News Media Section

# Who cares about yesterday's news?
# Use cases and requirements for newspaper digitization

**Clemens Neudecker**
Staatsbibliothek zu Berlin, Berlin, Germany.
clemens.neudecker@sbb.spk-berlin.de



## Abstract:

*Europeana Newspapers, Chronicling America, Trove - these are just a few of the major international newspaper digitization programs that have provided access to millions of pages of historic newspapers in recent years. But now that more and more vast newspaper collections are available digitally, what do people do with them? What are some typical examples that drive the usage of these collections? What are the questions that the scholarly community has for yesterday's news? What are the users' experiences and expectations?*

*This paper will explore a number of exemplary use cases for digitized newspapers from the scholarly community and creative industries, in an attempt to broaden and improve the understanding of the diverse use scenarios, their approaches and limitations. The selected initiatives are drawn from the experiences gathered in the collaboration of Europeana Newspapers with renowned researchers and subject experts and discuss areas of great relevance to newspaper digitization, including technical aspects (quality of digitization, formats and standards, tools and applications), curatorial aspects (which content is relevant, and to whom) and legal aspects (copyright, reuse, data harvesting/text mining). Following from this, general requirements are distilled and recommendations derived which can provide some guidance in the digitization of newspapers based on actual use scenarios and aimed at increasing usage of online historical newspaper collections.*

**Keywords:** Historical newspapers, Digitization, Digital Humanities, Use cases, Requirements.

## Introduction

Newspapers, especially historic ones, are a relatively complex type of content to digitize. Not only do they exemplify several characteristics which pose particular challenges in the digitization process, but also the copyright can be difficult to work out. On the other hand, due to their nature being broad and appealing to the general public, digital newspapers can attract a wide variety of audiences and spawn some particularly creative use cases.

**Challenges in newspaper digitization**

Mainly due to their size and fragility, newspapers are typically more challenging and costly to digitize than regular books or magazines. The large, bound newspaper volumes that lie in the stacks of libraries and archives often require particular care or restoration when being prepared for digitization. The binding can be so tight that it is almost impossible to open the volume 180° without severely damaging it.

An alternative can be to cut the binding and use a sheet feed scanner. Although this will require additional effort in the material preparation and a renewal of the binding following the scanning, it can yield significant increases in the digitization throughput. In addition, problems such as paper warping and curved text inside the beginning, which have a negative influence on later steps in the digitization workflow, can be avoided this way.

Fortunately microfilm copies are frequently available, which can be scanned instead of the paper originals. This can reduce costs and increase throughput of the digitization process significantly. The digitization of microfilms can therefore be regarded as one of the main reasons why large quantities of historical newspapers are now available digitally. On the other hand, this cost-efficient method relies on the availability of high quality master films that do not show signs of wear and tear, in order to arrive at a comparable quality of digitization as can be achieved when digitizing from paper originals.

Furthermore, to leverage the full potential of digitization, one must not stop at scanning and simply publishing digital facsimiles – additional, more complex processes must be included in the newspaper digitization process. Most importantly, the processing of the scanned images with Optical Character Recognition (OCR) produces electronic text that is currently used mainly for indexing and keyword search. But is has many more benefits, and can basically be seen as a stepping stone to enabling richer and more advanced functionalities in the online presentation.

However, OCR is a complex process composed of multiple steps. Prior to the recognition of characters and words, segmentation – i.e. the analysis and detection of layout features – is undertaken. Layout analysis serves to identify text regions and to separate them from e.g. illustrations or photos, so that subsequently paragraphs, lines, words and characters can be fed into the text recognition process.

Newspaper pages with their highly complex layout and structure are particularly challenging for layout analysis. The mix of multiple articles and headlines, with photographs and charts in between can confuse the layout analysis algorithm. Different fonts and sizes, advertisements and tables, differently styled paragraphs and multiple languages are only the most prominent barriers for an exact detection and classification of the layout structure. An analysis that was made in the Europeana Newspapers project identified a total of 86 different challenging and limiting factors for layout analysis methods.

A particularly characteristic problem in newspapers pages is the separation of individual articles. If columns are printed next to each other with very narrow margins, the layout analysis can fail to identify the separation of columns and thus recognize text belonging to different articles as belonging to one only. This has the additional negative effect that the correct reading flow is not retained in the digital text.

