

# Digitized Newspapers as Everyday Interdisciplinarity: The Transformation of Historical Scholarship

**Ian Milligan**

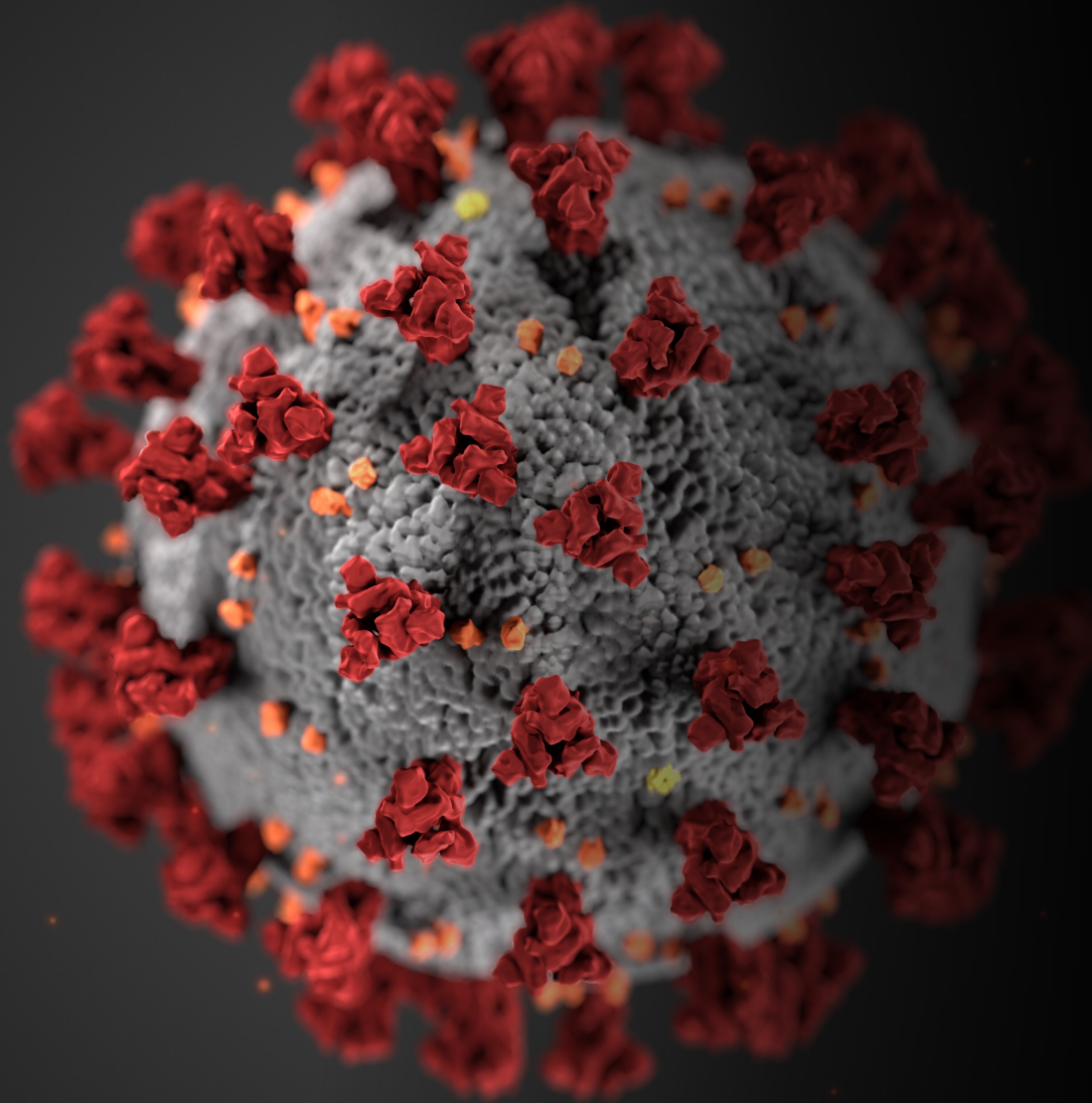
University of Waterloo, Canada







**All historians are  
*digital* historians**



---

**COVID is  
*accelerating*  
but not  
*inventing*  
trends**

---



A photograph of a person's hands resting on an open Bible. The person is wearing a grey knitted sweater and a ring on their left hand. A green arrow points from the text 'How historians imagine how they work' to the Bible. The Bible is open to a page with text in two columns, and there are some handwritten notes in the margins. The background is a wooden surface.

