Journal: CHR; Volume 94; Issue: 4
DOI: 10.3138/CHR.694

Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010

Ian Milligan

Abstract

[Abstract to come]

Keywords: [to come]

It all seems so orderly, advanced, and comprehensive. Instead of firing up the microfilm reader to navigate the *Globe and Mail* or *Toronto Star*, one needs only to log into online newspaper databases through a library portal. A keyword search for a particular event, person, or cultural phenomenon brings up a list of research findings. While date-by-date searching is also available, it seems clunky and slow; keyword searching, however, offers something new, something potentially transformative. Each result is broken down by date, newspaper page number, the section it appears in, and a further click brings you to the entire page, scanned at a decently high resolution, search terms highlighted for convenience. The surrounding context of the page, advertisements, and the original layout are all preserved. Previously impossible or implausible research projects can now be approached, especially when they involve wide swaths of social or cultural terrain.

Researchers cite what they find online. This is a problem, as this research process is built upon an often-misunderstood foundation. The problem matters because of the sheer increase in online-source citations. We can see this if we compare the ways historical newspapers were used before and after the introduction of two significant databases. Some examples are illustrative. In 1998, a year with 67 Canadian history dissertations in the ProQuest dissertation database, the *Toronto Star* appeared 74 times in that data set; by 2010, it appeared 753 times in a slightly larger data set of 69 dissertations. Controlled for sample size, this is a remarkable 991 per cent increase. The *Globe and Mail* saw similar growth: 58 to 708 (a 926 per cent increase). The *Montreal Gazette*, however, remained relatively stagnant (136 to 162, or 16 per cent increase), while the *Toronto Telegram* decreased slightly (31 to 19, or 72 per cent decrease). Neither is available through an online database. This is not simply confined to dissertations; a survey of articles published in the *Canadian Historical Review* demonstrates similar growth: the *Globe and Mail* went from being rarely cited between 1997 and 2002 to being by far the most cited newspaper between 2005 to 2011. What does this shift mean?

Canadianists have remained largely ignorant of the impact the two newspaper databases in particular have had on our profession. We are witnessing the application of

commercial optical character recognition (OCR) technology to our work, a process that takes an image, recognizes shapes that are in the forms of letters, and writes the output in plain text. These algorithms were originally and primarily designed for the efficient digitization of reams of corporate and legal documents, conventionally formatted. Applying these tools, initially designed for specific commercial applications, to historical documents yields mixed results.

The database and its interface offer seemingly complete, ordered access to historical information, a problematic promise because there is a lack of methodological reflection about how these databases work. It is an illusionary order. With other source materials, historians are relatively transparent about their research practices. Historical inquiry is based largely upon empirical research: archives, newspapers, oral interviews, quantitative research, and so on. Professional historical standards call for transparency: how were interviews conducted? How were archives produced? What historical evidence underpins our arguments?[1] We need to apply similar questions to how Canadian historians use databases. Only then can we begin to move from our individual silos of knowledge and practice toward a collaborative approach to these powerful tools.

In this article, I draw upon a comprehensive corpus of every Canadian history dissertation, made available through ProQuest and assigned the "Canadian History" subject code between 1997 and 2010, complemented by a secondary corpus of all *Canadian Historical Review* scholarly articles published during the same time. I make two arguments. First, online historical databases have profoundly reshaped the foundation upon which Canadian historiography is constructed. In a shift that is rarely – if ever – made explicit, Canadian historians have overwhelmingly embraced these online tools without explicit or even, it appears, implicit recognition. Second, search engines necessarily skew our research, as do Google, electronic books, and the *Toronto Star* online.[2] Yet the issues of poor OCR in these databases make this a very pressing and significant issue. A critical methodology is needed if historians are to use these tools responsibly.

Canada's Heritage Online: The Development of Pages of the Past and Canada's Heritage since 1844

The Toronto Star: Pages of the Past and the Globe and Mail: Canada's Heritage Since 1844 databases are significant. That these two newspapers were selected for digitization at an early date is not surprising: today the *Toronto Star* is Canada's most widely read and circulated newspaper, and the *Globe and Mail* holds second place in English Canada. Alongside the *Toronto Star* and the *Globe and Mail* was Paper of Record, a collection of smaller newspapers partially digitized and put online. There are

other online databases as well, especially (perhaps counter-intuitively, given the time frame) for early Canadianists. While the twentieth century has more potentially digitizable sources, much of this is mired in copyright and has thus run into digitization hurdles. I chose the *Globe* and *Star* as case studies for three main reasons: first, my data demonstrate that the *Star* and the *Globe* have become the two newspaper databases extensively used to the detriment of all others. Second, given overall trends in Canadian historiography, which have tended to favour later periods of study, we can have a much greater sample size: any dissertation studying the mid- to late nineteenth century onwards could use these two databases. This provides us with the large data set needed to assess overall trends in the use of online databases. Third, and just as importantly, these two databases give us ten years of data to analyze.

That said, an understanding some of the limitations is needed by users of all digitized primary sources, not just those of the *Globe* and the *Star*. Early Canadiana Online, for example, uses a similar OCR algorithm to render its full-text holdings searchable. Error rates are not easily accessible on their homepage. This foundational process needs to be first and foremost in the minds of people who use these databases. Indeed, if the *Globe and Mail* and *Toronto Star* present difficulties for OCR routines with their fairly well typeset pages, as will be discussed later, government documents and pre-twentieth-century holdings pose added challenges. Even if contemporary, cutting-edge digitization techniques were used, the issues that I outline here arise with other newspapers and typeset sources. Technology can do great things – I myself identify as a digital historian – but we are not yet at the "silver bullet" stage for easy, quick, and fully searchable digitization.

Databases such as Pages of the Past and Canada's Heritage since 1844 are part of a broader phenomenon of using established or emerging digital technologies for historical and humanistic research. Self-conscious scholarly communities have emerged in new and emerging subdisciplines, concretely theorizing about their impact, creating new research tools, and grappling with the implications of this new turn in humanities scholarship. These fall under the twin headings: the digital humanities, the broader exploration of how technology can be integrated into traditional scholarly activities and the creation of new forms of scholarship and media; and digital history, the application of digital methodologies and media to historical questions.[3] These sub-fields are beginning to emerge as established disciplinary realms with their own conferences, journals, and developing communities of scholars.

The emergence of these two sub-fields obscures the fact that many historians are already unwitting digital historians, in the sense that they are applying technology to

their historical work through database software and online source repositories. Yet these historians are often uncritical digital historians (even if they are, in other respects, rigorous in their professional practice). If we are all, or at least most of us are, digital historians, we need to subject our work to digital methodological criticism. This article aims to open up this conversation. If those of us who use online databases can begin to conceive of ourselves as digital historians, to some degree, we are on track to becoming good, reflective users of digital source repositories. If we are all to be digital historians, let us make sure that we are good ones.

The Quantitative Impact of Newspaper Databases on Canadian Historiography

How are these sources used in Canadian scholarship, and to what degree can we see a shift with the introduction of comprehensive newspaper databases? For this, I undertook a comprehensive review of English-language Canadian history dissertations. Dissertations are significant both as a symbolic entry point into the profession, and as the origins of many scholarly monographs. While representing the voice of the scholar, they also reflect the values and input of an examining committee and an external examiner as a check and balance to ensure degree and program integrity, as well as oversight from departmental and administrative representatives. Dissertations are scrutinized on several levels and represent original and striking contributions to historical knowledge.

Dissertations are also an extensive, comprehensive, and systematic data set, allowing for the useful comparison of large amounts of information. The more data that we can process, the more confidence we can have that we are seeing a trend and not simply the effect of a few outliers. In fact, the outlier issue is important, and will be explored later in this article. The larger the data set, in general, the more comfortable I am with my analysis. In 2010, for example, there were 69 dissertations, creating a data set of 24,750 pages. The sample size of the data was relatively constant over time, although occasionally larger: 1997 saw 87 dissertations with 33,382 pages; 2008, with only 63 dissertations (an outlier in terms of length) saw 51,862 pages (the average length evidently differed between those two years). In total, Canadian historical dissertations between 1997 and 2010 comprised 444,708 pages. One person cannot reasonably read this amount of data (not, at least, if that person wants do anything else as well). Given the size of the data set, which I consider an unparalleled comprehensive database of Canadian historiography, I approach the quantitative impact of newspaper databases through an automated approach.[4] While comparative information on the impact of databases on other national historiographies would be interesting, none is available at this time. Given differing national-level digitization schemes in countries like the United

States, the United Kingdom, and Australia, such comparisons would be a fruitful next step.

