
Borrowing from archival theory, the paper examines the application of More Product, Less Process (MPLP) on newspaper digitization projects. The paper includes an overview of the utilization of historical newspapers as research material and examines the complexities surrounding their preservation and discoverability. Analysis focuses on the strong and long-standing arguments for increased access to historic newspapers, using an exploratory, multi-method approach to examine the impact – potential and realized – of digitization technologies and initiatives on access to these sources.

Findings reveal the diversity of approaches to description and granularity in the United States, while also praising the standardization efforts of the National Digital Newspaper Program (NDNP) as well as NDNP’s leadership in facilitating access to historic newspapers. Findings demonstrate a need for increased collaboration in order to promote discoverability of digital newspaper collections created without NDNP funding, and suggest a need to de-stigmatize non-compliance with NDNP standards in independently funded projects.

Headings:

Access to information
Cost effectiveness
Digital libraries
Digitization
Library user satisfaction
Newspaper archives
ACCESS TO DIGITIZED AMERICAN NEWSPAPERS IN THE AGE OF MPLP

by
Jacqueline E. Chapman

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science.

Chapel Hill, North Carolina
November 2012

Approved by

_______________________________________
Jane Greenberg
# Table of Contents

- **Introduction: The Value of Yesterday’s News** ........................................ 2
- **Special Issues: Preservation and Bibliographic Control of Newspapers** ........... 5
- **History without Borders: The Promise of Digitization** ................................ 8
- **Are we there yet?: Serving Digital Natives and Digital Humanists** ................. 11
- **More Digitization, Less Segmentation: Wellesley, MA vs Brooklyn, NY** ........... 15
- **Do It Yourself Access: Co-Creating with Users through Crowdsourcing** ............ 18
- **Seek and Ye [Might] Find: Locating Newspaper Digitization Projects** ............... 22
- **Letters to the Editors: Recommendations** ................................................... 26
- **References and Works Consulted** .................................................................. 30
Introduction: The Value of Yesterday’s News

We are taught in secondary school that news is ‘history’s first draft’; this source is leveraged by teachers to help contextualize otherwise temporally distant events. It seems clear to us, then, that newspapers are useful for historical research. However, the literature demonstrates tension on the point of the historic acceptance of the inherent utility of newspaper as a resource for scholarly research.

Such deliberations over the appropriateness of newspapers as scholarly sources included thoughts as to “whether newspapers can be used to determine factual validity, if they are hopelessly biased and tainted, or if they can accurately represent public opinion” (Jones, 2005, p. 2). Yet, in his 1964 article, Wax states that “[c]ontrary to popular and professional belief, historians have collected and have used the American newspaper as prime historical source material since the birth of our nation” (p. 254), supporting the idea that many scholars held concerns on the matter, while maintaining that newspapers have always had historical value.

The evolution of attitudes towards newspapers as scholarly sources worthy of collecting (librarians) or utilizing for research (historians) took some time (Jones, 2005; Mills, 1981; Wax, 1964). Debate also reigned over whether newspapers should be considered primary or secondary resources – this ‘middle-ground’ format may have had implications on the medium’s treatment over time. For further information, Jones suggests The Newspaper and the Historian by Lucy Maynard Salmon, published in 1923, as the “definitive work on newspapers as reference sources” (Jones, 2005, p.3).
Wax describes the 1930s as the threshold of interest in and respect for the newspaper as a scholarly cultural artifact, stating that it was at this point that “the newspaper […] was fully appreciated and respected not only for its aid in other areas of research, but for its own study as well” (1964, p. 259). The growth in the appeal of newspapers as sources for scholarly pursuits is clear in a 2003 study in which Tibbo detected a stated interest in access to historic newspapers by historians. She reports that “many historians view newspapers contemporary to the events they are exploring as essential” and that, for some, period newspapers were the only source available for their area of research (p. 19). Jones also stresses the importance of the medium, stating “[n]o other reference source is as comprehensive in its coverage of daily life or as wide in its scope of possible content” (2005, p. 2). This breadth of scope in both format and content makes newspapers the ubiquitous resource, useable in a myriad of ways. Indeed, the newspaper is not reserved for the scholar or researcher, but rather has appeal to all members of the public through its unique reflection of the society in which it was produced (Jones, 2004, 2005). Historical newspapers represent a uniquely useful resource, appealing to the casual user, genealogical researcher, students, educators, as well as scholars conducing academic research across a wide range of disciplines.