## How historians imagine how they work

- Literature review, finding historiographical problem
- Identifying primary sources
- Analyzing primary sources
- Writing
- Revision
- Preliminary presentations (conference, article)
- Finished publication (dissertation, book)



## How historians *really* work

- Literature review, finding historiographical problem (**shaped by digitized/non-digitized secondary resources**)
- Identifying primary sources (**online finding aids**)
- Analyzing primary sources (**doing keyword searches in databases that you barely understand of documents that you don't know provenance of**)
- Writing/Revision/Publishing







---

## We know archives transform historical work...

---

- We are often vaguely cognizant of the role that archives play in shaping our histories, but we still tend to treat them as “neutral and unproblematic reservoirs of historical fact.” (Walsham)
- Archivists and historians share common origins, but we have diverged
- “Any visit by a historian to an archival institution is now an exercise in interdisciplinarity.” (Blouin and Rosenberg, *Processing the Past*)



A close-up photograph of a hand holding a black magnifying glass over a laptop keyboard. The magnifying glass is positioned over the keyboard, and the text is overlaid on the lower left portion of the image. A solid green horizontal bar is located at the bottom of the image.

**Just as archives mediate the  
past, so to do the workflows  
that we use in the digital age.**



**So, is using a digitized newspaper an exercise in interdisciplinarity?**






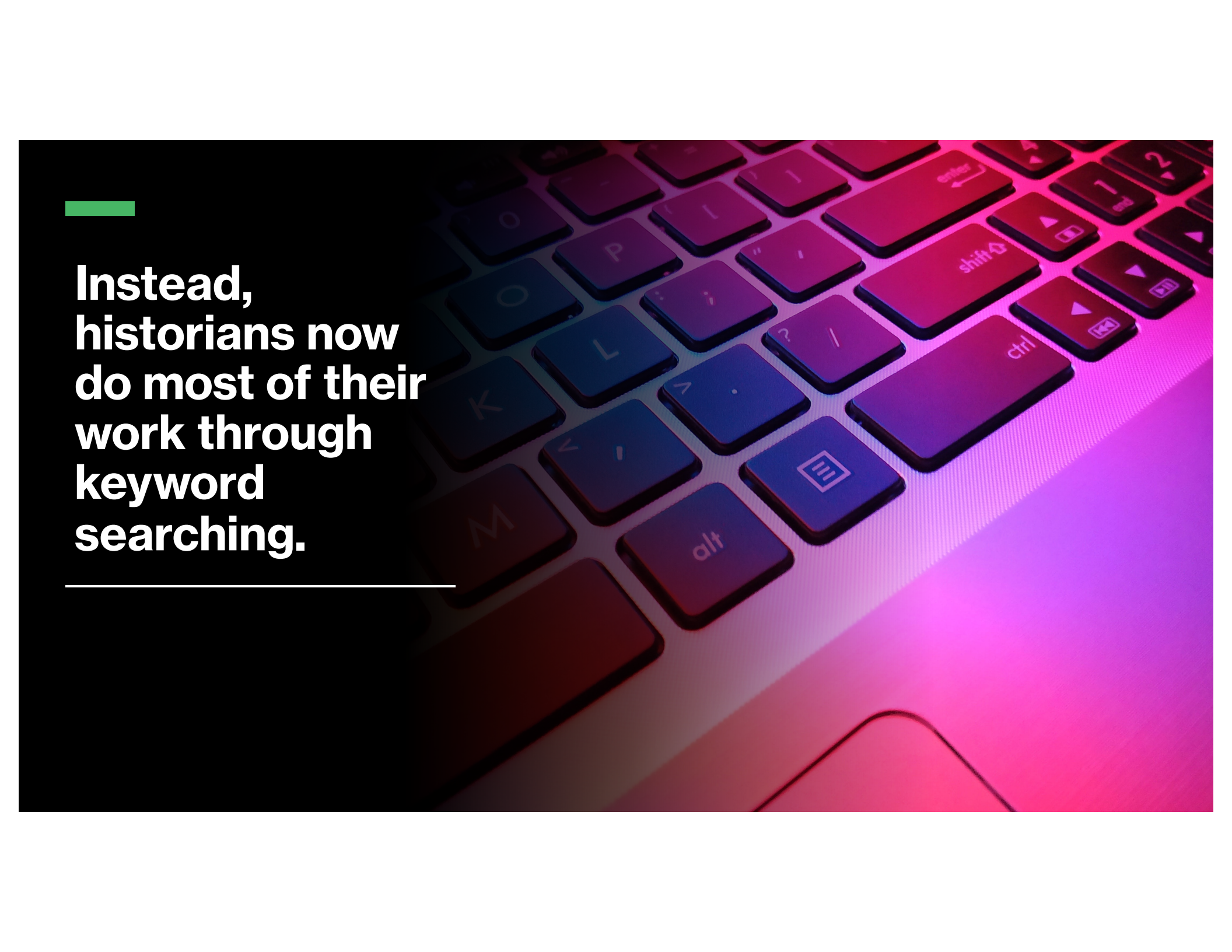
---

## How did historians previously use newspapers?

---

- Microfilm reels
- Sit in a dark basement, crank forward through pages looking for relevant documents
- Time consuming, boring (good chance to listen to music)
- But you learn a *lot* about context
  - i.e. a global event that happens, the tenor of advertisements, the relative placement of columnists or articles.





**Instead,  
historians now  
do most of their  
work through  
keyword  
searching.**

---





**But we  
uncritically  
use  
interfaces.**

---



# Let me use a Canadian example.

ProQuest  
UNIVERSITY OF WATERLOO  
ProQuest Historical Newspapers: Toronto Star

Basic Search | **Advanced Search** | Publications | Change databases

Your search for Kingston AND Roosevelt found 0 results.  
Please modify your search and try again. [Search tips](#)

**Advanced Search** [Command Line](#) [Recent searches](#) [Field codes](#) [Search tips](#) [University of Waterloo Library](#)

Kingston in Anywhere

AND Roosevelt in Anywhere

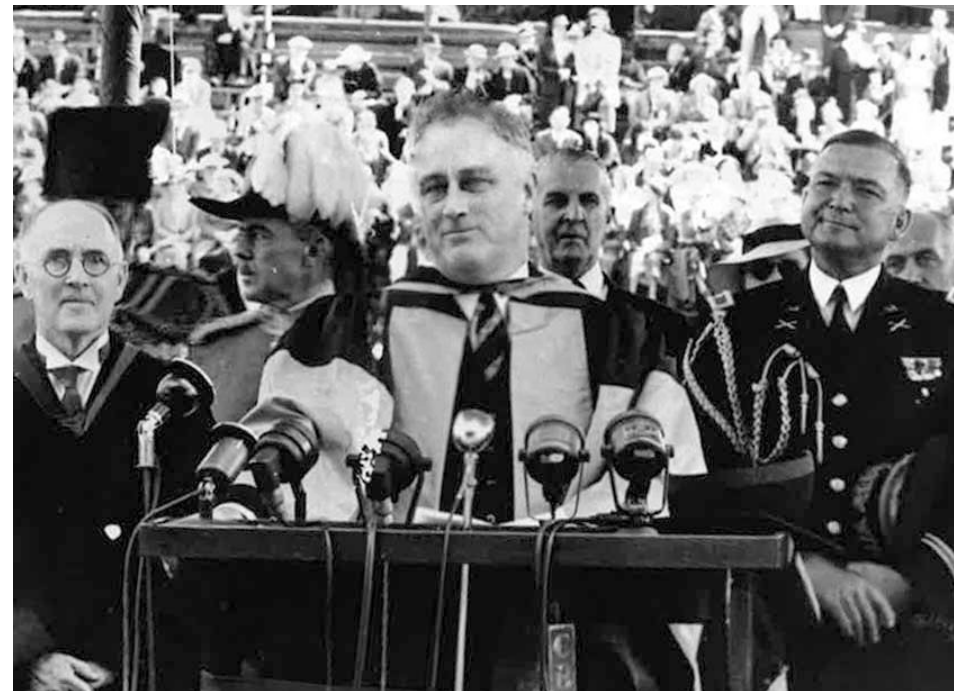
+ Add a row

Publication date: Specific date range...