At this point, then, it is worth pausing to discuss my conceptual and technological methodology.[5] Dissertations present several challenges as a data set, both conceptual and technical. Conceptually, the issue emerges of what constitutes a "Canadian history" dissertation. The ProQuest subject heading "Canadian history" captures dissertations defended in traditional history departments, as well as closely related works in departments of geography and sociology. It also occasionally includes more contemporary political and environmental studies. For the purposes of a reproducible database, the decision was made to adhere to ProQuest's classification system (which is based upon author self-declaration). The years from 1997 have the most comprehensive coverage. The timeframe of 1997 to 2010 was selected, resulting in 1,025 dissertations.

On the technical front, while electronic copies are made available to paying customers through the ProQuest dissertation database, these documents are rendered inaccessible to machine reading. The downloaded PDF files do not contain the normal elements of a healthy, robust PDF file: they are missing their cross-reference tables, as well as document trailers (normal components of a PDF file). In short, this means that while an end-user can view the image file, there is no text layer that would allow one to simply copy and paste textual characters. Reconstructing the text layer is difficult because of the nature of the files. Thankfully, while this is an issue with the ProQuest files, it is not so across all databases – most databases allow you to download PDF files with the text layer intact.

Each PDF was manually downloaded (ProQuest forbids the machine-reading of websites, which is the preferable way to execute a large-scale downloading project).[6] Specific file numbers were assigned and arranged by year. In order to automate the counting of citations – a forbidding prospect with such a large data set – each PDF was then "burst" into a series of individual single-page files (i.e., a 400-page document becomes 400 individual files). Each was then converted into a JPEG image file.[7] Finally, using Tesseract 3.0, each page had a text layer re-instituted on it. Tesseract is an open-source OCR program that opens an image, examines the arrangement of shapes and lines, determines what textual characters these shapes represent, and outputs the results to a text file. This made the project somewhat ironic: writing an article on the issues of OCR while adding that level of error to my own calculations. I admit that my calculations may be off by a small fraction; however, Tesseract 3.0 has the highest accuracy level of any similar program when dealing with formally typewritten and laid out documents.[8] Furthermore, the shift that, I argue, we see in the data is very substantial, far beyond OCR

error rates. This was a time-consuming process: each year took between three and four days of continuous twenty-four-hour computing.[9] For further reference, all of my source code is available online.[10]

With the ensuing plain text database, now fully searchable, I wrote a program that would count incidences of specific phrases: *Globe and Mail, Toronto Star, Ottawa Citizen*, and so forth.[11] This revealed information on how many dissertations used a given source, and how many times a source was referred to or referenced in a given dissertation. This allowed me to control for any outliers and to review quickly the keyword-in-context (in order to see if *Ottawa Citizen* was referring to the newspaper, for example, and not just somebody who resided in Canada's capital). While not a perfect methodology, apart from manually counting each and every citation (over hundreds of thousands of pages), it gives a robust sense of how often sources were used.

Approaching such a data set required some statistical awareness and finessing. The most important proviso to make is that each year will have several outliers. A dissertation may make extensive use of media sources (a comprehensive survey of newspaper coverage of the First World War, for example). Just one such source may throw the entire count off. An example can be found in my 2008 data set, where one dissertation accounts for more than half of the *Globe and Mail* mentions.[12] This is interesting, but we do not want it to unduly influence our argument concerning overall trends. Dealing with disparate data and outliers is a complicated undertaking, much studied by statisticians. One of the more robust measures to deal with this is a trimmed mean. On the basis of the shape of these data, and upon a review of the literature, I have elected to remove 10 per cent of outlying datapoints.[13] Otherwise, just one dissertation had the potential to distort our information. We want to see the trend.

My initial research hypothesis was that the introduction of databases would see a marked, if not overwhelming, increase in the use of the *Globe and Mail* and the *Toronto Star.* How long would it take for this increase to appear? Dissertations have a lengthy research and writing process. While time-to-completion data are notoriously difficult to obtain, as are most post-secondary education metrics, it is rare to see a dissertation completed in less than three years. The introduction of research tools would probably see some impact if they appeared mid-research, with their impact presumably becoming less apparent the further the doctoral candidate advanced through his or her writing stage. I thus hypothesized that I would probably see effects, if any, after three years following the 2002 introduction of the databases – therefore, starting in the 2005 ProQuest dissertation year. The chart is thus coloured in three bands: white to indicate the pre–Pages of the Past and Canada's Heritage Online; light grey when the databases were available but we

would not presumably see much impact in registered dissertations; and dark grey when I anticipated the impact – if any – would be felt.

For my first figure, I took the trimmed means of citation counts for the years of 1997 to 2010, laid the backdrop of the colours as noted above, and charted five newspapers along a simple line chart. The other newspapers, the *Toronto Telegram, Ottawa Citizen*, and *Montreal Gazette*, were chosen as controls for various reasons: the *Telegram* complements the *Star* well because it is an important regional newspaper, and the other two are similarly highly cited newspapers. While the *Citizen* and *Gazette* became available online to varying degrees, they did not have the high-profile impact of the two databases studied in this article.

The findings in a simple line chart of trimmed mean (10 per cent) frequencies exceeded my expectations (see figure 1).

FIGURE 1: Average number of newspaper appearances per Canadian history dissertations, with outliers removed, 1997–2010

The *Globe and Mail* and the *Toronto Star* were cited with considerably more frequency after the database was introduced, showing an initial burst but then continued high-citation levels. Before digitization, a newspaper like the *Ottawa Citizen* was roughly equivalent in historical usage to the *Toronto Star*, as one might expect, given their relative prominence in Canadian history. After the *Star* was digitized and made available, however, it became far more prominent. Given the wide range of this data set, these are robust findings. Remember that for each year we are looking at an average of approximately seventy dissertations, taken from the ProQuest subject heading dissertation database, and thus have a large enough data set to see reliable trends.

Of course, there are varying numbers of dissertations per year. If we take each year and divide by the number of dissertations, we see that the ensuing graph is nearly identical to the first one (see figure 2).

Journal: CHR; Volume 94; Issue: 4
DOI: 10.3138/CHR.694

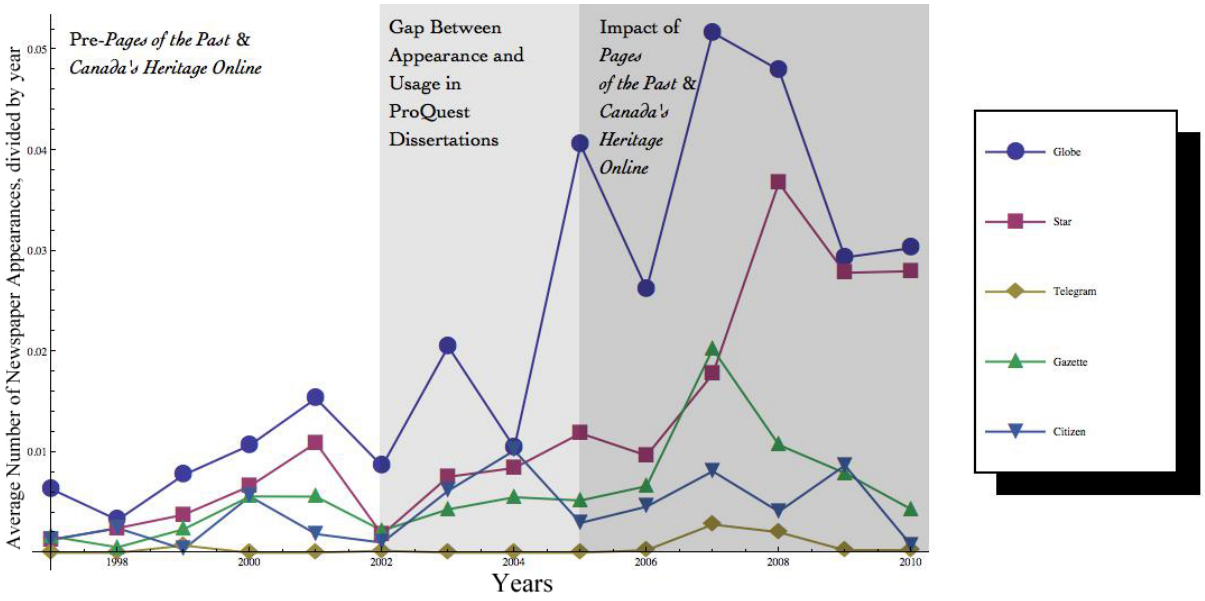


FIGURE 2: Average number of newspaper appearances in Canadian history dissertations, divided by year to normalize results, with outliers removed, 1997–2010

Again, the same trend is apparent: even when the number of dissertations is introduced as a control factor, we see the same overall rise and fall of citation counts.

