2017 IFLA
International News
Media Conference
LANDSBÓKASAFN ÍSLANDS
HÁSKÓLABÓKASAFN

Submitted on: 01.06.2017

Historical Newspaper Digitization on a Shoestring

Art Rhyno
Systems Librarian, University of Windsor, Windsor, Canada.
E-mail address: artrhyno@uwindsor.ca

Abstract:

Newspaper digitization can be very expensive for an organization. Since 2007, the University of Windsor in Ontario, Canada, and OurDigitalWorld/OurOntario (ODW) have been digitizing historical Ontario newspaper collections, often from microfilm and microfiche sources. With very little funding, the project has managed to assemble nearly 2 million pages of content. It has been necessary to find cost-effective strategies for every part of the digitization process, from scanning to Optical Character Recognition (OCR), right through to delivering newspapers online. This has been made possible by the richness and variety of Open Source solutions (particularly Tesseract for OCR, Olena for page segmentation, Hadoop for volume, and Solr/Lucene for indexing), and the cooperative nature of the project. Recently, ODW has tackled scanning directly, and constructed a prototype microform scanner using macro photography and a makerspace ethic.

Keywords: Newspapers, Digitization, Microform, OCR, Hadoop

Ontario Newspapers in Context

Ontario is a province in Eastern Canada that borders the United States, and has a newspaper history that dates back more than 200 years. The first newspaper in Ontario was called *The Upper Canada Gazette or American Oracle*, and was published 1793 in Niagara-on-the-Lake (where the Niagara River meets Lake Ontario). The premiere issue promised that the newspaper was destined to become "the Vehicle of Intelligence in this growing Province, of whatever may intend to its internal benefit and common advantage" [1].

Since that time, Ontario has become Canada's most populous province and is heir to a rich newspaper history. In addition to major daily newspapers such as *The Globe and Mail*[2], and *The National Post*[3], Ontario is home to many smaller community publications. The Ontario Community Newspaper Association counts some 300 member newspapers across the province, and almost every town and county in Ontario has seen a newspaper in operation at some point in its past[4].

Newspaper digitization in Canada has consisted largely of disparate and localized efforts. As Sean Kheraj has observed, "Canada's online historical newspaper archive is very limited, fragmented, and difficult to access". Kheraj agrees with the findings of a Higher Education Academy study that found that Canada lagged behind the US, UK, Australia, and New Zealand in the digitization of historical newspapers, with projects having no national initiative or consistent funding to draw on[5].

One of the organizations that has tackled newspaper digitization within these limitations is OurDigitalWorld[6] (ODW), a standalone non-profit with roots in long-standing regional digitization collectives in Ontario. ODW is the successor to a cultural heritage project called OurOntario, part of a provincial program coordinated by the Ontario Library Association (OLA). When the government funding that sustained OurOntario was discontinued in 2011, ODW was launched to continue and expand digitization efforts, both in Ontario and elsewhere. ODW inherited close to 1 million pages of historical newspaper content at the end of 2011, with the digitization of some titles dating back more than a decade, and has almost doubled this amount in the half-decade since becoming a standalone initiative.

Digitization Partnerships to Deal with Volume

Historical newspaper projects are often achieved using microform sources, typically microfilm. Scanning options for microfilm can allow for high volume, though the image quality is often compromised by the limitations of the media. Microfilm presents photographs of pages that may have been assembled decades ago, and these pictures can represent poor lighting conditions and highly variable legibility. On the other hand, digitizing even one reel of microfilm can typically result in 800 to 1000 quickly available page images.

ODW has relied extensively on partnerships to digitize newspapers and manage volume. The ODW offices are in the headquarters of the OLA in downtown Toronto, with web facilities close by at the University of Toronto and 70 kilometers away at Hamilton Public Library. Much of the processing associated with the newspapers takes place nearly 400 kilometers outside of Toronto via a Hadoop cluster at the University of Windsor.

The layout of ODW's newspaper service can be seen in Figure 1. Windsor uses mirrored backblaze[7] disk farms for image storage and processing data. The backblaze architecture allows for low-cost storage of newspaper images and derivatives, while master files are backed up to the Ontario Library Research Cloud (OLRC), a service run by Ontario's universities that automatically maintains three copies of every object in the cloud in three separate locations across the province for redundancy and reliable storage[8].

