

NEW BRUNSWICK HISTORICAL NEWSPAPERS PROJECT

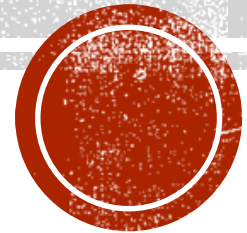
Leah Grandy (UNB Archives & Special Collections)

Mike Meade (Manager, Digital Imaging)

Jeremy McDermott (Senior Web Developer)

Jacob Sanford (Senior Technical Operations Manager)

Jeff Carter (Manager, Systems Group)



WHY NEWSPAPERS?



September - October 1972
Reels 2713 - 2718



NEWSPAPERS: SOMETHING FOR EVERYONE

- Students
- Academics and researchers
- Public



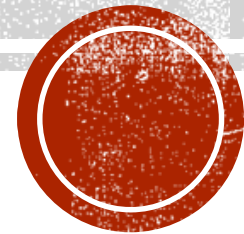
- UNB has the largest New Brunswick newspaper collection.
- UNB has been digitally preserving newspapers since 2010 from print and microfilm.

FORMATS

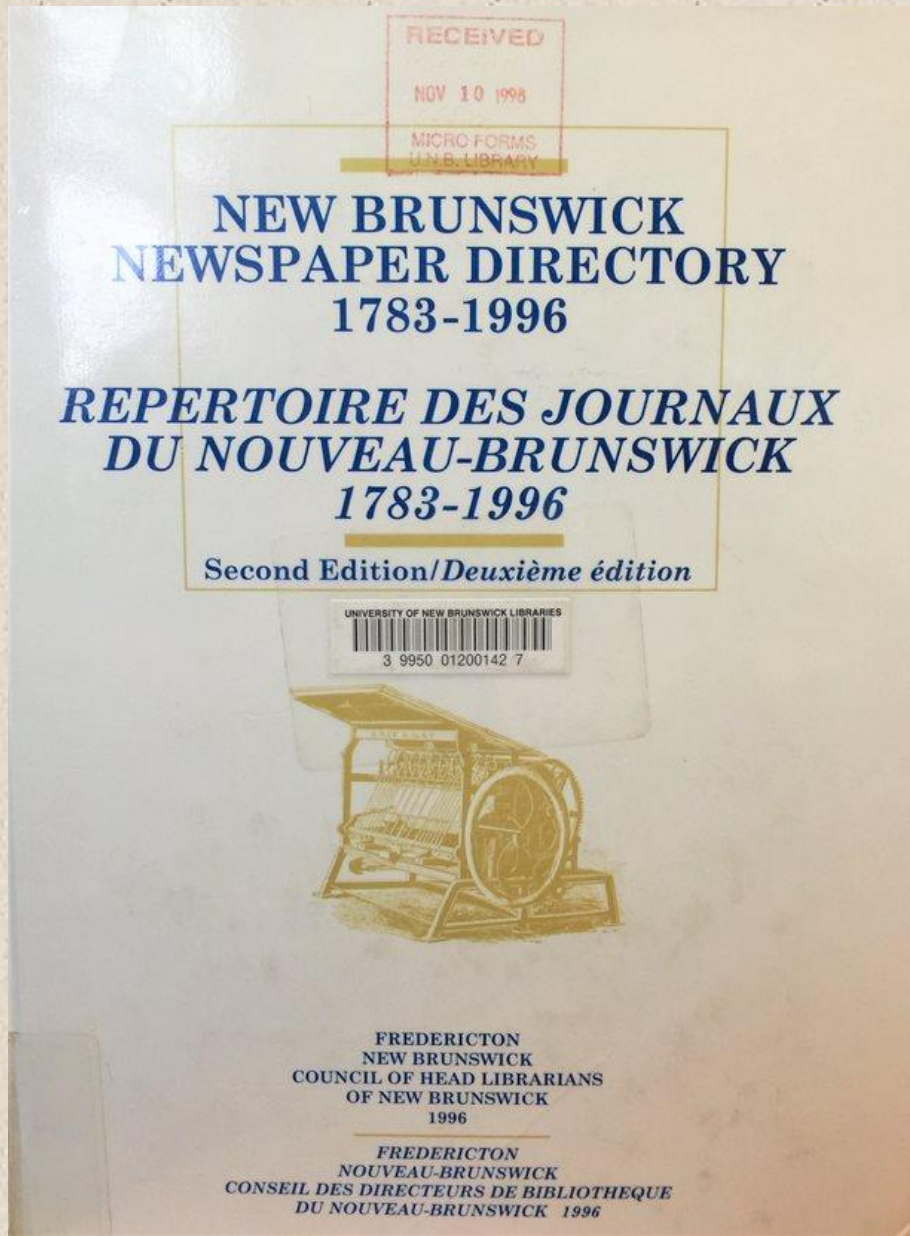




BACKGROUND TO THE PROJECT



- Provincial Archives of New Brunswick
- Council of Archives New Brunswick



Provincial Archives of New Brunswick

[Search](#) | [Exhibits and Education Tools](#) | [Research Tools](#) | [About PANB](#)