Start  
August 18 1938

End  
August 24 1938

[Search](#) [Clear form](#)





## Choose an issue to view

1938 ▾



✓ December

November

October

September

July

June

May

April

March

February

January



Dec 31, 1938 ▾

[View issue](#)

### Issue content

☐ Select 1-32

☐ [Page 1](#)

1 [Toronto](#)

[Details](#)

[Search within this issue](#)

0-1971); **Toronto, Ontario** [Toronto, Ontario]30 Ju





---

# Missing

---

- For some reason, August 1938 is missing (the 1 September 1938 issue has reference in the “letters to the editors” about articles written in the past week, so there were apparently issues)
- Crucially, you wouldn’t know that if you just did keyword searches.
- This was the first month I went looking for, because of FDR, what other gaps are there?



## Other limitations of search that historians may not know of

- The text that is being searched is created using **optical character recognition**, or OCR
  - ProQuest's implementation stems from *Pages of the Past*, an innovative project that saw the *Toronto Star* the first fully digitized newspaper in the world
  - But it's a commercial platform, so correcting OCR is difficult
  - Doesn't catch line-break hyphenation
  - **Cutting-edge OCR would have best-case scenario of 98%; even that leads to 50 incorrect characters on an average page of 500 words; word accuracy would be around 90%**

# Skimming is nearly impossible

The screenshot displays the ProQuest Historical Newspapers: Toronto Star interface. The header includes the ProQuest logo, the University of Waterloo access information, and navigation links for Basic Search, Advanced Search, Publications, and Change databases. The main section is titled "Publication Search" and features a search bar, a dropdown menu set to "In title", and a "Search" button. A sidebar on the left shows a "Publication date" filter with a bar chart for the range 1894 - 2010 (decades) and an "Update" button. The main content area lists "4 publications" with options to "View summary" or "View title only". The first three publications are detailed below:

Publication	Full text coverage	Citation/Abstract coverage	Publisher	Place of publication
1 Evening Star (1894-1900); Toronto, Ontario	Jan 2, 1894 - Jan 24, 1900	Jan 2, 1894 - Jan 24, 1900	Torstar Syndication Services, a Division of Toronto Star Newspapers Limited	Toronto, Ontario
2 Toronto Daily Star (1900-1971); Toronto, Ontario	Jan 25, 1900 - Nov 5, 1971	Jan 25, 1900 - Nov 5, 1971	Torstar Syndication Services, a Division of Toronto Star Newspapers Limited	Toronto, Ontario
3 Toronto Star (1971-2009); Toronto, Ontario	Nov 6, 1971 - Dec 31, 2009	Nov 6, 1971 - Dec 31, 2009	Torstar Syndication Services, a Division of Toronto Star Newspapers Limited	ISSN 0319-0781

