com/ (site visited 7.24.2004)
[19] See http://www.pagesofthepast.ca/Default.asp (site visited 7.28.2004)
[20] See http://www.ancestry.com/search/rectype/periodicals/news/main.htm?lfl=ttd. (site visited 7.27.2004)
[21] See http://www.newspaperarchive.com (site visited 7.24.2004)

There are a number of excellent digital newspaper projects that have been created by libraries or other institutions and are freely available through the Internet. While some have taken a regional approach by providing access to a large number of state newspapers, other projects have chosen to focus on individual newspapers. This section will provide a basic overview of these projects.

Perhaps the best collection that involves only a single newspaper is *The Brooklyn Daily Eagle*.[22] This project was produced by Brooklyn Public Library's Brooklyn Collection and was co-funded by the Brooklyn Public Library and the Institute of Museum and Library Services (IMLS). This site contains 147,000 pages in various formats. Currently the full text and page images of the paper from 1841 to 1902 are available online. The collection has been created using the ActivePaper Archive software (APA).

Another digital newspaper project that is powered by APA is the Colorado Historic Newspaper Collection.[23] This project has been created by a partnership of The University of Denver, the Colorado Digitization Program, the Colorado State Library, and the Colorado Historical Society. They digitized over 44 historic newspapers dating from 1859 to 1880 that were on microfilm already owned by the Colorado Historical Society and both the full text and images are available. Their goal is to eventually digitize over 200 historic newspapers if more funding can be obtained. The original project was funded by the IMLS and the Library Services and Technology Act (LSTA).

APA has also been used to create a number of smaller U.S. newspaper projects. One such collection is the Historical Missouri Newspaper Project.[24] The project has digitized about 11 newspapers with greatly varying date content. Several papers such as the *Liberty Banner* and the *Missouri Republican* have only one month digitized, March 1844 and July 1865 respectively. The project, unfortunately, has some broken links including those describing the project background and the participants. The content is much more limited than any of the other major regional projects. An even smaller newspaper project using APA is the Historical Digital Collegian Archive.[25] Pennsylvania State University has used APA to digitize its college newspaper the *Daily Collegian* from 1887 to 1940 and made it available through the library website.

Another major newspaper project supported by a different software package, CONTENTdm, is the Utah Digital Newspapers project.[26] The University of Utah Marriott Library in partnership with Brigham Young University has digitized 136,000 pages or a total of 17 Utah newspapers. This project was partially funded by the IMLS and LSTA. They recently received an additional grant that will support them through September 2005 and they are planning to add another 240,000 pages of digital newspapers. The newspapers in their collection range in date from 1858 to 1948 and cover almost all of the counties in Utah.

A variety of newspaper projects focused on non-English language newspapers or newspapers published by ethnic and racial minorities. The Hawaiian Language Newspapers project is a pilot project whose goal was to digitally scan selected newspaper articles and microfilm rolls of significant Hawaiian language newspapers that would be "pertinent to Hawaiian language and history courses." They indexed the images on a basic level and their final project ended up with

[22] See http://www.brooklynpubliclibrary.org/eagle/index.htm (site visited 7.27.2004)
[23] See http://www.cdpheritage.org/newspapers/ (site visited 7.19.2004)
[24] See http://newspapers.umsystem.edu/archive/Skins/Missouri/navigator.asp?skin=Missouri&BP=OK (site visited 7.21.2004)
[25] See http://www.libraries.psu.edu/historicalcollegian/ (site visited 8.10.2004)
[26] See http://www.lib.utah.edu/digital/unews/ (site visited 7.24.2004)

16 native Hawaiian language newspapers published between 1870 and 1920. They provide GIF images of the newspapers and some transcribed articles.[27]

A more sophisticated project with similar content is the Hawaiian Nupepa Collection.[28] This site offers a collection of fully searchable Hawaiian language newspapers covering the period 1834-1948. It has been built with the open source software Greenstone Digital Library, which has been used to create numerous digital library collections such as the Maori Niupepa Project, which will be described later. The nupepa collection includes 120,000 news pages taken from 100 separate periodicals and is the "product of the Hawaiian Language Newspapers Project, operated by Alu Like, Inc., through its Native Hawaiian Library and its Hawaiian Language Legacy Program."

Another example is the Georgia Historic Newspapers Database, a project that is still currently in development. This database is part of Galileo, Georgia's Virtual Library, an initiative of the Board of Regents of the University System of Georgia to develop an extensive virtual library for Georgia.[29] It is also an outgrowth of the Georgia Newspaper Project which is part of the USNP. Currently it includes three newspapers: *The Cherokee Phoenix* from 1828 to 1833, *The Colored Tribune* from 1876, and *The Dublin Post* from 1878 to 1887. The *Cherokee Phoenix* was a newspaper published for Native Americans and *The Colored Tribune* was an African American newspaper. Facsimile images of all the pages are available as PDFs for viewing.

The only current project found that digitizes a Spanish language newspaper is the digitization of the *El Clamor Publico*, the first Spanish language paper in California after the revolution that was published from 1855 to 1859.[30] The completely searchable digital facsimile of this newspaper was created as part of the Digital Archive at the University of Southern California's Archival Research center.

Several other individual newspapers have also been digitized and made searchable. To begin with, the early twentieth century newspaper the *Morning Leader* from 1902 to 1903 has been digitized by the Port Townshend Public Library in Washington.[31] This project is hosted by the University of Washington Digital Libraries Collection and is another CONTENTdm based project. The final U.S. digital newspaper project reviewed here is the digitization of the *Stars and Stripes* by the Library of Congress for the American Memory project. *The Stars and Stripes* was a U.S. military newspaper published from 1918 to 1919 and this entire run has been digitized. The military would later use this same title for their service paper again in World War Two and has been publishing it ever since. Users can view full images of every issue of this newspaper and search the full text as well.[32]

One other digital collection that is also worth mentioning, although it is not a digital newspaper project, is The Valley of the Shadow project. This extensive website serves as a "digital archive of primary sources that document the lives of people in Augusta County, Virginia and Franklin County, Pennsylvania during the era of the American Civil War."[33] Historic newspapers are an integral part of this primary source collection. This website provides an excellent overview of how to read a nineteenth century newspaper and also provides extensive historical information about each newspaper, a type of information many of the other sites did not contain. It also contains information about the politics and viewpoint of each newspaper, its basic layout and

[27] See http://128.171.57.100/hnp/index.shtml (site visited 7.22.2004)
[28] See http://nupepa.org/cgi-bin/nupepa (site visited 7.18.2004)
[29] See http://www.galileo.usg.edu/express?link=zlgn (site visited 7.22.2004)
[30] See http://www.usc.edu/isd/archives/arc/digarchives/elclamor/ (site visited 7.25.2004)
[31] See http://wlo.statelib.wa.gov/ptpl/morning_leader.htm#From%20microfilm%20to%20Internet (site visited 7.27.2004)
[32] See http://memory.loc.gov/ammem/sgpsasquery.html (site visited 7.20.2004)
[33] See http://valley.vcdh.virginia.edu/newspapersp2.html (site visited 7.22.2004)

what could be found on each page. The newspapers used in this collection were the *Republican Vindicator* (1859-1867), *Staunton Spectator* (1857-1867), *Franklin Repository* (1859-1867), *Valley Spirit* (1859-1867), *Semi-Weekly Dispatch* (1861-862) and the *Village Record* (1863)

In addition to these freely available websites, there are several U.S. newspaper digitization projects that are still in either the proposal or development stage. The California Newspaper Digitization Project (CDNP) presents an excellent website that details the complete findings of their first phase, a feasibility study.[34] It includes an overview of the project goals and the request for proposals they sent to potential vendors. This site also includes links to the different vendors test sites that were set up as beta versions of a fully searchable California digital newspaper collection using one newspaper, the *Alta California*. This website provides an excellent opportunity to view a variety of different search interfaces and display options for historic newspapers.

Another proposal website is that of the Pacific Northwest Ethnic & Special Audience Newspapers Proposal.[35] This website describes a plan to eventually digitize a large number of ethnic and minority newspapers from the Pacific Northwest and have them be fully searchable. This site does not contain any digital images but does include a preliminary list of target newspapers.

Major Digital Newspaper Projects from Around The World

One of the largest European collections of historic newspapers available online is at ANNO or Austrian Newspapers Online.[36] The website interface is in German and the collection is made up of about twenty newspapers with varying date coverage. While the majority of the digitized newspapers are from the nineteenth century, some do go back to the late eighteenth century. The newspapers are viewable either as TIFF images or PDF files. Another major digital collection of nineteenth century newspapers and periodicals is "Australian Periodical Publications, 1840-1845".[37] This website is part of the Australian Cooperative Digitisation Project and contains more journals than newspapers. All items are viewable as multi-page PDF files.

Although several major U.S. newspaper projects used APA, only two international projects utilized this software. The major collection using this product is the British Library Online Newspaper Archive.[38] The newspapers available include London's *Daily News*, *The News of the World*, *The Weekly Dispatch*, *The Manchester Guardian*, and the *Penny Illustrated*. While the dates available for each paper vary, there are typically five distinct years for each paper, most commonly 1851, 1856, 1886, 1900, and 1918. These dates were chosen due to important historic events occurring in these years, and the project designers wanted people to be able to contrast and compare newspaper coverage on each event. The other international newspaper project supported by APA is the *Palestine Post*: 1932-1950, a website sponsored by Tel Aviv University.[39] Users can search various months of this publication for every year between 1932 and 1950. This site seems to be a test or pilot, however, since many of the links don't work and there is no information provided about the project.

Another minor or pilot project was carried out by the Waterford City Public Library in Ireland. They designed a pilot project site called "Waterford in Wartime" to test the feasibility of future

[34] See http://cpc.stanford.edu/CDNP/ (Site visited 8.03.2004)
[35] See http://www.lib.washington.edu/mcnews/ethnicnewspapers/ (site visited 7.28.2004)
[36] See http://anno.onb.ac.at/ (site visited 7.22.2004)
[37] See http://www.nla.gov.au/ferg/ (site visited 8.01.2004)
[38] See http://www.uk.olivesoftware.com/ (site visited 07.22.2004)
[39] See http://kipp.tau.ac.il/Archive/skins/Palestine/navigator.asp (Site visited 7.25.2004)

scanning projects, and for this pilot they digitized all of the available issues of the *Waterford News* from 1915 to 1917.[40] While Waterford has chosen not to make the full images of the entire newspaper issues available, they have provided full text and page images for all major articles.

The only major Canadian digital historic newspaper collection is the Early Alberta Newspaper Collection, which contains both daily and weekly Alberta newspapers.[41] They have over 650,000 page images in their collection ranging from 1885 to 1985. This project is part of "Our Future, Our Past: the Alberta Heritage Digitization Project."

One major European project that is still in the production stages is the Lambrakis Press Archives.[42] Currently this website is only available in Greek and it simply provides information about the collection. Their plan is to create a newspaper collection of over 1,000,000 pages from 1890 to the present. According to the main architects of the project, the creation of this site "aims both at the salvation of endangered material (paper) and at the creation of digital library services that will allow full utilization of the archives by all interested parties." [43]

A number of excellent digital collections are available from the New Zealand Digital Library. The first of these is the Maori Niupepa Collection, which consists of historic newspapers published primarily for a Maori audience between 1842 and 1932.[44] The project was created through the work of the Alexander Turnbull Library and New Zealand Micrographic Services. This collection includes facsimile images of original pages, bibliographic commentaries and English abstracts for many of the newspapers. Like the Hawaiian Nupepa project this site was built with the open source software Greenstone, and includes both an English and a Maori interface.

