



## The Digitally-Literate Historian: A Workshop

Ian Milligan (University of Waterloo)

# Static Slides for this Workshop

- <https://www.ianmilligan.ca/files/queens-workshop2.pdf>
- **I will drop them in the chat!**



# Who am I?

- BA, Queen's (2006); MA, PhD York (2007; 2012)
- My chronology spans a few different periods
  - Microfilm in the basement of Stauffer Library
  - Archival photography was forbidden during my first visit to Library and Archives Canada as an undergraduate; common by the time I did my PhD
  - From social history to digital methods (and back again!)
- I use infrastructure as a historian but also **build infrastructure.**



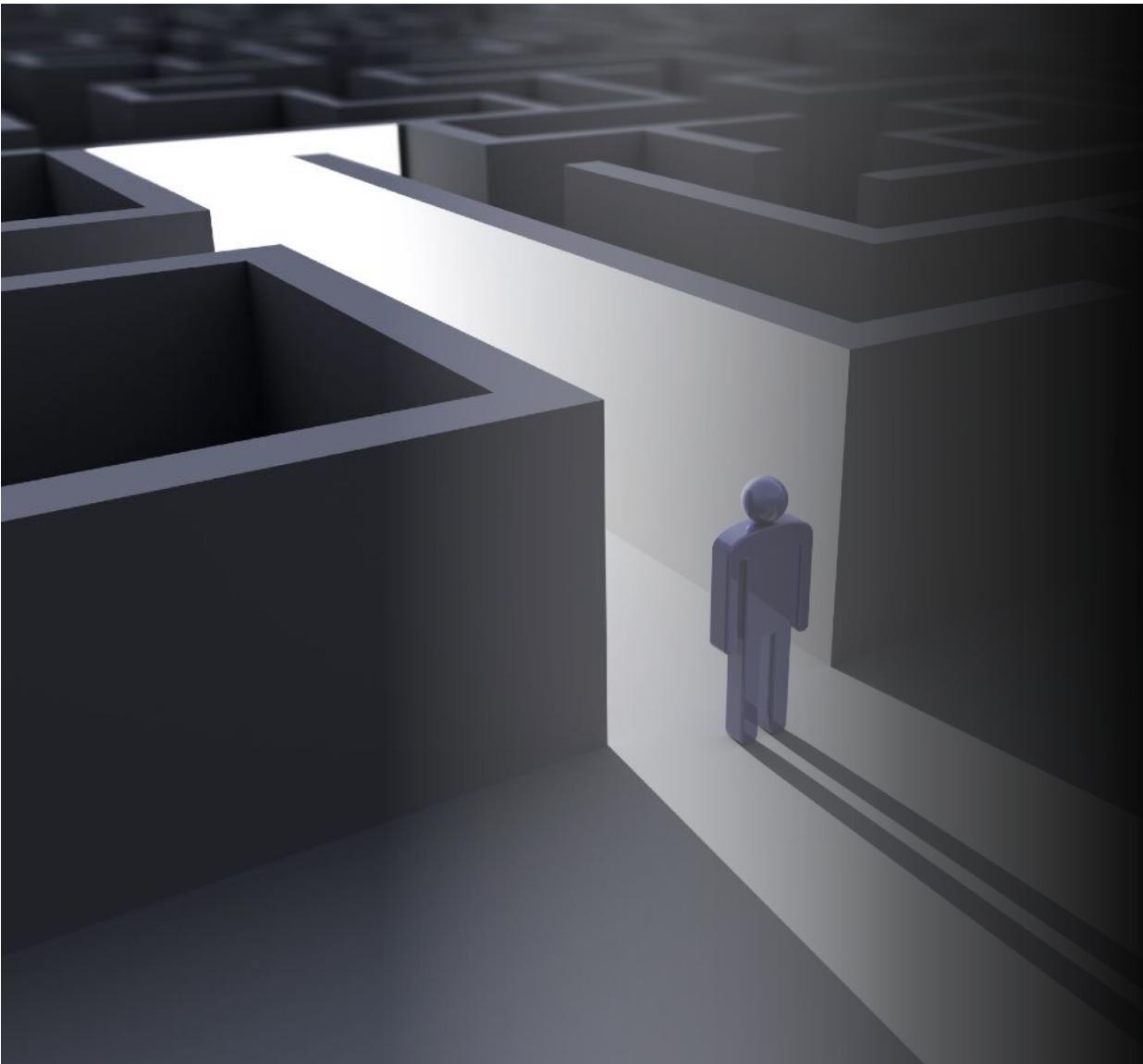
---

# Who are you?

---

- I see lots of familiar faces (note I am speculatively writing this assuming that some of you came back!)
- For the new folks this week!
  - Name, place in the program
  - How many primary sources have you read on the Internet?





## Plan for the Workshop

---

- A series of conversations and exercises to explore the **core question** of “how is historical knowledge mediated and accessed in the digital age?”