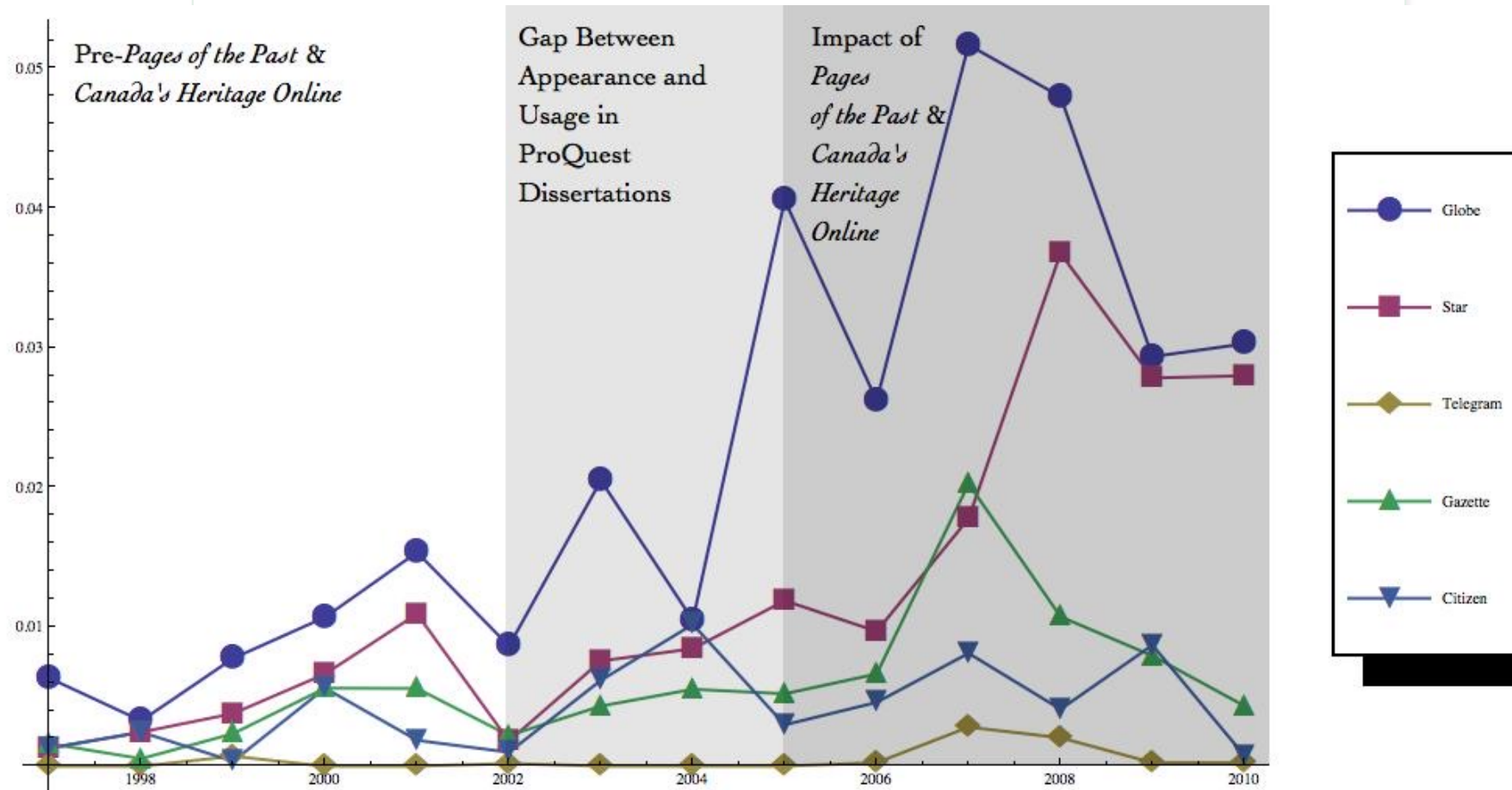


The background is an abstract, textured surface. It features a mix of dark grey, black, and muted blue tones, with lighter, almost white, speckled areas. The texture appears rough and uneven, similar to a weathered wall or a close-up of a mineral surface. The lighting is somewhat uneven, with darker patches on the right and lighter patches on the left.

**And  
digitization is  
uneven**



# In Canada







## In other words

The more something is digitized the more it is used (i.e. the *Toronto Star* and the *Globe and Mail* are used far more than before; the *Toronto Telegram* is almost never used)

**The mediation of a source impacts its use**

# The Impact of this Medium Shift

- We now interact primarily through keyword search (i.e. the system forces us more or less to do this)
- We don't fully understand the construction of this database.
- The text is inaccessible to do transformative digital scholarship with.
- **Yet we still cite it all the same: Pages of the Past, ProQuest, Clipping File, Microfilm; yet each system dramatically impacts our work and the way we understand the source.**





**If using an archive is an exercise in interdisciplinarity, perhaps we should think of using online platforms the same way?**