If we inquire further into how these newspapers are being used, we see that more researchers are citing them, but that the most critical increases are a result of the fact that dissertations published during the post-database period are using them more than ever. Figure 3 provides a visualization of how many dissertations have at least one reference to any of the newspapers:

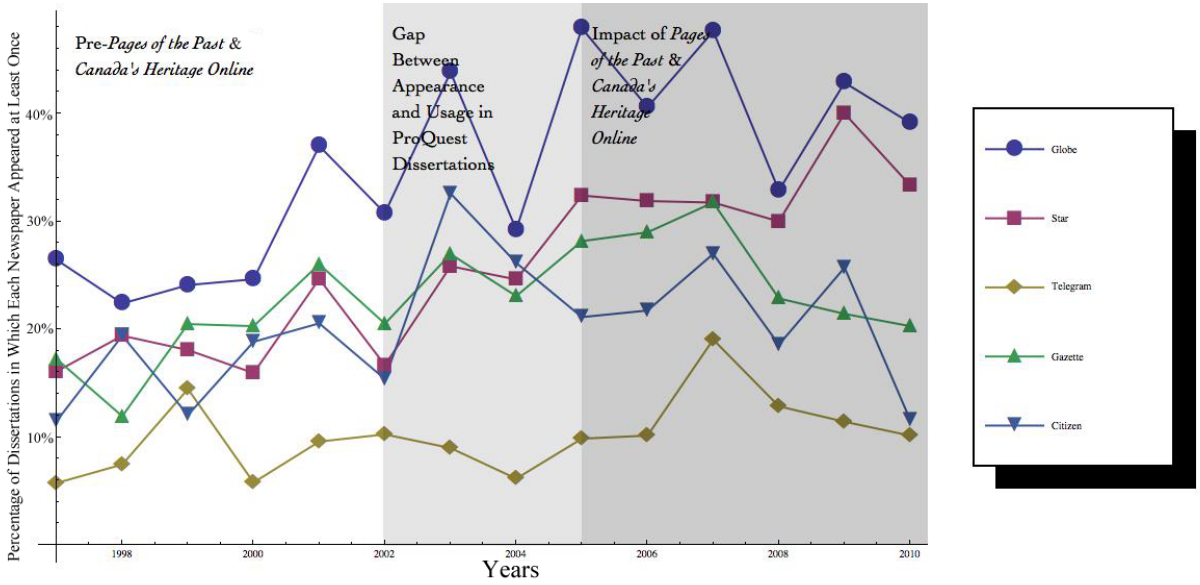


FIGURE 3: Percentage of dissertations in which each newspaper appeared at least once, 1997–2010

There has been a fairly pronounced statistical shift in the number of dissertations that have used the *Globe* (between 1997 and 2002, an average of 28 per cent of dissertations drew on the *Globe* at least once; between 2005 and 2010, the average was 42 per cent). The *Star* has similarly seen an increase (from 18.5 per cent between 1997 and 2002 to 33.0 per cent between 2005 and 2010). Other newspapers did not see similar shifts. Newspapers readily accessible online are being used more frequently. They are also being used in a more sustained manner, as demonstrated in the following set of visualizations.

Above, we can see that the introduction of databases had a profound impact on the use of newspapers in dissertations. Even if outliers are incorporated, we can see the increasing densification of citations to the *Toronto Star*. Again, the dark grey represents the period of time between 2005 and 2010 when we can expect to see the impact of the databases; the light grey the three-year gap that one expects in the lengthy dissertation process, and the white the pre-database period. Consider figure 4.

Journal: CHR; Volume 94; Issue: 4
DOI: 10.3138/CHR.694

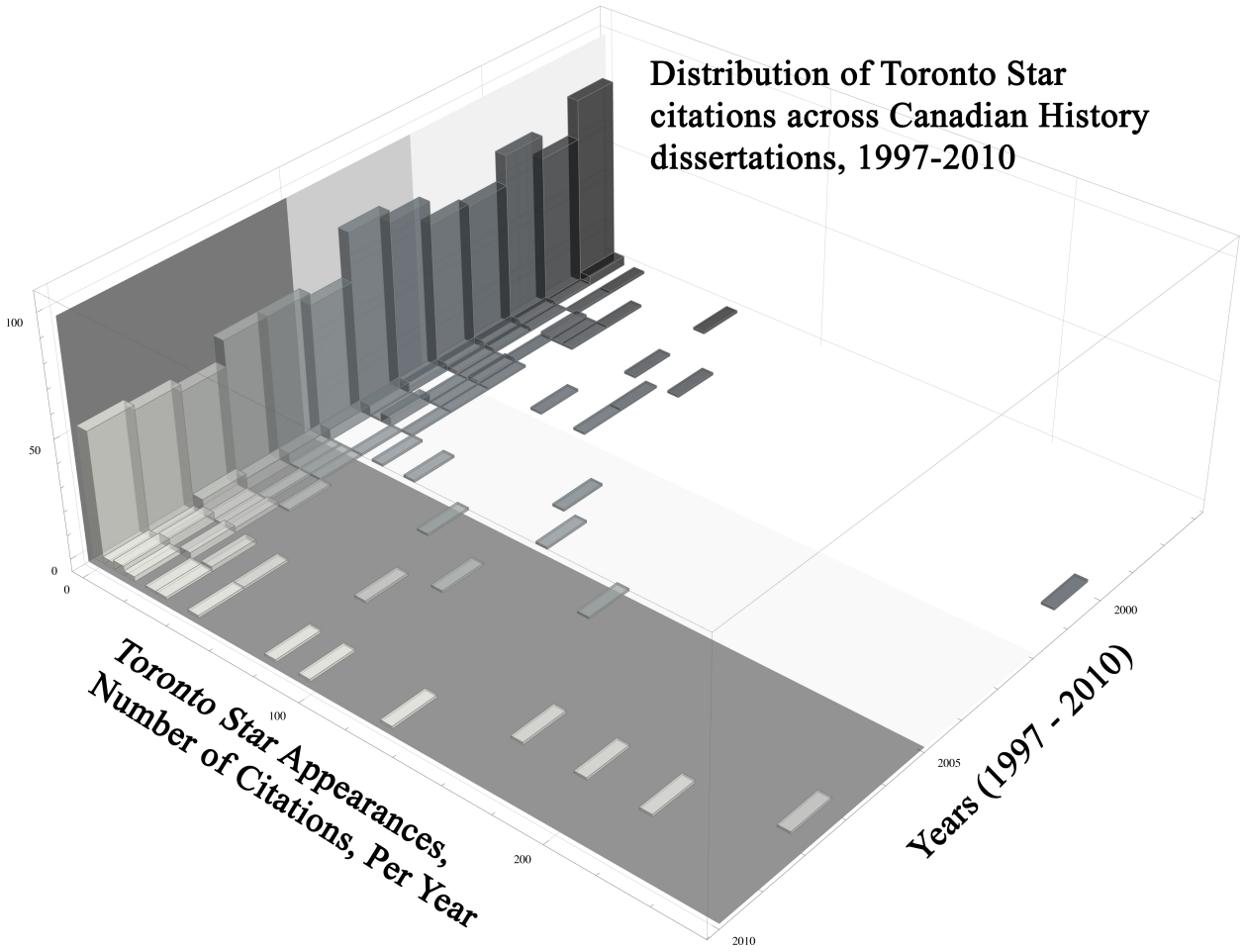


FIGURE 4: A 3D histogram *of Toronto Star* citations. Each bar represents a "bucket" into which the numbers of citations are placed. This way, we can plot outliers and see how data are distributed. Note that this graph includes outliers that were removed from figures 1 and 2.