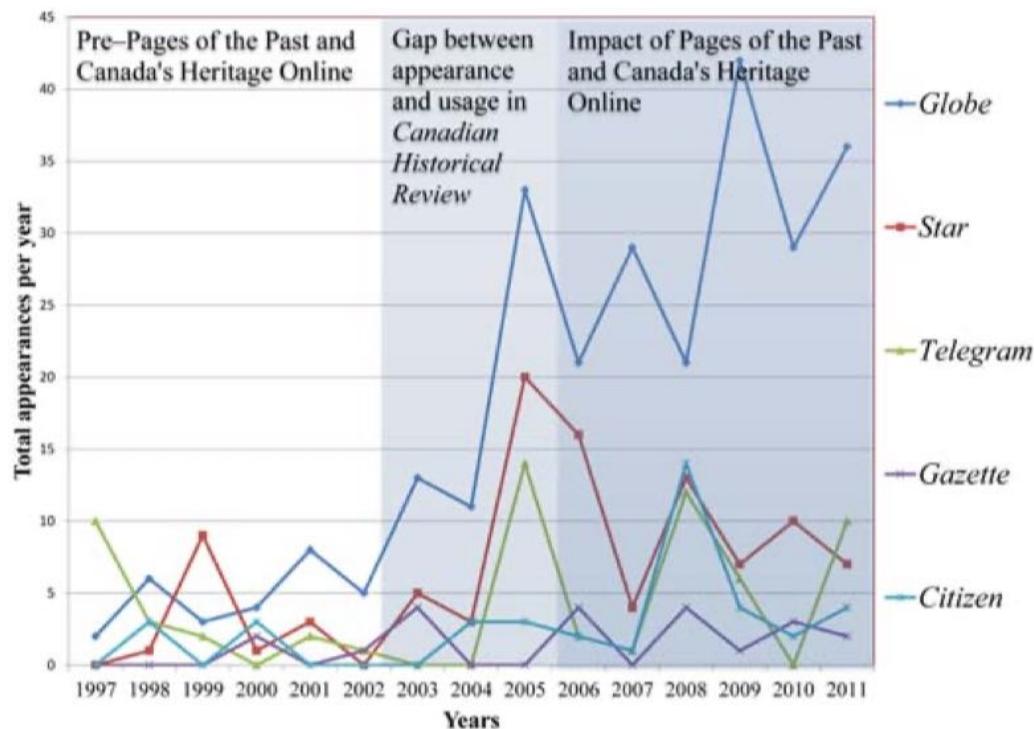
# Readings

- Tim Hitchcock, “Confronting the Digital: Or How Academic History Writing Lost the Plot,” *Cultural and Social History*, vol. 10, issue 1 (2013): 9-23.
- Lara Putnam, “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast,” *American Historical Review*, vol. 121, issue 2 (April 2016): 377-402.
- Ian Milligan, “We Are All Digital Now: Digital Photography and the Reshaping of Historical Practice,” *Canadian Historical Review*, vol. 101, issue 4 (December 2020): 602-621.

# Primary Sources



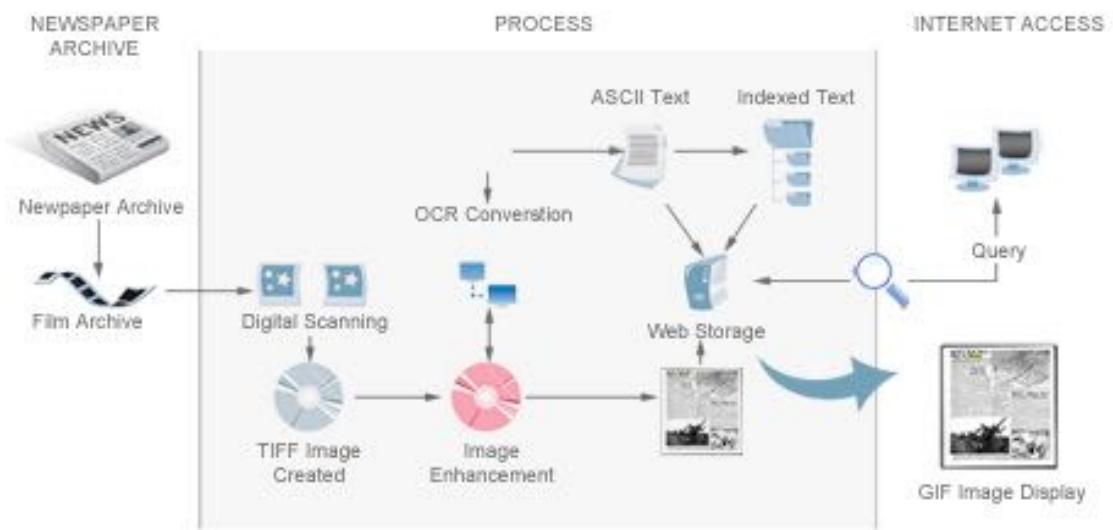
# Digitization



- Some of you read my 2013 article “Illusionary Order” in the *Canadian Historical Review*
  - (Wow, I think, my writing style has improved since 2013)
- As things are digitized, and other things aren’t, our citation practices change accordingly.

## Anatomy of a Digitized Newspaper: *The Toronto Star*

- Digitized by Cold North Wind's "Paper of Record" project
  - "Conceived by electronic publishing and web pioneer, R.J. (Bob) Huggins in a local Ottawa, Mexican restaurant in 1999" (their official history)
  - Forward thinking: in 1999 broadband wasn't widespread, but was necessary in an environment to access scanned GIFs