The second major newspaper digitization project from a New Zealand library is the "Papers Past" website.[45] This website contains digital images of over 600,000 pages from both regional and national nineteenth century New Zealand newspapers and periodicals. It has been created by the National Library of New Zealand and the Alexander Turnbull library.

The final international digital newspaper collection we will review here is the Tiden project or the Nordic Digital Newspaper Library.[46] This distributed collection contains Nordic newspapers from 1640 to 1860 and was formed by the participation of four major libraries: The Royal Library of Stockholm, the National Library of Norway in Mo, the State and University Library of Argus, and the Helsinki University Library, which coordinated all efforts. Each collection is maintained separately and all of the projects can be accessed from the main project website.

The Finnish Historical Newspaper Library is part of the Tiden project and when completed will contain the full text of all the Finish newspapers published between 1771 and 1890.[47] This database has both an English and Finnish interface and currently contains about 70 titles or 413,000 pages. Another Tiden project collection, the Norwegian Portal for Digitized Newspapers, has a website that is only available in Norwegian.[48] This collection contains the full

---

[40] See http://www.askaboutireland.ie/pilots/three/index_main.html (site visited 7.25.2004)
[41] See http://www.ourfutureourpast.ca/newspapr/ (site visited 7.27.2004)
[42] See www.lpa.gr/homegr.html (site visited 7.22.2004)
[43] Gatos, B. et. al. "A Digital Library from Newspaper Archives." ACM Digital Libraries 2000: A Demonstration Submission. <http://www.lpa.gr/acm/p1.htm > Accessed 7.27.2004.
[44] See http://www.nzdl.org/cgi-bin/niupepalibrary?a=p&p=about&c=niupepa&l=mi&nw=utf-8
[45] See http://paperspast.natlib.govt.nz/ (site visited 7.24.2004)
[46] See http://tiden.kb.se/ (site visited 7.27.2004)
[47] See http://digi.lib.helsinki.fi/index_en.html (site visited 7.28.2004)
[48] See http://www.nb.no/avis/ (site visited 7.28.2004)

text of ten digitized newspapers with varying date ranges, many of which go back until the early 19th century. The Swedish Historical Newspaper Archive does not have a comprehensive website detailing its collections, but according to the Tiden project website it is concentrated on "a collection of newspapers called "Posttidningar" and contains newspapers from the period 1645 to 1721. In order to use this database, the user has to log onto a Retrievalware database and follow a complicated set of instructions posted on the Tiden website. The final Nordic project listed on the Tiden website is the Danish historical newspaper archive, which apparently contains three fully browsable Danish newspapers published between 1759 and 1865. The links to this project, however, led to the main page for the State and University Library of Arhus. No links to the collection could be found from this page.

Project Evaluation-Comparison of Content and Goals

One of the major weaknesses of many newspaper projects was a lack of background information about the newspapers. Since newspapers, particularly those of the 19th century were often highly partisan in nature, information regarding their origins, editorial policy and any possible political orientation would have been extremely helpful. As the creators of "The Valley of the Shadow" indicate in their essay about Civil War era newspapers, "newspapers of the nineteenth century were overtly partisan, in the belief that a newspaper's mission was to promote the line of a particular political party…in the eighteenth and into the early nineteenth century, it was common practice for a newspaper to receive funding directly from a political party."

The *Brooklyn Daily Eagle* was one of the few major American projects that provided extensive editorial and historical information about its chosen newspaper.[49] It includes an extensive timeline about the newspaper that is also matched up against major Brooklyn historical events. While the Colorado Historic Newspaper project provides a list of titles but no information about any of the newspapers, the Utah Newspaper project includes some basic descriptive information about each newspaper title. The Georgia Historic Newspapers database offers catalog information on each newspaper such as publisher, dates of publication and subject headings.[50] Perhaps the least amount of information could be found at the Missouri Historic Newspaper project. It does not supply a comprehensive list of newspaper titles or information, which makes it more difficult to use their collection. The British Library newspaper pilot also offers no information about any of their chosen newspapers.

The smaller collections often provided better information and context regarding their newspapers. The *El Clamor Publico* project offers a brief but useful history regarding the political orientation and editorial stance of its newspaper. One notable site is the *Morning Leader.* They provide some excellent background regarding this newspaper, the town in which it is published, and some brief information about the larger national context during the scope of the collection. It helps the user to better understand this newspaper and the role it served in the small shipping community of Port Townshend, Washington. In general, it seems that these smaller sites that focused exclusively on just one or two newspapers were more likely to provide detailed information about their contents and history.

One of the most extensive and well done sites was *The Stars and Stripes* from American Memory. Not only do they include the full image and text of the newspaper's entire run, there is also a detailed section called "A Closer Look at The Stars and Stripes."[51] This section includes

[49] See http://www.brooklynpubliclibrary.org/eagle/index.htm. For example, choose the FAQ or the "Eagle History and timeline"
[50] See for example, http://io.gsu.edu/cgi-bin/homepage.cgi?style=&_id=82402189-1154469285-0195&_cc=1
[51] See http://memory.loc.gov/ammem/sgphtml/sashtml/sp.html. (Site visited 8.13.2004)

descriptive information about the paper's content with full text examples from the paper, a detailed history on the editorial staff, information about military censorship, a full staff list, and other useful links. There are also major historical links to help put the newspaper in its context, including a timeline and a historical map. Both the *Brooklyn Daily Eagle* and the *Stars and Stripes* serve as excellent examples of the types of added materials that make a digital newspaper collection both more interesting and accessible.

Another website that contained a wealth of useful information was the Maori Niupepa project. Scholars have created individual bibliographic commentary for the majority of niupepa which includes publishing details, background and subject matter, and information on the likely contents of the paper.[52] In contrast, the Hawaiian Nupepa project offers no information about any of the newspapers in its collection. The website *Australian Periodical Publications 1840-1845* contains descriptive information on each of its titles including publication information, previous titles and the subjects the periodical covered.[53] Similarly, the Papers Past website includes both a descriptive and bibliographic entry about each newspaper that also contains useful information about its political orientation.[54] The Finnish Historical Newspaper Library provides basic catalog information about each newspaper, such as dates published, publisher and place of publication.[55] Finally, the Norwegian Portal for Digitized Newspapers includes a brief history and description for each newspaper.[56] The vast majority of international projects, however, seemed to contain limited if any information about the newspapers in their collection.

The commercial sites varied in the amount of information they provided about each newspaper. Paper of Record, the commercial site from Cold North Wind, offers useful history about a number of its newspapers. When you choose to browse a particular paper you are provided an optional link to view the history of the newspaper; this brief history includes editorial policy, publisher and notable content. This is a very useful feature when available. The other commercial sites, Ancestry.com and Newspaperarchive.com seem to make the assumption that their researchers will already know what papers they wish to search, or that the history of the newspaper is unimportant due to the nature of their research needs.

Another element missing from many of these newspaper collections was important structural information about the newspapers such as typical page layout, type of information contained on each page, and various sections likely to be found. As described earlier, the way in which older newspapers were organized and laid out poses challenges not just to their digitization but also to effectively finding information with the newspaper. The *Brooklyn Daily Eagle* was one of the few digital newspaper projects to give major structural information about the newspaper itself. This website offers an extensive summary of the newspapers structure, such as what information can be found within advertisements, birth announcements, and various features such as human interest stories or society and entertainment pieces. A timeline is given as to how the newspapers structure changed over time including detailed sections that explain how the structure of the paper changed in five year increments, such as what sections changed or moved within the paper. It also provides an overview of what information would likely be found on each page of the newspaper during these time periods. Similar if more limited information is given by *The Stars and Stripes* and *The Morning Leader* projects. Useful information is also given by The Valley of the Shadow

---

[52] See for example, http://www.nzdl.org/cgi-bin/niupepalibrary?gg=text&e=d-0niupepa--00-0-0-014-Document-text---0-1l--1-en-50---20-about---001-001-0utfZz-8-0&a=d&gg=text&d=16commentary.
(Site visited 8.14.2004)
[53] See for example, http://www.nla.gov.au/ferg/bfull/14403668_bfull.html. (Site visited 8.14.2004)
[54] See for example, http://paperspast.natlib.govt.nz/data/NZA/description.html. (Site visited 8.15.2004)
[55] See for example, http://digi.lib.helsinki.fi/digi/aurora/nimeketiedot.jsp?p_tunnus=0355-6913&p_tyyppi=kausi&p_kieli=en. (Site visited 8.17.2004)
[56] For example, see http://www.nb.no/avis/aftenposten/. (Site visited 8.12.2004)

project as to the layout and contents found within each of their newspapers, including an explanation of the column layout on each page. Providing this type of information makes browsing a much more fruitful experience for the user. It can lessen the amount of time needed to find particular information, such as if a user wanted to quickly locate the "death" or "obituary" section.

Another important difference between many of the projects was their ultimate goals or the reasons for which they were created. While some projects were limited due to their restricted funding, other projects were designed specifically to meet certain goals. For example, some projects served as pilots from which to study the success of digitization for future efforts. Consequently the designers picked either limited runs of newspapers, or target time periods for a number of newspapers. For example, the British Library Newspaper Pilot did not utilize all of the advanced features of APA because it was intended as a test project on which to base future digitization projects. Their main goal was to digitize selected years of several major newspapers so people could compare historic events, and see how successful APA would be at providing full access to their collection.[57] Another pilot project was done by the Waterford City Library, who chose to digitize four years of their city newspaper the *Waterford News* from the World War One era. They chose to do such a limited run because their main goal was to create a successful test project that would help them develop guidelines for further digitization projects as well as create justification that such projects could be feasible for public libraries.[58]

The various libraries involved in the TIDEN project or the Nordic Digital Newspaper Library also had different objectives. While the goal of the Norway and Denmark libraries was to get as much of their collections online as possible with minimal search capabilities such as browsing by title and date, Finland and Sweden decided to include full text search options even if that meant adding more limited content. Consequently, the amount of materials available from each participating library differed greatly. In addition, for all of the libraries involved in the Nordic Digital Library, the content that they ultimately chose to digitize was determined by the importance of the newspapers as well as copyright issues. The Royal Library of Sweden chose to digitize *Post-och Inrikes* from 1640 to 1721 while Finland chose to "build a digital platform depicting the day to day life in the 18th and 19th centuries." Denmark chose to digitize one newspaper *Adresseavisen* from 1751 to 1890, and, finally Norway chose to include a number of newspapers from the 19th and 20th century. [59]

The Australian Cooperative Digitization Project which created "Australian Periodical Publications, 1840-1845", selected materials based on Ferguson's Bibliography of Australia and confined to a critical six year period in Australian history. They chose a small time period not only to make their project more manageable, but also because this time period was extremely important in Australian history.[60] In addition, the creators of this project were most concerned with increasing access to the materials rather than developing a robust variety of searching options, which is one reason their website is currently only browsable. Similar concerns were echoed by the creators of the Papers Past website at the National Library of New Zealand. They called their project "part of the library's long term commitment to digitisation as a primary way to increase access to the library's collection."[61] The creators of the Maori Nupepa collection had a

---

[57] Deegan, et. al. "The British Library Newspaper Pilot."
[58] Fitzgerald, Emer. "Newspaper Digitisation Pilot Project." Waterford City Library, 3.30.03. < http://www.askaboutireland.ie/wcityreport.pdf> Accessed 8.14.2004.
[59] Bremer-Laamanen, Majlis. "The Nordic Digital Newspaper Library." NORDINFO-NYTT. February 2001. < http://www.nordinfo.helsinki.fi/publications/nordnytt/nnytt2_01/bremer.htm>. Accessed 7.16.2004.
[60] Thompson, John. "Electronic Alchemy: The Australian Co-operative Digitisation Project, 1840-1845." < http://www.nla.gov.au/ferg/jthomp.html>. Accessed 7.27.2004.
[61] "About Papers Past." National Library of New Zealand. <http://paperspast.natlib.govt.nz/about.html>

slightly different focus. While they wanted to make their collection accessible to researchers and students of the Maori language, in particular they wanted to create a significant Maori language resource for "Maori medium education, where there is paucity of Maori medium materials"[62]

The original goal of most projects funded by the IMLS was to serve as "demonstration" projects and help discover the best practices and ultimate viability of creating a historical digital newspaper archive. Several of the projects are still actively adding content depending on their funding status. Both the Colorado Historic Newspaper Collection and the Utah Digital Newspapers project have recently received new grants to further add to their collections. According to the Colorado website, it is "the intent of the project partners that Colorado's Historic Newspaper Collection would eventually include papers through 1923, a total of 1,640,000 pages." The Utah project has a similar goal of ultimately digitizing additional newspapers. They also wish to begin administering a training program to other academic and historical institutions who wish to launch similar initiatives. The Missouri Historical newspaper project did not provide any information about their project or any future plans.