In the above visualization, we see the distribution of *Toronto Star* citations from 1997 through 2010. Note that the decreasing number of dissertations that have no citations (the ones right at the back), and the increasing number of dissertations with growing numbers of *Star* citations (they creep out away from 0 and toward the outlier at 1,000). The impact is undeniable. Almost as if on cue, as soon as we reach the 2005 calendar year, the *Toronto Star* is cited far more often. Similar findings appear with the *Globe and Mail*. Compare this to a similar visualization of the *Montreal Gazette* (figure 5).

Journal: CHR; Volume 94; Issue: 4
DOI: 10.3138/CHR.694

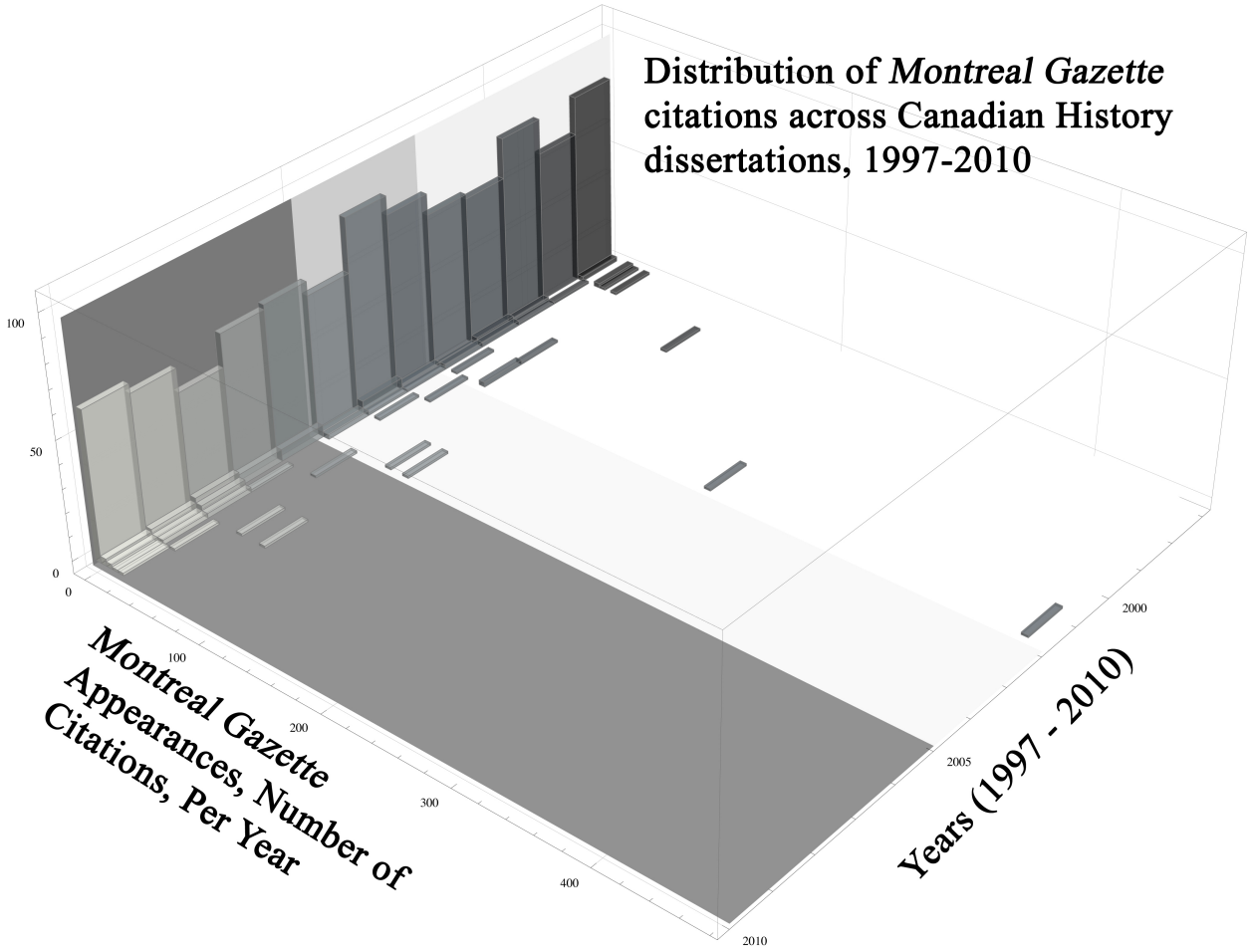


Figure 5: A 3D histogram of *Montreal Gazette* citations

While there have been some shifts, the change is minimal; indeed, the outliers are less pronounced as time goes on. Similar findings appear for the *Citizen* and the *Telegram*. The impact of online newspaper databases on Canadian history dissertations is undeniable: spread across such a large data set, with outliers controlled, we see a discernible shift toward the *Globe and Mail* and the *Toronto Star*, to the detriment of other sources. Can we see the same behaviour elsewhere?

The *Canadian Historical Review* presents an interesting comparison. Whereas dissertations represent the beginnings of historical work, the *CHR* is a prestigious venue with contributions from historians of all levels. A similar methodology was employed: all articles published between 1997 and 2011 were downloaded and word frequency calculated. Using the same three colours as before (white for pre-database, light grey for the interregnum, and dark grey for the impact period), the results are displayed in figure 6.

Journal: CHR; Volume 94; Issue: 4
DOI: 10.3138/CHR.694

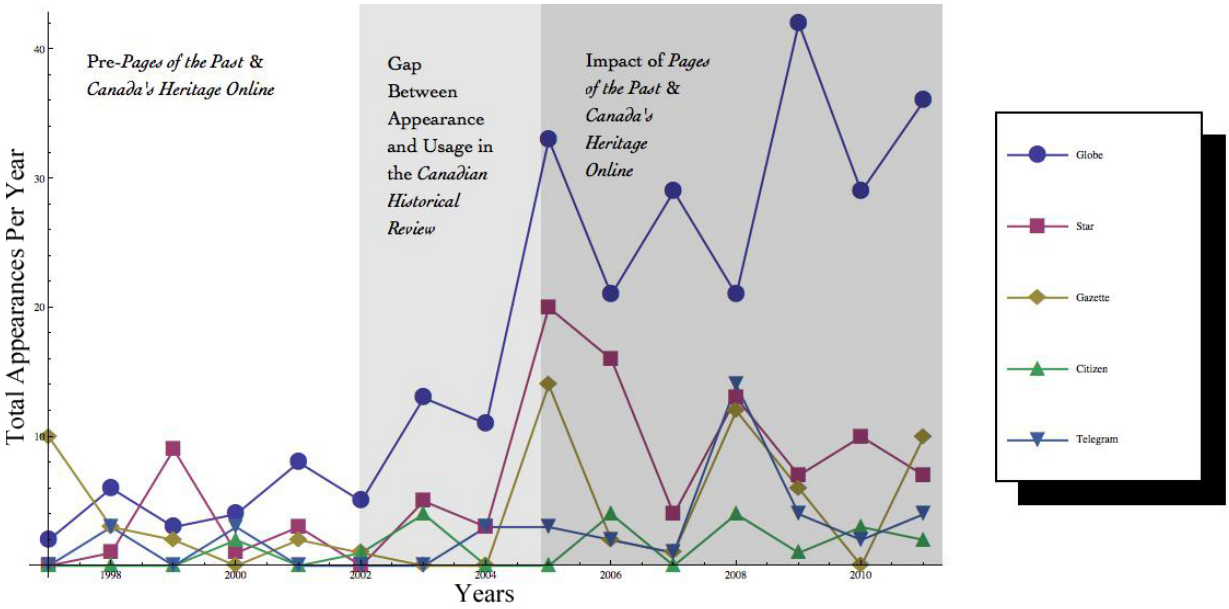


FIGURE 6: Average number of newspaper appearances in *Canadian Historical Review*, all articles per year, 1997–2011

The impact of the *Toronto Star*'s Pages of the Past is not as big as the *Globe and Mail*'s Heritage Online, which befits the national breadth of the *CHR*. For the *Globe*, however, we see a marked and sustained increase in citations in the post-2002 period. The impact of the database begins to manifest itself vividly by 2005. What is also notable, however, is the general increase in newspaper citations. There is a sustained move toward slightly more newspaper research, but this is beyond the scope of this paper. Usage of the *Globe* increases by an order of magnitude more than any of the other newspapers studied.