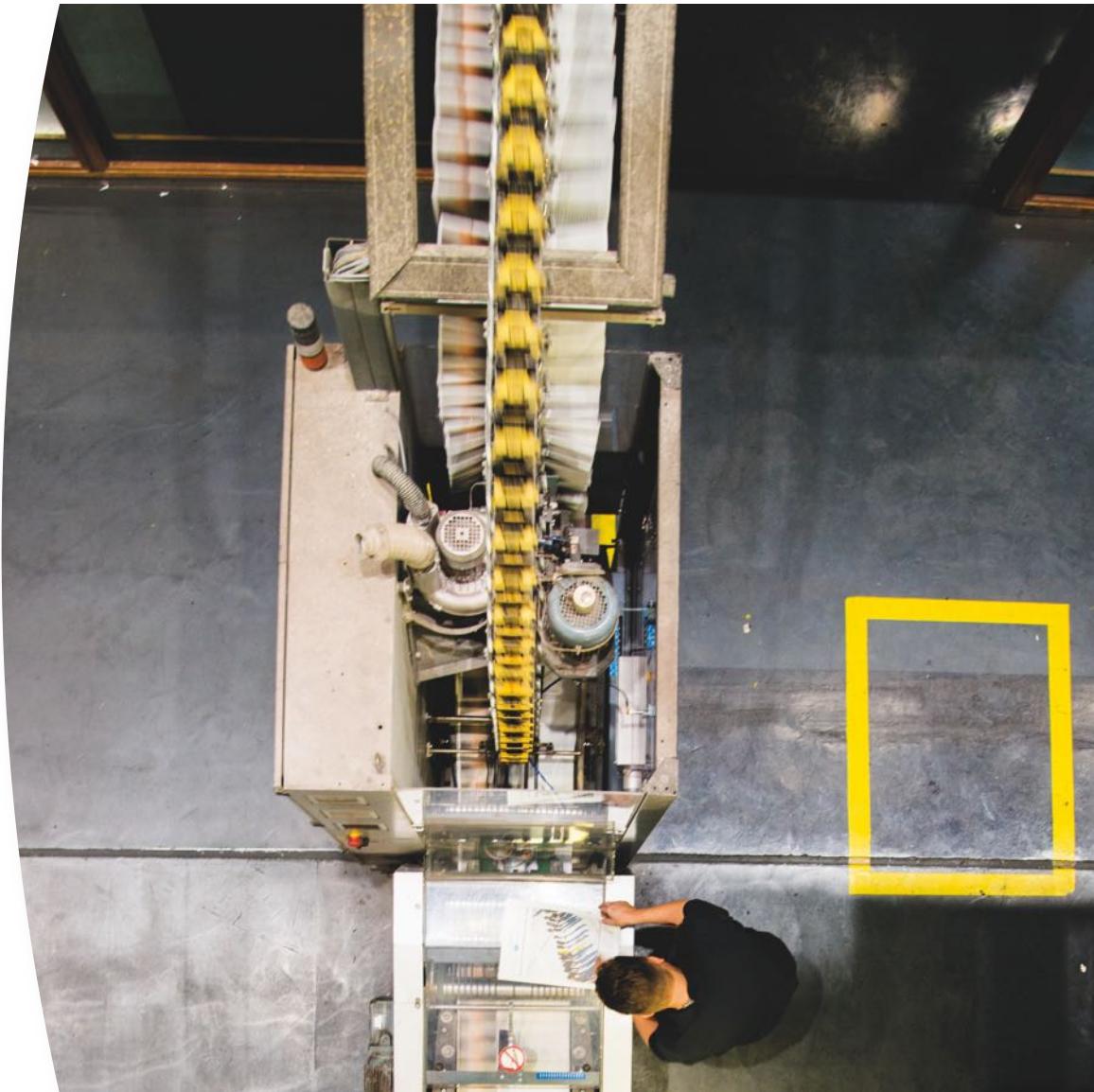


<https://paperofrecord.hypernet.ca/default.asp>

# First movers!

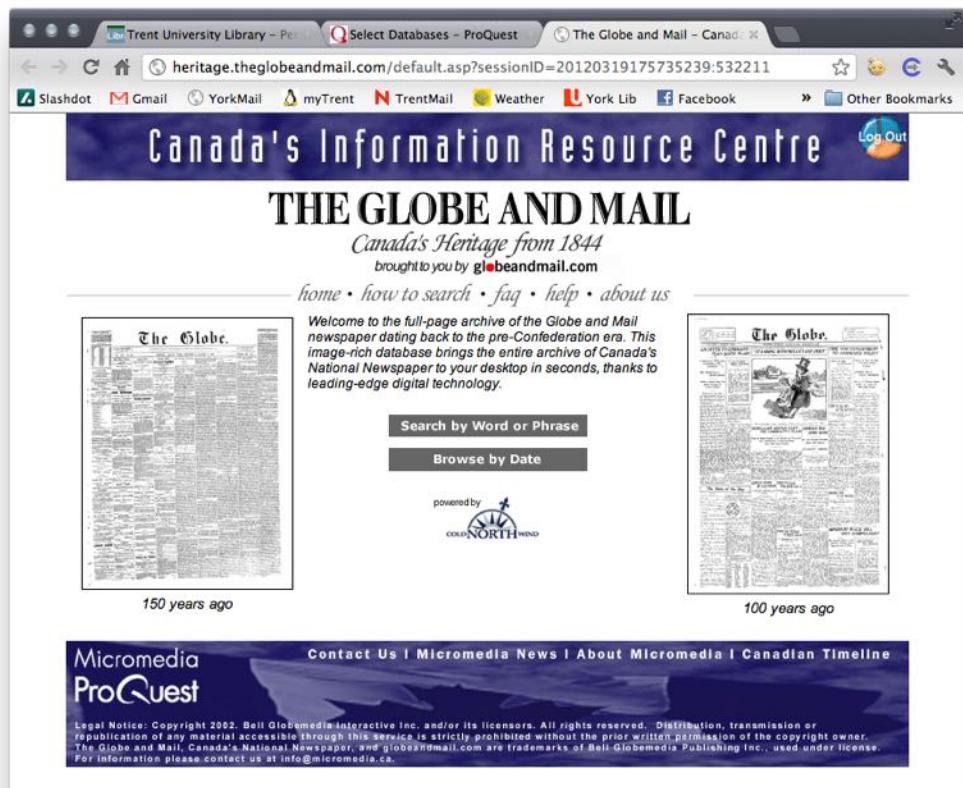
---

- To the best of my knowledge, the *Toronto Star* is the first newspaper in the world to be digitized in its entirety; the *Globe and Mail* was second and *New York Times* third.



## Anatomy of a Digitized Newspaper: *The Toronto Star*

- Paper of Record has troubles
  - ProQuest initially a big investor, then becomes a “Former” investor;
  - Partnership with LAC falls apart in 2003;
  - Google finances them and uses it as launchpad as part of the **Google News archive** (2008-2011)
    - Google gave publishers free access to their scanned files as part of its collapse
- Now accessed via **ProQuest** (which is itself an evolution of University Microforms, and is now sadly neglecting its microfilm collection)



(I couldn't find an old *Toronto Star* interface screenshot!)

# Some initial thoughts

- **OCR**
  - Developed initially for corporate discovery, struggles with newspapers
  - The original *Toronto Star* and *Globe and Mail* digitization was carried out with fairly rudimentary OCR, probably giving an 80-90% accuracy.
  - Even with cutting-edge OCR on pre-1950 documents
    - i.e. if you had 98% character accuracy – sounds great – but suddenly your word accuracy is somewhere between 95-98%.
  - Struggles with line hyphenation



**But more importantly...**

## Changes the relationship of a scholar to the newspaper

- As I showed in the Nugent lecture, skimming is difficult on these interfaces.
- Search + retrieval.
- Proprietary interfaces lock the data away. **In fact, this to me is the saddest consequence of this privatization of our cultural heritage.**

The screenshot shows a web browser window for the ProQuest Historical Newspapers: Toronto Star. The URL is [search-proquest-com.proxy.lib.uwaterloo.ca/hnptorontostar/publications](http://search-proquest-com.proxy.lib.uwaterloo.ca/hnptorontostar/publications). The page is titled "ProQuest Historical Newspapers: Toronto Star". It features a search bar with dropdown options for "In title" and a "Search" button. Below the search bar is a "Publication date" filter set to "1894 - 2010 (decades)" with an "Update" button. To the right, there are four publication entries:

Rank	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	Toronto, Ontario

## Virtual Records and Real History

Ronald W. Zweig

*Computer technology is changing the way people generate documents and create records. A growing proportion of transactions in the decision making process only exist digitally. Social and economic historians have been using numeric datasets as primary source material for years, but almost no attention has been paid to the impact of machine readable textual records on historical writing. This article considers the advantages and disadvantages for the historian of the shift from paper records to electronic documents, and suggests a number of approaches to historical research made possible by the new technology. Historians will have to deal with new sorts of 'documents' – records that only exist virtually and are integrally 'linked' to other documents and data sources. The concept of 'provenance' of sources will be transformed in the environment of electronic archives. It will be possible to trace the decision-making process in the wider context of other transactions taking place at a given time, allowing researchers to avoid the 'tunnel-vision' created when specific subjects are pursued in conventional archives. The task of tracing the decision-making process would be greatly simplified if electronically generated documents also included a record of document usage, as well as content. Finally, electronic records will help the researcher identify 'nodes' of policy-making, and the topics that occupied the decision-makers.*

'Government is every where antecedent to records, and letters seldom come in amongst a people till a long continuation of civil society has, by other more necessary arts, provided for their safety, ease, and plenty. And then they begin to look after the history of their founders, and search into their original, when they have outlined the memory of it. For it is with commonwealths, as with particular persons, they are commonly ignorant of their own births and infancies: and if they know any thing of

their original, they are beholden for it to the accidental records that others have kept of it.'<sup>1</sup> New technology is transforming the way people work, and the manner in which they generate records of the work they do. These records are preserved on different media from conventional records (magnetic and optical media as opposed to paper), and present unique problems of preservation and accessibility. Archivists must already deal with entirely different kinds of records from those they traditionally handle. This is of obvious importance to the historian, who will soon be faced with technical challenges that scholars have never previously faced. However, the revolution in office and archival practices goes far beyond the newness of the machinery which creates electronic records and of the media on which they are stored. The very nature of the 'documents' and the working environment in which they are created, are being transformed in a manner that has major implications for the methods of conventional historical research.