The Utah project was the only major IMLS project to describe in detail how they selected their content for digitization. They concluded that adding any large daily newspapers from Salt Lake County would quickly exhaust their funds because of the high page volume. They thus sought input from several major Utah historians. They asked them which counties other than Salt Lake County and those with newspapers already represented in their collection would be of the most historical significance if added. After getting a list of the most historically significant counties, they then chose newspaper titles that were available on microfilm from those counties [63]

Some of the smaller American newspaper collections also provided some information about their goals or collection focus. The stated goal of the Georgia Historic Newspaper Database is to "convert every Georgia newspaper to digital format and to make this resource available free of charge as a searchable text database in GALILEO."[64] So far they have digitized only thee newspapers, and no plans have been announced to digitize more. The Hawaiian Language newspaper project also had a fairly limited plan. Its main goal was to make a large number of Hawaiian language newspapers accessible for classes to use. Their goal was not to create a full text searchable collection but rather to increase access to their unique newspaper collection.

To truly add value to the research experience, digital newspaper collections should include relevant historical and editorial information about their newspapers. Knowledge of a newspaper's political orientation, major editors and any potential biases are important when assessing their content. Structural information should also be provided about each newspaper's layout when possible. This is greatly helpful to those users who wish to quickly browse a newspaper to find specific content. Another factor that is important in determining the success of a newspaper collection is what level of searching is supported. The next section shall examine in detail the kinds of search option provided.

---

[62] "Breaking the Browsing Barrier for Historic Searching of Newspaper Text."
< http://www.cs.waikato.ac.nz/~tetaka/Browsing%20Barriors.htm>
[63] Arlitsch, Kenning and John Herbert. "Digitalnewspapers.org: The Digital Newspapers Program at The University of Utah." *Serials Librarian*, 47, (1+2) Nov 2003. <http://www.lib.utah.edu/digital/unews/serials_librarian.html>
[64] See http://io.gsu.edu/cgi-bin/homepage.cgi?style=&_id=82402189-1154474748-8500.

Project Evaluation-Search Methods Available

The newspaper projects surveyed offered a variety of different searching options. This section discusses how the projects differed in the options they offered by considering the following questions:

- What full text search methods are available?
- Are there any advanced search options available such as Boolean searching? Proximity operators? Are these options prevalent?
- What type of date limitations are offered? Is browsing by date supported?
- Can you save a list of items for later viewing?
- How common is the assigning of categories to articles? Is this prevalent across the different systems?

For a table comparing general search features please go to Appendix One.

Commercial Projects

In general the major commercial projects such as the London Digital Times Archive and ProQuest Historical Newspapers tend to offer the most robust searching options, particularly in the ability to build more sophisticated queries using truncation, proximity operators and field searching. They also offered more advanced date searching options, such as limiting search results between specific dates, or before or after a particular date. Another major feature that these two databases offered was advanced category searching. In Gale's Digital Times Archive, categories were assigned to the metadata of every article, so that results could be limited to particular types of items such as advertising, law reports, birth notice, marriage announcement and obituaries to name a few of the many options. Every item that can be searched on and viewed has been assigned a category, rather than a subject to facilitate searching. In the help section, you can find out what type of news is included within each subcategory. ProQuest Historical Newspapers also allows searching by article type or category. There are over 15 categories which can be searched against ranging from birth notices and comics to lottery numbers and real estate transactions. Another convenient feature of these two databases is the ability to save search history or keep track of what searches you have already done.

The more minor commercial operations were quite varied in their search features. Ancestry.com offers the special feature of searching by personal name, not just by keyword, reflecting the fact that their product is targeted toward genealogists. Both Ancestry.com and Newspaperarchive.com also offered the ability to limit their keyword search to newspapers in a particular geographic location such as city state or country. The website Paper of Record does not offer this feature and also has a major disadvantage in that you cannot search all of the newspapers at once.

ActivePaper Archive Projects

All of the APA projects support browsing the newspaper by date or doing full text searching. Most of the projects chose to offer both a "Date Search" screen where you choose a newspaper issue through a calendar interface and a "Keyword Search" screen. One useful feature that all of the APA projects share is the ability to sort search results by either "score, title, date, word count, date ascending, date descending, section or publication." This is a feature that is not supported by either CONTENTdm or Greenstone. Another useful common feature of all of the APA projects

was that the user could keyword search in all of the publications or just one newspaper. In addition, all of the projects offer a category search option to search for just articles, pictures or advertisements.

The *Brooklyn Daily Eagle* offers the broadest range of advanced search features including the use of proximity operators and all of the Boolean commands. The Colorado, Missouri and British Library projects did not offer these features. One unique search feature offered by the *Eagle* is the "Selected Subjects" tab. As described on their site, this option "consists of links to news articles, illustrations, and photographs organized by topic and then chronologically within each topic section. The list of topics is linked to each section, and within each section you'll find the headline of the article, illustration or photograph, followed by the date of the issue and the page number. Each headline or title is linked to the actual item." The subjects are broadly based such as Advertisements or Humor. The Colorado Historic Newspaper Project includes a similar but more limited "Featured Topic Section" which currently includes topics and articles linked to "Colorado Statehood." The British Library Newspaper Pilot also has a similar feature called "Collections'. This area includes a number of collections such as "The Great War" and "The Crystal Palace", "which are handpicked groups of articles that have a shared topic." Another excellent feature shared by both the *Brooklyn Daily Eagle* and Colorado websites is a section advising genealogists on how to best search the newspaper, such as where the obituaries are found and some good keywords to use. All three of these projects also contained excellent FAQ's which provided detailed information about how to search.

One area of searchability also offered by all APA projects was limiting one's search by date. The Colorado and Missouri projects, however, have a problem with the "date range" option of the "Keyword Search". When you pick this option, a calendar comes up that allows you to pick dates starting from 2004. The problem is that if you pick a particular paper with limited dates offered such as 1833 to 1834, the calendar option does not automatically limit you to searching within these dates. To figure out which years are available for a newspaper, you have to go back to the "Date Search" screen. The British Library Pilot interface helps solve this problem by having both the date and keyword search features found on one screen. By not having a separate screen for browsing and searching the newspaper, it makes it easier for the user to select a date range when searching a particular newspaper.

Some other common search features offered by APA are the "Last Issue Viewed" and "Last Search Results" tabs. By clicking on the "Last Issue Viewed" tab the user can see the last full page they viewed, while clicking on the "Last Search Results" displays the most recent search results you viewed and is activated only if you have done more than one keyword search. There is no functionality to save multiple lists of search results. The Historic Missouri Newspapers project and the Historic Digital Collegian did not have "Last Issue Viewed" feature while the *Palestine Post* and the British Library Newspaper Pilot did not have either feature enabled.

The article saving feature utilized by APA projects is the MyCollection function. When you are viewing an article you can choose to add it to MyCollection which serves as a personal clippings file. According to the help section from *The Brooklyn Daily Eagle* website, this "virtual clippings file is stored on your computer's hard drive in the form of a cookie in a temporary folder containing links to selected articles. There is no limit to the number of articles that can be saved to MyCollection; however, these cookies are not permanent, and can be erased if your temporary files are deleted." Only the Colorado Historic Newspaper project had a feature where you could get to "My Collection" at any time by having it as a button on the side. The British Library project did not support this feature at all.

ContentDM projects

The ContentDm interface used with the Utah Digital Newspaper Project allows most of the same search features as those supported by the APA. The user can browse newspapers by date, search just one newspaper or the entire newspaper collection. Users are also able to keyword search against four specific categories: births, marriages, deaths and article title. This category searching can only be done within individual papers, however, not the entire collection. Users can also browse over a map of Utah to get a list of papers for that county. To do more complicated keyword searches, there is a separate advanced search screen. In this advanced search mode, you can use various options such as "any of the words" "all of the words" and "none of the words." Although search terms are not highlighted in selected items, secondary search capabilities are possible because the PDF images contain hidden text, which makes secondary searching available from both the Adobe toolbar and a special search box. This type of searching was not possible with any other of the newspaper projects which have chosen to offer newspaper images as PDF.

One special feature of the Utah project is available from the individual newspaper search screen. If you click on the "Go" button next to any of the births, marriages or death keyword boxes while leaving the search box empty you can get a list of all births, marriage or death notices classified for that newspaper. This newspaper project was unusual in that it had staff assign birth, marriage and death notice classifications to articles by reading the headlines of many items. On the other hand, a major weakness of the CONTENTdm search interface is the inability to limit by date.

Although the *Morning Leader* project also uses CONTENTdm, it offers a very different and simpler search interface. The user is presented with four search boxes to search for a variety of items such as keyword, vessels, building names, etc. While phrase searching is available, there are no other advanced searching options. Nonetheless this project did make searching by item category a significant feature, and users can limit their keyword search to over ten different categories including such items as "buildings" and "named individuals."

The CONTENTdm system projects also provide an article saving feature called "MyFavorites." The user can add a particular document or a particular page to this file and create a customized search portfolio. One nice feature is that the user can get to the "MyFavorites" page at any time while searching the digital library collection. The user doesn't have to be viewing an individual article to access this option as with APA. The "My Favorites" page can also be saved as a HTML file which can then be shared with others.

Greenstone Projects

In general, Greenstone projects allow for accessing collections by date, by series (or publication ttle) and by full text searching. While the New Zealand Niupepa Project offers more advanced searching options, the Hawaiian Nupepa project offers only basic keyword searching on either "all" or "some" of a number of terms. With the New Zealand project, you can set "preferences", such as advanced search mode using either a ranked or Boolean query. The advanced search mode also allows you to choose other options such as displaying search history, turning on case sensitivity or forced truncation. The user can also change the language of the search interface. In addition, the user can do a full text search of all content or of just the newspapers, commentaries or abstracts. Searching is done against a page level index, so lists of search results provide the page number on which a search term is found.

Browsing both of these Greenstone collections by date involves picking a year and then a publication. Browsing by series allows you to pick a title from a list and then the date of the issue

you wish to browse. One problem with both of the Greenstone projects reviewed here was that they did not allow their keyword searches to be limited by date. The only way to search by date was to browse. Additionally, while neither of the niupepa projects assigned article types or categories, the software can be customized to allow such a feature. A sample newspaper database set up for the California Newspaper Digitization Projects by ByteManagers and iArchives using Greenstone shows how this might be done.[65] They have allowed searching by article categories such as birth announcements and weddings. One major different wit this software is that it does not have a comparable item saving feature such as "MyCollection" or "MyFavorites."