What effects have these changes had on the topics, time periods, and regions studied by Canadian historians? The small size of the *CHR* database led me to focus on dissertations when addressing this question, as we could reasonably infer some long-term trends in the larger dissertation data set. A major issue, of course, would be establishing causation rather than correlation; the reason to study a given time period or topic is affected by far more than simply the availability of digitized sources, although it would be reasonable to infer that it is one factor among many. The first step was to take a case study approach of dissertations in four years (1997, 1998, 2009, and 2010) to see if those that used newspapers diverged notably in topics. The results were at best inconclusive. On the face of it, topics at the macro level do not seem unduly affected by digitization. However, if we used computational techniques, could deeper, more subtle shifts be detected in the historiography?

Journal: CHR; Volume 94; Issue: 4
DOI: 10.3138/CHR.694

To establish the broad contours, I began by several simple counting measures: the relative frequency of provincial/territorial names, major city names, and dates in the texts. These results do not show a shift to the same degree as newspaper citations. While there is fluctuation between usage of different provinces, cities, and so forth, no patterns emerge. After 2005, there was considerably more usage of post-1960 dates, but it is hard to establish a causal relationship. The data are interesting, and will be made available online, but unhelpful in establishing any possible trends.

Luckily, we have one more trick up our sleeve: topic modelling, which is a novel digital method for tracking changes over time. Using MALLET (MAchine Learning for LanguagE Toolkit), a command-line based program developed at the University of Massachusetts at Amherst, I was able to pick out frequently occurring clusters of words (or topics). With this corpus, after some experimentation, I began with picking out the top fifty topics that appeared. Subsequently, I narrowed them to topics that showed sustained and significant statistical shifts. In figure 7, we see the topics, their numbers, and their frequency over time presented in sparkline form.[14]

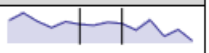
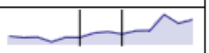
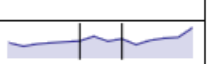
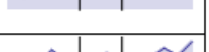
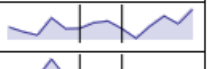
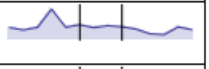
Topic No.	Mallet Topic Models, Cdn History Dissertations 1997-2010	Sparkline
2	great toronto men faire opposition nac plans range	
5	british toronto canada field fonds modern officers international james politics box chinese star jones space argues postwar centre election	
8	toronto history time canada change years act west south western association archives community national immigrants women living press	
9	canada owner press university war british social history north early land	
16	disease wife international simply required bay progress scientists cleark female provided relation site emigration	
18	canada history cultural university british canadians world aboriginal	
22	deletion endings government ymca emergent	
29	bureau francais paper tile companies preservation trading pulp universities managers valley ourage beaver usage	
31	persons operational cleaning guiding regiment migration paint settlers latin emigration carter printing township	
34	religious irish missionaries music policy church methodist police labor welfare trade executive	
45	economic east production nova saint canadians wages	

FIGURE 7: Selected topic models in Canadian history dissertations, 1997–2010. The first horizontal line line in the right column represents 2002, the second horizontal line represents 2005.

A few sustained shifts occur abruptly with the advent of digitized sources, as opposed to ongoing trends (for example, the reduction in “great Toronto men” as a topic

has been ongoing with fits and starts since 1998): greater emphasis on postwar politics and attending ethnicity (topic 5); a sharp reduction in topic 16 starting initially in 2005 and continuing thereafter; and a precipitous decline in "eastern economic history" (topic 45, a consistently popular topic until 2006 when it declines steadily).

From this traditional and computational analysis, what can we learn? The databases have not changed topics at the macro level: that is, we cannot establish causality between their availability and the topics that are conceived. However, we can trace smaller changes at a more micro level: in the topics that make up the broader projects, as well as the areas and sources consulted to a smaller extent. One of the most common "topics" of words, *economic eastern history*, has declined considerably since 2005, whereas *social history* has increased dramatically after a fallow year in 2006.

A few notes are necessary. First, topic modelling is not a magic bullet. These topics are imperfect and do not match seamlessly to trends. The ambiguity of the topics above, such as *economic eastern history*, is unavoidable at this level, and their utility is thereby limited. Second, when discussing broad historiographical shifts, one must take care to not confuse correlation with causation. Many different currents and themes are at play in the topics that historians write about, particularly young scholars – notably a growing movement toward transnationalism and internationalism, as well as continued debates about the relative importance and potential complementariness of political, military, social, and cultural histories.[15]

While one factor among many affecting the topics studied may be the availability of digital resources, the more trenchant point, and interesting from the perspective of the impact of online databases, is that the way topics are researched has changed without much notice. Again, the shift toward digital collections is not happening in a vacuum. Here, the availability of this technology has been coupled with institutional shifts in both Ottawa and in various university administrations to arguably speed up these trends.

A significant issue at play here may be the changing level of in-person access to archival resources over this period. Library and Archives Canada (LAC) has been advancing an agenda of ostensible "modernization," which implicitly and in some cases explicitly prioritizes online access over in-person (witness the relatively recent debate over changing hours of access).[16] Indeed, the Canadian Association of University Teachers (CAUT), the Canadian Historical Association (CHA), and others have mounted campaigns against this attention paid to a minority of digitized priority sources as opposed to massive on-site holdings.[17] Digital holdings are not immune to austerity, however, as even 50 per cent of digitization staff have been cut in recent rounds. In any case, the slow cutting of on-site service at LAC may account for some of the increasing

shift to online sources. We may see some of this already in my data set, and this may accelerate the trend detected to date. With the termination of LAC's Inter-Library Loan program, I anticipate that we will see a further shift toward online sources and away from increasingly harder-to-obtain traditional resources.[18] Furthermore, while post-secondary education financial data are notably hard to obtain on a sector basis, the number of graduate students being enrolled has often outstripped financial support for research travel. Governments have also put pressure on universities to increase time-to-completion rates, encouraging graduate students to finish doctoral degrees in four or five years. This may also be encouraging students to move toward greater numbers of online sources. They are certainly more cost effective than distant traditional sources.

Whatever the reason, it is clear that the introduction of Pages of the Past and Canada's Heritage Online has quantitatively affected our research. Their introduction has led to a noticeable increase in citations to those newspapers, in both a comprehensive dissertation database and as in the *Canadian Historical Review*. In the next section, I will demonstrate how databases have a qualitative impact on our research, through the issue of poor OCR and inherent unreliability and unpredictability. Despite the illusionary order and comprehensiveness offered by these databases, the reality is far more complicated.

Can we trust the black box? OCR and Databases

In several reviews of forthcoming and current research databases, a common assumption is that historians will be reluctant to use technology that does not meet exacting professional standards. In a 2010 *Journal of Victorian Culture* article, Richard Deswarte argued that "if a resource or dataset contains inaccuracies, is misleading, incomplete or poorly designed, scholars will avoid it or will raise concerns as they do of any problematic traditional paper archive source."[19] Historian Alexander Maxwell similarly noted, "Researchers do not care about searching sources whose accuracy they do not trust."[20] Formal reviews of newspaper databases often highlight unreliable OCR, even if this is seen as being offset by the sheer amount of information now available, or the democratizing effect of databases (accessible from home, often with only a public library membership).[21]

Can historians, without a reasonable doubt, trust databases such as those provided by the *Toronto Star* and the *Globe and Mail*? It first helps to have an understanding of how such databases are constructed. The *Toronto Star* was the first newspaper in the world to be digitized in its entirety. As its official history on its website notes, Cold North Wind was forward looking: they conceived of the project in an era of largely dial-up Internet services, limited storage, and other technical limitations, yet the final product was usable in the technically overbuilt late and post-dot-com boom.[22]

Indeed, the rhetoric on their webpage indicates that they foreshadowed the explosion of interest in big data, a phenomenon that has recently spread to the American White House and beyond.[23] From an audacious beginning with the large *Toronto Star* archive, stemming from 1894 until the then-present, Cold North Wind and their newspaper division, Paper of Record, has grown to now include 21 million images. The *Toronto Star* was subsequently followed by the *Globe and Mail*: Canada's Heritage since 1844, which put that newspaper online stretching back to 1844. These Canadian newspapers were part of a much broader North American and European phenomenon. The digitization and deployment of the *New York Times* collection back to 1851 in July 2002 led to much scholarly notice; as Barry Popik wrote in the *Journal of English Linguistics*, "No more needle-in-a-haystack microfilm searching. Throw away that index. Just type in a keyword, wait a second or two, and out pops your search results!"[24] While some inherent problems of keyword searches were occasionally noted, more common were concerns around the cost of accessing these troves.