It may be due to this high amount of challenges in layout analysis that only 36% of the digital newspaper collection holders surveyed in a 2012 study by the Europeana Newspapers project have actually performed any form of layout analysis on their digitized newspapers.[1]

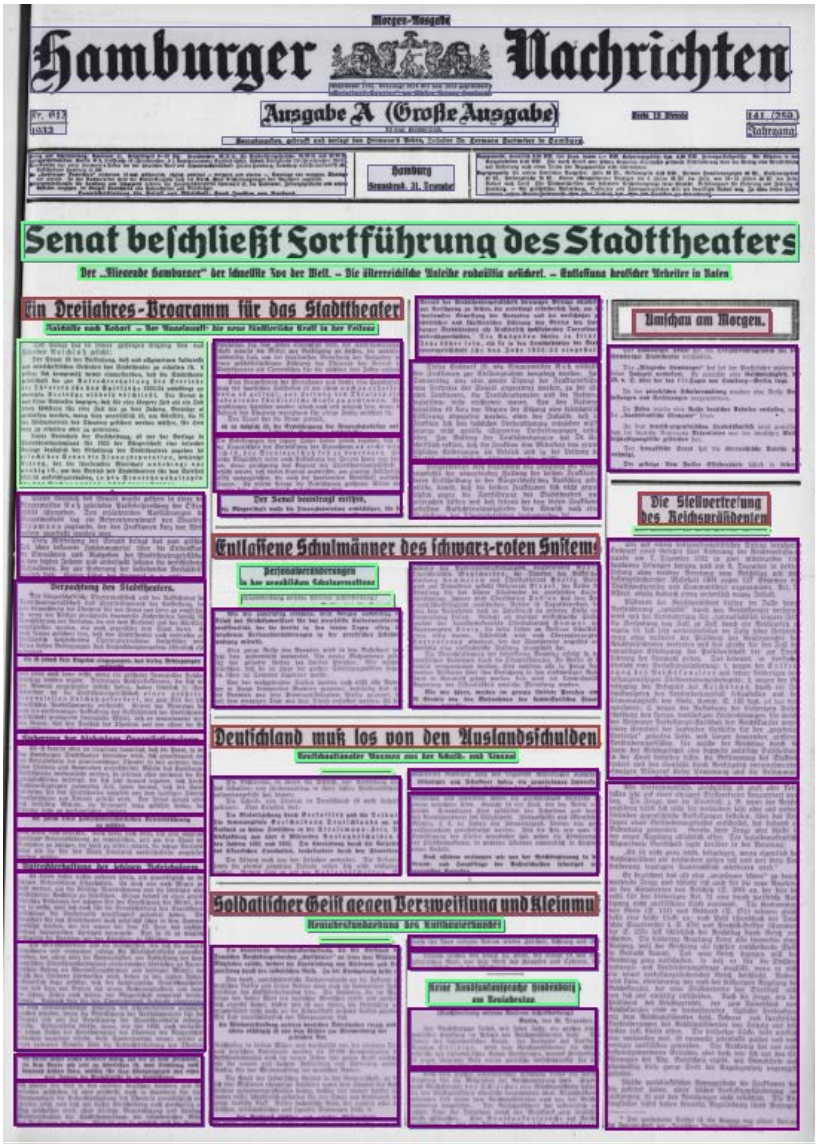


Figure 1: Example of newspaper layout analysis result

Alongside the biennial International Conference on Document Analysis and Recognition (ICDAR), a competition on newspaper layout analysis is organized. Thanks to the availability of an open performance evaluation dataset from the Europeana Newspapers project,[2] the 2013 competition set its focus on the performance of layout analysis for historical newspapers.[3] The results of the competition, in which both academic and commercial technologies were evaluated, provide a good overview of the state-of-the-art in historical newspaper layout analysis.[4]

[1] http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/ENP-Deliverable_4.1_final.pdf
[2] http://primaresearch.org/datasets/ENP
[3] http://www.primaresearch.org/HNLA2013/
[4] http://www.primaresearch.org/publications/ICDAR2013_Antonacopoulos_HNLA2013

**The state of newspaper digitization**

Regardless of these challenges, recent years have seen a renaissance of newspaper digitization, with large-scale newspaper digitization projects or programs being pursued across the globe.