Given the importance of historic newspapers, it is of no surprise that they should often be one of the first materials subjected to new technologies suited towards increasing the broader accessibility of large numbers of documents and records. The adoption of digital technologies in providing increased access to historical materials provides scholars with new areas of research hitherto unseen. In the United States, digital technologies have been used on historic newspapers since the early 2000s - now is an appropriate time to
look over what has been accomplished, what lessons have been learned, and where to
direct research in the future. This paper aims to get a sense of the state of newspaper
digitization in the United States, and to look at how changing user needs and expectations
in this ‘Digital Age’ may influence newspaper digitization projects in the future.
Special Issues: Preservation and Bibliographic Control of Newspapers

While newspapers are uniquely useful, they are also inimitably difficult to work with. In print form, even today’s users can struggle with the over-sized pages of modern issues acquired during the daily commute. These large papers in ever thicker bound volumes are expensive to ship through interlibrary loan programs, difficult to store due to their sheer number, and difficult to preserve – somewhat notoriously so, thanks in a large part to the work of Nicholson Baker¹. Newspapers are ephemeral in nature, and newsprint was not designed to last. The result is thin, fragile, acidic paper that holds up reasonably well in bound volumes with little handling, but will ultimately deteriorate through use and the ‘inherent vice’ of its medium.

Yet, preservation is not the immediate difficulty in working with newspaper; instead, researchers and information professionals alike have been preoccupied with the sheer size of these collections as a hurdle to working with them. As Wax noted in his 1964 exploration of available newspaper resources, “[t]he vastness and scope of files in some states almost defy description” (p. 270). Interestingly, Wax was writing before the vast majority of newspapers had been indexed nation-wide; one can extrapolate from this the extent of newspaper holdings today. Researchers cannot hope to work through the amount of resources available to them, and struggle to find the ‘gems’ among this sea of text. Librarians are also concerned with the size of newspaper collections – the papers

take up a great deal of space in the stacks, they can be difficult to page as well as store, and establishing any kind of intellectual control over their contents is practically impossible. Adding to this, newspapers presents a specific bibliographic challenge in that they were renamed, bought, sold, and dissolved numerous times over the past hundred years. Attempting to gain any kind of intellectual and bibliographic control over newspapers is vital to their becoming useful historical resources. Librarians identified this need early, and dedicated a large amount of time and money to this pursuit. The United States, like many countries, took on this problem as one of national concern, funding this work at the federal level.

In 1972, the Library of Congress along with the National Endowment for the Humanities initiated the United States Newspaper Program\(^2\), “a nation-wide, cooperative endeavor to locate, catalog, and preserve on microfilm newspapers published in the United States from the eighteenth century to the present so that these newspapers will be available to researchers” (Hedin & Leader, 1998, p.308). This project was able to produce the single most comprehensive resource concerning the library holdings of newspaper collections nation-wide, and in the process microfilmed volumes upon volumes of newspaper. These microfilmed newspapers were much easier to share through interlibrary loan, and multiple copies could exist in different libraries simply by ordering a copy of the film. A preservation copy was kept such that the content of the papers might always be accessible.

The comprehensive inventory created from this project is vital to historical research, and forms the backbone of digitization projects today\(^3\). However, this program

---

\(^2\) See: [http://www.neh.gov/us-newspaper-program](http://www.neh.gov/us-newspaper-program)

\(^3\) See: [http://chroniclingamerica.loc.gov/search/titles/](http://chroniclingamerica.loc.gov/search/titles/)
was not without its faults, largely on the fate of the original copies of these papers after they were filmed. Additionally, we know that not every newspaper made it to a library, and not every newspaper was ultimately bound and/or microfilmed. In many cases, we don’t know what we lost. The USNP inventory is the closest the field is likely to get to a complete inventory, and is serves its purpose well – but the inherent problem of newspapers lingers on: how to make this infinitely useful and complex resource discoverable, useable, and accessible?
History without Borders: The Promise of Digitization

That digitizing unique and rare materials provides unprecedented access to these sources, is well known to the field. Digitization means unprecedented levels of access to materials, available anytime from anywhere with an internet connection. Digitization has the power to reunite geographically disparate materials, to allow multiple people to use the same resource at the same time, and to dramatically increase usage without subjecting the original object to an increased level of contact. Digitization also functions as public relations or marketing, in that its outputs lend themselves to use in social media and to sharing. Digitization increases discoverability to the public at large and opens up new avenues of research: “[d]igitization of historic newspapers may lead to new and unexpected user groups and help redefine the ways traditional user groups have utilized them” (Jones, 2005, p. 2). The potential, it would seem, cannot be overstated.

However, what is meant by ‘digitization’? Often, scholars do not intend to focus on the reimagining of an analogue document alone, but refer also to enhancements that allow for discoverability. An image of a newspaper is just that – a photograph. In order to find it, a user must rely on metadata applied to and associated with the digital object.