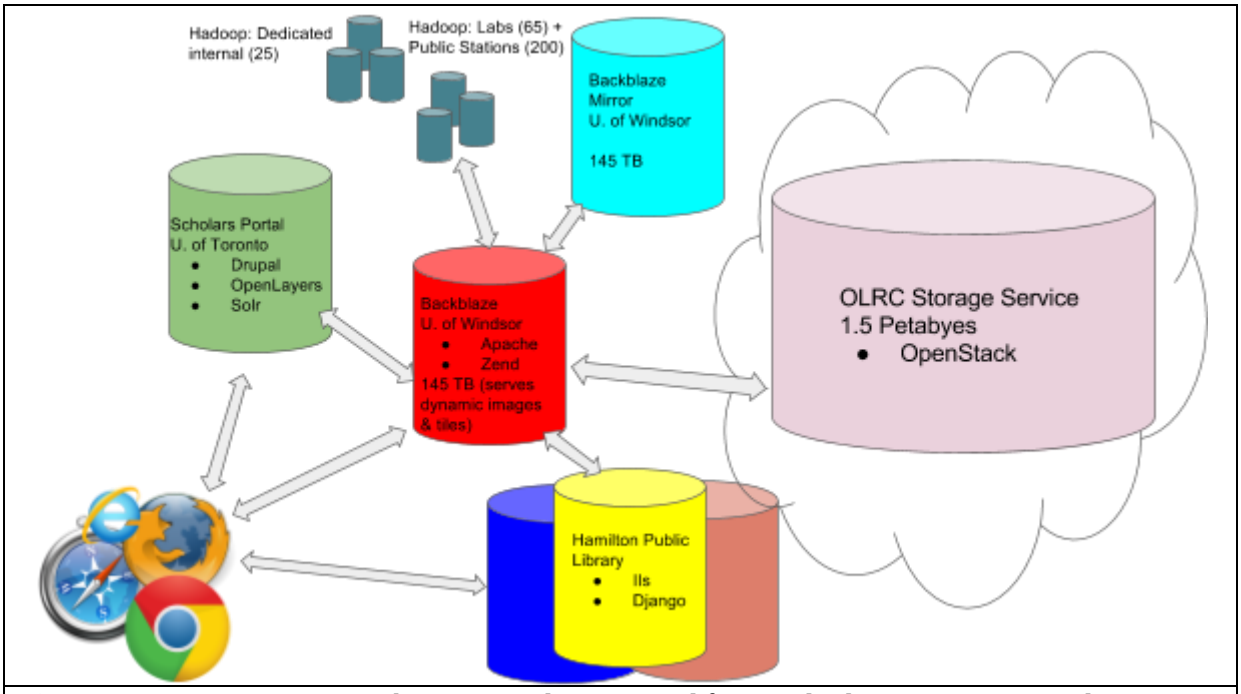


Figure 1 - ODW newspaper architecture. Titles presented from multiple sources, processed using Hadoop and backed up to OLRC.

ODW's partners for newspapers include:

- Scholars Portal[9] (SP), a service of the Ontario Council of University Libraries that provides a shared technology infrastructure and shared collections for all 21 university libraries in the province. Housed at the University of Toronto, SP provides infrastructure for web hosting and OLRC storage services.
- Leddy Library, University of Windsor[10]. The main library at the University of Windsor was a strong supporter of OurOntario and continues to contribute to ODW. Essex County sits alongside the border between the United States and Canada on the Detroit river, and is home to a vibrant newspaper history, including the publication of two abolitionist newspapers in the 1850s (*Voice of the Fugitive* and *The Provincial Freeman*), and several French language titles from the area's long-standing Francophone community.
- Hamilton Public Library[11]. Like the University of Windsor, Hamilton Public Library was deeply involved with OurOntario and continues to support ODW by hosting several servers.

Newspaper OCR - the role of Tesseract and Olena

Optical Character Recognition (OCR) is a common technology used to extract text from images. For historical newspapers, OCR is one of the key enablers for identifying and retrieving content. Yet newspapers can pose enormous challenges for effective OCR, particularly if scanned from compromised sources like microfilm and microfiche (see Figure 2).