Other Projects

Just because projects were not created with a major digital library software package does not mean that they do not provide robust searching options. The Georgia Historic Newspapers project offers excellent searching features. The user can do a basic search and limit that search by date, article type, or publication. There are also separate proximity and Boolean search screens with multiple search boxes. Another good example is the *El Clamor Publico* project. They offer searching in both English and Spanish. In addition, they support searching with Boolean operators, proximity operators and the use of both wildcards and truncation. Similarly, the *The Stars and Stripes* provided excellent searching capabilities. This newspaper offers browsing by date or full text searching. A variety of advanced searching features are offered including phrase searching and Boolean searching. The user can also specify a date range that they wish to search or search just a specific day. One interesting feature is that they allow users to limit their search to a particular newspaper page. Futhermore, the *El Clamor Publico* was the only minor project that allowed users to display their search history.

Of all of these projects, only the Georgia Historic Newspapers database allowed searching by different types of article category. The creators of this database actually read through much of the material and assigned categories such as advertisement, fiction, and local news. This level of indexing was unusual in general.

Non U.S. projects also varied as well as to their search and browsing options. In general, none of the major digital newspaper collections allowed any full text searching options, with the exception of the Maori Niupepa project and two of the Tiden projects. The ANNO project, Australian Periodical Publications 1840-45, Alberta Newspapers and Papers Past: New Zealand Papers and Perodicals, all only allow browsing of their collections by date. In general, to access these collections you first choose a title to browse and then a date. While Australian Periodical Publications lists a "Search" feature, it is used to search for the titles of journals or their descriptive entries, such as to find out what digitized journals or newspapers are available that were published in Sydney or are about the temperance movement. The Alberta Newspapers collection allows access to its collection by having the user create a list of newspaper titles by either date or place of publication. All of these sites have sought to improve the user's browsing experience and access opportunities by allowing browsing to be done in a digital environment rather than in an archive.

One minor European project, "Waterford at Wartime" offers very limited search features in addition to browsing options. Users could search the collection by keyword and use commands to include or exclude terms. The site also provides a complete list of articles available in the newspapers in chronological order.

---

[65] See http://64.90.195.24/gsdl_cnp/cgi-bin/cnp?a=q. (Site visited 7.28.2004)

The search features offered by the various Tiden newspapers varied greatly. The Finnish Historical Newspaper Library supports browsing by date and full text searching. From the basic keyword search, fuzzy searching logic is used. The user can search just one newspaper or all of the publications at once. It also offers robust date searching features such as searching a specific day, between specific dates, or before or after a certain date. In the advanced search mode, Boolean searching is available instead of fuzzy keyword searching which allows for the use of proximity operators. The project also includes a browsable subject index in Finnish/Swedish that was compiled in the 1890s. The Norwegian Portal For Digitized Newspapers offers the user the ability to browse by date or do a full text search. There are advanced searching features such as truncation options, searching "all of the words" or "just this word." It also offers limiting searches by date, although this feature was not working at the time of this writing. All of the searches seemed to pull up more modern newspapers.

Since most of the non U.S. projects were browsable rather than searchable, most did not offer any advanced search strategy or article saving options. Only the Swedish Historic Newspaper project allowed users to see a list of their previous searches. None of them offered a "MyCollection" or "MyFavorites" function. In addition, none of the international newspaper projects surveyed here made any article or category level searching available except for those projects that used APA as their software such as The British Library Newspaper Archive and the Palestine Post.

The different newspaper projects showed a great deal of variety in terms of the searching features they supported. Many collections were only able to be browsed, and most of the collections that offered robust full text searching features were either commercial or developed with a specific software package. The next section shall discuss the variety of ways that newspaper project chosen to present their newspaper images, and the viewing methods provided.

Project Evaluation-Viewing Methods and Displaying Options

While almost all projects allow for browsing newspapers as well as some kind of search capabilities, the viewing experience can be quite different. This section will compare the various ways the projects have chosen to display search results and allow users to view and manipulate the actual digital newspapers by considering the following questions:

- Can an article be seen in the full context of the newspaper page?
- How easy is it to navigate an individual page and between different pages of a newspaper issue?
- How common is article or item zoning, or the ability to read and manipulate individual page objects?
- How are search results displayed? How much bibliographic information is provided?
- How common is it to have search terms highlighted (or to be able locate hits in context)?
- What are the various printing and emailing options?

Please see Appendix Two for a table comparing various displaying options.

Commercial Projects

The various commercial vendors tend to offer an excellent variety of viewing options and high quality newspaper images. In the London Digital Times Archive by Gale, browsing by date provides a list of thumbnail images of each page on the left with a hyperlinked list table of

contents on the right. The user can either scroll over the page image and each item title will appear in a popup bubble or scroll over a link at the right and the corresponding article will highlight in yellow on the page image. This enables the item to be viewed in context of the page. Pages can be viewed as either enlarged images or as PDF files. To go the next page in the same issue just scroll further down the screen, or use the arrows at the top. The keyword search results list gives a full bibliographic citation with title and date information as well as article type for each item. When you pull up a list of items from a keyword search, you can choose to look at just the article, the whole page with the article outlined in red, or a PDF file of the page. If you choose the page option, the article will then be highlighted on the page. The search terms are also always highlighted within the article.

With ProQuest Historical Newspapers the list of keyword search results is presented as a list of hyperlinks, each with full bibliographic information such as item title, newspaper date, and page location. By simply clicking directly on the hyperlink, the chosen item will display as a PDF file and give a full bibliographic citation. Search terms are not highlighted anywhere in the item. From the list of results, there is also the option to look at the full page by clicking on "Page Map." This choice provides an entire page image so the user can see their item in the context of the page's layout. The chosen item is not highlighted or outlined on the page as with Gale and APA, which makes it more difficult to find. From the "Page Map" view you can also jump to any other page in that issue. If you scroll over the "Page Map" image a pop up bubble appears to give you the title of any discrete item. Clicking on it than opens it as a PDF file. It is very easy to navigate around the page and between pages

In general, the commercial projects offer the most extensive printing and emailing options. The London Digital Times Archive allows a text version of both individual items and lists of items to be emailed to oneself. Individual items or entire pages can also be printed or saved as PDF files. The same basic options are available with ProQuest Historical Newspapers. The user can email a list of results with the added option of emailing an individual item as a PDF to themselves. All items can also be printed.

With the smaller commercial operations, the amount of information provided in the search results varies greatly. For Paper of Record a keyword search brings up a results list that includes the specific title, date, section, page number and number of hits. The results display for Ancestry.com is more basic and provides only the title of the newspaper and the number of occurrences or hits for the search term. No date information is provided. With Newspaperarchive.com the list of search results are presented as two to three line snippets of information around the search term along with the title and date of the newspaper. Both Paper of Record and Newspaperarchive.com display entire pages as PDF files, articles cannot be viewed individually. Paper of Record presents a nicer display and also provides arrows at the top to navigate quickly through the whole newspaper. To view images at Ancestry.com it is recommended to download their advanced viewer. By using the Advanced Viewer, hits are highlighted on the page image in yellow. The advanced viewer also offers navigation options similar to those of Adobe Acrobat and allows zooming in on the page image and moving around the image.

In terms of printing and emailing options, only Paper of Record provides a direct link at the top to email the item. All of these commercial sites allow users to save or print their items. While Ancestry.com also allows subscribers to save their page images as JPEGs or to print them, the other services allow newspaper pages to be saved or printed as PDF files.

ActivePaper Archive

The viewing options provided by APA are excellent. For every issue a full facsimile of each page is presented or if the user prefers a PDF file of an entire issue can be downloaded and then saved or printed. There are arrows at the top of the page image that allow for easy navigation. Scrolling over the page image causes each individual article to be highlighted. With the *Brooklyn Daily Eagle* and the British Library project there was the additional option of opening individual sections of articles highlighted in blue. The user chooses an article by clicking anywhere on it and it will open and display in a separate window.

When you do a keyword search in APA projects, the list of results display sorted by whatever feature you have chosen. A full bibliographic citation for each item is given including a thumbnail image, word count, byline, page number and section. There are also three different viewing options provided: a preview, opening the article, or viewing the full page. If you choose to open the article it will open in a separate window that floats above the main newspaper page and search terms will be highlighted in purple. Once an individual article window is open the item can be saved to "MyCollection." The item can also be printed or a link to it can be sent to your email. The user cannot download a PDF of just the individual article or email the actual item. Entire newspaper issues can also be printed if they are downloaded as PDF files.

CONTENTdm

With the CONTENTdm system used in the University of Utah project, a keyword search creates a results list with a thumbnail image of the item, bibliographic information and the date. When you choose an item from the results list, a PDF image of the chosen item appears presented in a frameset. A hyperlinked table of contents to the entire newspaper issue displays next to the chosen item, with the current item's title highlighted in the table of contents. This display allows you to see what other articles are on this same page in the issue and on every other page. The user can simply click on another article title to view it or choose to view the entire page itself if you want to view an item in the context of the whole page layout. The entire issue can be browsed in this fashion. One problem with this interface is that search terms are not highlighted within the text of either the page or the chosen item, although there is a search tool provided to find terms.

The CONTENTdm system also offers several different views within the same frameset. If you click on "Document Description" there is a full bibliographic description of the item. There are also options to view a "Page Description" or to view "Page & Text", which pulls up the full page image of the selected article with a transcription of the text next to it. In addition, there are a variety of printing options. Individual items, pages or entire issues can be printed. In order to email an item you need to first add it to "MyFavorites" and then save the whole file as a HTML page. This page can than be sent as an email.

The *Morning Leader* serves as an example of how this software can be customized for a smaller collection. Searching on a term causes results to display as thumbnail images with the date and the first few words of the article listed. Clicking on the image displays the entire page as a GIF image. The user can then scroll down to the desired article, the headline or title of which will be listed at the top of the page. One useful added feature is the "Description" feature. By clicking on this link, you can view full bibliographic information for the article plus all of its keywords.

Greenstone Digital Library

Greenstone digital library projects also offer customization in terms of the viewing options available. For both of the niupepa projects search results display as lists of hyperlinks with

specific title, date and page information for each hit. There are a variety of viewing options. Both projects offer the ability to view the full text of a chosen item with the search term highlighted. The Hawaiian project offers the user the ability to view a miniature preview image of the page, a full screen page image, or to view the entire page as a PDF file. Since you can't search for your term in either the PDF or page image it is not always easy to locate the item of interest on the page. The New Zealand Project offers these same viewing options with the exception that pages cannot be viewed as PDF files. With both of these projects there is no article zoning so viewing can only be done on the page level. The scanning quality for both of these projects is excellent, however, and the text is highly readable. Individual pages can be printed, but the only option to email would be with the Hawaiian project. A user can save a PDF file of a page and then email it as an attachment.

Other Projects

MrSID

Many of the newspaper projects used other systems to support the display and viewing of newspaper images. A customized image viewer called MrSID and created by LizardTech is used with the *El Clamor Publico.* Keyword searching creates a results list of all the issues that contained your term. Clicking on "View" takes the user to a screen with thumbnail images of the newspaper and a description of each page's contents. To find one's search term the user has to read the page descriptions and make note of the page and column number of their item for search terms are not highlighted. The MrSID viewer enables the reader to zoom in on the image, change the size of the viewing window, and navigate around both the page and the entire issue. While this tool is fairly well suited for viewing large images, it is somewhat difficult to use when trying to focus in and find an individual item of interest. The user cannot search within the image for their term nor is the search term highlighted. They must note the location of the article from the search results page and then use the MrSID technology to close in on the selected target. The MrSID technology, while very advanced was not well suited to trying to close in a newspaper item especially a newspaper with no clear breaks between stories and columns. The only printing and emailing options are to email or print a copy of the page description.