New Brunswick Newspaper Directory

[Foreword](#) [Preface](#) [Introduction](#) [Notes](#) [Illustrations](#) [Search](#)

This directory contains **697** newspaper listings.

View newspaper

Acadie dimanche, Moncton



Go

Or click a link below to view an index:

- [View index by Place](#)
- [View index by Publisher](#)
- [View index by Chronological Order](#)

Documents available:

-  [Publishing History](#)

This *PDF file shows the publishing history of the newspapers arranged by place.



New-Brunswick Court, published every Thursday... TREASURY BONDS... NOTICE...

NOTICE... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

FOR SALE... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

SHERIFF'S SALES... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

REMEDIES, DRUGS, A.C. PERFUMERY, SPICES, PAINTS, OILS... THE SOUTHERN TRADING CO. LTD. GENERAL MANAGERS...

PROJECT STATUS

- Currently the most accurate index of New Brunswick newspapers. Over 1,050 New Brunswick newspaper titles indexed with linking to free (and pay wall) online newspapers. 350 new titles have been discovered and added. Approximately 120 locally digitized, searchable titles with digitization is ongoing. New partnership with Canadiana/CRKN to share digitized content.

NOTICE... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

FOR SALE... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

SHERIFF'S SALES... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

REMEDIES, DRUGS, A.C. PERFUMERY, SPICES, PAINTS, OILS... THE SOUTHERN TRADING CO. LTD. GENERAL MANAGERS...

NOTICE... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

FOR SALE... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

SHERIFF'S SALES... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

REMEDIES, DRUGS, A.C. PERFUMERY, SPICES, PAINTS, OILS... THE SOUTHERN TRADING CO. LTD. GENERAL MANAGERS...

NOTICE... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

FOR SALE... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

SHERIFF'S SALES... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

REMEDIES, DRUGS, A.C. PERFUMERY, SPICES, PAINTS, OILS... THE SOUTHERN TRADING CO. LTD. GENERAL MANAGERS...

NOTICE... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

FOR SALE... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

SHERIFF'S SALES... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

REMEDIES, DRUGS, A.C. PERFUMERY, SPICES, PAINTS, OILS... THE SOUTHERN TRADING CO. LTD. GENERAL MANAGERS...

NOTICE... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

FOR SALE... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

SHERIFF'S SALES... THE BUILDING formerly occupied by the late JAMES WHITE... NOTICE...

REMEDIES, DRUGS, A.C. PERFUMERY, SPICES, PAINTS, OILS... THE SOUTHERN TRADING CO. LTD. GENERAL MANAGERS...