With its close in March 2015, the Europeana Newspapers project had processed close to 10 million pages of historical newspapers with OCR, and another 2 million pages with OLR (Optical Layout Recognition), i.e. enhanced layout analysis including article segmentation.[5] In October 2015, the Chronicling America portal reached 10 million digitized newspaper pages online.[6] Since 2010, the number of newspaper articles available in the Australian National Library's Trove portal increased from 15 million to almost 200 million in March 2016.[7] The survey conducted by the Europeana Newspapers project in 2012 highlights the fact that nearly 130 million pages or 24,000 titles were already digitized in Europe. Moreover, out of the 47 respondents to the survey, 85% (40) offer access to their digitized newspapers free of charge.

A more comprehensive analysis of the global state of newspaper digitization[8] was performed by ICON, the International Coalition on Newspapers program, and presented at the ICON Summit alongside the 2015 IFLA International News Media Conference at the National Library of Sweden.[9] Based on the information gathered, the authors make a conservative estimate of the total number of digitized newspapers held in major repositories in the U.S., the UK and Europe as exceeding 30,000 titles, and possibly well over 45,000 titles when taking all titles digitized worldwide into account.

Moreover, major new digitization programs are in the pipeline in several countries. For example in Denmark, the National Library started to digitize 32 million newspaper pages.[10] In Germany, a national newspaper digitization program is currently under consideration by the German Research Foundation following a pilot project completed in early 2016.[11]

Alas, despite the great achievements already accomplished, it is important to remember the fact that even the impressive amount of pages already digitized nevertheless only constitutes a minor fraction of the (growing) total newspaper holdings available in libraries and archives. The number of unique newspaper pages available in libraries worldwide has been estimated roughly to be over 1.5 billion. This makes the 130 million pages digitized in Europe appear almost negligible – it constitutes less than 0.001% of what has ever been published.[12]

It must be noted though, that due to the absence of exact statistics for newspaper digitization globally, the above figures can only serve as a rough approximation, and must be taken with a grain of salt.

---

[5] http://europeananewspapers.github.io/

[6] https://blogs.loc.gov/digitalpreservation/2015/10/extra-extra-chronicling-america-posts-its-10-millionth-historic-newspaper-page/

[7] http://trove.nla.gov.au/system/counts?env=prod&history=y

[8] http://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf

[9] http://www.crl.edu/events/framing-common-agenda-newspaper-digitization-and-preservation-icon-summit

[10] http://en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/

[11] http://www.slub-dresden.de/ueber-uns/projekte/juengst-abgeschlossene-projekte/zeitungsdigitalisierung/

[12] http://de.slideshare.net/alastairdunning/representation-and-absence-in-digital-resources-the-case-of-europeana-newspapers

**Use cases**

The use cases that are driven by the availability of large quantities of historical newspapers are manifold. Contrary to books or magazines, newspapers were always mainly meant to address the general public. Therefore the variety of content provides numerous perspectives for the newspaper content to be used and analyzed.

As part of the Europeana Newspapers project, 10 interviews were conducted with researchers that use digitized newspapers in their research.[13] In the following, we will try to summarize some of the main findings from the interviews.

Not surprisingly, 7 out of the 10 scholars interviewed chose newspapers as their research topic mainly because of the particular view on the "small" stories of daily life that can be found there - the stories that will make it to a newspaper, but would likely never appear in a history textbook. To quote from the interview with Leon Saltiel: "Newspapers are crucial to trying to figure out the daily life".[14] Out of the remaining three, the research interest of 2 scholars can be roughly characterized as being mainly in the linguistics domain, with 1 scholar predominantly researching political issues/sentiments in the newspapers. One scholar was already using digitized historical newspapers with students in class as part of the regular curriculum.