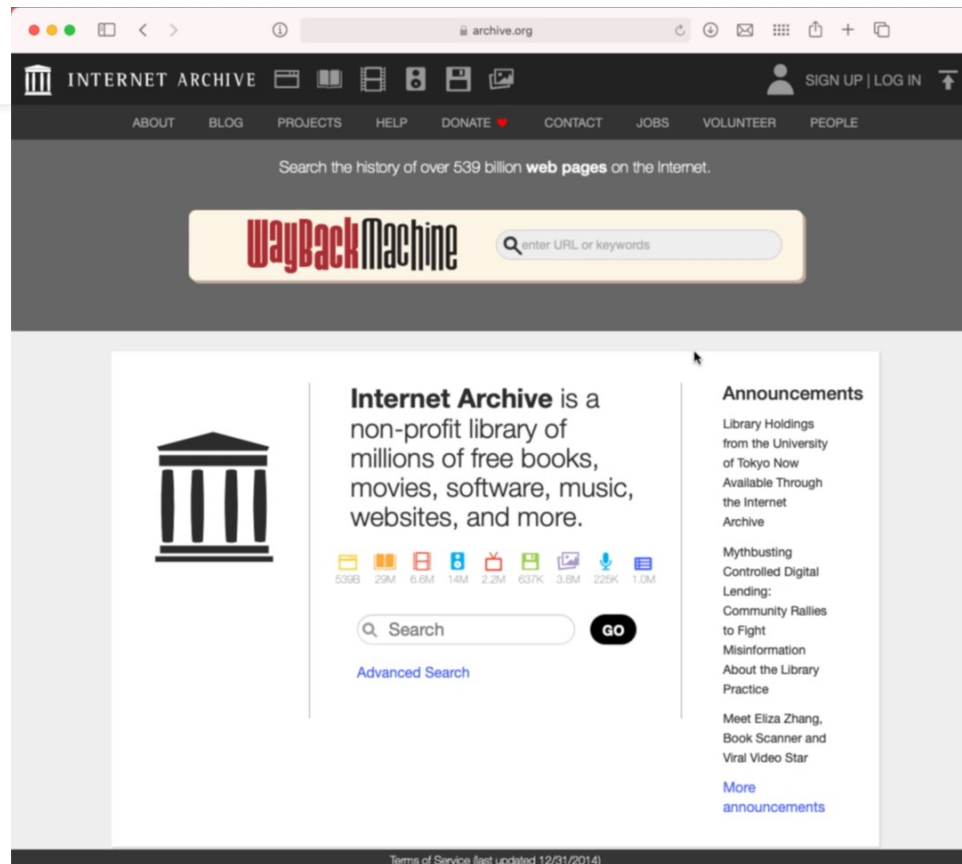
**So what can we do?**





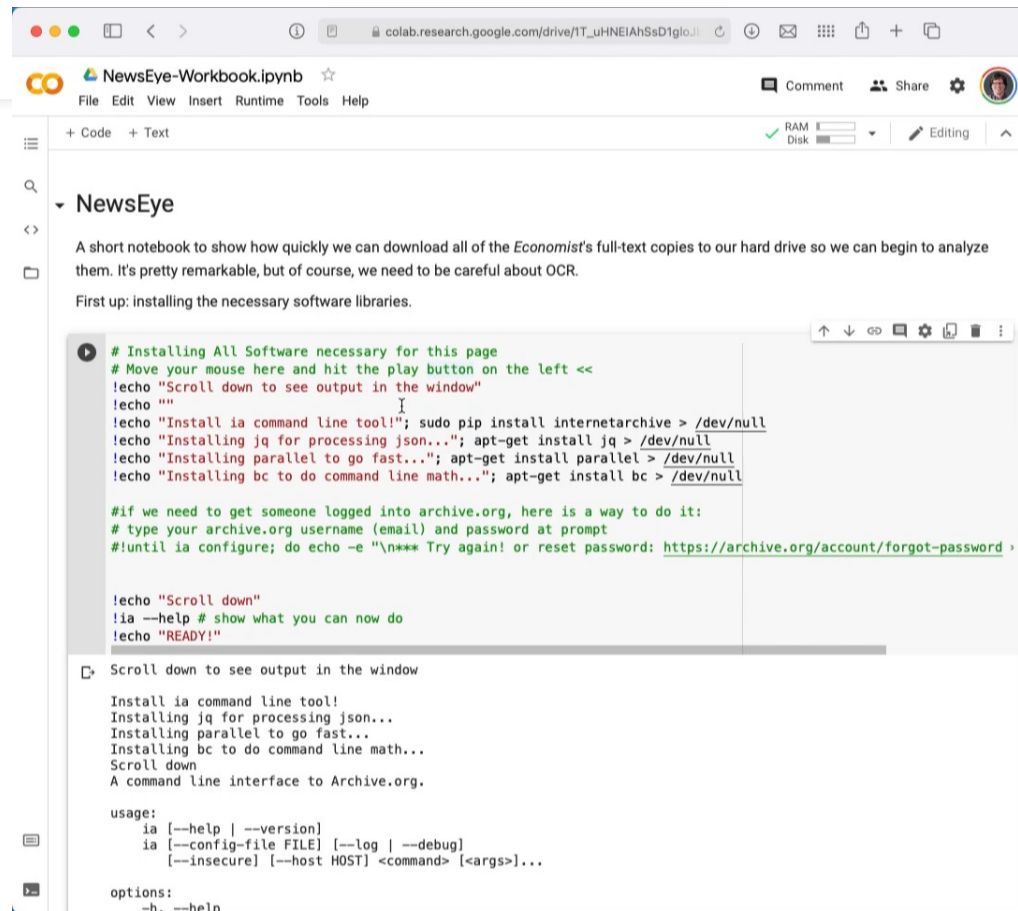
**Come to the NewsEye  
Conference!**

# Embrace Open Platforms and Tools





# Embrace Open Platforms and Tools



The screenshot shows a Google Colab notebook interface. The browser address bar at the top indicates the URL: `colab.research.google.com/drive/1T_uHNEIAHsD1gloJ...`. The notebook title is "NewsEye-Workbook.ipynb". The menu bar includes "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". On the right, there are buttons for "Comment", "Share", and a user profile icon. Below the menu bar, there are tabs for "+ Code" and "+ Text", and status indicators for "RAM", "Disk", and "Editing".

The notebook content is titled "NewsEye" and includes the following text:

A short notebook to show how quickly we can download all of the *Economist's* full-text copies to our hard drive so we can begin to analyze them. It's pretty remarkable, but of course, we need to be careful about OCR.

First up: installing the necessary software libraries.

```
# Installing All Software necessary for this page
# Move your mouse here and hit the play button on the left <<
!echo "Scroll down to see output in the window"
!echo ""
!echo "Install ia command line tool!"; sudo pip install internetarchive > /dev/null
!echo "Installing jq for processing json..."; apt-get install jq > /dev/null
!echo "Installing parallel to go fast..."; apt-get install parallel > /dev/null
!echo "Installing bc to do command line math..."; apt-get install bc > /dev/null

#if we need to get someone logged into archive.org, here is a way to do it:
# type your archive.org username (email) and password at prompt
#!until ia configure; do echo -e "\n*** Try again! or reset password: https://archive.org/account/forgot-password ,
```

Below the code cell, there is a scrollable output area showing the results of the commands:

```
Install ia command line tool!
Installing jq for processing json...
Installing parallel to go fast...
Installing bc to do command line math...
Scroll down
A command line interface to Archive.org.

usage:
  ia [--help | --version]
  ia [--config-file FILE] [--log | --debug]
  [--insecure] [--host HOST] <command> [<args>]...


options:
  -h, --help
```



# What does this mean?

- If working with an archive we don't really understand is an “exercise in interdisciplinarity,” **isn't working with a newspaper that we don't really understand one too?**
- **Open platforms as much as possible** (I know copyright is a thing!)
- **Finding ways to translate knowledge;** I think historians can get hit by a blast of the obvious when they realize that they need to think about their platforms. But they rarely do.



A dark, textured wooden surface serves as the background. A light gray rectangular sticky note is placed on the left side, featuring the word "thanks!" written in a black, cursive script. To the right of the sticky note, a black marker with white text is positioned vertically. The marker has "4mm Tombow ABT", "Acid Free", and "N15" printed on it. A small green horizontal bar is located above the main text on the right.

*thanks!*

**Looking  
forward to  
the  
discussion!**

---