REMOVED... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

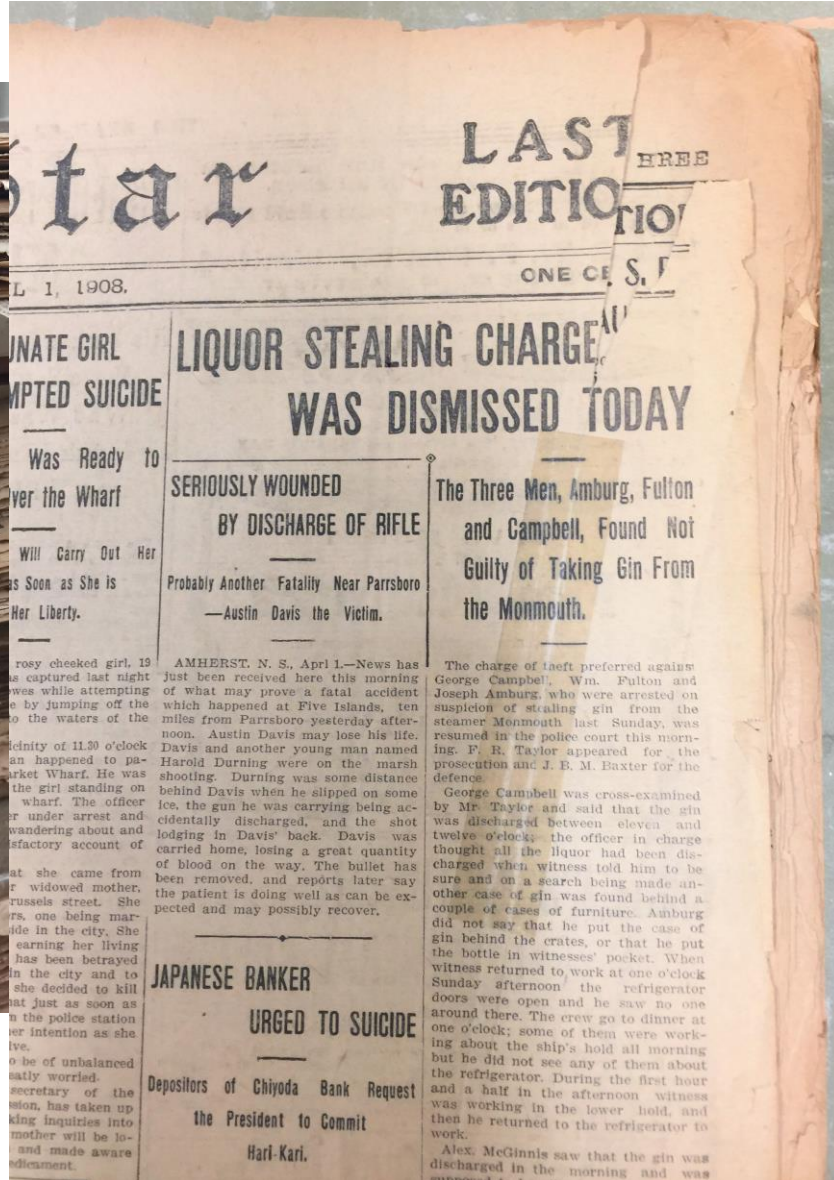
REMOVED... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

REMOVED... THE PERSONS having any demands against the Estate of ANDREW JENKS... NOTICE...

NEWSPAPERS.LIB.UNB.CA



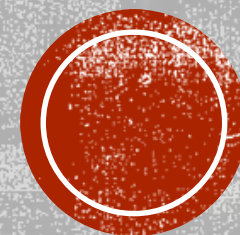
MATERIAL CONDITIONS

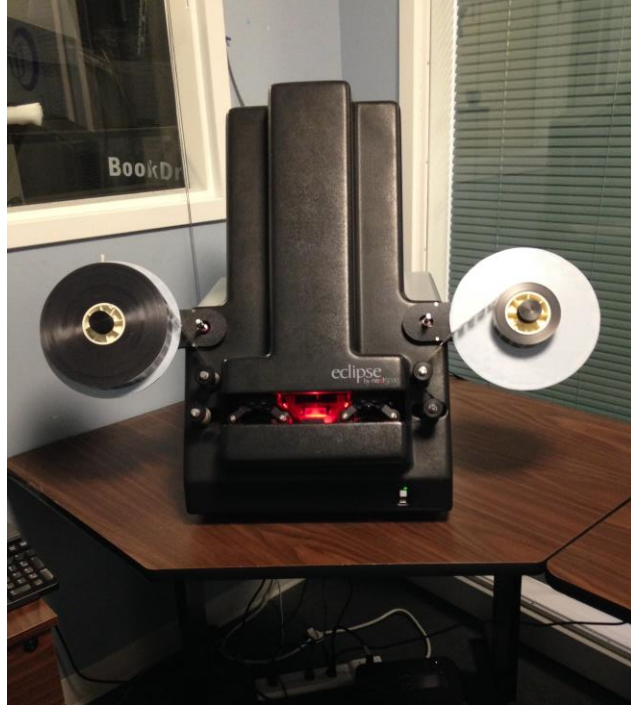
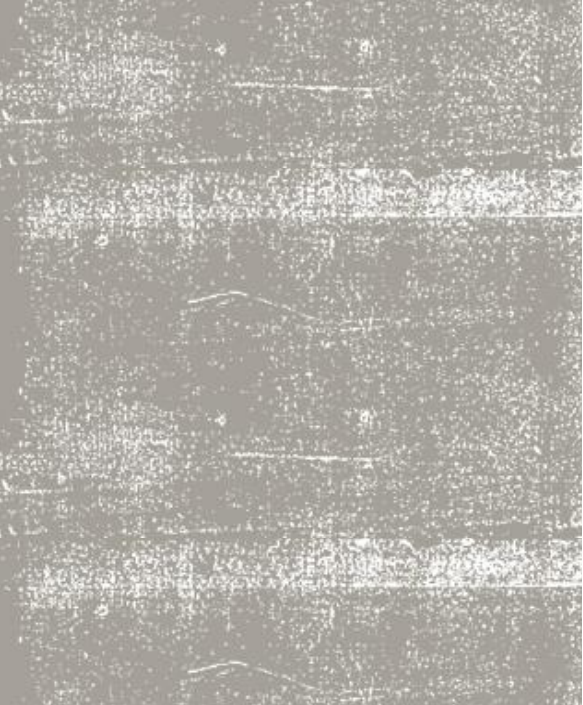