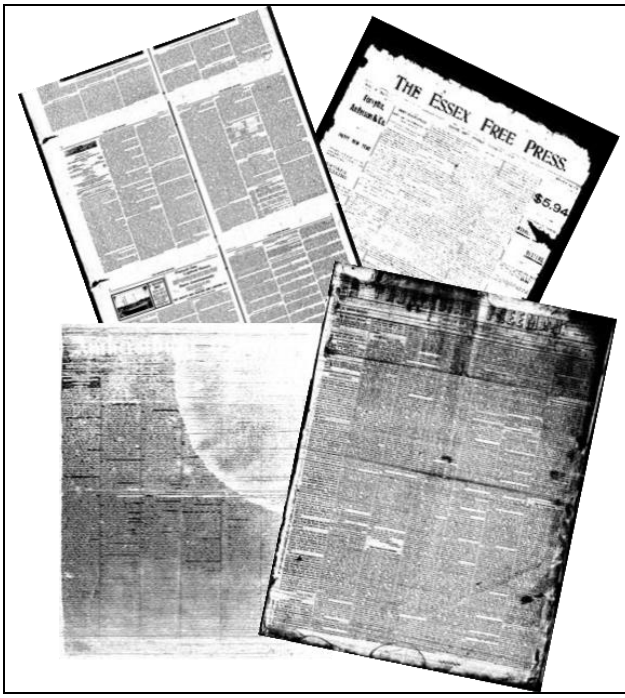
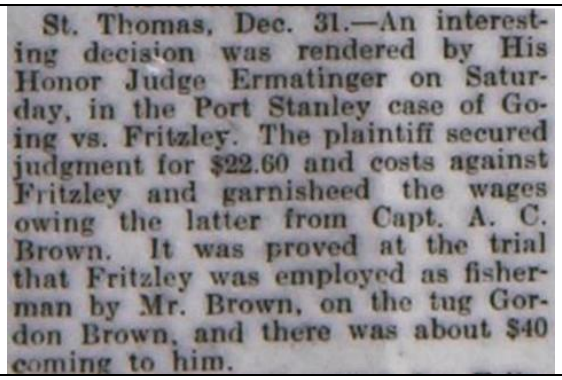


Figure 2 - A sample of just some of the hazards waiting in microfilm sources. ODW has dealt with microfilm with very poor lighting conditions (*The Amherstburg Courier* and *The Provincial Freeman*), titles with as much as 8 to 16 pages on a single camera shot (*The Marine Record*), and many of the titles have sections missing from pages (*The Essex Free Press*).

One of the most utilized OCR applications for historical newspaper collections is Abbyy[12], a well-respected and highly capable OCR solution. In addition to impressive accuracy rates on low quality images, Abbyy has a cluster processing option that provides additional capacity for high volumes. ODW has transitioned most of its OCR processing from Abbyy to Tesseract[13], a long-standing OCR application that was released as open source in 2005, and has had development sponsored by Google since 2006.

At this point, over half of the pages in the newspaper collection have been processed with Tesseract. Although ODW made use of Abbyy's clustering capabilities, the option was limited to Windows at the time, and it was felt it would be easier to monitor processing and manage capacity in a Linux environment. And, of course, there was interest in trimming any budget line where costs were incurred.

In 2011, a major effort was undertaken to define a path that would allow Tesseract to be a comparable replacement for Abbyy in newspaper processing. In testing different scenarios for OCR deployment, it was observed that Tesseract's accuracy seemed to benefit greatly from preprocessing newspaper-based images in order to increase the contrast of the text (see Figure 3). Tesseract does have some internal routines for cleaning up images, but this seemed to be an area where Abbyy had a strong advantage.


St. Thomas, Dec. 31.—An interesting decision was rendered by His Honor Judge Ermatinger on Saturday, in the Port Stanley case of Going vs. Fritzley. The plaintiff secured judgment for $22.60 and costs against Fritzley and garnisheed the wages owing the latter from Capt. A. C. Brown. It was proved at the trial that Fritzley was employed as fisherman by Mr. Brown, on the tug Gordon Brown, and there was about $40 coming to him.