Social and economic historians faced the transformation of their source material years ago, and today very few of them would admit to being unfamiliar with the techniques of handling computerised numerical data. Political and diplomatic historians, however, have been able to delay dealing with the new archival sources because very few textual records

*Ronald Zweig is Senior Lecturer in Jewish History and Chairman of the Computing Committee of the Humanities Faculty of Tel Aviv University. He is author of *Britain and Palestine: the Second World War and German Reparations* and the *Jewish World* and editor of *David Ben-Gurion: Political Leadership in Israel* and the journal *Studies in Zionism*.*

# The Transformative Potential of Newspapers Has Been Lost

• *History and Computing*, vol. 4, no. 3 (1992)

• **Ronald W. Zweig, "Virtual Records and Real History"**

- "This situation is quickly changing as the first machine-readable textual records deposited in archives are being opened to research."
- Thinking about sophisticated research types: digital text would be good for "sophisticated search and retrieval," and keyword searching might be too overwhelming but "if they are combined with an understanding of linguistic equivalences, proximity and Boolean searches, and other techniques used in text retrieval, it will be possible to control the results of a search and to improve its quality."

# So, with 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.**



**Is there a better way?**

A screenshot of a web browser displaying the Internet Archive homepage. The address bar shows "archive.org". The header features the "INTERNET ARCHIVE" logo with a classical building icon, followed by navigation links: ABOUT, BLOG, PROJECTS, HELP, DONATE, CONTACT, JOBS, VOLUNTEER, and PEOPLE. On the right, there are "SIGN UP | LOG IN" and "UPLOAD" buttons. A central search bar contains the text "Search the history of over 525 billion web pages on the Internet." Below it is the "Wayback Machine" logo and a search input field with placeholder text "enter URL or keywords".

A screenshot of the Internet Archive homepage. On the left, there is a large black icon of a classical building with four columns. The main content area features a large text block: "Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more." Below this are several small icons with their corresponding statistics: 525B (books), 28M (movies), 6.5M (software), 15M (music), 2.2M (websites), 627K (images), 3.8M (audio), and 224K (documents). There is also a "992K" link. At the bottom of this section are a search bar with placeholder "Search" and a "GO" button, and a "Advanced Search" link. To the right, a sidebar titled "Announcements" lists recent news items: "Thank you Ubuntu and Linux Communities", "Internet Archive's Modern Book Collection Now Tops 2 Million Volumes", and "Community gathers for an online celebration of Michelson Cinema Research Library". A "More announcements" link is also present.

A screenshot of the Internet Archive homepage showing "Top Collections at the Archive". It displays four dark rectangular boxes with blurred images of historical documents, books, and artifacts. At the bottom of the page, a footer bar includes the text "Terms of Service (last updated 12/31/2014)".

# What to do with newspapers?

- Use them! But be **conscious**. A checklist:
  - Can you be self-reflective about why you chose the newspaper you did? (If it is because of accessibility, that's good; but say it)
  - Should you cite the way that the source has been mediated?
  - What potential impact might this process have had on what you were looking for? Are there supplemental keywords that you could look for? A version of a document on Archive.org that facilitates skimming? Some sampling or digging deeper?
  - Any holes that may have thrown off your analysis?