Newspaper digitization is both simple and complicated. Let us use the case of the *Toronto Star* as a pertinent example. At a speed of roughly one million pages of newspaper per month, digitization was carried out from microfilm originals. From the microfilm, each individual page was subsequently produced as a decently high-resolution PDF document, averaging approximately 700 KB. Every page was put through an OCR scanner, producing a text file of the text found; users who enter a search term are searching their query against the text file. Once a match is found, the PDF is made available to the user.

Why the focus on keywords? While digital date-by-date searching has its advantages over traditional analogue microfilm searching – the ability to consult sources from home without travelling to a library – it does have its disadvantages: skimming via these databases is more tedious than microfilm and far slower. Indeed, without a large and high-resolution monitor and high-speed Internet connection, and sometimes even with, microfilm offers a superior browsing experience. Keywords, however, offer additional functionality not found in analogue sources. Arguably, while undoubtedly some users are skimming occasionally, the main reason for the increased use of databases is keyword searches. This enables large-scale media searching: representations of a specific word, for example, activities of a group, or evolving cultural conceptions of a term.

This is not without its downsides: the browsing model can lend itself to useful contextualization of a research project, learning about topics seemingly unrelated to your specific queries, getting a sense even as images skim across your screen of the zeitgeist

of the source or time, and gaining a comprehensive rather than focused survey of the past. Similar concerns have been voiced by scholars on the transition from traditional books found on library shelves to e-books.[25] For example, we may see a decline in spinoff projects emerging from the serendipity of the unrelated newspaper or archival find. Furthermore, we could also see scholars adhering closer to their initial projects (and thus attendant keywords).

Whether or not they should, historians are using keyword searching – a methodology with issues stemming from its underpinning system. Historians need a deeper understanding of OCR and what it means. OCR is concerned primarily with commercial markets and users: the massive digitization and transcription of large arrays of typewritten documents, often in corporate, legal, and governmental settings. The application of this technology to historical documents brings with it an initial step of difficulty then, as it attempts to complete a closely related yet not identical task. In a comprehensive article, a team of three researchers has outlined the major problems facing OCR routines as they tackle historical documents and newspapers.[26] I want to draw on the key points they provide with some further elaboration:

- *Non-standard fonts*: Historical newspapers cannot be relied upon to use standard typefaces, and OCR routines are seldom trained on the specific corpus.
- *Printing noise*: As a conventional historian, I love to find the imprint of the printer's hand on the actual paper. For instance, well into the postwar era, manually typeset documents can betray small errors. These are interesting to a historian, but anathema to an OCR routine.
- *Line and word spacing*: Spaces between characters in a manually laid out document are not universal and can lead to some words being split in the middle or, alternatively, being run together.
- *Line-break hyphenation*: While easy to fix today, many early OCR processes did not include line-break hyphenation. As a result, if a word transgresses a column – once even more frequent than today, given the narrower columns in historical newspapers – that word is forever lost to keyword searching (and thus, perhaps history).
- *Medium transformations*: These documents began as physical, paper copies. They were then microfilmed, and then digitized. Every time this transformation occurred, a few more data were lost.

Another issue that arises with newspapers is multi-column text: newspapers are often laid out in a series of vertical columns, which can throw off an OCR process, which

may not know when the spaces indicate a gap between words or a gap between columns. Other scholars have argued that historical applications are "very different" and that the "'black box' nature of commercial OCR software means it is not easily customizable to suit varied historical collections." They are now beginning to develop open-source academic alternatives.[27]

In the worst-case scenario, that of seventeenth- and eighteenth-century digitized collections, we can see an accuracy rate in the 40 per cent ballpark; others have estimated that in these conditions, perhaps even more than half of the information may be missed through keyword searching.[28] Digitization of modern newspapers, such as the large project conducted by the ProQuest Historical Newspapers collection, sees higher success rates: for the main body of articles, unedited text that has been run through OCR sees a success rate in the 80–90 per cent range. As these figures were obtained around 2002, based on discussions with ProQuest, we can assume that similar technology was used in the *Toronto Star* and *Globe and Mail* digitization.[29] As already noted, newspapers themselves raise additional issues for documents read by OCR that are not present in the commercial data sets that these programs are designed for. As Blanke, Bryant, and Hedges wrote in their 2012 *Journal of Information Science* article – over ten years since the initial digitization of these Canadian data sets – contemporary technology continues to choke on magazines and newspapers: "Handling of non-text elements also had a major bearing on the results for the magazine and newspaper samples, although in these cases page segmentation and layout analysis were also a considerable factor. The often challenging layouts used in these scans produced a very high variability in error rates in all the applications we tested; even the two commercial tools were frequently flummoxed (often by images which the other handled correctly)."[30]

The technology to deal specifically with historical documents is not here now, and it was certainly not there ten years ago when the *Star* and *Globe* were digitized. The most expensive OCR routines, finely honed on legal and corporate documents, choke on unique layouts. Note that this is before considering the junk data incorporated by microfilm streaks, issues of reproduction, and other artifacts intrinsic to early-twentieth-century and older documents.

Perhaps the best study of OCR and historical documents can be found in Simon Tanner's report on OCR feasibility. Tanner, a senior manager with the Centre for Computing in the Humanities at King's College London and the founding director of King's Digital Consultancy Services, lays out the feasibility case in easy-to-understand language. Even with cutting-edge OCR technology on pre-1950s documents, an accuracy

rate of 98 per cent would be the best-case scenario (probably higher than the actual effects). What does this mean for research?

> For example: [take] a page of 500 words with 2,500 characters. If the OCR engine gives a result of 98% accuracy this equals 50 characters incorrect. However, looked at in word terms this could convert to 50 words incorrect (one character per word) and thus in word accuracy terms would equal 90% accuracy. If 25 words are inaccurate (2 characters on average per word) then this gives 95% in word accuracy terms. If 10 words were inaccurate (average of 5 characters per word) then the word accuracy is 98%.[31]

Even in a best-case scenario, OCR technology will necessarily produce a less-than-ideal situation in search integrity. The obvious way to improve accuracy is with direct human intervention. The Old Bailey Online proceedings, for example, the basis of the Big Data Criminal Intent project (a product of the first round of the international Digging into Data grant), are correctly described as the "largest bodies of accurately transcribed historical text."[32] However, this was achieved through manual typists: for the roughest text, between 1674 and 1834, the texts were independently transcribed by two individuals; their results were then compared by computer. For the 1834 to 1913 data set, one typist typed while an OCR routine also did this; results were similarly compared.[33] This painstaking process reduced the estimated error rate to one in every 3,000 characters. Given the already onerous problems of database costs, the sheer size of a newspaper corpus and other logistical concerns, we are not likely to see a perfect text-readable collection of many other historical documents in the near future.

How does the black box of Canadian newspaper databases work in practice? I call this the illusionary order of online databases. The problem most obvious to researchers is the "false positive," or mistaken hit. If an erroneous article is flagged for the researcher's attention, she or he will likely notice it and discard it as a "false positive." We are used to this. Missed articles ("false negatives"), however, elude the researcher's gaze. Say a search term is entered, such as the "artistic woodwork" strike of 1973, a useful example as it occurs during the period in which OCR quality should be the highest and allows us to search over a discernible time frame (September until December 1973). The database executes the search, and an ordered list of results is pulled up: a few hundred hits, each of which arranged by full date, section, page number, and rough metadata relating to the content (News, Opinion/Editorials, Business, for example). Clicking on each link brings a full-text PDF of the article.

Many articles are revealed, the "hits," far more quickly than the same results could be achieved with microfilm. Comparing the microfilm results with those from OCR, however, quickly demonstrates a true negative "miss." Indeed, a search for *artistic woodwork* will miss a featured, front-page, above-the-fold article on the strike. In this case, the phrase is simply missed: probably due to an OCR artifact error. The search phrase appears only once, as subsequent references are simply to "artistic," and so our original search would completely miss this article. This would be a serious issue. A featured article such as this is indicative of the level of attention the strike receives, runs with a fascinating picture of a United Church minister being pushed by a police officer, and includes a list of arrests.[34] It is, however, obscured by the illusionary order of the database. As I will note, we need to use multiple search terms in order to find maximal results.