Digitization is access to materials but also a sort of enhanced access to the materials – a value-add that allows researchers to search for people, locations, and events specifically named in the text, search newspaper items by format (obituaries, comic strips, editorials,
and so forth), or search text by subject (politics, weather, and other subjects). Newspaper digitization projects hold much promise in these areas:

“Growing interest in primary source material, an increasing impatience (or occasionally, disdain) for manual searching, space savings, the need to preserve and archive in multiple formats, and the quickly changing expectations of a new generation of researchers accustomed to immediate access to electronic resources are driving an expanding need for newspaper digital archives” (James-Gilboe, 2005, p.156).

Herein lies the trouble, and the complexities of newspapers reemerge. Though the USNP inventory provides an excellent foundation for digitization projects, it cannot ‘solve’ the inherent qualities of newspaper that cause such difficulties in making the paper and microfilm forms difficult to disseminate, describe, and utilize.

Describing digitized newspapers is not easy. The same challenges that faced researchers and librarians when dealing with a large amount of print resources, and then a large amount of microfilmed resources, still remain: missing and damaged items, the ephemeral nature of the physical materials, a lack of consistency by the original editors, poor printing quality of the original paper, and other factors all stand in the way of providing end users with the most accurate and useful digital collections. Historic papers are frequently mislabeled with inaccurate dates, volumes, issues, and so forth, and are sometimes mislabeled twice when a former librarian attempted to correct for the editor’s mistakes. Newspapers boast irregularities in length, with a weekly paper sometimes running four pages, other times six, eight, or more.

The most obvious need in newspaper digitization is to be able to read the text. Standards in place for accurate and detailed digital imaging ensure that the human eye can detect most text. However, these are still images of texts, or photographs of newspapers – to render the materials truly useful, and to take greatest advantage of the
technologies available, full-text searching of the papers should be enabled. Optical Character Recognition (OCR) helps to provide full-text access to the printed word by automating the transcription process. Various software companies have created programs that can automatically detect letters, numbers, and symbols in a digital image. These programs place a layer over the file, mapping the resulting text to the location on the page in which it was detected, which can then be searched. OCR, metadata provides hitherto unseen access and discovery ability, also improved inventories which lead to an improved ability to identify and fill gaps by targeting known missing items.

Clearly, newspaper digitization (and the processes that render the text searchable and aid in making various materials findable) is time consuming and can be cost prohibitive. Yet, the rewards spur researchers and librarians alike – the sheer amount of information, much of it impossible to access in any other form, make newspaper an invaluable part of the historical discourse. It is clear that “newspaper digitization is exploding” (Herbert & Estlund, 2008), and this interest continues to expand.

Yet, it is clear what the ideal digitized newspaper collection should look like: “[i]n order to realize their full potential, historical digital newspapers should support keyword searching of both the entire newspaper and individual news objects while also supporting the ability to browse the newspaper by date” (Jones, 2005, p. 37). This paper seeks to determine the relative state of newspaper digitization projects in the United States today as compared with this ideal.
Are we there yet?: Serving Digital Natives and Digital Humanists

If by ‘Are we there yet?’ we mean to ask ‘Do we have the ideal digital library of all American newspapers?’ the answer, of course, is a resounding no. But then, digitizing all of the newspapers in the United States has not been established as the goal of any institution – though some statewide projects, such as Utah’s Digital Newspaper Program, have implied that full state coverage is their ultimate goal (Herbert & Estlund, 2008). At the same time, digitizing as many newspapers as possible would seem to be the implied objective of these digitization programs in aggregate.

Determining what progress has been made towards digitizing the entirety of the United States historical newspaper collections requires extensive data gathering, beyond the scope of this paper. However, determining rates and costs of digitization per page alongside the number of pages digitized and an estimate of the number of pages remaining would give a fairly accurate prediction of how much money and how much time it would take to get closer to providing such a resource. However, this data is not readily available. The lack of transparency between organizations is somewhat surprising. NDNP is very open about the amount of the grants given to each organization – grants are awarded for 100,000 pages of newspaper, and the costs are generally in the area of $300,000. However, the cost varies greatly – from the mid-$100,000s to the mid-$300,000s. Additionally, no indication of the amount of in-kind donations accompany
the grants. Through NEH is transparent about its granting process, the numbers are not sufficient for determining a ‘cost-per-page’ digitization rate. It is worth noting that many institutions utilize somewhat different standards in their processes, so comparing between organizations may be disingenuous.

However, the Library of Congress does list the number of titles documented in the USNP inventory: 151,814\(^5\). As of October 19, 2012, NDNP had successfully digitized a select number of (but not all) issues in over 800 titles\(^6\). The difference between these numbers is stark. Given that the first grants were awarded in 2005\(^7\), a very rough estimate might suggest that 6 years might produce 800 titles worth of newspaper (though, not the entire run of each title). At that rate, every known title will not be represented in NDNP for another 189 years – and even then, each issue of each title would not be represented in the NDNP collection\(^8\). These rough numbers indicate that, if the goal is to provide online access to a great deal more of this (still imperfect) resource in the lifetime of anyone graduating with a B.A. in History today, something needs to change.