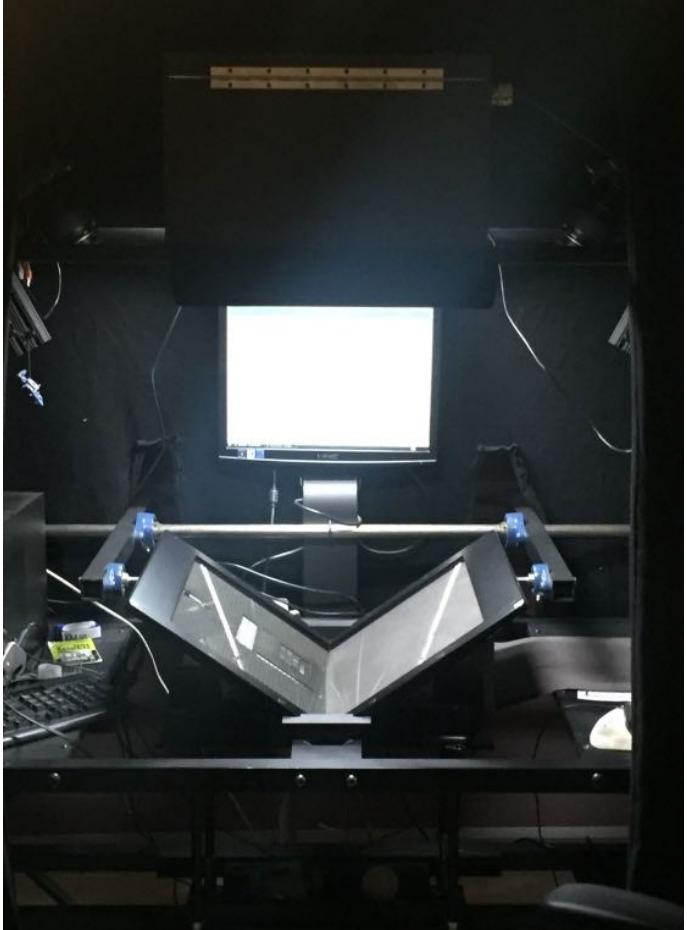
INTE



# What about digitized archives?

---

**Internet Archive Scanning Centre**, Robarts Library,  
University of Toronto



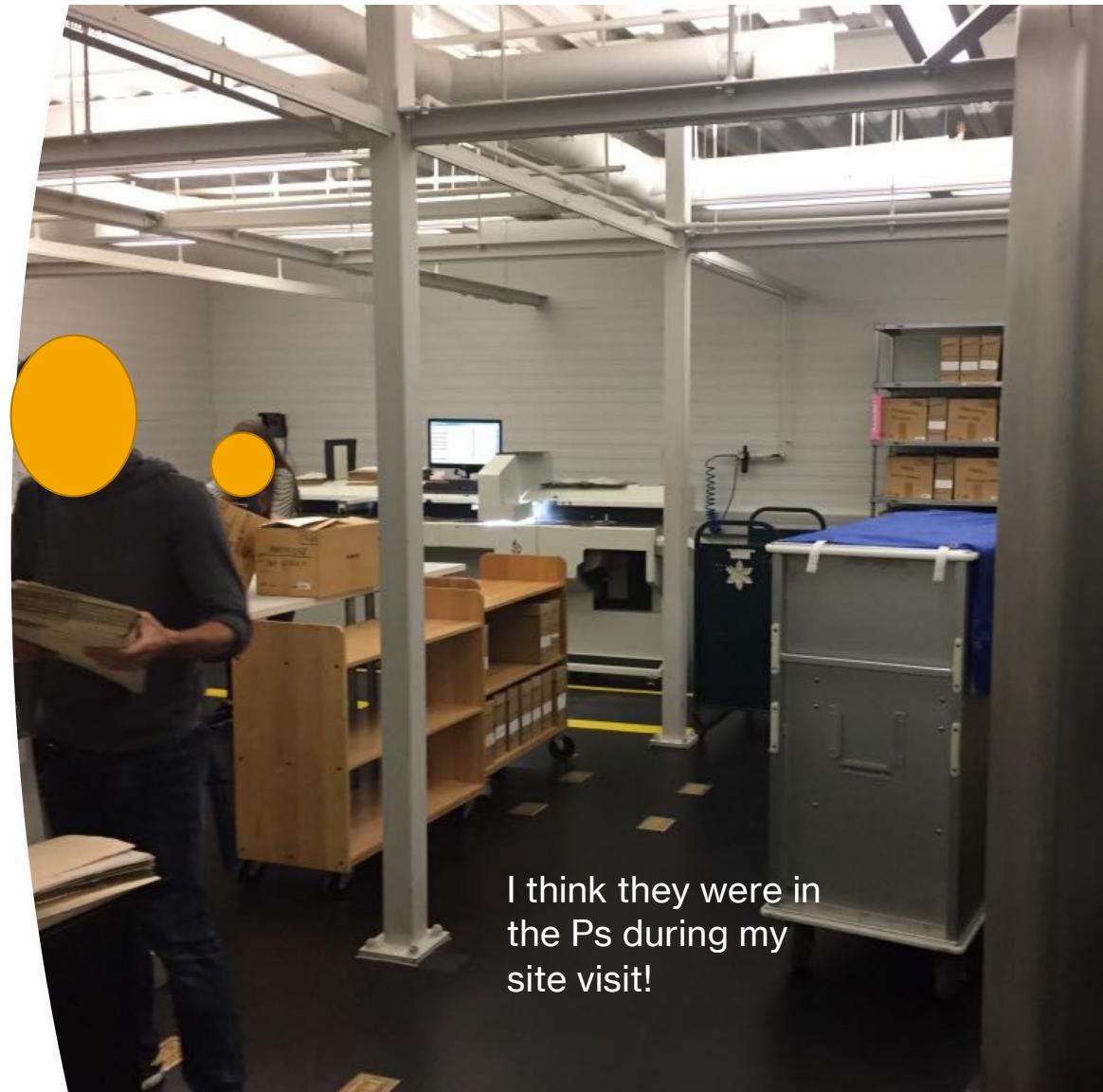
# Digitizing is Expensive

---

- **Several cents per page.** There is a range (all I can say, damn service roles).
- **Metadata is labour intensive:** but if you don't describe the material that you digitize it isn't too useful.
- **Donors may have mixed feelings:** it is one thing to provide access to people in a reading room; but decontextualized access via web search is a different kettle of fish.

# So, what gets digitized?

- In some cases, **microfilm** can be more readily digitized (i.e. Library and Archives Canada)
- **Popular collections**
- **Collections with particular synergy around institutional priorities** (i.e. Truth and Reconciliation, EDI, or other strategic institutions; this may or may not sync with researcher demand)
- **Anniversaries** (i.e. WW1 service records; major mega-project at LAC using repurposed cheque scanners)



I think they were in the Ps during my site visit!

## What gets digitized?

- Pulling our gaze back then from big institutions
  - Geographic unevenness in digitization;
  - Big institutions have the resources + overhead to do this (in most cases);
  - Smaller institutions won't have the resources, training, expertise to do this; OR will do it in an unsustainable manner due to uneven training/national infrastructure



## Increasingly being done via partnerships

- **CRKN/Canadiana.ca**  
(Monographs, Periodicals,  
Government Documents)
- **Internet Archive** (contacting  
with institutions; i.e. University  
of Waterloo)
- Discovery interfaces can  
often be better at scanning  
partner than on the original  
institution.

The screenshot shows a web browser window displaying the Internet Archive's search results for 'University of Waterloo'. The URL in the address bar is `archive.org/details/uwaterloo`. The page header includes the Internet Archive logo, navigation links for ABOUT, BLOG, PROJECTS, HELP, DONATE, CONTACT, JOBS, VOLUNTEER, and PEOPLE, and options to SIGN UP | LOG IN, UPLOAD, and Search. The main content area features the University of Waterloo logo and the text: 'In the heart of Waterloo Region, at the forefront of innovation, the University of Waterloo is home to world-changing research and inspired teaching. At the hub of a growing network of global partnerships, Waterloo will shape the future by building bridges with industry and between disciplines, institutions and communities.' Below this, there are two tabs: 'ABOUT' and 'COLLECTION'. The 'COLLECTION' tab is selected, showing 61 results. A sidebar on the left provides filtering options for 'Search this Collection' (Metadata or Text contents), 'Part Of' (Canadian Libraries), 'Media Type' (texts), and 'Year' (1995, 1994, 1993, 1992). The main content area displays four book covers from the collection, each with a title, author, and a small preview image. The titles include 'GUELPH AND WELLINGTON COUNTY: A Bibliography of Settlement and Development 1800-1900', 'WATERLOO COUNTY COUNCILLORS: A Collective Biography', 'CANADIAN INDUSTRY IN 1871', and 'FOUNDING FAMILIES OF WATERLOO TOWNSHIP, 1800-1850'. Each book entry includes a small image, the number of views (e.g., 1,733, 3,081, 220), and a rating star icon.

---

# What to do with digitized archives?

---

- Ask questions
  - Why was this digitized?
  - What wasn't digitized?
  - What biases might this introduce into my work?
- When doing “sideglances” (as Putnam puts it) reflect on what’s happening, and how you’re looking at the tip of an iceberg – just like somebody from another field would be if they only relied on digitized sources in *your* field.