Daeja Image Viewer

The Papers Past project requires only allows browsing and requires that users download the viewONE image viewer from Daeja image systems. They chose to use this system so viewers would not have to download any plug-ins, and it would also "allow for delivery of TIFF format scanned from microfilm without any reformatting to an alternative format." This viewer allows the user to zoom in and out and manipulate the image in various ways while also allowing the user to navigate easily between the pages of an issue. A user can print individual pages or the entire issue. There are no emailing options, except to save an individual page as a TIFF file and email it as an attachment. Individual articles cannot be viewed separately.

The Norwegian digital newspaper project also requires you downloading of the viewOne Image Viewer. After you choose a newspaper to browse, you are taken to special search screen with date and title browsing features on the left and an image of the chosen page on the right. The user can then use the viewOne tools to manipulate the image. Printing of the image is available but there is no capacity to email or save the page image. From this same screen, the user can also easily page through an entire issue, go to another issue for this newspaper, or choose another newspaper entirely. Doing a text search pulls up a very basic list of search results including title, date and page information. Clicking on a result takes you to the same image viewing method as with browsing by date, and the page on which your term was found displays in the window on the right. Search terms are not highlighted and there is no way to view an individual article.

PDF Images Only

The Georgia Historic Newspapers database has chosen a very simple presentation style. Keyword searching creates a results list of all the issues that contained the term including article title, article type, and the page and column on which the hit appears. The user must note this location and then choose the appropriate page for that issue. Each page displays as a PDF image and to get to the chosen item the user can use Adobe Acrobat to zoom in on the corresponding section. It is not possible to display individual articles and search terms are also not highlighted. Although newspaper pages are in PDF format the zoom tool can't be used to search for specific words. This can make finding a particular item challenging.

The browsable collection "Australian Periodical Publications, 1840-1845" also only allows page images to be viewed as PDF files. The user browses the periodicals by first choosing a title and then picking a date. Each issue contains a hyperlinked list of all of its contents and clicking on a link takes you to a PDF image of that page or series of pages. The normal Adobe Acrobat tools can then be used to manipulate the image.

Multiple Image Display Options

The *Stars and Stripes* offers more sophisticated options for viewing and displaying its newspaper. A keyword search results list simply displays the issues with the search term. Clicking on any link will take you to a customized viewer with a navigator box on the left and a viewer box on the right. A full page image is presented in the navigator box, and a red box outlines the area where the term was found. In the zoom view on the right, there is a close up of that area of the newspaper highlighted by the red box in the navigator. The search term is highlighted in blue not only in that box but in every other place it occurs on the page. To view another area of the page, simply click on another area of the newspaper in the navigator box. There are a number of zooming and resizing options which can be used to refocus the page. The user can also choose to view the page in other formats such as PDF, TIFF or MrSID. The entire newspaper issue can also be viewed as a PDF. In addition, the user can use arrows to page through the entire issue. Individual articles cannot be viewed in separate windows and there are no specialized printing or emailing options.

The Finnish historical newspaper allows users to both browse and keyword search, and the display options are slightly different for each method. After choosing a title to browse, the user can view the full page as a TIFF or GIF file or download the image as a PDF. Links are provided that allow easy navigation through entire issues. The search results from a keyword search include bibliographic information such as the date, title, and a paragraph that surrounds your search term with the term highlighted. Clicking on one of the results pulls up the full page where your reference is found but the search term is not highlighted. The user then has to navigate around the page to find their term. There is no way to search the individual page image at the display level and also no way to display individual articles.

Other Display Options

The Waterford Public Library project displays search results as a list of hyperlinks with the newspaper date and the first line or two of one's hit. By clicking on this link, the user is taken to a page with a transcription of the story on the left and a thumbnail image of the article on the right. By clicking on the image, the user can pull up an enlarged image but there is no way to see the article in context of its full page.

The Austrian Newspaper Online project only allows individuals to browse newspapers by date, so it does not have a keyword search results display. After choosing an issue, all of the pages for that

newspaper will display as small images at the bottom of the screen. Clicking on any thumbnail pulls up a full image of that page. The page can be moved up and down by using the arrows on the side and bottom of the screen, and the image can be also be resized. There are arrows provided at the top of the screen to go to the next issue of a newspaper. Another viewing option is to download the entire issue as a PDF file, which allows for basic printing options.

A very simple presentation style is offered by the Alberta Newspaper Collection. The creators have simply scanned in entire runs of newspapers from microfilm and the user browses entire issues at a time by using scrolling bars on the side and bottom of the screen. There is no way to focus in on a particular article or to print or email individual items from the paper.

Overview of Major Digital Library Software Providers

As indicated in the previous section, there were three major software products used to create digital newspaper collections found in this survey: Olive Software's ActivePaper Archive, CONTENTdm and Greenstone Digital Library Software. Their major features and differences will be briefly discussed here along with a brief overview of some other major digitization providers. Most of the information provided here is from individual product websites, articles about projects using the specific software and promotional literature from the Online Computer Library Center (OCLC).

Olive Software's ActivePaper Archive

The product ActivePaper Archive has been developed by Olive Software specifically to digitize and archive historic newspapers.[66] This product has been promoted extensively by OCLC and was used by many of the projects surveyed in this paper. In fact, in OCLC's position paper on newspaper digitization, they praise the software for it "offers users a hitherto impossible experience of accessing historic newspapers given that it gives access to different diachronic and synchronic views of content: a single title can be accessed sequentially by date, as with paper and microfilm, a complete issue can be viewed, a complete page, or an individual component."[67] This product has been designed to support a stand alone newspaper collection, rather than support a digital library collection of varying media and materials like Greenstone or CONTENTdm.

This software offers a variety of benefits such as creating highly readable and searchable "full text image maps of text-intensive documents." Viewing newspapers through APA is a very user friendly experience which feels similar to browsing a print issue. In addition, by providing access to the "digital surrogate" long-term preservation of the historic newspapers is supported.
It also requires no downloads of specific software to access the collection, making it particularly desirable to a public library audience. This software also offers copyright management tools; it can block displays of anything an organization does not have permission to display. It also offers multiple language support and can be used with dozens of languages not only in its interface but for indexing and searching. Long term preservation is also enhanced by using open standard TIFF imaging and XML formatting which can make the digital information easily available for later migration.

---

[66] This discussion of Olive Software's ActivePaper Archive draws on the following documents: "Olive: Historical Newspaper Collection Software."http://www.oclc.org/services/brochures/olive.pdf; "Olive at a Glance." http: www.oclc.org/olive/about/default.htm; "OCLC's Solution for Historical Newspaper Access" (April 2004) <http://CDNP.cdm.oclc.org/solution.pdf> and the company website www.olivesoftware.com

[67] "OCLC Position Paper on the Digitization of Historic Newspapers." <http://digitalcooperative.oclc.org/digitize/digitalnewspaper.html>

APA offers its users advanced searching features and "XML Result Ranking" that provides search results based on the importance of XML tags. It has also created patented SmartScroll software for efficient page and article presentation. For the designer, APA provides unlimited metadata indexing that includes an "intuitive workgroup based interface" so staff can add "unlimited metadata for each information entity." Automatic indexing allows for easy search and retrieval of articles, pictures or advertisements with the search terms highlighted. According to product literature, it uses "pixel-based indexing, textual content is indexed by word patterns in digital scans, not individual letters." In addition, it uses Verity, the "world's leading XML based search engine."

This software has a number of patented technologies designed specifically to do OCR on historic newspapers for greater accuracy in keyword searching. The article, "Digitizing Historic Newspapers: Progress and Prospects" that was published in the August 15, 2002 issue of *RLG Diginews* offers an excellent discussion of some of these patented features.[68] According to this article, APA wants to promote both readability, or "the user's capacity to view and comprehend historic text" and searchability, "the user's capacity to reach relevant content through provision of search criteria." They conclude that both of these make up accessibility, which should be key goals of any digitization effort.

This article argues that APA supports readability by allowing the "user to read directly from images instead of the OCR generated text." This is achieved by an image processing technique called "segmentation" "which breaks the page down into smaller information units (articles, pictures, ads etc.), identifies them, and infers the relations between them." APA then uses artificial intelligence and a "patented bitmap indexing and image search technology" to overcome poor image quality and complex page layouts. To provide searchability, APA relies on OCR generated word patterns that are stored in XML format. They also use a patent pending fuzzy logic search technology called "Adaptive Probability Fuzzy Search" or APFS to "compensate for text inaccuracies by applying fuzzy logic according to the probability for error in each word pattern." According to this article, fuzzy logic is applied only when needed in addition to other "special OCR techniques." Their OCR technology also produces "word patterns" instead of straightforward ASCII text conversion.

The next step according to APA is to link searchability and readability through a technique called Bitmap Indexing, which "allows for indexing of each meaningful group of pixels (containing a page element like an article title, a body text word patterns or a picture) on the page image" This process allows for the manipulation of image clips rather than just the whole page image. In addition, this also allows search hits to be highlighted in an article image and search results pages to display images of article titles not just OCR text.

The final step is to convert all digital files into ActivePaperXML, which ties XML to the actual newspaper images themselves. Their process uses three XML layers, "one based on the NewsML/NITF standards, one on the Dublin Core, and a third on PRML, or Preservation Markup Language. PRML maps the newspaper's layout, recording each piece of text and each page object." PRML was developed by Olive Software and they are currently working with OCLC to standardize the first two layers with industry standard tags to make sure that their archives are based on an open platform. Ultimately it is the unique PRML tags that lay the basis for Bitmap indexing and APFS. The newspaper archive created by APA functions as an XML repository. All the results of image processing are organized into "logical file system hierarchy." This allows for

[68] Deegan, Marilyn. et. al. "Digitizing Historic Newspapers: Progress and Prospects." Ibid.

flexibility because the archive can be distributed over multiple hard drives or storage media and avoids the use of a database system.

This software has the advantage of having been created specifically to address the many issues of a major newspaper digitization project. The projects created with this software display newspaper images beautifully and have excellent searching features. At the same time this very fact could make this software unattractive to those who wish to include access to historic newspapers simply as part of a larger digital library collection. Another major problem with this software is that it is most likely too costly for smaller libraries and cultural institutions who wish to digitize newspaper collections and its high number of patented technologies make its interoperability with other systems less likely in the future.

CONTENTdm

CONTENTdm is a digital collection management software that is sold by DiMeMa, Inc, and unlike APA is has not been designed exclusively to manage historic newspapers. This software has nonetheless been promoted extensively by OCLC as a solution for creating digital newspaper collections. Despite this fact the only major newspaper collection to use this software is the Utah Digital Newspapers collection. It has been used, however, by dozens of libraries and cultural institutions to manage digital library collections.[69]

This software package offers a number of benefits: the major one being its relative affordability. Its pricing is based on the number of digital objects stored and as the collection grows the library can pay for additional objects. Various organizations can also share a license by distributing client workstations with up to 50 users. The University of Utah suggests this software is superior because its price is affordable for small institutions who want to launch project of their own and add items incrementally. The smallest license that can be purchased is for $6,000 and allows for a collection of up to 8,000 images. This software also features an image rights tool to display watermarks or trademarks, a query builder tool, and OAI compatibility.