This access issue is not just about availability, it is about the way in which these resources are used and the way in which historical research is conducted today. Evidence shows that lack of online presence means that a resource is not available to most members of the public, and that the scholarly record may be affected by the loss. If historians favor electronically available texts, those texts will carry more weight than

---

\(^4\) See: List of National Digital Newspaper Program grants, 2010-2012 [https://securegrants.neh.gov/publicquery/main.aspx?q=1&a=0&n=0&o=0&k=0&f=0&s=0&p=1&pv=208 &d=0&y1=1&yf=2010&yt=2012&prd=0&cov=0&prz=0&wp=0&pg=0&ob=year&or=DESC](https://securegrants.neh.gov/publicquery/main.aspx?q=1&a=0&n=0&o=0&k=0&f=0&s=0&p=1&pv=208 &d=0&y1=1&yf=2010&yt=2012&prd=0&cov=0&prz=0&wp=0&pg=0&ob=year&or=DESC)

\(^5\) See: [http://chroniclingamerica.loc.gov/search/titles/](http://chroniclingamerica.loc.gov/search/titles/)

\(^6\) See: [http://www.loc.gov/ndnp/news/index.html#20121024_1](http://www.loc.gov/ndnp/news/index.html#20121024_1)

\(^7\) See: [http://www.loc.gov/ndnp/about.html](http://www.loc.gov/ndnp/about.html)

\(^8\) Of course, this very rough math doesn’t take into account the potential for new technologies that might increase the rate of digitization, nor does it consider fluctuations in funding, nor does it imagine the number of newspapers titles growing.
their off-line counterparts. This is particularly troubling, as newspapers are seen as a potential antidote to ‘archival silences’ where documentation of various aspects of society has not survived or was never created. Newspapers can be helpful in filling some of the gaps through documentation of events that, in aggregate, may aid in providing descriptive information concerning past events, locations, and individuals.

This was true in print, true with microfilm, and it is even truer with digitization. For example, “[t]he New York Times was the first newspaper with an index, and many scholars and librarians believe that this was what helped lead to the initial and continuing over-utilization of this newspaper as a historic source” (Jones, 2005, p. 22). The effects of selection are longstanding. Jones goes on to describe a study where a research project had to be “specifically designed around the limits of microfilm” and another which was limited due to difficulties in “locating newspapers for the given period, especially in Southern states” (Jones, 2005, p. 36). Newspaper research, she summarizes, “privileges proximity” (p.36). Newspapers that are not available online are most certainly used less often, and of those, newspapers with better OCR rates may soon be favored in research across disciplines.

In their 2012 article, Lorang and Zillig found that their first experiment with working with OCRed newspapers “raised fundamental questions about newspaper digitization procedures and issues of scalability, computational infrastructure, and methodology that influence the electronic analysis of historical newspapers” (p. 320). This echoed the findings of Maxwell in 2010, who argued that “[a]n html version of a digital document enables useful text searches, of course, but researchers do not care about searching sources whose accuracy they do not trust” (Maxwell, 2010, p.28). The outputs
of digital newspaper programs are not meeting the needs of increasingly technologically-savvy scholars, and nor are these projects meeting the expectations of the general public (or the newest generation of students) when it comes to the ‘searchability’ of these resources.

In a recent study, Holley looked into these OCR standards to determine if there were an accepted, universal (or, at least, national) best practice. While no official rates have been determined by prominent institution(s), Holley concludes that ‘Good’ OCR accuracy is between 98-99%, while ‘Average’ rests between 90-98%, with ‘Poor’ at 90% or lower (where 10% or more OCR is incorrect). In the same article, Holley finds accuracy rates between 71% to a little over 98% in her sample of digitized newspapers. (Holley, 2009a, p. 5). This closely aligns with other studies in the last decade, which look at the OCR accuracy rates of digitized newspapers (R. Allen, Waldstein, & Zhu, 2008; Arlitsch & Herbert, 2004; T. Blanke, Bryant, & Hedges, 2011; T. Blanke, 2012; Booth & Gleb, 2006; Lorang & Zillig, 2012).