Surveying dissertations demonstrates several examples of bad practice. However, I have made the decision to speak in generalities. While dissertations are important sites of historical practice, the authors (especially in the critical post-2005 period) are often in tenuous employment situations as adjunct or contract faculty, are postdoctoral fellows, are engaged in the critical process of preparing dissertations for publication, or in the best-case scenario are tenure-stream faculty without the security of tenure. These dissertations were also written under time pressure and without a comprehensive understanding of these databases.

That said, fairly significant issues have emerged:

. *Authority through numbers*: The spectrum of citations has increased in several different respects, with many dissertations now using a few citations where they typically used none before, others using dozens where they might have used one or two citations, and a few now deploying two or three hundred citations where they may have used thirty or thirty-five in the past. If we can accept the premise that the overall quality of historical work has not increased, we can then infer that these numbers do not increase validity on their own. They do, however, move us away from analytical generalization toward exhaustive citation; adding clout to our findings by virtue of many, many more citations than had been previously possible. Given the problematic OCR that these data are founded on, there is a good chance that we are now ascribing to ourselves greater authority than is warranted.

. *Decline of third-party non-digitized newspapers*: While, as noted above, the *Toronto Telegram* was not commonly cited, the gulf between the *Telegram* and the *Star* and the *Globe* has considerably increased. Comprehensive

media surveys are now more common than ever (i.e., hundreds of citations of the *Globe* and the *Star*), backing up social histories of Toronto-area groups, Ontario politics, and even national political trends. The *Telegram* is rarely included. This has the effect of concentrating historical attention on the same two sources, excluding other ones.

. *Inappropriate use of place*: The increased use of the *Globe* and *Star* has had a slight effect in concentrating study on Toronto, as demonstrated in some of my analysis, and a slight increase in Ontario. More problematic, however, is the use of these newspapers to study events that have taken place near Toronto: events in cities near and within the Golden Horseshoe that have local newspapers. While these newspapers often carry similar syndicated content, local narratives and columnists (where they exist) can be lost.

These three points are not critical on their own, but when coupled with the overall trustworthiness of the databases in question, they bear deeper analysis. Indeed, this section opened with professional opinion that historians would not use tools that they do not fully understand, and certainly that they cannot fully trust. These databases cannot be completely trusted, at least not without critical methodological reflection that needs to be front and centre in introductory material. An archive with a similar element of error would warrant explicit recognition.

While Canadian historians slept, our historiographical foundations have been profoundly altered by online databases. What does this quantitative shift mean for Canadian historiography? As British historian Tim Hitchcock has put it, we have a serious issue:

> We read online journal articles, but cite the hard copy edition; we do keywords searches, while pretending to undertake immersive reading. We search “Google Books,” and pretend we are not.
>
> But even more importantly, we ignore the critical impact of digitisation on our intellectual praxis. Only 48% of the significant words in the Burney collection of eighteenth-century newspapers are correctly transcribed as a result of poor OCR. This makes the other 52% completely un-findable. And of course, from the perspective of the relationship between scholarship and sources, it is always the same 52%. Bill Turkel describes this as the Las Vegas effect – all bright lights, and an invitation to instant scholarly riches, but with no indication of the odds, and no exit signs.[35]

Using Pages of the Past and Canada's Heritage since 1844 uncritically is akin to using a collection of the *Canadian Historical Review* as your primary source, a collection that has 5–10 per cent of the pages randomly ripped out, without rhyme or reason, and critically with no way to tell if they were there or not. This is not a technical knock at these two databases; if an OCR routine could obtain 99 per cent reliability on historical documents, that would be a technical feat – yet for qualitative, rigorous historians, 99 per cent would provide only illusionary coverage. These technologies may be better than what has come before, but we need to recognize these issues. Digital sources are mediated in ways different from traditional ones.

This problem is compounded by our lack of transparency. Several dissertations and a few historical monographs do mention the use of online newspapers such as the *Toronto Star* and the *Globe and Mail* in bibliographies or appendices, but for the most part, their use remains at best implicit (however, given the high use of online databases, perhaps one needs to be explicit about consulting the analogue version instead). As the preceding section demonstrated, their availability has shaped and will continue to shape how professional historians engage with the past. I have shown how OCR skews research, leads to missed hits, and lays a flawed fundamental layer beneath much of our history.

It also tends to concentrate attention toward a handful of newspapers. First, the *Globe and Mail* was always the most cited newspaper, as befits Canada's allegedly national newspaper, but its citation counts have increased dramatically. This tends to highlight a centrist, Central Ontario perspective of the news and of historical events. Second, the *Toronto Star* has increased in many dissertations from a source that was akin to others of similar stature (the *Ottawa Citizen* or *Montreal Gazette*) but is now one of the most cited sources in Canada. The *Star* appears throughout the Canadian historiography in several unlikely places: being used to recount events throughout Ontario, or even beyond, by virtue not of its significance as a source but its online availability.

Conclusions, or What Can We Do?

In conclusion, what can historians do to improve their treatment of the past? First, the most important factor is recognition. Historians need to be cautious about the sources they use and be transparent with them. We need to recognize the actual sources that we are using: a comprehensive survey of a microfilmed *Toronto Star* is not equivalent to using a searchable database processed through OCR. One is not superior to the other, to be sure, but the advantages of the latter need to be posited against the possible downsides for any given topic. Second, we need to consider an array of best

practices for professional historians. Right now historians are operating in separate silos: developing our own best practices for interacting with databases, but not beginning a critical discussion of the next steps. We need to break out and begin a discussion – among researchers and pedagogically among those of us who teach undergraduate and graduate students.

As noted above, while the *Globe* and the *Star* offer useful case studies, these lessons apply to other digitized sources as well. Sources that did not begin their lives as digital documents have undergone digitization, and it behooves us to ask questions about it. While early Canadianists, for example, may notice OCR problems more easily than postwar researchers using newspapers, because of the higher error rate, the general lessons remain true for all of us. As we increasingly become digital historians, if not self-consciously at least in our use of digitized source materials, these are lessons that require serious attention.

First, for event-based history, comprehensive skimming is necessary: for example, if you are looking for coverage of a specific event in 1973, depending on your research question and goals, it may behoove you to read all of the op-eds during that period, or feature articles, or whatever section is most relevant to your work. Such an approach melds the undeniable convenience of online research with the potential comprehensiveness of traditional, analogue research. Second, just as when we used to use card catalogues and other research tools that employed subject headings, multiple searches are necessary to find the information you are looking for. Try searching along the lines of pluralized versions of a text, alternate spelling, abbreviations, synonyms, and so forth. This will lead to duplication of results but allow you to have a slightly higher confidence in the robustness of your data.

Third, by subsequently being up-front about our research findings, how certain searches worked, how others did not, we can – as a professional community of historians – collectively enhance our experience and outcomes with these search engines. At present, there is an absence of documentation on experience with either of these two databases. This has important historiographical consequences for our profession and our conceptualization of the past. If the online version is being used, it should be cited as such. This is more honest about our sources and the past. Methodological considerations do not need to take up an undue number of journal pages, but they do need to be there.

Fourth, we can also ask for better features, at least for future digitization projects. Historians need to work with database creators and managers to mitigate some of the potential shortcomings. Open projects that allow participant action are a model to emulate and look toward. One of the best examples is the Australian Newspapers

Digitisation Program. They are up-front about both the problems of OCR and have an avenue to continually improve the material – as their website in May 2012 declared, "9000+ members of the public have corrected 12.5 million lines of newspaper text so far. A big thank you to all those people! The electronically translated text is often poor quality and needing improvement. If you improve the text it makes the text search better for everyone. Help join in with this activity. It is easy, fun and addictive."[36] Given the degree of enthusiasm with which participants have flocked toward collaborative and continual-improvement projects such as Wikipedia, StackExchange, and so forth, I do not believe that the Australian statement is one of hyperbole. As noted earlier, comparative studies of how different national digitization schemes have influenced scholarship would shed further light on how to address potential problems.

Research tools such as Pages of the Past and Canada's Heritage Online have profoundly reshaped the basis on which Canadian historiography is being constructed. They should continue to be used, as they enable new forms of expansive scholarship in tighter time frames, a necessity in this era of ever-shrinking budgets. As cuts to Library and Archives Canada's on-site access and Inter-Library Loan programs demonstrate, online sources will become increasingly important for accessing sources in any way for many of our graduate students and cash-strapped faculty members. Given this acceleration, these research tools need to be used – and problematized – like any other finding aid. We need to be self-conscious about possible bias, with an understanding of the underlying OCR technology, in order to ensure that we are as true as possible to the past. While a search term brings up a seemingly well-ordered list of research findings, we must take care that we are not being sucked into an illusory order of comprehensiveness.