In the German newspaper digitization pilot researchers were also asked for their opinion on what they deem the most important factors in the digitization of newspapers. Their answers show the broad interests of scholars in this particular media type. The following (unordered) list summarizes the main requirements that were mentioned:

- Typological coverage, i.e. the main newspapers from major cities and hubs
- Newspapers with a very long publication history or wide outreach and reputation
- Newspapers that published texts of special historical importance (e.g. prominent editors, writers)
- Innovators, i.e. newspapers that brought about important changes or innovations to the newspaper as a media type
- Regional newspapers that are/have been of particular relevance in their local communities
- Coverage of the full political spectrum and the main publications of historical political sentiments/movements (e.g. workers press)
- Newspapers directly connected to a particular historical event (e.g. exile press)
- Coverage of language and linguistic features across newspaper publication history

Next to these diverse examples of what researchers are interested in seeing digitized, there are others which are (perhaps) more easily satisfied. When it comes to large-scale text and data mining, volume is often the main key requirement. A good example of this is the Viral Texts project, which is concerned with reprinting networks in 19th century newspapers and magazines.[15] Starting with the content available in Chronicling America, the team has since expanded to include also newspaper collections from Europe (collaborating with several partners in the Europeana Newspapers project) and languages other than English.

[13] http://www.europeana-newspapers.eu/category/interviews-with-researchers/
[14] http://www.europeana-newspapers.eu/qa-with-newspaper-researchers-leon-saltiel/
[15] http://viraltexts.org/

The Viral Texts project aims to identify what "went viral" in the past by evaluating the similarity of newspaper articles and thus reprinting of viral articles using a natural language processing technique called "shingling". Shingling identifies document similarity with the help of n-grams, which makes this approach particularly robust when dealing with smaller variations in the text, as typically occur within collections that have been processed with OCR. It also has the added benefit that the algorithms used can be rather easily transferred to different languages. Accordingly, the quality of the OCR and the language of a text are aspects that are not so relevant for the achievement of the project's goals as is the availability of a vast amount of newspaper pages to be analyzed, so that representative conclusions can be drawn.

Other, particularly creative examples of reuse of digitized newspapers have been presented by Tim Sherratt, Manager of Trove, in his talk "Digitized newspapers and the varieties of value" at the closing event of the Europeana Newspapers project at the British Library in November 2014. [16] The newspaper collection within Trove is famous not only due to its scale, or the success of their crowd-sourced OCR correction activities, but also due to the many and varied examples of reuse that really help to demonstrate the many ways the content made available there is valuable to someone. The website *Trove Traces*[17] collects links that refer to items in Trove, and indeed there are some rather unexpected examples found within.

*Ravelry*, a website for knitters and crocheters, lists 978 knitting patterns that were extracted from newspapers within Trove.[18] The "Elegant Elephant" from the 1959 November 25 issue of "The Australian Women's Weekly" has been produced 53 times based on the pattern found in Trove's newspapers. [19]

*Eyes on the past* is another intriguing example of the creative ways to explore newspaper content in Trove – the site is based on about 1.800 pictures of human faces that were extracted from the newspapers where an eye could be detected using face-recognition technology. Upon clicking an eye in the website's landing page, one is taken to the newspaper article where this image was found. This serves as a nice example to remind us that people have always been at the center of history, and provides a serendipitous entry point to explore the otherwise overwhelmingly vast newspaper collection.

Similarly to the above, simple, game-like exploratory websites such as *Headline roulette*[20], *Query pic*[21] and *The front page*[22], all of which were conceived and built by Tim Sherratt using the Trove API, offer new ways to dive into the newspapers, much different and often more fun than the standard keyword search interface.

Also Chronicling America maintains a list of many interesting examples that show how their content is being reused by others. [23]

[16] http://de.slideshare.net/wragge/digitised-newspapers-and-the-varieties-of-value
[17] http://trovespace.webfactional.com/traces/
[18] http://www.ravelry.com/patterns/sources/trove/patterns
[19] http://trove.nla.gov.au/newspaper/article/43016316
[20] http://wraggelabs.com/shed/headline-roulette/
[21] http://dhistory.org/querypic/
[22] http://dhistory.org/frontpages/
[23] http://www.loc.gov/ndnp/extras/#reuse

Another great example of the innovative use of historical newspaper content by a national library in Europe is provided by the KB National Library of the Netherlands. In cooperation with a small enterprise they developed "Hierwashetnieuws!" ("Here was the news!"), an app for mobile devices that connects the digitized historical newspapers of the KB with GPS coordinates. Due to this, users can easily find articles from the past relating to their current location. The app, made available freely for Android and iOS, was downloaded more than 20,000 times within just one week following its initial release.