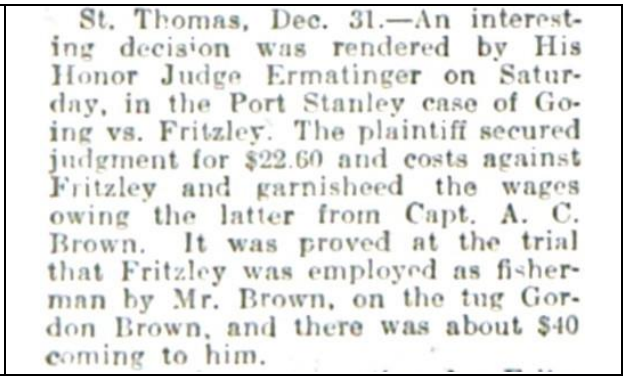

St. Thomas, Dec. 31.—An interesting decision was rendered by His Honor Judge Ermatinger on Saturday, in the Port Stanley case of Going vs. Fritzley. The plaintiff secured judgment for $22.60 and costs against Fritzley and garnisheed the wages owing the latter from Capt. A. C. Brown. It was proved at the trial that Fritzley was employed as fisherman by Mr. Brown, on the tug Gordon Brown, and there was about $40 coming to him.

Figure 3 - An example of image preprocessing, in this case, applying a *Difference of Gaussians* filter to the image on the left in order to produce the image on the right. This helps isolate the text and generally provides more accurate OCR. Sample from *The Windsor Evening News*. January 3, 1908.

At least one technique had accidental origins. In testing different image techniques for OCR processing, it was common to extract a region of the newspaper image for manipulation, such as an individual paragraph or perhaps one column. This was done for convenience, it is simply faster to apply something like an image filter to a smaller image than to a larger one, and the OCR response time is similarly shorter.

In comparing accuracy rates for Tesseract, it was observed that a newspaper page would often score better if it was parceled up and the OCR was applied to smaller regions than if the OCR was done on the entire page at once. Abbyy did not seem to be nearly as vulnerable to multicolumn and complex text layouts as Tesseract.

This led to a search for solutions to automatically segment newspaper pages into paragraphs. The Olena image platform[14] was identified as a possible candidate technology to augment Tesseract's workflow. Olena's tools had fared well in historical page analysis competitions[15], and Olena supports an XML-based layout format (PAGE[16]) to identify regions, which meant the results were easy to fold into ODW workflows. Olena also had tools to visualize its layout analysis (see Figure 4).

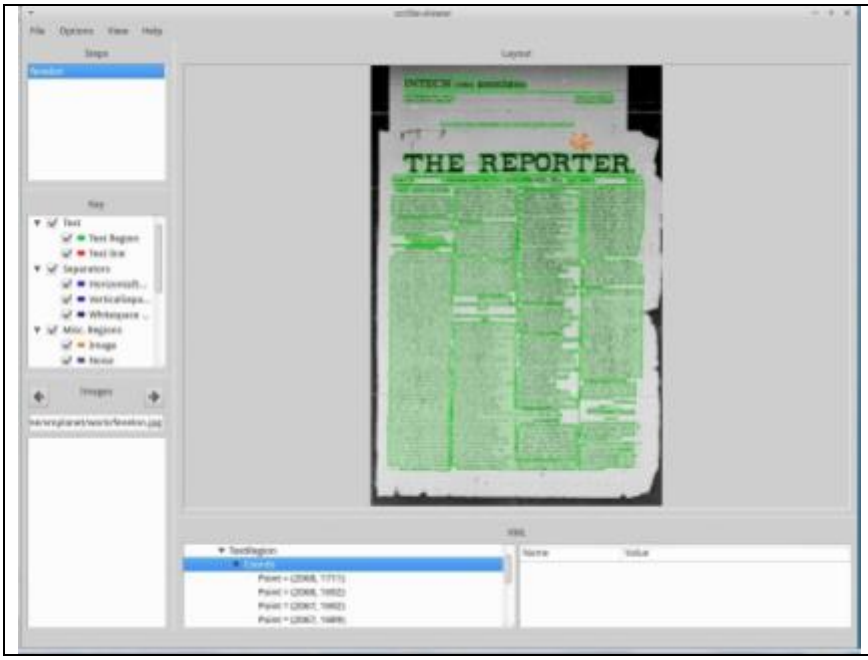


Figure 4 - Olena's *scribo-viewer* utility. The green sections designate text on the page.