Digital humanists and digital natives alike require full, accurate text transcriptions of a large swath of cultural history. It is clear that “[t]he best online newspaper archives permit fast, precise searching over vast content pools, but users are still able to browse through the paper page by page” (James-Gilboe, 2005, p.156). Archives and libraries have limited funds, and almost despite the large, distributed budgets offered by the Library of Congress and the National Endowment for the Humanities through NDNP, these outputs are created slowly and imprecisely. The question, of course, becomes how to get more newspapers ‘out there’ for users while meeting these demands.
More Digitization, Less Segmentation: Wellesley, MA vs Brooklyn, NY

In the US, the generally accepted standard is for digitized newspaper to be described at the issue level. NDNP prescribes issue-level descriptive and structural metadata for all of its funded projects. As such, NDNP has influenced not only its grantees, but also other organizations seeking to emulate this ‘best practices’ model. Indeed, scholars have noted that “[e]ven with the variety of efforts and the range of invested parties, historical newspaper digitization projects are remarkably similar” (Lorang & Zillig, 2012, p. 306). This similarity, Lorang and Zillig note, extends beyond issue-level digitization, and certainly includes inaccurate OCR.

However, when newspaper digitization was in its infancy, boutique digitization projects reigned. The *Brooklyn Daily Eagle*\(^9\) is an excellent example of this work. The paper was digitized at the page/issue level, and many specific articles (but not all) were cropped, isolated, and given their own subject metadata. Multiple essays and contextual information is provided, and the site has a curatorial feel to it. A user can search by ‘topic’ in order to locate relevant materials. However, the OCR accuracy is low. The *Richmond Daily Dispatch*\(^10\) is another frequently cited digital newspaper project, which also goes ‘above and beyond’ the NDNP standards. It also segmented its articles into separate files, but this project was more focused on OCR accuracy (Crane & Jones, 2006;\(n\)o references provided.)

---
\(^10\) See: [http://dlxs.richmond.edu/d/ddr/index.html](http://dlxs.richmond.edu/d/ddr/index.html)
Lorang & Zillig, 2012). Both of these projects diverge significantly from the NDNP model, and neither were funded through NDNP. However, each provides enhanced access points that users may find useful.

More recently, the Wellesley Free Library decided to digitize the Wellesley Townsend on its own initiative. *The Wellesley Townsend* is now available online\(^{11}\), for free, with what appears to be average OCR. What is most interesting about this is that the newspaper has been placed online with very little metadata. The paper has been divided up by year, such that a user cannot search by date, or even month. This deviates strongly from NDNP, yet it mimics the user experience of a bound volume of newspapers or a microfilm reel. Deviation from the standards set by NDNP may be appropriate in the interest of increasing digital access to a larger number of papers, if it is cost effective to digitize at the year rather than issue level.

In recent years, there have been technological advances that allow for further segmentation of newspaper pages to article level and ad level. This brings added value through increased access points. OCR is becoming more and more accurate, and programs that allow for in-house corrections to OCR mistakes are now available. All of this requires effort on the part of the digitally responsible institution. Findings seem to indicate that users would prefer more and more robust metadata associated with historical newspapers online, and also more digitized newspapers. A compromise has been fashioned using issue level segmentation and once-over OCR: an evaluation of the utility of this in light of changing user expectations and skill is outside of the scope of this paper, but is a potential area for future research.

\(^{11}\) See: http://www.wellesleyfreelibrary.org/digitized-content.html
However, the most interesting finding in looking at these various digitization projects is the discovery that, despite Lorang and Zillig’s contention that newspaper projects are largely the same, there is a great deal of variety in the approaches taken by pre-NDNP projects and, more recently, independent projects. NDNP projects are, by nature, NDNP compliant – those that deviate do so in order to provide increased access points or accuracy (such as in the case of Utah’s program). The Wellesley example represents an interesting turn, and is seemingly in keeping with the goal to make more newspapers available online quickly. It remains to be seen if other organizations will follow suit.

More Product, Less Process (MPLP) has become a staple of archival theory in the less than 8 years since its publication. In that time, the theory has fine-tuned itself, and has become a philosophy of archival life, rather than an approach to processing backlogs. Indeed, Greene’s 2010 article “MPLP: It’s Not Just for Processing Anymore” indicated MPLP’s applicability far beyond the processing area. Is Wellesley’s approach to digitizing historical newspaper in keeping with MPLP-like philosophy? Greene states that when enacting such a philosophy, “[t]he goal is to work smarter, not harder; to do things ‘well enough’ rather than ‘the best way possible’ to accomplish more with less (or the same) resources” (p.199). What does ‘working smarter’ look like for newspaper digitization? Like Wellesley? Brooklyn? Utah?
Do It Yourself Access: Co-Creating with Users through Crowdsourcing

Article segmentation and classification are only one way in which digitized newspaper content can be rendered more useful to the end-user. Optical Character Recognition (OCR) correction is a large part of the processing time for many digital projects, and developing new and innovative ways of increasing OCR accuracy is a highly anticipated and heavily pursued area of research. As previously discussed, simply providing digital images is not enough, and low-accuracy OCR is not sufficient for most users:

“The full text search transforms the way scholars interact with historical sources, and means that keyword searches have limited value in the era of the full-text search. A full text search engine requires error free html text versions of digital sources, and this may be expensive in conjunction with scanned images” (Maxwell, 2010, p. 33).