By installing CONTENTdm on a workstation, the collection creator can import all kinds of digitized materials and build collaborative collections. Items can be added singly or in batches. This software has extensive metadata editing features like allowing the user to create metadata templates to "speed and standardize the entry of descriptive and administrative metadata." Metadata entry can also be done at up to 50 remote locations, allowing people at multiple institutions to access and update a collection. CONTENTdm also uses structured metadata fields and offers the "flexibility to add descriptive, structural and administrative metadata to meet local needs." Furthermore, common data standards are used and metadata fields are fully configurable. It offers a "controlled vocabulary for consistent, uniform metadata entry" and is shipped with the LOC Thesaurus for Graphical Materials. The collection creator can also import their own controlled vocabulary or use Dublin Core Qualifiers. Finally, it is fully compliant with OAI-PMH, OAI Protocol for Metadata Harvesting, and also supports export of metadata descriptions using XML.

To use CONTENTdm, the collection creator can either purchase licensed software or the OCLC hosted version which can be accessed from the Internet. The search clients can run on any system.

---

[69] The discussion that follows has been based on the following product literature: "CONTENTdm: Digital Collection Management." http://www.oclc.org/services/brochures/contentdm.pdf/; "CONTENTdm at a glance." http://www.ococ.org/contentdm/about/default.htm/; "OCLC's Solution For Historical Newspapers." Ibid; and "CONTENTdm Features." http://contentdm.com/products/features.html

The main acquisition station must be downloaded to the creator's station and all scanners and input devices must be TWAIN compliant. A PowerPoint plug-in is also required. For the user, CONTENTdm supports a web-based user interface that allows flexible searching with Boolean operators, advanced searching by defined fields, across all fields, across one collection or across many collections. It also supports the use of a My Favorites file to save items for reference and presentations. Several features that CONTENTdm does not support are ranking of search results except by date and it also does not highlight search terms in chosen articles.

A major advantage of this software for building digital libraries is that it supports multiple types of data and media: audio, video, PDF, image files and more. If one wanted to utilize newspapers as part of a larger digital collection, one could easily integrate the digitized newspaper collection with other digital collections. It allows for metadata harvesting and thus easily supports distributed collections across the Internet. According to the creators of the Utah Digital Newspaper Project, two of the major advantages of CONTENTdm are that it can "present a unified interface for all digital collections at a given institution" and that "distributed collections can have their metadata harvested and centrally aggregated to present to the user what appears to be a single collection for searching."[70] If the ultimate goal is to integrate a historic newspaper collection with other digital collections, CONTENTdm can do an excellent job.

Greenstone Digital Library Software

Greenstone is "a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM." It is an open-source multilingual software that is issued under the terms of the GNU public license.[71] This software is produced by the New Zealand Digital Library Project at the University of Waikato. In a 2003 article published in D-Lib Magazine, Ian Witten, one of the designers of Greenstone, states that it "is intended to lower the bar for construction of practical digital libraries, yet at the same time leave a great deal of flexibility in the hands of the user."[72]

The Greenstone system operates under UNIX, Windows, and Mac OS/X and also works with standard web servers. This software has been used for a variety of digital library collections and is not just specific to newspapers. Collections can contain text, pictures, audio, and video as well as other formats. In addition, Unicode is used throughout Greenstone to allow any "language to be processed and displayed in a consistent manner." Greenstone collections can be served locally from a library's own web server or remotely on a shared digital library host. Collections are typically distributed over the Internet but can also be placed on CD-Roms. There are literally hundreds of international digital library collections that have been built using this product.

There are some standard options that come with each Greenstone designed collection. The front page of each collection contains a statement of purpose and an explanation of how the collection is organized. Greenstone comes with extensive documentation and is highly customizable. The main website provides links to an extensive FAQ and a series of user manuals. In general, all of the features that Greenstone supports are customizable to fit the designers' wishes.

The basic structure of a Greenstone collection is determined at the initial set up. Some of the basic steps involved in setting up a collection include specifying the format of source documents, deciding how to display them on screen, determining what the source of metadata will be, choosing what full text or browsing search features will be enabled, and outlining how search and

---

[70] Arlitsch and Hebert, "Digitalnewspapers.org" Ibid

[71] For extensive documentation see, http://www.greenstone.org.

[72] Witten, Ian H. "Examples of Practical Digital Libraries: Collections Built Internationally Using Greenstone." *D-Lib Magazine.* March 2003. <http://www.dlib.org/dlib/march03/witten/03witten.html>

browse results should be displayed. The next major step is specifying the collection, or the source data that comprises the collection. The user is prompted to indicate where the source documents are located. Source documents can come in a variety of formats and are converted into standard XML form for indexing by "plug-ins." Plug-ins exist for a number of different formats such as plain text, HTML, WORD and PDF. New ones can also be written for different document types by following detailed instructions provided by the Greenstone Digital Library Developer's Guide. Plug-ins also perform metadata conversion, while modules called classifiers build browsing data structures from the metadata.[73]

After specifying the location of the source documents, a directory structure is created that includes "subdirectories to receive, index, and present the source documents." The next step is the actual conversion of the documents into a standard XML format. The construction and presentation of all collections is controlled by specifications in a configuration file. All appropriate plug-ins must be specified in the configuration file. After all documents are converted into XML, the full text searching indexes and browsing structures specified in the collection configuration file are automatically created from the source document text and supporting files.

A collection can be designed so that it has an "index of full documents" "an index of paragraphs" or an "index of titles." It can support metadata-driven browsing indexes, or if desired, full text searching indexes can be built from the metadata. Source documents can be hierarchically organized into logical sections and their corresponding metadata can be associated with either whole documents or with individual sections. For example, one could choose to associate metadata with an entire newspaper page rather than to all of its individual articles. Since this metadata is the raw material for the indexes, and it must either be explicitly provided in attached documents or it can be derived from the source documents themselves.[74]

The basic Greenstone interface offers users several different ways to find information. As discussed earlier, the basic interface allows users to do full text keyword searching for particular words, browse documents by title or by subject, while a more advanced searching mode can be enabled from the "Preferences" page. The user selects indexes that have been built from different parts of the full text or from the metadata. In addition, browsing also has users examine "data structures created from metadata" such as lists of authors, titles, dates or hierarchical classifications." The creator of a Greenstone collection specifies what structures they want for searching and browsing by creating specific instructions in the configuration file. The configuration file can be simple or complicated, it lists the indexes such as "document text," all the appropriate plug-ins such as "ZIPP" plug, and also lists the appropriate metadata collection by writing a plug-in such as OAI or Dublin Core. After the collection is built, it can be installed and viewed.

As the creators of Greenstone conclude in one of their papers, this software's greatest strengths are that it is widely accessible, multiplatform, and highly extensible since new plug-ins can be written to accommodate new document types, and new classifiers can be written to create new kinds of browsing indexes based on metadata. It supports multi-language collections on a large scale, and this is why many of the international collections that use it include millions of documents.

The Do it Yourself Option-Other Software and Methods Used

---

[73] Witten, Ian H et. al. "Greenstone: Open-Source Digital Library Software." D-Lib Magazine, Oct 2001. 7 (10) http://www.dlib.org/dlib/October01/witten/10witten.html. Most of the discussion of this software that follows comes from this article and the documentation on the website.

[74] Witten, et. al. ibid.

There are also a variety of U.S. digital newspaper collections that were not created through the use of specific digital library software such as CONTENTdm, APA or Greenstone Digital Library. A variety of projects used other software or created their own "in-house" digitization process. Typically the creators of these newspaper collections would do their own scanning, use an "off the shelf" OCR project, and do all of the tagging and transcribing in-house. The most common display method chosen was to rely on PDF files. Example of such project include the Georgia Historic Newspapers, the Hawaiian Language newspapers and the Alberta Historic Newspaper project.

There were also two different viewing software that were used with a number of projects. The MrSid software from LizardTech was used in the *El Clamor Publico* Project. The other major software used was the Daeja Image Viewer, which was used with the Papers Past project and the Norwegian Digital Newspaper project. In addition, the only OCR software package that was named as the OCR of choice was Abbyy Finereader which was used by the Waterford City Public, the New Zealand Niupepa Project and the Nordic Digital Library.

Other Digital Content Management Software Companies

The companies listed below are just a few of the many companies that provide digital content management services to libraries. Brief information about their services has been gleaned from their websites and is listed below.

Byte Managers is a company that provides "content conversion, cleansing and extraction services."[75] They provide these services to publishers, academic institutions and corporations. This company focuses on coding in a variety of formats such as SGML, HTML, XML, XPAT, EAD and TEI Lite and also has experience creating digital images from a variety of input formats such as paper documents, photos, microfilm and microfiche. They offer a variety of scanning services, OCR conversion and various output formats such as images or PDFs with searchable text. Although their website indicates that they have done projects that include the digitization of newspapers, they do not have links to work done or samples. A sample project of theirs can be viewed, however, through the CDNP portal, for which they used Greenstone to create the interface. For this project, they teamed up with IArchives.[76]

Another company that provides a variety of digital library services is Digital Divide Data, a company based in Cambodia that aims to provide high quality data-entry and digitization services while also providing for the social, human & economic development of their Cambodian staff.[77] They "scan, key, OCR and tag archives from a variety of source formats and languages, including English, French and Latin." This includes double keying texts, performing OCR and OCR cleanup services, and providing tagging for a variety of DTDs in several coding formats such as SGML and XML. Digital Divide Data also provided an evaluation newspaper site for the CDNP.[78]

Yet another company is Luna Imaging, which was formed by J. Paul Getty Trust and the Eastman Kodak Company.[79] This company provides a variety of digital imaging services including scanning services, a variety of image enhancement options, and creating image metadata and text records. Their Insight Software Systems has been used by a great number of libraries and

[75] See www.bytemanagers.com
[76] See http://64.90.195.24/gsdl_cnp/cgi-bin/cnp2.
[77] See http://www.digitaldividedata.com/
[78] See http://california.medarch.digitaldividedata.com/.
[79] See http://www.lunaimaging.com

museums to digitize and manage their visual collections. There were no specific examples, however, of this software being used to support a major digital newspaper collection.

The company Sirsi also has a long history of working with the library community and offers a variety of software products such as the Hyperion Digital Archive and Unicorn Integrated Library Management System. One product that they have recently released is the Sirsi Digital Heritage Room.[80] This product is designed to support "the digitization and presentation of a set number of images or documents from the library." Sirsi provides the scanning services, loading of images, and metadata creation. If a library chooses to continue developing its digital collections they can purchase additional "CollectionPaks." All metadata created is supposedly one hundred percent searchable. They also offer an optional feature of indexing all of the textual material in a collection and enabling full-text search capabilities. Since this software was just released this spring, there have yet to be any major digital library collections designed with it.

The company with the longest record of providing imaging services to academic, commercial and industrial markets is Northern Micrographics, a company which has been in business for over 50 years.[81] They provide a variety of binding, micrographics and digital imaging services. One of the most prominent digital library projects on which they have worked is the Making of America project, where they provided digital preservation services for both Cornell and the University of Michigan. Among the services they offer are the scanning of microfilm in bitone, grayscale or color, a variety output formats such as TIFF, GIF, JPEG, and PDF. They also do image post processing, indexing, and the development of file directory structure development, and metadata development. They have done a number of newspaper digitization projects including digitizing the student newspaper the *Northern Iowan* for the University of Northern Iowa and the *Argus* for Illinois Wesleyan University.[82]

Best Practices and General Conclusions

Building a successful digital newspaper collection can require a great deal of time, money and staff commitment. One theme throughout the literature regarding the building of digital newspaper collections is the importance of planning every step from the technology that will be used to the type of images that will be provided. Despite some general agreement on the digitization process, there are still no formal best practices or standards for digitizing historic newspapers. There have been several major conferences that have discussed the variety of issues involved and offered some general suggestions.