ODW now relies exclusively on a 100% open source stack for OCR processing. OSS image modification tools like ImageMagick[17] are used for preprocessing images with significant blur or imperfections. Olena divides complex page layouts into smaller component images for Tesseract. In turn, Tesseract supplies OCR text and positional information (word coordinates).

It should be noted that not all newspapers require OCR. ODW has a small but significant collection of newspapers that have been "born digital" and the PDFs that have been used for publishing have been supplied to the project. The text and positional information in these

cases are usually extracted using the Apache PDFBox library[18], though the PDF documents themselves can contain raster images and sometimes there is no text layer to work with.

Hadoop for Batch Processing

ODW often receives large sets of newspapers reels that need quick turnaround for digitization. The microfilm reels may be borrowed from the collection of a public library or an organization may be digitizing newspapers as part of a grant that has a requirement for immediate online availability. Hadoop[19], the open source cluster framework for running applications on commodity hardware, helps improve the throughput of large sets, particularly for OCR. The use of Hadoop allows existing library hardware to be used for processing. Libraries often have extensive investments in public computing infrastructure that often sits idle during closed hours. In an academic library, computer lab space may also be readily available between semesters and during holidays.

The University of Windsor maintains a Hadoop cluster that is a mixture of legacy equipment that has been rotated out of public service, and virtual machines on public-facing workstations. Hadoop can be run on many operating systems, but ODW uses Ubuntu on a closed network, in part because Hadoop uses many different non-standard ports and it is easier to handle the network requirements in this setup. The older equipment can be accessed at any time while the virtual machines are utilized when the library is closed.

Hadoop streaming, a utility in Hadoop that allows scripts to be used for processing, provides a method of coordinating processing tasks in popular scripting languages. ODW uses Python scripts to apply Olena segmentation and Tesseract OCR to page images, this allows processing to be configured with familiar technologies, and to be submitted and tracked with standard Hadoop tools (see Figure 5).

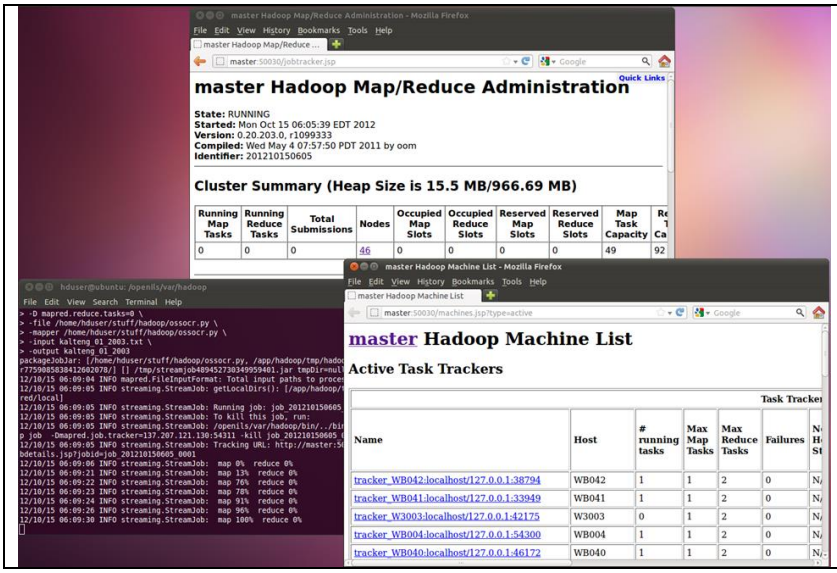


Figure 5 - Hadoop's console and web interface allow a processing job to be tracked and managed.

Hadoop streaming is also used for more general purpose tasks in newspaper digitization. For example, creating derivatives of newspaper pages. Virtually any command that can be executed at the command line can be folded into a Hadoop script and the results can be transferred to a common storage area.