**... and finally,  
digital  
photography.**

# Digital Photography in Archives

- You've read the piece, but my topline takeaways are:
  - We all **take** thousands of photographs;
  - It seemed obvious, but went "academic viral" when I wrote about it.
- Different histories will be written because we can't (a) do follow-ups; (b) we select the sources we're going to use at the very moment in our PhD (for example) when we know the least; and (c) we've started thinking about "collecting" and "writing it up/thinking about it" as even more discrete stages.

The screenshot shows a web browser displaying an article from The Atlantic. The header includes the site's logo, a search bar, and navigation links for 'Popular' and 'Latest'. The main title 'The Way We Write History Has Changed' is in large, bold, black font, with a subtitle 'A deep dive into an archive will never be the same.' Below the title is the author's name 'ALEXIS C. MADRIGAL' and the date 'JANUARY 21, 2020'. The article features a painting of an elderly man with a long white beard, wearing a red robe, sitting at a desk and reading a book. A caption below the image reads 'PRINT COLLECTOR / GETTY / FARNKNOT ARCHITECT / SHUTTERSTOCK / THE ATLANTIC'. To the right of the article is a sidebar with an advertisement for Databricks, featuring a thumbnail image of a document titled '8 Steps for a Developer to Learn Apache Spark with Delta Lake' and a 'DOWNLOAD NOW' button.

**TECHNOLOGY**

## The Way We Write History Has Changed

A deep dive into an archive will never be the same.

ALEXIS C. MADRIGAL JANUARY 21, 2020

PRINT COLLECTOR / GETTY / FARNKNOT ARCHITECT / SHUTTERSTOCK / THE ATLANTIC

History, as a discipline, comes out of the archive. The archive is not the library, but something else entirely. Libraries spread knowledge that's been compressed into books and other media. Archives are where collections of papers are stored, usually within a library's inner sanctum: Nathaniel Hawthorne's papers, say, at the New York Public Library. Or Record Group 31 at the National Archives—a set of Federal Housing Administration documents from the 1930s to the '70s. Usually, an archive contains materials from the people and institutions near it. So, the Silicon Valley Archives at Stanford contains everything from Atari's business plans to HP co-founder William Hewlett's correspondence.

**RECOMMENDED READING**



**Every stage of primary  
source collection has been  
transformed by the digital.**

# Let's discuss.

- What do you think we should do when thinking about **the mediating role of technology in databases?**
  - How do you know what's included? What's not?
  - What have you found?
- What do you think about the **mediating role of cameras in the archive?**
  - What was your experience?
  - Do you wish you'd taken more photos? Less?
  - What do you wish you knew then that you know now after working with them?
  - Would you share them?



# Secondary Sources

## Secondary Sources are Mediated Online As Well

- The HathiTrust Emergency Digital Library
  - What books are present?
  - What books in your field aren't?
  - What biases might we see?
- The holdings of Michigan/Illinois as the foundation
- **And, more importantly, copyright begins to rear its head.**

The screenshot shows the HathiTrust Digital Library website. At the top, there's a navigation bar with links for Home, About, Collections, Help, and Feedback. The HathiTrust logo is on the left, and a yellow 'LOG IN' button is on the right. A sidebar on the right says 'Want to get the most out of HathiTrust? Log in with your partner institution account to access the largest number of volumes and features.' It also has a link for guests. The main content area has a search bar with options for Full-text, Catalog, and Full view only. Below the search bar are links for Advanced full-text search, Advanced catalog search, and Search tips. A large orange circle on the left says 'The Library is Open! HathiTrust Response to COVID-19'. To the right are three boxes: 'BROWSE COLLECTIONS' (with a book icon), 'READ BOOKS ONLINE' (with a computer monitor icon), and 'DOWNLOAD BOOKS\* & CREATE COLLECTIONS' (with a lock icon). The bottom of the sidebar notes that guest login requires institutional login.

# Secondary Sources are Mediated Online

- The **Internet Archive** lends books based on having one copy in its possession; i.e. they have a “physical archive” of old books and periodicals, so you can check out a book under DRM conditions if they own it.
- During COVID-19, they allowed multiple loans without tying allocation to the physical holding, citing an “emergency.” Lawsuit ensued.



The screenshot shows a news article from The Nation. The title is "Publishers Are Taking the Internet to Court". The subtitle reads, "In a lawsuit against the Internet Archive, the largest corporations in publishing want to change what it means to own a book." The author is Maria Bustillos, and the date is September 10, 2020. Below the article are social media sharing icons for Facebook, Twitter, and Email. To the right of the article is a photograph of Brewster Kahle speaking at the Internet Archive offices in San Francisco in 2010.

Brewster Kahle at the Internet Archive offices in San Francisco, 2010. (Beatrice Murch / CC BY 2.0)

# Google Snippets Driving Citations?

- If I was a betting person, I would wager that the availability of a book via Amazon LookInside! and Google Books snippet view would have a positive impact on its citation impact – both amongst undergrads but also amongst faculty colleagues
- But we would never confess in a citation that we only read the snippet.

The screenshot shows a Google Books search results page for the query "Ian Milligan". The results are filtered to show books. On the left, there's a snippet of the book "History in the Age of Abundance?: How the Web Is Transforming Historical ...". The snippet includes the book cover, a star rating of 0 reviews, and a "Write review" link. Below the snippet are links for "About this book", "My library", "My History", and "Books on Google Play". A red "VIEW EBOOK" button is visible above the snippet. On the right, the first page of the book's introduction is displayed, with the title "INTRODUCTION" at the top. The text of the introduction discusses the collective cultural heritage and its challenges in the digital age for historians. At the bottom of the snippet area, it says "Pages displayed by permission of McGill-Queen's Press - MQUP. Copyright". The browser interface shows the address bar with "books.google.ca/books?id=qXCMDwAAQBAJ&printsec=bl", the search bar with "Ian Milligan", and various navigation buttons like back, forward, and search.

INTRODUCTION

Our collective cultural heritage, the legacy that we leave behind for the generations to come, faces a serious problem in the digital age. We used to, as a rule, forget. Now we have the power of recall and retrieval at a scale that will decisively change how our society remembers. For historians, professionals who interpret and bring shape to narratives of the past, this is a dramatic shift. The digital age brings with it great power: the prospect of a more democratic history and of more voices included in the historical record, a realization of the social historian's dream. Yet it also brings significant challenges: what does it mean to write histories with born-digital sources – from websites written in the mid-1990s to tweets posted today? How can we be ready, from a technical perspective as well as from a social or ethical one, to use the web as a historical source – as an archive? Historians with the training and resources are about to have far more primary sources, and the ability to process them, at their fingertips. What will this all mean for our understanding of the past? How can these sources be used responsibly? Finally, if historians cannot rise to the moment, what does this mean for the future of our profession?