In 2002, OCLC released a position paper on the digitization of historic newspapers. The concepts discussed in this paper grew out of a workshop series called *All The News That's Fit to Scan* held in 2001 and 2002, and out of the experiences of those who helped to develop the British Newspaper Library Pilot and Forced Migration Online. The workshops were held in England and the United States and involved librarians, archivists, publishers and software developers. They offered a number of conclusions regarding content selection, cooperation, copyright issues, costing, revenue generation and technical questions.[83] A major Canadian discussion was also undertaken in October of 2002. The National Library of Canada, in association with the Canadian Initiative on Digital Libraries, the Canadian Newspaper Association and the Association of Canadian Studies hosted a convention entitled "Canadian Newspapers Online: A National

[80] See http://www.sirsi.com/Sirsinews/20040224digitalheritageroom.html.
[81] See http://www.normicro.com
[82] See http://www.iwu.edu/library/services/argus1.htm.
[83] "OCLC Position Paper on the Digitization of Historic Newspapers."
http://digitalcooperative.oclc.org/digitize/digitalnewspaper.html

Consultation."[84] The convention included about 80 participants from both the academic and commercial realms, and this consultation aimed "to explore cooperative strategies to strengthen, on a national basis, online access to contemporary and historical newspaper content for Canadians." On a smaller scale, the Waterford City Public Library of Ireland also released a position paper suggesting a national newspaper digitization scheme for Ireland. This was the only position paper released by a public library and illustrated how public libraries shared similar concerns with those of national organizations.[85]

A number of questions remain regarding the best technical solutions, what kind of content to provide in addition to the newspapers, and what level of searchability to support. This final section provides an overview of the standards and best practices that have been suggested by these and other groups and makes additional recommendations for how to create an "ideal" digital newspaper collection.

Choosing Content

In terms of selecting newspapers to digitize, there was general agreement that newspapers of record should be selected for beginning projects, particular those with national or wide geographic coverage. OCLC suggests that the beginning content should be selected from national and regional newspapers in countries that already have national projects established. They also emphasize the importance of ultimately digitizing specialist publications of minorities within populations. For the selection of regional newspapers, they suggest that decisions could perhaps be made by local librarians in collaboration with local history groups. The Canadian initiative also made some suggestions regarding content, arguing that for retrospective digitization projects complete newspaper issues and comprehensive back-runs should be digitized and that a piecemeal approach should be avoided. They thus suggest that local dailies and newspapers with limited geographic coverage may not be feasible targets for digitization, even though these are the newspapers often most prized by genealogists and local historians.

Cooperation & Coordination of Digitization Efforts

Another important theme seen in the literature is of the growing need for national and international cooperation. OCLC argues that there is a great need for cooperation between all kinds of libraries, educational institutions, technology providers, library organizations, public sector bodies, end users, and standards organizations to name a few. They suggest that one of the most important goals is for "good information networks about what is being selected for digitization" to be established, "with perhaps an online register being kept that can be checked by a library before selecting a title." The Canadian initiative called for an inventory of "all online newspapers and current digitization projects, including those of libraries, historical societies, genealogical and other associations."

Unfortunately there is currently no central online catalog to search for digital newspaper collections or to find digitization projects in process. To find the projects listed in this survey, a variety of tools were used including Internet search engines, a student created bibliography, and several digital library portals. Many of the largest digital newspaper collections are listed in OCLC's Worldcat, but there is no easy way as yet to create a list of such collections in this database. While the NEH has just begun the National Digital Newspaper Project to try and create a national digital resource that will provide free access to historically important newspapers from all over the United States, it is only in the very beginning stages.

---

[84] "Canadian Newspapers Online: A National Consultation." Libraries and Archives of Canada, Ottawa, October 7-8, 2002. http://www.collectionscanada.ca/obj/r3/f2/06-e.pdf

[85] Fitzgerald, ibid.

One good example of successful cooperation would be at the University of Utah. In 2002 the J. Willard Marriott library began supporting the digital needs of smaller cultural heritage institutions by providing fee-based scanning services and space on its CONTENTdm server. They helped create the Mountain West Digital Library, which supports digital aspirations of institutions throughout Nevada and Utah.[86] A good example of international cooperation would be that of the Tiden project, where several major libraries worked together in planning a larger Nordic Digital Newspaper Library.

Copyright & Intellectual Property

The question of copyright and intellectual property rights was also briefly touched upon by OCLC and the Canadian initiative, although the main conclusion among all groups was that the best approach currently is to select older materials for digitization that are not affected by copyright. Both CONTENTdm and APA have added in features to their software to address these issues by blocking access to restricted content.

Funding/Costing/Marketing a Digital Project

The problem of funding projects and generating revenue to sustain them was a major point of discussion by the various organizations involved in creating newspaper projects. The OCLC briefly discussed the issue of creating a costing model with the main conclusion being that more effective models need to be developed since projects are almost never done on a small scale and the total costs can be quite high. On a positive note, the OCLC found that the unit costs per page are now modest and affordable. When planning a budget they also provide a list of things to consider such as the purchase of software, hiring of new staff, time of staff, premises, overhead, costs of long term preservation, on-going support. Similar suggestions were made by the Waterford City Public Library in terms of budgeting.

The Waterford City Public Library also argued that an effective means of cost sharing needs to be developed. They proposed the creation of a national digitization bureau or regional centers and that these centers could digitize jobs from participating libraries. This would provide a much more effective means of cost sharing, rather than each individual library having to buy equipment or obtain more skilled staff. They also suggest that these digitization centers should be responsible for the promotion and marketing of newspaper digitization, since public library staffs in general do not have time for such projects. The Canadians offered a similar proposal to that of the Waterford City Public Library. They propose the establishment of a Canadian Newspaper Centre to prepare a business case for the digitization of Canadian newspapers promote their value and support various projects.

One important point made throughout the literature is the need to effectively promote digital newspaper collections and projects in order to secure funding. The OCLC suggests that early on in any project "delivery should be designed to capture public imagination in order that support and revenue for future phases could be secured." The creators of the Utah Digital Newspaper Project launched a broad publicity campaign to announce their project which has received consistent funding to expand.

Despite the vast literature regarding the need for effective cost models and budgeting advice, relatively little pricing information could be found. The California Newspaper Digitization Project (CDNP) reported that many of the price estimates they received from vendors were

[86] Arlitsch and Hebert, "Digitalnewspapers.org" Ibid.

confidential but that they could provide some guidance on cost estimates. For a one million page project, their "estimates for digitization and image processing (including OCR) ranged from about $400,000 to $2,600,000."[87] They concluded that the more specialized the searching and retrieval features provides the more expensive the cost. Another cost item that had to be factored in was whether or not hosting services would be needed. An additional important point they make is that the costs of "preservation repository services" are still largely unknown. There are few preservation providers and decisions need to be made regarding security, back up, and copying between mediums.

OCLC states that in terms of pricing, the two most critical elements are quality of newspaper or the quality of microfilm. They give detailed pricing information for IArchives/ CONTENTdm and APA in a brochure they prepared for the CDNP in May of 2004.[88] For a breakdown of these costs please see the table in Appendix Three.

Technical Issues & Digitization Processes

The process of going from microfilm or hardcopy newspaper to online digital surrogate can be fairly complex, although the level of cost and technological expertise needed to conduct such a project largely depends on the desired result. The digitization process involves a number of common steps, however, whether the ultimate goal is simply to provide browsing access to one year of a newspaper or a completely searchable digital archive. OCLC has argued that "digital access to microfilm content is usually achieved by scanning to TIFF, then converting to text using Optical Character Recognition Software (OCR). The problem with this process is that low-image quality (often the case in historical material) prevents OCR software from generating readable text."[89] Nonetheless, many libraries often chose this option when digitizing their newspaper collections. Others chose to use a customized software product such as APA, CONTENTdm or open source software such as Greenstone.

Digital newspaper projects chose a variety of ways to go from newspaper to digital image. While some chose to scan from microfilm others chose to scan from paper. Some performed a variety of the tasks in-house, some outsourced the entire process while others chose to outsource some tasks while doing others themselves. In general it was agreed upon that scanning is best done from 35mm microfilm when possible at a resolution of 300 to 400 dpi at the original size of the newspaper. While TIFF images were the most common results output, there was less agreement in this area. A variety of OCR methods were used depending on the desired results. One other major area of agreement was the need for standard naming conventions when it came to saving digital files.

The creators of the digital newspaper collection at the University of Utah wrote an article specifically detailing how the process they developed to digitize their collection could be adapted successfully by other libraries. In the March 2003 issue of *D-Lib* magazine, they suggest a two pronged approach.[90] The first major step is to contract with a local digitization bureau or provider that can scan, process newspapers, produce open source image files and XML tagged metadata. The next step is to choose a database system such as CONTENTdm in order to successfully maintain and access the collection. The authors also argue that it is much cheaper to outsource scanning rather than do it in-house. Finally they propose that future projects need to use newspaper scanning and processing results that are non-proprietary and that any databases used to

[87] See http://cpc.stanford.edu/cndp/recommendations.html.
[88] See "OCLC's Solution for Historical Newspapers Access." April 2004. <http://cndp.cdm.oclc.org/solution.pdf
[89] "Olive At A Glance." OCLC Promotional Literature. <http://www.oclc.org/about/default.htm
[90] Arlitsch, Kenning. Et. al. "The Utah Digital Newspapers Project." March 2003. *D-Lib Magazine.* 9 (3). http://www.dlib.org/dlib/march03/arlitsch/03arlitsch.html. Accessed 7.27.04.

present the newspapers should utilize open source images and metadata fields. This question of open source images and metadata was seen throughout the literature, because without open source standards the future migration of collections could prove difficult. In addition, the use of open standards could ultimately promote sharing between collections and allow for metadata harvesting.

Technical Issues & Metadata Standards

Perhaps the greatest challenge that has yet to be surmounted is the development of uniform technical specifications and metadata standards for digitizing historic newspapers. OCLC discussed this problem at length during their conference. One of the biggest problems they conclude is that there are a still a number of unanswered questions regarding the long term preservation of digital content and what should be considered the preservation object, the microfilm, the digital file, or the XML repository. Another important point they make is that finding a way to make systems interoperable is crucial so that online newspaper archives are cross searchable with each other. To accomplish this, designers must use the same software and the same standards. The OCLC believes that XML should be the text markup language choice, and that a number of DTDS will need to be catered for such as METS, EAD and TEI. The Canadian initiative also concluded that it was important to determine long term preservation solutions and national metadata standards for online newspapers.

Ultimately, to ensure success creators of digital newspaper projects will need automated processes for metadata creation and markup to ensure interoperability. Currently in order to search a digital newspaper collection, the user must go to each individual website and search it directly. There is no way to search across multiple collections, even those all supported by the same software such as APA. Without the ability to ultimately search across multiple collections simultaneously, the full research promise of digitized historic newspapers cannot be realized.

The types of metadata needed to support newspaper digitization projects have also been discussed by a variety of national organizations such as the LOC, NEH, and OCLC. The major software products, APA, Greenstone and CONTENTdm all support various established metadata schemes. One point of contention has been whether to support page level or article level metadata. The NEH in its recent proposal announcement has chosen to support only page level metadata because its initial access interface will be "based on a fully automated approach to text conversion without article-level segmentation or article level metadata."[91] In contrast, most of the freely available U.S. newspaper projects, whether based on CONTENTdm or APA, support article level segmentation. The majority of projects from outside the United States, however, were browsable only and did not support any type of article level metadata.