Lucene/Solr as Indexing Backbone

It is hard to imagine a more successful search technology than Lucene[20], and a more effective platform for utilizing Lucene than Solr[21]. The scalability of Lucene/Solr is well recognized among digital library projects. Indexing millions of pages of newspaper text requires such robust solutions but a key capability for ODW is Lucene's support for merging, the ability to build and maintain separate indexes for specific newspaper titles. This means that an index can be built for a title as soon as the OCR is available and then merged into a master index for web searching (see Figure 6). Every newspaper title processed by ODW has its own index so that the sets can be redone and merged as desired. This allows a newspaper title to be processed more than once if desired, for example, if higher quality images of the newspaper pages become available.

```
$ java -cp ./core/lucene-core-6.5.0.jar:./misc/lucene-misc-6.5.0.jar
org.apache.lucene.misc.IndexMergeTool new_index old_index new_title
```

Figure 6 - Lucene's index merging tool as used at the command line. In this case, the directory "old_index" contains the existing index, and the directory "new_title" contains the index of the newspaper to be added. The merged index is then placed into the "new_index" directory.

ODW has also found that the open format of Lucene indexes is an advantage for modifying indexes with java programs. This is sometimes done when a field needs to be remapped, or the content for indexing benefits from a java interaction, such as adding an additional field (see Figure 7).

```
int numDocs = reader.numDocs();
System.out.println(numDocs + " to process...");

//step through the index
for ( int i = 0; i < numDocs; i++) {
    Document doc = reader.document( i);
    doc.add(new LongField("ds_mod",pubDate.getTime(),
        Field.Store.YES));
    //do more field work
```

Figure 7 - Java snippet showing the process of working through the documents in an index.

Solr is the delivery point for the indexes assembled for the newspapers. Both front-end newspaper services offered by ODW (http://news.ourontario.ca and http://ink.ourdigitalworld.org) utilize Solr for discovery, and it provides a consistent and multiplatform solution for large collections.

Microform Scanning and the Quest for the $1K Scanner

Inspired by recent DIY and makerspace memes popular in the library community, ODW has developed a first generation macro photography prototype for microformat digitization. Macro photography is the process of taking photos of small objects, for example, insects or flower petals. Typically, a special lens is used to zoom in on the object at a level that allows intricate details of the object to be captured.

Macro photography is also sometimes used to retrieve images from archives of 35mm film. The method usually involves placing the film on a light table and using a macro lens to zoom in on the frame containing the photo of interest[22]. There is even a commercial unit for retrieving photos via this approach which uses cellphones instead of traditional cameras[23].

In testing various lens on newspaper microfilm, there does not seem to be a discernable difference between a high-end macro lens, and a lower cost unit, such as a Raynox lens[24]. In fact, with enough patience, a standard magnifying glass can achieve almost the same results for 35mm film. Successful macrophotography of film seems to be dependent almost as much on the underlying light as the macro lens.

ODW's prototype (see figure 8) employs a LED work light (about 800 lumens) with a sheet of white plexiglass. The original camera was an entry level Canon DSLR (Rebel T6[25]) but the number of photos required for a typical reel of newspaper microfilm has caused concern about the internal mirror used on DSLR cameras. The mirror is typically attached with a thin piece of plastic and is the most likely part to wear out, though this has not been the experience with digitizing via this method so far[26].