For this use case it was particularly beneficial that the KB National Library of the Netherlands used articles segmentation in all of their nearly 10 million pages digitized, so that only the article clippings could be presented. Otherwise it would not have been possible to display the relevant articles on the small screens found in mobile devices without severely limiting its usefulness.

Last but not least, a more familiar but major industry use case for historical newspapers is genealogy and family history: enterprises and initiatives such as FamilySearch[24] or Ancestry.com[25] exploit the digitization of historical newspapers for their service offerings. While FamilySearch is a volunteer-based, free service, run by the Genealogical Society of the Church of Jesus Christ of Latter-day Saints, Ancestry.com is a privately held company that requires subscribers to pay a monthly fee for the use of their genealogical records. As of 2012, Ancestry.com also operates the Newspapers.com website, claiming to provide access to more than 100 million pages from 4.000 historical newspapers via the site.[26]

FamilySearch operates a crowd-sourced indexing service where volunteers transcribe names from scans or microfilm. First, three people check each record, then two people index the same record at separate times before an arbitrator compares the two records and corrects any overlooked mistakes or errors. Thanks to the information about individuals in articles and obituaries that may not be recorded elsewhere, historical newspapers form a particularly valuable source.[27]

If full-text is already available, Named Entity Recognition (NER) can be an exceptionally valuable enhancement for this use case. NER is used to identify the names of persons, locations and organizations in texts. In a subsequent step, the entities can be (semi-automatically) disambiguated and for example mapped to identifiers available in other online information resources, such as authority files or DBpedia/Wikidata, thereby effectively creating linked data. Furthermore, the information revealed by NER can be used to enhance retrieval, e.g. by allowing users to browse names and places found within a particular title, issue or article, or to improve relevance ranking of search results. A case study analyzing the query log files of the digital newspaper portal at the National Library of Wales hints at the fact that, contrary to other digital collections, a vast majority of searches performed were for person or place names rather than keywords.[28] The Europeana Newspapers project made available open source software and training data for NER in historical newspapers in the three languages German, French and Dutch.[29]

[24] https://familysearch.org/
[25] http://www.ancestry.com/
[26] https://www.newspapers.com/about/
[27] https://familysearch.org/wiki/en/Digital_Historical_Newspapers
[28] http://dharchive.org/paper/DH2014/Paper-310.xml
[29] https://github.com/EuropeanaNewspapers

**Requirements and recommendations**

For the majority of the use cases presented here, the availability of full-text is a requirement. Thanks to advances in technology and experience, the challenges entailed in the process must no longer be a barrier stopping newspaper digitization projects from performing OCR and layout analysis. [30] The Europeana Newspapers project has published a number of recommendations and best-practices.[31]

Interestingly though, while the demand for better quality OCR remains prominent amongst users and use cases, the presence of errors and noise in the OCR results is not necessarily a show-stopper for any of these. In fact, in those cases where large volumes of data were used, the importance of the OCR quality decreases. In other cases, OCR errors can be instrumental in attracting users to engage with the material and, if properly managed, help improving its quality.[32]

Accordingly, newspaper digitization projects should always aim to include at least OCR, and preferably also advanced layout analysis, such as article segmentation, in their workflow.

Copyright and licensing can be a real minefield when trying to open up access to historical newspapers. What makes newspapers particularly tricky in this regard are the many different individuals that contribute to a single newspaper issue. Not only is there a multitude of authors who wrote the articles, but also illustrators, photographers or content that has been included from other sources, like press agencies and so forth. To add to this, even if the original rights holders have transferred copyright to the newspaper publisher, and memory organizations enter into according agreements with the publishers that allow them to digitize this content, even the publishers may not always hold the right to grant redistribution as digital resource to others. Dr. Lucie Guibault, an associate professor at the Institute for Information Law of the University of Amsterdam who specialized in international and comparative copyright and intellectual property law, pointed out that typically contracts between publishers and journalists that were established before 1993 do usually not include any digital rights.[33]

To be on the safe side, the online presentation of digitized newspapers by cultural heritage organizations typically ends with what Tim Sherratt coined as the “copyright cliff of death”. This can be seen very clearly in the visualizations of the availability of digitized newspaper in the aforementioned study by ICON, where a steep decline in the availability of newspapers published past 1923 in the US and past 1944 in Europe can be observed.