The reasons for this are simple. User expectations have changed, but no matter the user, the power of full-text is undeniable. Technologically literate and illiterate users alike expect searches to function like Google, and return the most relevant ‘hits’. They want to trust the returns they get in the same way that they have come to trust libraries and rely upon search engines. Search engines on a library site seem worthy of trust.

In addition to students, genealogists, and casual researchers, digital humanists will want to have access to the full text of the original documents, rather than the digital images alone or low-accuracy OCR. For digital humanists to utilize the vast amounts of data available in these newspaper archives, they will need access to the mark-up of the
OCR, and they will need that OCR to be as accurate as possible in order to support their research.

Access to information is one of the aspects of work that information professionals pride themselves on, and here they do not fail. Where Loran and Zillig sought text or xml files, from libraries and for-profit entities alike, they were met with open arms. Yet, it appears that the corner-cutting on OCR correction is a problem that the information profession simply has not addressed. Loran and Zillig state that, for the kind of digital humanities research touted in the literature “access isn’t the problem—the accuracy of the transcriptions is the key issue” (Lorang & Zillig, 2012). Not only is the corpus of digitized newspaper poorly ‘transcribed’ (or, “interpreted” as Loran and Zillig point out the NDNP portal ‘Chronicling America’ so euphemistically puts it), but, according to Loran and Zillig, archivists, librarians, and their compatriots appear to be purposefully misleading their users as to the accuracy of the ‘transcriptions’ of their digital files!

Unfortunately, there is little denying the charge. While the professional literature discusses OCR accuracy ad nauseam, “issues of OCR accuracy and a lack of disclosure about accuracy rates” prevail (307). Loran and Zillig express much concern over the fact that NDNP Guidelines do not provide benchmarks for the accuracy rates of OCR or OWR (Optical Word Recognition). They point to a misleading disclaimer on many newspaper digitization websites (including Chronicling America and the Brooklyn Daily Eagle) which states that OCR is not 100% accurate – an insufficient explanation for the uninformed consumer. Loran and Zillig call for a “frank conversation about the way digital newspaper resources are created, their embedded values, and what they mean for research” (p. 305).
The literature indicates that professionals involved in newspaper digitization projects would do well to explore methods of increasing OCR accuracy. Although NDNP has introduced standardization to much of the newspaper digitization projects in the United States, “[t]he limitations of the NDNP newspaper collections include the inaccuracy inherent in OCR vs. key-stroked text” (Allen & Sieczkiewicz, 2010, p. 2). It appears that many projects, including NDNP, have seen correcting OCR to 100% or near 100% as an ideal, but an unnecessarily costly one in terms of expense and time. It is known that “…only correction and validation by human beings can ensure that the final result will be close to 100 percent accurate as compared to the original content” (Neudecker & Tzadok, 2010). The complexities around character and word recognition have not yet been mastered by the best software. However, given the importance of accurate full-text to scholars and the general public, OCR correction should be more heavily considered by professionals seeking to enhance or begin a newspaper digitization project.

Given the difficulty and expense of OCR correction, it comes as no surprise that most projects are reluctant to engage in the tedious and time-consuming task. However, the potential uses of accurate full-text are persuasive. On this point, Neudecker and Tzadok conclude:

“When one looks at the challenges presented by automated text recognition of historical printed material, it becomes apparent that the huge amounts of digital resources can only be transformed into highly accurate resources by using this potential and involving volunteers in the correction of bad OCR results” (2010, p. 123).

The implementation of robust crowd-sourcing initiatives as a method of addressing massive historical transcription projects is beginning to gain some traction in the archival
world, with successful projects such as “What’s on the Menu?”¹² and “Old Weather”¹³ serving as case studies. The application of these ideas to newspapers seemed daunting due to the volume of papers. However, the National Library of Australia decided to run a pilot test to determine feasibility. Their project has been widely successful, and should serve as a model to other newspaper projects.

Crowd sourcing transcription work exponentially increases the accuracy of available full-text, provides a forum for communication between researchers and with the hosting institution, and can lead to future engagement with users through fundraising opportunities or additional volunteer opportunities. Crowd sourcing does have its own costs – developing or purchasing the necessary software, monitoring and managing the volunteer base, advertising the project, and keeping detailed statistics for the purposes of assessment. The pros vastly outweigh the cons in the case of crowd sourcing. By incorporating users into the solution, these organizations are also encouraging community engagement in their project and increasing the number of stakeholders in their institution.