The problem can be summed up in something as innocuous as a personal homepage, hosted on the free GeoCities.com service. GeoCities.com, founded in 1994, provided free websites to anybody who wanted to create one. A user would visit GeoCities.com, enter an email address, and receive a free megabyte (later two, then ten) to stake an individual space on the burgeoning Information Superhighway. These sites took many shapes and sizes: a Buffy the Vampire Slayer fan site, a celebration of a favourite sports team, a family tree, even a young child's tribute to



**Many of the factors driving  
online primary sources are  
also driving secondary ones.**

# Sharing our Research?



social media



Facebook



Twitter



Instagram

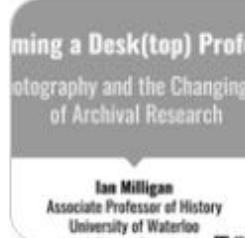
# As so much is digital...

- ... so too is the impulse to share.
- From conference presentations
  - **Example of my 2020 AHA Presentation.** A Sunday morning in New York City, after the Saturday night parties, a decently full but small-sized conference room. (don't ask me about my 2018 one where panelists = audience size)
    - Great reception.
    - Decided to share my slide deck.
    - Went viral.
    - *Atlantic* article.



Dan Cohen   
@dancohen

Eye-opening survey by [@ianmilligan1](#) reveals the changing practice of historians: Over 90% now use research trips to take photos of archival documents that they will examine when they get home; on average, they take 1,000 photos; 40% take over 2,000 photos.



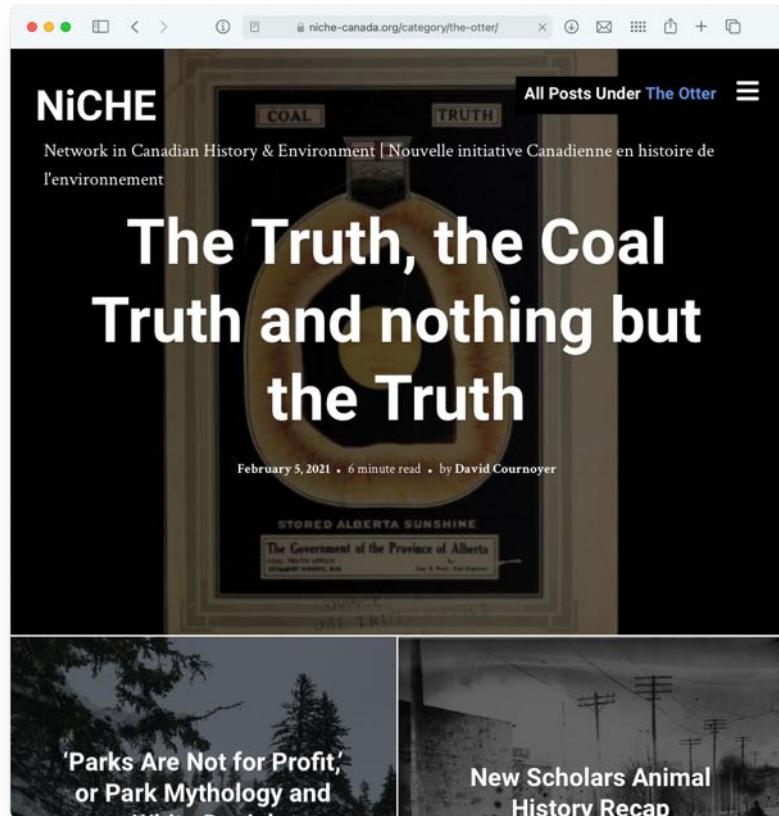
AHA 2020 - Becoming a Desk(top) Profession  
Becoming a Desk(top) Profession Digital Photography and the Changing Landscape of Archival Research [Read title + ..](#)  
 [docs.google.com](#)

10:08 AM · Jan 6, 2020 · Micro.blog

**351** Retweets   **183** Quote Tweets   **820** Likes

## ... pros/cons

- Great to have exposure!
- That said, most people in those 183 Quote Tweets and seemingly hundreds of other discussions/replies I would wager did not read the slide deck
  - i.e. I was accused of being hostile to digital technology, not caring about children (when like a key point is “hey, this lets me spend time with my two little kids”), etc. etc. It was fun but lost some respect for click-bait colleagues.



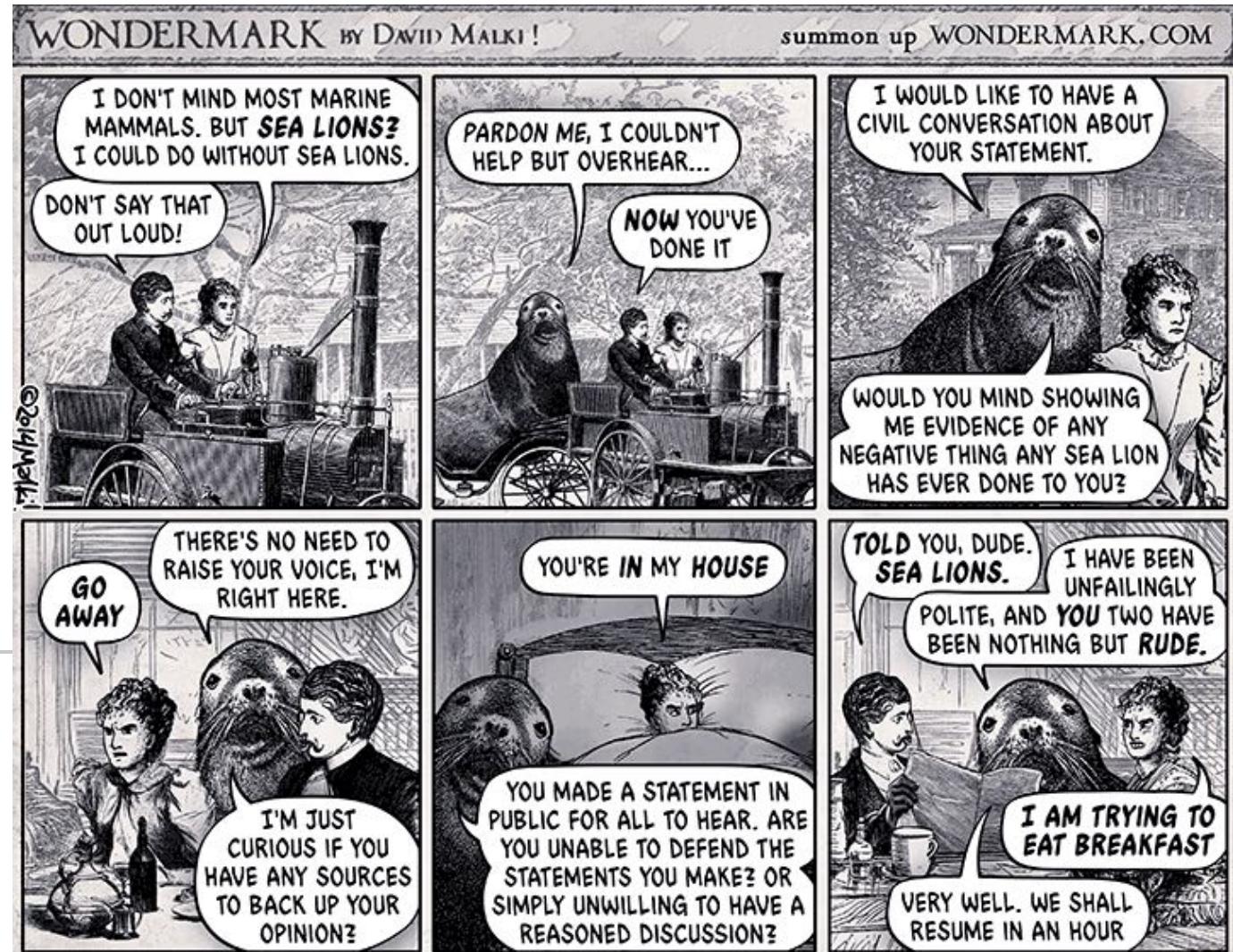
## ... similarly, with sharing thoughts!