In order to avoid any legal issues, the Europeana Newspapers project only focused on those newspapers which are in the public domain. Thanks to this, it was possible to release the OCRed full-text produced in the project as a free dataset available for download via Europeana Research. [34] This has already led to a massive response by the research community[35], and a new interactive visualization of Europe’s historical newspapers.[36]

[30] Cf. http://zs.thulb.uni-jena.de/servlets/MCRFileNodeServlet/jportal_derivate_00202173/j11-h1-auf-2.pdf
[31] http://www.europeana-newspapers.eu/public-materials/deliverables/
[32] http://www.dlib.org/dlib/march10/holley/03holley.html
[33] http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D6.2.3_Report_on_Final_Workshop_London.pdf
[34] http://research.europeana.eu/itemtype/newspapers
[35] http://research.europeana.eu/blogpost/found-researchers-wanting-to-use-historic-newspapers

Meanwhile, it is also important not to introduce new access barriers as part of the digitization process. E.g. some public-private-partnerships for digitization projects have led to new embargos preventing unrestricted access to historical material that should nowadays be available free of charge.

While exceptions are possible for some use case such as e.g. reuse in educational resources, and work is currently undertaken to enable similar exceptions for text and data mining for scientific purposes, numerous valuable use cases cannot easily be categorized as falling into either category. Therefore, in order to exploit the full range of reuse possibilities, newspaper digitization should focus on unproblematic material first, and release this under clear open licenses (e.g. Creative Commons).

Having clear and simple, preferably machine-readable, rights statements associated to each digital object is also an essential prerequisite when providing programmatic access to digital newspapers via API. The work between Europeana and the Digital Public Library of America on a common recommendation for standardized international right statements can serve as a valuable stepping stone in this area.[37]

The examples of creative reuse mentioned above make particularly heavy use of APIs in order to provide new and innovative pathways into the collections with more functionality than is typically offered in the historical newspaper portals provided by libraries and archives. Most large-scale data mining project do require APIs for automatic harvesting and analysis of the digital content. Apps and novel applications providing serendipitous access to digitized newspaper collection can only function if the data is made available in a way that programmers can easily understand and use. Accordingly, the importance of having an API for a digital newspaper collection can in many ways be regarded nearly as similarly important as having a classic portal for searching and browsing.

## Conclusion

Newspapers are a unique resource in many ways, and remain so also in the digital sphere. There are a great variety of challenges to encounter when digitizing historical newspapers on a larger scale. Nevertheless, the results that can be achieved when leveraging the full potential of newspaper digitization are equally great. The richness and variety of content in newspapers can attract more diverse users and use cases than other, more specialised resources. Newspapers have always been published mainly for the general public, and therefore nearly everyone can find back something of his own interest in digital newspaper collections.

Mass-digitization and online availability of historical newspapers by cultural heritage organizations should therefore not be seen as a threat to the relevance and existence of the press, but rather as a great opportunity and renaissance of the value and appeal to the general public.

[36] http://research.europeana.eu/blogpost/developing-an-interactive-visualisation-of-europe-s-historic-newspapers
[37] http://rightsstatements.org/

**References**

Apostolos Antonacopoulos, Christian Clausner, Christos Papadopoulos, Stefan Pletschacher: ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR2013), Washington DC, USA, August 2013, pp. 1486-1490, DOI=http://dx.doi.org/10.1109/ICDAR.2013.293.

Günter Mühlberger: Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Strukturerkennung (OCR). ZfBB 58 (2011) 1, pp. 10-18.

Clemens Neudecker, Willem Jan Faber, Lotte Wilms, Theo van Veen: Large scale refinement of digital historical newspapers with named entity recognition, Proceedings of the IFLA 2014 Newspaper Section Satellite Meeting, Geneva, http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf.

Simon Tanner, Trevor Muñoz, Pich Hemy Ros. Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive, D-Lib Magazine 15(7/8), July/August 2009, http://www.dlib.org/dlib/july09/munoz/07munoz.html.