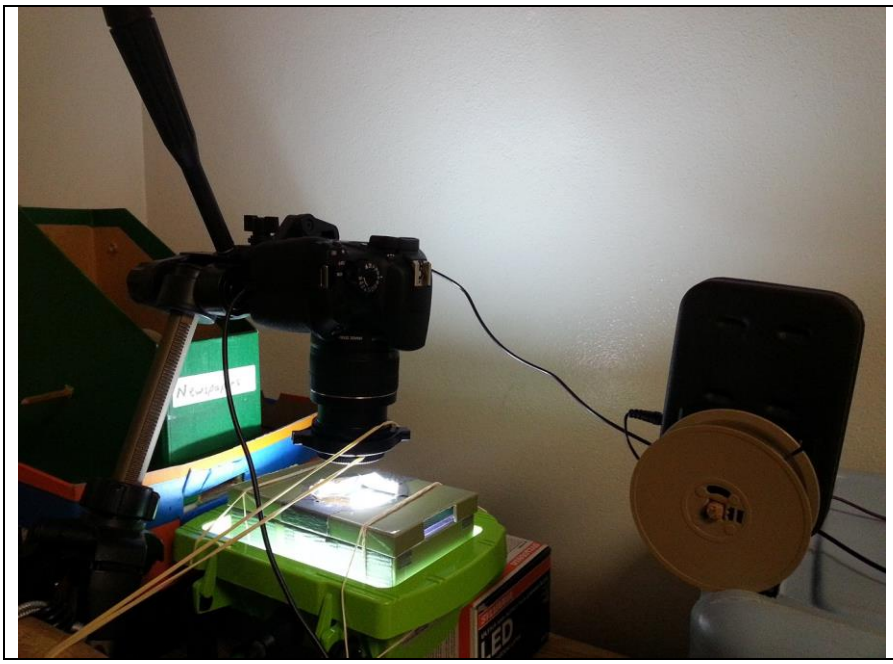


Figure 8 - Microfilm scanning prototype, note use of elastic to keep macro lens from shifting.

Given the rapid decline in the cost of mirrorless cameras, efforts are underway to evaluate mirrorless cameras with a similar price point. ODW has applied for a series of grants to fully flesh out a blueprint for building affordable microfilm/microfiche scanners with a target price point of $1000 US for associated materials, excluding the cost of a workstation for driving the process.

Open source software has been essential to make the prototype viable. OpenCV[27] has been used to identify portions of pages that require more than one camera shot (see Figure 9). The portions are then merged using Enbend[28]. OpenCV is then used once more to crop the resulting newspaper page.

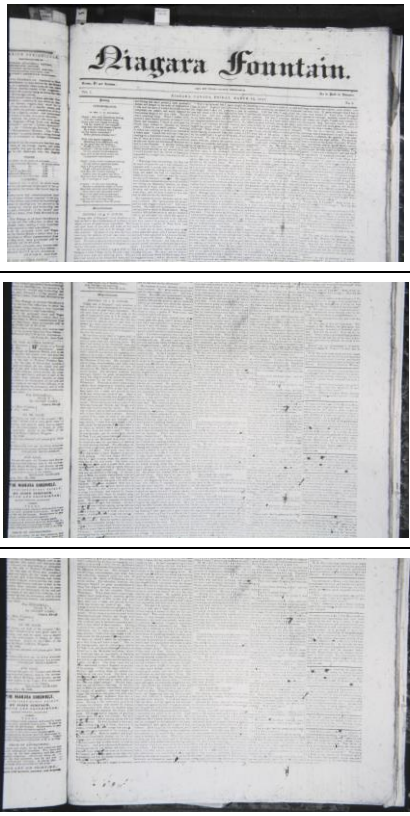


Figure 9 - a newspaper page across three camera shots. In this case, OpenCV will detect the starting and ending images of the sequence, and Enblend will then be used to merge the images together.

Possible Future Directions

ODW, via its predecessor, OurOntario, was a very early adopter of Solr, and it has been a stable part of ODW's digital library infrastructure for over a decade. Despite this, it would be worthwhile to evaluate Elasticsearch[29] to possibly replace or augment Solr. Elasticsearch is also highly regarded, and its JSON-centric approach might be a strong fit for interacting with newspaper indexes.

Olena has been invaluable for OCR processing, but it might also have a role for image presentation. Changing color levels for improving the legibility of text often has the consequence of degrading the graphics and photos on a newspaper page. In addition to identifying text blocks, Olena also attempts to delimit image segments. This would make it possible to perform one action on text segments and to exclude or perform a different action on image segments (see Figure 10).

Figure 10 - Brightness/contrast adjustment on left image to brighten text has the unfortunate consequence of darkening photos and graphics. On the right, Olena's coordinates for detected non-text regions on the page are leveraged to exclude original photos and graphics from adjustment.

Similarly, it would be possible to run OCR on PDF documents where images are intermixed with accessible text. This is very common, for example, with advertising, where the newspaper publisher has dropped in an image for an ad, but it is surrounded by text blocks. Currently, an assessment is made on whether to use OCR or text extraction, but the best results might lay in a combination of both.

Acknowledgments

The author wishes to thank the University of Windsor for being a strong supporter of local newspaper digitization as well the staff and volunteers who help make ODW a success. A special thanks to DiscoveryGarden and the University of Prince Edward Island for supporting the author in researching OCR during a sabbatical in 2011.

References

1 As quoted in Raible, Chris. *The power of the press. The story of early Canadian printers and publishers*. Toronto: James Lorimer & Company Ltd. 2007. p. 31.