- ActiveHistory, personal blogs, CHA blog, NiCHE/The Otter, etc. all places to reach a larger audience.
  - But as, we discussed last week, there are costs – comment sections, etc.
- In an ideal world, I sometimes tweet -> resonates -> blog post -> resonates -> put together a journal article (both of my *CHR* papers actually came out of that workflow)
- People are meaner today, though, and less charitable. (I had some stupid takes in my grad student days, but I don't think people cared as much)

# My own advice, for what it's worth...

- **A personal page is increasingly a necessity**, i.e. for book review editors, etc. to help find you for opportunities.
- **Institutional pages are good**, but they're transient and you might want to build up your own “brand” (yeah yeah)
- Twitter is good, but beware:
  - Don't make fun of people, because I am amazed how many professors “lurk” on Twitter but don't have accounts.
  - Jokes that work within in-groups don't always work within out-groups.

# Sea lion

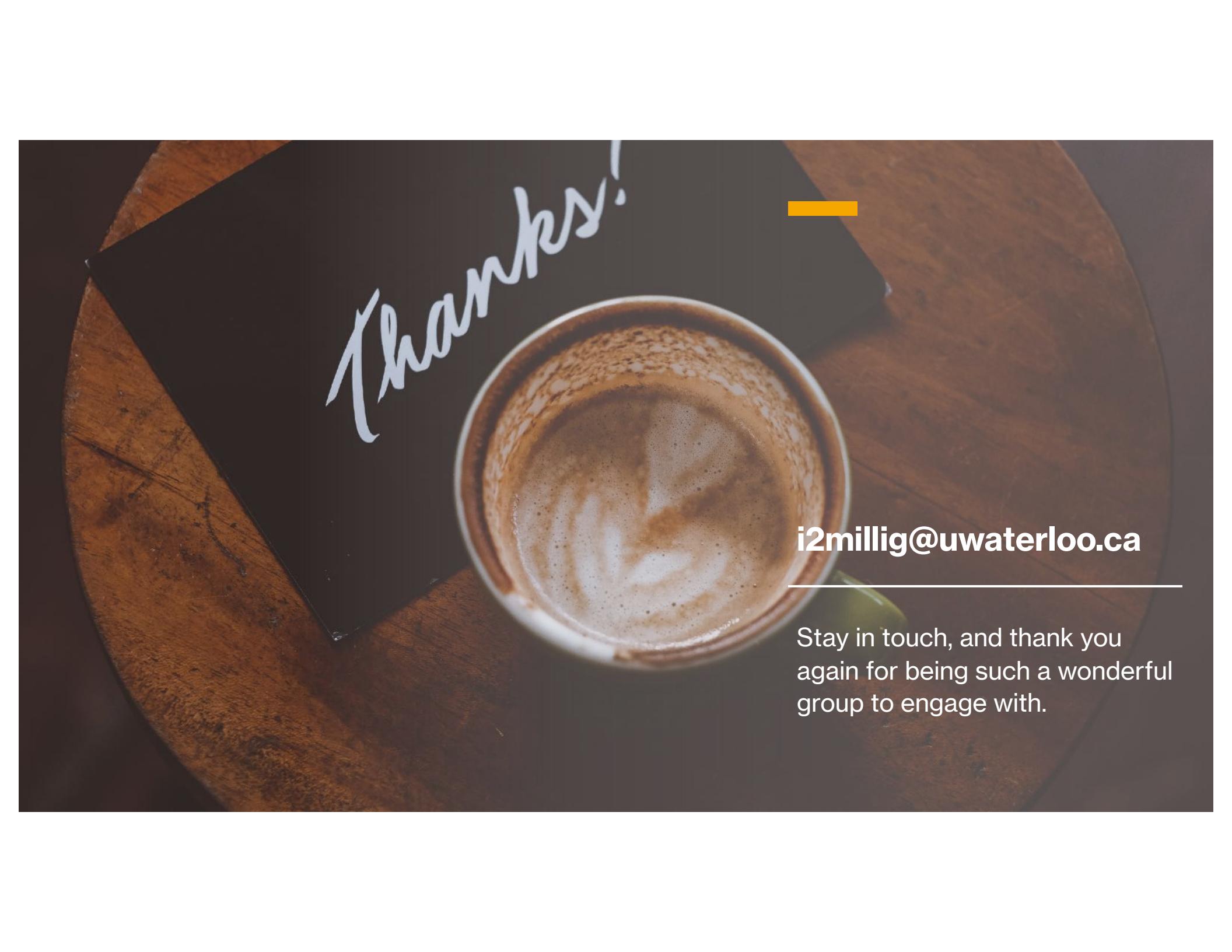


# Some discussion

- What have your experiences been engaging online?
- Do you share your research? Why or why not?
- What do we gain from sharing? What do we lose?
- What would you recommend to your peers about things that have worked?
- Things that we should stay away from?

# Conclusions

- We are all digital now: in how we do our primary research, our secondary research, and disseminate it all online.
- I sometimes think of:
  - **Digital History (capital D, capital H)** as the subfield we spoke about last week
  - And **digital history (lowercase d, lowercase h)** as the transformation that we are seeing us all.
- It has been such a pleasure to be here with you all, and look forward to staying in touch.



Thanks!

[i2millig@uwaterloo.ca](mailto:i2millig@uwaterloo.ca)

---

Stay in touch, and thank you  
again for being such a wonderful  
group to engage with.