¹² See: http://menus.nypl.org/
¹³ See: http://www.oldweather.org/why_scientists_need_you
Seek and Ye [Might] Find: Locating Newspaper Digitization Projects

That the United States cannot boast a comprehensive, centralized portal for every newspaper ever made on its shores is not surprising, nor is it a reasonable expectation. Instead, libraries, archives, museums, newspaper and media conglomerates, for-profit corporations and government entities across the country have all participated in adding to the nation’s digital newspaper availability. As previously noted, these efforts have striking similarities. Yet, they are also, by nature, scattered and decentralized – and some of the for-profit examples are hidden behind paywalls.

If an institution were to decide to leverage some budgetary amount towards digitizing newspapers (guided by NDNP principles or not), how would a researcher be able to find this resource? If newspaper digitization rates in the United States were to exponentially increase organically, outside of the auspices of the NDNP or any other centralized organization, how would these resources become discoverable? Right now, the answer is unclear.

Where the United States Newspaper Program was commendable in bringing together the most comprehensive collection of bibliographic information related to American newspapers in history, the National Digital Newspaper Program does not promise similar results. NDNP builds off of its parent program’s bibliographic information to help identify potential candidates for digitization. It does not, however, update those records to link to NDNP digitized holdings. While it does host these files in
a centralized system as part of the Chronicling America site, it does not endeavor to list
all of the potential sources of digitized newspapers on the web. NDNP requires its
grantees to research what work has already been done, and NDNP is willing to ingest
newspapers digitized without NDNP money so long as the materials come from a grantee
and the materials have been digitized and described in compliance with NDNP standards.

To be sure, documenting all of these newspaper digitization projects would be
complex and difficult. Digitization projects such as these are constantly cycling in and
out of existence, dependent on soft money funding, and the whims of donors and for-
profit entities. As a result, it can be very hard to track these projects, especially if they are
one-offs (like in the Wellesley example). Funding is highly volatile. Given the current
economic climate, it is entirely possible that NEH, LoC, or any other agency or
organization funded by the federal government may soon have to face difficult decisions
concerning which functions to cut. Other funding sources may disappear altogether, as if
on a whim: Google got into the newspaper digitization business for a while, but left after
only two years – which, in and of itself may testify to the complexities and high cost
(financial and personal) of newspaper digitization. While keeping track of all of these
disparate projects might be difficult, it would still be immensely useful.

The need for a service of this nature is highlighted by this year’s “Best Free
Reference Websites” list\textsuperscript{14}, reviewed and compiled each year by the Emerging
Technologies in Reference Section (MARS) of the Reference and User Services
Association (RUSA) of the American Library Association (ALA). On the list was not
one, but two free reference services that helped to connect users to digitized newspapers:

\textsuperscript{14} See: http://www.ala.org/rusa/sections/mars/marspubs/marsbestfreewebsites/marsbestref2012
NewspaperCat\textsuperscript{15} and Newspaper Map\textsuperscript{16}. NewspaperCat’s review contained this observation: “[l]arge numbers of historical newspapers are digitized every year by libraries, archives, historical societies and other organizations but they remain underutilized because they are virtually buried in the web.” It is clear that resources like NewspaperCat and Newspaper Map are being created in order to provide a much needed service, and that reference librarians are finding them useful.

However, NewspaperCat was funded by a one-time internal grant (run out of the University of Florida), and does not appear to be updated frequently. Meanwhile, Newspaper Map admits that “the historical layer is poorly updated and not searchable” – Newspaper Map is more suited towards current papers than historical ones, but it is also global, and its focus appears to be European (due, in part, to the fact that the service is based in Sweden). ‘Crowdsourcing’ the problem has also not provided a perfect answer, but appears to be more robust than either NewspaperCat or Newspaper Map – the Wikipedia entry for ‘List of Online Newspaper Archives’\textsuperscript{17} is very active and contains papers that are not listed on either service. Yet, there is no comprehensive and authoritative listing for all of the digitized newspapers in the United States, though there is a clear need for one.

While locating newspaper titles and holdings has vastly improved since 1964, thanks to the USNP, it is clear that locating digitized newspapers has now become problematic. Wax’s thoughts on locating newspaper holdings are, today, better suited to the researcher’s experience locating digitized newspapers and determining the quality of any full-text associated with the digital images: “Despite all the guides, one must

\textsuperscript{15} See: http://ufdc.ufl.edu/hnccoll
\textsuperscript{16} See: http://newspapermap.com/
\textsuperscript{17} See: http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives
continue to use the talents of a Sherlock Holmes, the doggedness of a Dr. Watson, and the wiles of a James Bond to achieve anything remotely resembling completeness in this field” (p. 270).
Letters to the Editors: Recommendations

It seems that the potential for research using newspapers cannot be overstated. Yet, the complexity of the medium and the realities of today’s funding resources will necessarily stunt growth in this area. The sheer amount of newspaper available is stunning, and present rates of digitization give little reason to expect a comprehensive digital historic newspaper collection in the United States anytime in the next few decades, barring some truly dramatic change in technology and/or funding resources.