2 "Home - The Globe and Mail". <http://www.theglobeandmail.com/> Web. 13 April 2017.

3 "National Post| Canadian News, Financial News and Opinion". <"http://www.nationalpost.com/index.html"> Web. 13 April 2017.

4 "OCNA - About". < http://www.ocna.org/about> Web. 13 April 2017

5 "Canada's Historical Newspaper Digitization Problem, Part 2". N.p., Feb. 13, 2014. <http://activehistory.ca/2014/02/historical-newspaper-digitization-problem>

6 "OurDigitalWorld". <https://ourdigitalworld.net> Web. 13 April 2017

7 "Backblaze - Home". <https://www.backblaze.com> Web. 13 April 2017

8 "Ontario Library Research Cloud". <https://cloud.scholarsportal.info> Web. 13 April 2017

9 "Scholars Portal". < http://scholarsportal.info> Web. 13 April 2017

[10] “Leddy Library|University of Windsor”. <http://leddy.uwindsor.ca> Web. 13 April 2017

[11] “Home | HPL - Hamilton”. <http://www.hpl.ca/> Web. 13 April 2017

[12] “OCR, PDF, Text Scanning Software and Solutions - ABBYY”. <https://www.abbyy.com> Web. 13 April 2017

[13] “tesseract-ocr”. <https://github.com/tesseract-ocr> Web. 13 April 2017

[14] “Olena - LRDE”. <https://www.lrde.epita.fr/wiki/Olena> Web. 13 April 2017

[15] See the EPITA (The École Pour l'Informatique et les Techniques Avancées - OLENA comes from the EPITA Research and Development Laboratory) competition entry in A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher. “Historical Document Layout Analysis Competition”. *Proceedings of the 11th International Conference on Document Analysis and Recognition* (ICDAR2011), Beijing, China, September 2011, pp. 1516-1520

[16] See Stefan Pletschacher and Apostolos Antonacopoulos. 2010. “The PAGE (Page Analysis and Ground-Truth Elements) Format Framework”. In *Proceedings of the 2010 20th International Conference on Pattern Recognition* (ICPR '10). IEEE Computer Society, Washington, DC, USA, 257-260. <http://dx.doi.org/10.1109/ICPR.2010.72>

[17] “Convert, Edit, Or Compose Bitmap Images @ ImageMagick”. <https://www.imagemagick.org/script/index.php> Web. 13 April 2017

[18] “Apache PDFBox| A Java PDF Library”. <https://pdfbox.apache.org> Web. 13 April 2017

[19] “Welcome to Apache™ Hadoop®”. <http://hadoop.apache.org> Web. 13 April 2017

[20] “Apache Lucene - Apache Lucene Core”. <https://lucene.apache.org/core> Web. 13 April 2017

[21] Apache Solr. <http://lucene.apache.org/solr> Web. 13 April 2017

[22] A good explanation of the process is given at “Scanning without a Scanner: Digitizing Your Film with a DSLR”. <https://www.bhphotovideo.com/explora/photography/tips-and-solutions/scanning-without-scanner-digitizing-your-film-dslr> N.p., n.d. Web. 13 April 2017.

[23] “Lomography Smartphone Scanner - Lomography Shop”. <https://shop.lomography.com/en/smartphone-scanner> Web. 13 April 2017

[24] .”DCR-250 Super Macro conversion lens for D-SLR cameras, 4K and HDV Camcorders”. <http://www.raynox.co.jp/english/dcr/dcr250/indexdcr250eg.htm> Web. 13 April 2017.

[25] “EOS Rebel T6 EF-S 18-55mm IS II Kit”. <https://www.usa.canon.com/internet/portal/us/home/products/details/cameras/dslr/eos-rebel-t6-ef-s-18-55mm-is-ii-kit> Web. 13 April 2017.

[26] About 150 reels have been processed with the prototype at the time of this writing.

[27] “OpenCV library”. <http://opencv.org> Web. 13 April 2017.

[28] “Enblend/Enfuse”. <http://enblend.sourceforge.net> Web. 13 April 2017.

[29] “Open Source Search & Analytics · Elasticsearch”. <https://www.elastic.co> Web. 13 April 2017.