In an article published in 1999, "Metadata and Data Structures for the Historical Newspaper Digital Library", it was suggested that each part of a newspaper page should be a news object and that standard descriptions were needed of both "the logical structure and the physical layout of newspaper content."[92] In addition, the authors believe that detailed metadata about page images and about all of the basic objects on the page including "text objects" and "graphical objects" are needed. The authors believed that an ideal historical newspaper digital library should allow

[91] "The National Digital Newspaper Program (NDNP): Technical Guidelines For Applicants." <http://www.loc.gov/ndnp/ndnp_techguide.pdf>

[92] Allen, Robert B. and John Schalow. "Metadata and Data Structures for the Historical Newspaper Digital Library." Conference on Information and Knowledge Management **:** Proceedings of the eighth international conference on Information and knowledge management. Kansas City, Missouri. 1999. 147-53. ACM Portal: The Guide to Computing Literature. Accessed 8.13.2004.

searching at the article or "news object" level to truly enhance their research use. It remains to be seen which kind of metadata will most frequently be supported in the future.

Levels of Searching Support

Another important question that was considered was what level of searching should be supported. The OCLC believes that a full range of advanced searching options are important as is the ability to easily read or browse the newspaper, which is why they have chosen to actively promote both CONTENTdm and Olive's ActivePaper Archive as solutions for digitizing historic newspapers. In their survey of public librarians, the Waterford City Pilot project discovered that an overwhelming majority felt that the most important benefit of newspaper digitization would be to enable full text search capabilities.

The CDNP also came up with a number of recommendations. Several of their vendors supplied them with evaluation sites or beta newspapers databases that were then evaluated by a number of users who were asked to rank what features they thought were essential and which were important. The CDNP ultimately concluded that a number of searching and browsing features were essential for a successful project. The found that users wanted articles which were easy to read and could also be seen in the context of the original newspaper page. The most important advanced search commands were being able to do phrase searching, use Boolean commands, and keyword search against both the full text of the newspaper and article titles. Users also wanted to be able to specify a date range when searching. In terms of lists of search results, evaluators wanted publication name, date, and a link to the article. Once results were opened, users wanted to be able to locate their hits in context, such as by having the terms highlighted in the text of the article. It terms of browsing, users wanted to be able to easily navigate within the page, between the pages of an issue, and between issues. Browsing among dates and being able to easily identify online holdings for each title were also ranked as highly essential.[93]

Evaluators were also asked to list those features that were considered important but not essential. Searching by article type was considered highly useful. The Canadian initiative came to similar conclusions. In terms of searching, they argued that all audiences would benefit from sophisticated search capabilities such as full-text searching, some structured search options, and the ability to view the original newspaper page image so that articles could be seen in their original context. As seen by this discussion, although not all digital newspaper collections supported full text searching capabilities, they are generally considered of the greatest importance by users.

Conclusions: An Ideal Digital Historic Newspaper Collection

The great variety of digital historic newspaper collections available online illustrates the fact that there are currently no definitive standards when it comes to their creation. Project websites varied on the additional content they included, on the kinds of browsing and searching options supported and on the display options offered. Despite this variety, in this author's opinion, there are a number of features that should be displayed by an "ideal" newspaper project.

Digital historical newspaper collections should not just provide access to their newspapers, but should also provide information about those newspapers. A full list of the content available and

[93] See http://cpc.stanford.edu/cndp/evaluation.html

date coverage for each newspaper should be provided. Historical and general background information should also be given about each newspaper in the collection whenever possible. Such information might include political orientation, information about the editors, and general areas of subject coverage. Such information is important for a fuller understanding of the newspaper collection and any potential biases in the information they contain. Information should also be provided, such as what types of content could be found on each page or how the structure of the newspaper changed over time. Explanations of nineteenth century newspaper terminology and organization, such as the fact that deaths were not listed in a separate obituary section but often under a section simply termed "deaths" can be of great aid to genealogists.

In addition, some information about the larger historical context should also be provided. If a newspaper collection is limited to a specific geographic region or time period, information should be given about the general history of that area at that time. While a full dissertation need not be provided, links to relevant historical websites, bibliographies of relevant reading, biographies about important local people, lists of place names found in the newspaper are just a few of the additional types of content that could make a newspaper collection a much fuller research experience.

In terms of searchability, full text searching should be supported whenever possible and should support a number of advanced options, including phrase searching and Boolean searching. More advanced options like truncation and wild card searching are nice features but not necessarily essential. Being able to search by article categories such as birth notices, advertisements or articles can also be very important, particularly to genealogists. It makes sifting through huge amounts of newspaper content much more efficient. In addition, when there is more than one newspaper in the collection, the user should be able to search just one newspaper or the entire collection.

Perhaps the most important full text search feature that should be supported is that of being able to search against article level metadata. Although supporting such a feature greatly increases the costs of a project, being able to search on an article level is one of the major advantages of digitization. It enables researchers to quickly compile lists of relevant hits. Another highly important feature to support is the highlighting of search terms. One of the major problems with many of the smaller digitization projects was that they did not enable users to see their terms in context, so even though the user could do a keyword search it was often not easy to figure out why a particular newspaper page was pulled. Finally, a number of date limitation options should be offered to make searching even more effective.

In terms of browsing options, newspaper images should be highly readable and articles should be able to be viewed both individually and in the context of the larger newspaper. For many scholars, being able to see where an article was placed can be considered as important as the content of the article. Browsing by date is highly important for those researchers who wish to peruse a specific newspaper issue, see how a social or political idea changed over time, or to track coverage about a specific event. In addition, easy navigation should be provided between the pages of individual newspaper issues and between different issues of the same newspaper through the user of arrow, or another intuitive system.

When it comes to displaying the newspaper images, the option to view and download images as PDF files should be supported. The PDF file has become so ubiquitous that it is likely that most users will be able to view newspaper images or issues in this format. In addition, these PDF files should have hidden text embedded in them so that they can be searched. Adding this second level of searchability would greatly increase the ease of use of most currently available digital

newspaper collections. In terms of customized viewers, the Daeja Image viewer worked very well for newspapers while the MrSID viewer was difficult to use. Any customized image viewer use should allow for easy image manipulation, the ability to zoom in and out, and the ability to refocus the image. Easy printing and emailing options should also be supported whenever possible.

In conclusion, many of these features are not easy to support or to design. Most of these features are found in either large commercial projects or those projects that used a major software package. Nonetheless, many of them are necessary to make a digital historic newspaper collection truly user friendly and a tool that will aid researchers of all levels.

Appendix One: Table Comparing Major Search Feature

The following table compares the major searching features of the different newspaper projects. The various projects supported by ActivePaper Archive, CONTENTdm and Greenstone have all been subsumed into one column each in order to illustrate general features of the software. In addition, a number of the projects are not listed in this table because they only support browsing by date and not full text searching. These projects include Australian Periodical Publications, 1840-1845, Austrian Newspapers Online, Papers Past, and Alberta Historic Newspapers.

	London Digital Times	ProQuest Historical Newspapers	Paper of Record	Newspaper archive.com	Ancestry. Com	Active Paper Archive	CONTENT dm	Greenstone Digital Library	Georgia Historic Newspapers	Stars and Stripes	Finnish Historical Newspapers	Norwegian Newspaper Project
Boolean operators and or searching	X	X	X	X		X	X	X	X	X	X	X
Phrase Searching	X	X	X	X		X	X	X		X	X	X
Proximity Operators	X	X				X			X		X	
Wildcard Searching	X	X				X						
Truncation	X	X						X	X	X		
Browsing By Date	X	X	X	X	X	X	X	X		X	X	X
Limit By Date	X	X	X	X	X	X			X	X	X	X
Save Search History	X	X				X		X				
Category Searching (e.g. article type advertisement)	X	X				X	X		X			
Index or Field Searching (Author, etc)	X	X										
Change Language Interface		X						X			X	
Save items	X	X				X	X					
Search all newspapers	NA	NA		X	X	X	X	X	X	NA	X	X
Search one newspaper only			X	X	X	X	X		X		X	
Ranking/Sorting Search Results	X	X				X				X		

Appendix Two: Table Comparing Display Options

	London Digital Times	ProQuest Historical Newspapers	Paper of Record	Newspaper archive.com	Ancestry. Com	Active Paper Archive	CONTENT dm	Greenstone Digital Library	Georgia Historic Newspapers	Stars and Stripes	Finnish Historical Newspapers	Norwegian Newspaper Project
Bibliographic Information Provided in Search Results	X	X	X	X		X	X	X	X	X	X	X
View Entire Newspaper Page As Image	X	X	X	X	X	X	X	X	X	X	X	X
Easy Navigation Through Newspaper Issue	X	X	X	X	X	X	X	X		X	X	X
View articles individually or in separate window	X	X				X	X					
Page Display Options												
PDF	X	X	X	X		X	X	X	X	X	X	
TIFF								X		X	X	
GIF											X	
Special Image Viewer Used?					X					X		X
Search Terms Highlighted	X		X		X	X		X		X		
Email items	X	X	X		X	X	X					
Print Individual Item	X	X				X	X					
Print Full Page	X	X	X	X	X	X	X	X	X		X	X
Download Individual Article	X	X										
Download Full Page or Issue	X	X	X	X	X	X	X	X	X	X		

Appendix Three: Table Comparing Software Costs

	APA	CONTENTdm	Total Cost for estimated million page project
Analyzing Collection (for 1,000,000 pages minimum 10 days)	$1200 per day	$1200 per day	$12,000
Travel & Lodging Expenses	Varied	Varied	Varies
Scanning & Cleaning pages			
Bitonal	$0.40 to $0.80 per page	$0.40 to $0.80 per page	$600,000 for a million pages (assumed medium price of .60)
Greyscale	Not provided	$ 0.70 to $1.00 per page	
Profiling publication (noting language, layout for AI)	First profile is included in per page cost $500 each additional profile	Not included service	$5000 (assuming 10 publications with 100,000 pages each)
Identify page components. Page segmentation to client specs Key in various metadata and article classification according to client specs Machine read text, index each article Combine OWR text & XML metadata with image of each page and article	APA breaks these costs out different	$1.27 a page for all these services	$1,270,000
Segmentation OCR Bitmap Indexing Create PDF files XML tagging & mapping Highlighting element coordinate on page Indexing all items & words on each page automatically	$1.50 to $2.00 per page based on film quality and volume	Corresponding service priced above	$1,750,000 (assuming medium price $1.75)
Import image & XML metadata into CONTENT dm server software	N/A	$.15 per page (is this cost only if you go with their hosting service?)	$150,000
Ship tape or DVD to collection owners	Varies by size of deliverable	Vary by size of deliverable	Varies
License Costs Group	$80,250 (includes 1 test server CPU &	$36,000 (for one production server)	$80,250 for APA $36,000 for CONTENTdm

	1 production server.*)		
Maintenance First year Annual Maintenance	 Free $13,642	 Free $6,000	 $13,642 for APA $6,000 for CONTENTdm
Installation & Training	$6,000		$6,000 for APA
Customized Interface	$ 1000 minimum (at $75 an hour)		
Total cost			APA-$2,453,250 (not including travel or shipping) CONTENTdm $2,074,000 (not including travel or shipping)

* Performance requirements determine number of servers. One Olive APA server CPU can serve at least 500,000 pages to five simultaneous users