It cannot be said that cultural heritage institutions have been ignoring the importance of this vast resource, and it is clear that various stakeholder organizations will continue to produce digitized newspapers at a fairly rapid rate. However, despite the well intentioned design and standards behind the work thus far, it is irrefutable that the result has altered the historical discourse and restricted access by not providing digital surrogates of all newspaper collections in our hands and by not prioritizing OCR accuracy.

It is clear that organizations, such as NDNP, have determined that issue-level segmentation of newspaper with low quality OCR is the foundation upon which other organizations may improve as they see fit. This is in many ways cost effective, in that the Library of Congress and the National Endowment for the Humanities do not limit production to a select group of newspaper projects executed to perfection. This and other considerations inherent to the NDNP guidelines have undoubtedly resulted in a wider variety of newspapers freely available online. These choices seem to reflect the archival
processing method MPLP, in that resource allocation is a major factor in determining the level of granularity of description, and that lack of professionally-applied access points do not prevent the object from being available to the public.

Yet, many users – digital humanists and digital natives in particular – find this lack of granularity and accurate, reliable access points frustrating and misleading. This paper has provided a cursory examination of the cost effectiveness of some institutions adopting article-level segmentation and increased description in newspaper digitization projects when the bulk of the United States’ historic newspaper content is not yet digitized. Research found inconsistency in reporting progress and project costs both in the literature and on relevant websites of projects and project funders. Accurate determinations of rates is not possible in a project of this scope. However, findings indicate that such a study is needed.

Issue level segmentation may prove to be less vital than high-accuracy OCR in the utility of digitized newspapers – as such, non-NDNP compliance should not bear a stigma, so long as choices are weighed and then documented. Enhancements such as article level segmentation, name and location authority control, and other improvements continue to be improved upon and may one day become standard in all digitization projects. Yet, there is an access bottle neck around high-quality OCR. While a digital humanist, genealogist, group of students, or any member of the public could, feasibly, manually transcribe or run OCR on available digital files, they are entirely without recourse if the papers are not digitized at all. But the users cannot digitize the newspapers – the papers and the appropriate equipment are in the possession of the professionals.
Archival theory’s MPLP states that no collection is the same, and none should receive the same treatment as a matter of course – the level of processing should be determined based on the perceived need. Digitized newspaper projects were varied to begin with, but have since standardized in the U.S. in order to achieve compliance with NDNP funded projects. The effect of this standardization has been largely positive, however, the digitization is occurring at a slow rate, when the amount of available newspapers are taken into account. Establishing an OCR accuracy level for NDNP is not wise at this time, but NDNP should provide more robust disclaimers on *Chronicling America* and require the same of any site run by an NDNP grantee.

The number of newspapers produced in the United States is staggering – it has also necessarily limited the number of papers digitized. Advisory boards for each digitization project have been careful in their selection processes, weighing all necessary aspects. Yet, scholars have repeatedly pointed to the ‘gaps’ in the record due to the choices that each project has made. The collective appearance of these individual projects is largely unknown to the individuals and institutions working so diligently to make these materials available. Preferential treatment of certain papers chosen by advisory boards cannot and should not mean no treatment at all for others. Professionals decided it was all worth keeping, work indexing, worth microfilming – surely, it is all worth sharing. Professionals should be encouraged to weigh breadth of access as well as ease of use in making decisions concerning newspaper digitization workflows and outputs. This kind of awareness can only be achieved with a collective, contextual view – through an accurate measure of what has already been done.
Additionally, it is vital that an organization come forward to serve as a digital portal to digitized newspapers. The Library of Congress would appear to be a logical home for such a service. Professionals across the country should strive to collaborate in order to promote discoverability of digital newspaper collections created without NDNP funding. Locating such resources is currently difficult, as each project is isolated by its funder, and at times duplicated. The differences in production quality should be noted, such that those in need of near-100% OCR accuracy might know which papers to turn to. Transparency on the part of the information profession is paramount to educating the public as to the need for increased funding for digitization projects, ensuring that researchers are aware of the limitations of digital tools, as well as advocating for and building trust in the information professions as a whole.

In his 2009 book, The Pleasures and Sorrows of Work, Alain de Botton describes a train journey, in which he is surrounded by commuters reading newspapers. His description of this experience of newspaper reading in this context is perhaps more poignant when applied instead to the researcher attempting to make sense of all of the of millions upon millions of digitized newspaper pages: “To look at the paper is to raise a seashell to one's ear and to be overwhelmed by the roar of humanity” (p. 237). It is the duty of information professionals to simultaneously increase the volume of this roar, but also to help users find the individual voice or voices they seek.
References and Works Consulted