DIGITIZATION PROCESS





DIGITIZING FROM MICROFILM

- 2014 - Nextscan Eclipse microfilm scanner part of BNI contract
- completed 3 million pages of Brunswick News International material in just over 4 years



QUALITY CONTROL

- Image quality enhancements
 - Cropping, rotating and scaling
 - Surrogate generation
 - Can be a bottleneck
- Directory/File structure for ingest
 - Parent folder for titles
 - Unique folder for each issue
 - Images named with page numbers



IMAGE METADATA

```
define('ISSUE_DATE', mktime(0, 0, 0, 20, 7, 1878));
define('ISSUE_TITLE', 'The Penny Dip');
define('ISSUE_VOLUME', '1');
define('ISSUE_ISSUE', '28');
define('ISSUE_EDITION', '');
define('ISSUE_SUPPLEMENT_TITLE', '');
define('MISSING_PAGES', '');
define('ISSUE_ERRATA', '');

define('ISSUE_CONTENT_MODEL_PID', 'newspaperIssueCModel');
define('PAGE_CONTENT_MODEL_PID', 'newspaperPageCModel');
define('ISSUE_LANGUAGE', 'eng');
define('SOURCE_MEDIA', 'print');

// ** Save this file as metadata.php
```

- Issue level metadata:
 - Publication date
 - Title
 - Volume/Issue number
 - Alternative titles (supplement or section)
 - Errata

The background is a collage of three images. On the left is a close-up of a green printed circuit board (PCB) with various electronic components. In the center is a close-up of a computer keyboard, focusing on the keys and the keyboard's frame. On the right is a close-up of a person's hand holding a pen, with the pen tip pointing towards the left. The overall color palette is dark, with shades of blue, green, and black.

PROCESSING THE IMAGES

For display and fulltext searching

POST-SCAN INGEST PROCESS

Custom Processing Script

- PHP / Robo Framework
- 'Systems Toolkit'
 - <https://github.com/unb-libraries/systems-toolkit>

Steps:

- Issue Discovery
- Page OCR
- Issue Creation, Page Ingest
- DZI Tile Generation
- Issue Audit

1. ISSUE DISCOVERY, OCR QUEUEING

- Search path for all subfolders that contain a metadata file
- Search for all images in that folder and consider them to be each a page from that issue.
- Check if each image has an associated tesseract .hocr file generated. If not, the image is queued for OCR with Tesseract in a later step.
 - This is done to optimize core usage - some issues have a small number of pages
- Convert the metadata to JSON format for assertion to the Drupal REST API.



2. PAGE OCR

- Operates on all queued images to generate OCR for each page.
- No issue creation is done in the REST API until all issues have OCR complete.
- Tesseract 5.x, run in Docker
 - <https://github.com/unb-libraries/docker-tesseract>



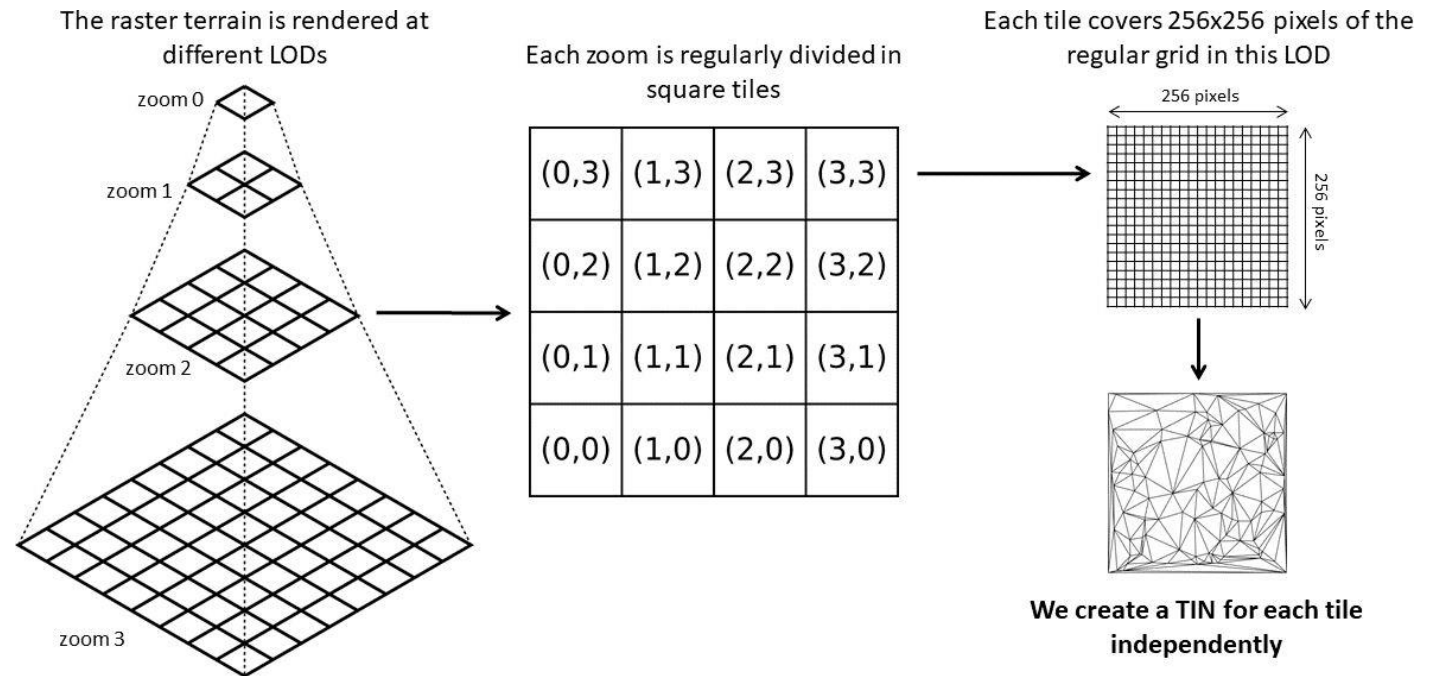
3. ISSUE CREATION, PAGE INGEST

- Once OCR is complete, the issue is then created.
- Iterate over folders and creates each an issue by POSTing to the Drupal REST API
- After the issue is created, create the pages and attached to the issue (via the REST API)
- Add OCR output to page.
- Writes a marker file inside the subfolder to indicate the issue has been ingested.



4. DZI TILE GENERATION

- Page viewer, **OpenSeadragon**, can leverage a set of 'tiles' to minimize the bandwidth for zooming.
- These tiles are then generated.



5. ISSUE AUDIT

- Audits all issues in the import
 - Verifies metadata
 - Hashes page files and compares to local
 - Checks page numbering



PROCESSING TIMES

- Processor: Core(TM) i9-12900 (16 Cores)
- Memory: 64GB
- Multi-threaded processes set to use 12 threads.

- Average total processing/ingest/audit time:
 - 30-40s per page

- Can scale horizontally to accommodate volume.



OCR / ACCURACY

Cropped			Uncropped			Delta			
Word Count	Sum Confidence	Confidence / Word	Word Count	Sum Confidence	Confidence / Word	Word Count	Word Improvement	Sum Confidence	Confidence / Word
2658	237761	89.45109105	2184	187381	85.79716117	474	21.70%	50380	1.036539299
2287	192983	84.38259729	985	82442	83.69746193	1302	132.18%	110541	0.6851353601
3658	332761	90.96801531	976	83715	85.77356557	2682	274.80%	249046	5.194449735
1416	121264	85.63841808	1211	47583	39.2923204	205	16.93%	73681	46.34609768
2486	201635	81.10820595	484	35442	73.22727273	2002	413.64%	166193	7.880933226
2233	199091	89.15853112	2653	191529	72.193366	-420	-15.83%	7562	16.96516512
2891	244375	84.52957454	2857	186810	65.38676934	34	1.19%	57565	19.1428052
3155	271258	85.97717908	2770	235405	84.98375451	385	13.90%	35853	0.9934245682
1821	166493	91.42943438	714	65792	92.14565826	1107	155.04%	100701	-0.7162238866
2638	245602	93.10159212	2666	247954	93.0060015	-28	-1.05%	-2352	0.09559061486
1728	161021	93.18344907	1816	167250	92.09801762	-88	-4.85%	-6229	1.085431453
2924	271011	92.68502052	2945	272447	92.51171477	-21	-0.71%	-1436	0.173305749
2906	234395	80.65898142	1026	79268	77.25925926	1880	183.24%	155127	3.399722158
2200	201292	91.49636364	2668	241502	90.517991	-468	-17.54%	-40210	0.9783726319
2007	171088	85.24564026	1685	142542	84.59465875	322	19.11%	28546	0.6509815054
2495	226225	90.67134269	2554	229649	89.91738449	-59	-2.31%	-3424	0.7539581905
2979	263733	88.53071501	3033	265995	87.70029674	-54	-1.78%	-2262	0.8304182691
3573	328250	91.86957739	3446	314514	91.26929774	127	3.69%	13736	0.6002796494
308	27199	88.30844156	307	27522	89.64820847	1	0.33%	-323	-1.339766911
1167	102274	87.63838903	2009	177146	88.17620707	-842	-41.91%	-74872	-0.5378180365
2779	241767	86.99784095	3129	243836	77.92777245	-350	-11.19%	-2069	9.070068499
2071	179268	86.5610816	2575	222905	86.56504854	-504	-19.57%	-43637	-0.003966940599
1816	153920	84.75770925	2334	200410	85.86546701	-518	-22.19%	-46490	-1.107757758
1210	100270	82.8677686	2338	197378	84.42172797	-1128	-48.25%	-97108	-1.553959378
1732	150003	86.60681293	1823	156242	85.70597916	-91	-4.99%	-6239	0.9008337778
1511	119603	79.15486433	1887	147104	77.95654478	-376	-19.93%	-27501	1.198319548
2939	247202	84.11092208	1372	108972	79.42565598	1567	114.21%	138230	4.685266106
3440	290874	84.55639535	2479	205376	82.846309	961	38.77%	85498	1.710086353
1456	132064	90.7032967	1468	133650	91.04223433	-12	-0.82%	-1586	-0.3389376291
1961	177574	90.55277919	310	27735	89.46774194	1651	REMOVED	149839	1.085037259
340	30462	89.59411765	326	29579	90.73312883	14	4.29%	883	-1.139011187
3035	274914	90.58121911	3024	272415	90.0843254	11	0.36%	2499	0.4968937136
2801	257798	92.03784363	2773	255800	92.24666426	28	1.01%	1998	-0.2088206353
1924	178742	92.9012474	6	424	70.66666667	1918	REMOVED	178318	22.23458073
1199	104726	87.34445371	1855	165614	89.27978437	-656	-35.36%	-60888	-1.935330655
2795	252021	90.16851521	2790	251509	90.14659498	5	0.18%	512	0.02192022365
						Mean Extra Words	307.25	33.71%	3.87%

Slight improvement if images are cropped to page content, but can be very labour intensive.





New Brunswick Historical Newspapers Project

New Brunswick Historical Newspapers Project provides researchers with unified access to UNB Libraries' current and historical newspaper collections in all formats, from New Brunswick and across the world. Search and discover print, microform, and selected digital newspaper titles available from UNB Libraries.

Fulltext searching is available for our growing list of digitized titles. Learn more about the New Brunswick Historical Newspapers Project or contact us with questions.


For more worldwide digital newspaper content, consult UNB Libraries licensed electronic Newspaper collections. UNB/STU login required.

Title Search Fulltext Search

Search for Newspaper Titles

Search by title, location, publisher, notes, description or combination, i.e. Moncton 1932

Include newspapers published outside of New Brunswick




Featured Publication

Evening Times
Feb 9, 1905 - Mar 3, 1910
Saint John NB
Canada

Daily (Every evening except Sunday)
Some errors in issue numbering.

This Day in New Brunswick History



The Christian Visitor.

NBHP Partnerships



WEBSITE

- Drupal 9
- Custom entities
- Solr search
- Custom UNB Libraries Bootstrap theme
- [newspapers.lib.unb.ca](https://github.com/unb-libraries/newspapers.lib.unb.ca)
 - <https://github.com/unb-libraries/newspapers.lib.unb.ca>