Thanks are due to the Social Sciences and Humanities Research Council, which provided the funding that made this work possible. Thomas Peace and Jennifer Bleakney gave insightful comments on earlier drafts of this piece, as did *ActiveHistory.ca* readers when it was still in blog form. Thanks also to the *Programming Historian* team, and in particular William Turkel, who have helped make accessible the sorts of techniques that made this research possible.

1 In this, I am inspired in spirit by Bruce Curtis, *The Politics of Population: State Formation, Statistics, and the Census of Canada, 1840–1875* (Toronto: University of Toronto Press, 2001), which raised crucial methodological questions about how Canadian censuses should be employed by rigorous historians.

2 For a very recent discussion of e-books and their effects on historical research, see Kim Martin and Anabel Quan-Haase, "Are E-Books Replacing Print Books? Tradition, Serendipity, and Opportunity in the Adoption and Use of E-Books for Historical

Research and Teaching," *Journal of the American Society for Information Science and Technology* 64, no. 5 (2013): 1016–28, Wiley-Blackwell, http://authorservices.wiley.com/bauthor/onlineLibraryTPS.asp?DOI=10.1002/asi.22801&ArticleID=1049091.

3 For more on this, see Daniel J. Cohen and Roy Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web* (Philadelphia: University of Pennsylvania Press, 2006), Center for History and New Media, http://chnm.gmu.edu/digitalhistory/. A fascinating state-of-the-field discussion can also be found at Daniel Cohen, M. Frisch, P. Gallagher, S. Mintz, K. Sword, M.T. Kirsten, A.M. Taylor, W.G. Thomas III, W.J. Turkel, "Interchange: The Promise of Digital History," *Journal of American History* 95, no. 2 (Sept. 2008): 442–51, http://www.journalofamericanhistory.org/issues/952/interchange/index.html.

4 To my knowledge, the only other major automated historiography is David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *ACM Journal on Computing and Cultural Heritage* 5, no. 1 (Apr. 2012): 1–19, Princeton University, http://www.cs.princeton.edu/~mimno/papers/a3-mimno.pdf. This article uses topic modelling (a methodology briefly used in this paper) to analyze a hundred years of several classics journals to discern overall trends.

5 It is my belief that in order to ensure reproducible results, scholars must share their methodology. In an era of programming languages and the digital humanities, this may need to be more explicit than a traditional methodology of providing detailed archival citations.

6 The prohibition stemmed in part from the chill that fell over digital scholars in the wake of Aaron Swartz's 2011 arrest for massively downloading JSTOR files.

7 Bursting is not necessary, but reduces memory load and, more importantly, allows for an easier restarting of the OCR process should something go awry.

8 Tobias Blanke, Michael Bryant, and Mark Hedges, "Ocropodium: Open Source OCR for Small-Scale Historical Archives," *Journal of Information Science* 38, no. 1 (2012): 76–86, Sage Journals, http://jis.sagepub.com/content/38/1/76.abstract.

9 My own work is carried out in the Mathematica programming language, an integrated platform for technical computing that allows me to process, visualize, and interact with large arrays of historical information.

10 My code is shared through GitHub, an open-source programming repository website, at https://github.com/ianmilligan1/Illusionary-Order.

11 I did so by creating a new dataset: one that contained all of the bigrams and trigrams for a given word. Bigrams, trigrams, and so on, are *n*-grams: sets of recurring two- or three-word phrases in this case. A similar approach was used in the Google *n*-gram corpus. It is a standard building block of Natural Language Processing (NLP), a growing academic field. For more on NLP and *n*-grams, see Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval* (Cambridge, UK: Cambridge University Press, 2009), http://nlp.stanford.edu/IR-book/.

12 The dissertation in question is Stacey Jo-Anne Barker, "Feeding the Hungry Allies: Canadian Food and Agriculture during the Second World War" (PhD diss., University of Ottawa, 2008).

13 For more on robust statistics, see Rand R. Wilcox, *Fundamentals of Modern Statistical Measures: Substantially Improving Power and Accuracy*, 2nd ed. (New York: Springer, 2010), esp. 129–45.

14 Sparklines are discussed by data-visualization expert Edward Tufte in his *Beautiful Evidence* (Cheshire, CT: Graphics, 2006), 47–63. They are "small, high-resolution images usually embedded in a full context of words, numbers, images (47), which let readers quickly see the rise and fall of datapoints over time.

15 See, for published overviews, the essays by Chris Dummitt, Adele Perry, and Katie Pickles in *Contesting Clio's Craft: New Directions and Debates in Canadian History*, ed. Christopher Dummitt and Michael Dawson (London: Institute for the Studies of the Americas, 2009). Debates about the direction of Canadian history are now as likely as not to be taking place online, waged on blogs as various as ActiveHistory.ca, Christopher Dummitt's *Everyday History* (http://christopherdummit.blogspot.ca), Christopher Moore's *History News* (http://christophermoore.blogspot.ca), Sean Kheraj's *Canadian History and Environment* (http://www.seankheraj.com), and Andrew Smith's *The Past Speaks* (http://pastspeaks.com). Apologies to the many able bloggers, both junior and senior, who have been left off the list for reasons of brevity.

16 See "LAC Begins Implementation of New Approach to Service Delivery," *Library and Archives Canada website*, Library and Archives Canada, http://www.collectionscanada.gc.ca/whats-new/013-560-e.html. This was tackled in Bill Curry, "Visiting Library and Archives in Ottawa? Not without an Appointment," *Globe and Mail*, 1 May 2012, http://www.theglobeandmail.com/news/politics/ottawa-notebook/visiting-library-and-archives-in-ottawa-not-without-an-appointment/article2418960/.

17 For the main campaign, see "Save Library & Archives Canada," http://www.savelibraryarchives.ca. The Canadian Historical Association's comment is available at "CHA's Comment on New Directions for Library and Archives Canada," http://www.cha-shc.ca/en/News_39/items/19.html.

18 See "End of Interlibrary Loan Service," Library and Archives Canada, http://www.bac-lac.gc.ca/eng/Pages/end-ill-service.aspx.

19 Richard Deswarte, "Growing the 'Faith in Numbers': Quantitative Digital Resources and Historical Research in the Twenty-First Century," *Journal of Victorian Culture* 15, no. 1 (Aug. 2010): 285, Taylor-Francis Online, http://www.tandfonline.com/doi/abs/10.1080/13555502.2010.491665?journalCode=rjvc20&#.UiY32hZCooY.

20 Alexander Maxwell, "Digital Archives and History Research: Feedback from an End-User," *Library Review* 59, no. 1 (2010): 27, Emerald, http://www.emeraldinsight.com/journals.htm?articleid=1839344&show=html. He also notes, "Digital transcriptions, however, hold little interest for researchers because the information is mediated and unreliable" (24).

21 The downsides of OCR as being generally balanced by the accessibility of information is implicit in Jean-François Mouhot, "Archival Review: ProQuest Historical Newspapers," *Contemporary British History* 24, no. 1 (Mar. 2010): 131–4, Taylor-Francis Online, http://www.tandfonline.com/doi/abs/10.1080/13619460903553867?journalCode=fcbh20#.UiY4khZCooY. This vein also appears in several other articles referenced in this chapter. The democratizing element of these sorts of databases is a provocative argument made in Sandra Shoiock, "The Return of the Armchair Scholar," *Journal of Scholarly Publishing* 36, no. 2 (Jan. 2005): 49–57, Project Muse, http://muse.jhu.edu/login?auth=0&type=summary&url=/journals/journal_of_scholarly_publishing/v036/36.2roff.html.

22 For background on the project, please visit Cold North Wind's relatively detailed website, "About Paper of Record," https://paperofrecord.hypernet.ca/default.asp.

23 Big data is a difficult term to define, but essentially refers to amounts of data that are beyond conventional software programs and methodologies. For my own textual work, if I could load it into a conventional text editor to work with it, I do not consider it big data; once I need to turn to specialized information-retrieval solutions, I consider it big data.

24 Barry Popik, "Digital Historical Newspapers: A Review of the Powerful New Research Tools," *Journal of English Linguistics* 32, no. 2 (June 2004): 114, Sage Journals, http://eng.sagepub.com/content/32/2/114.

25 Martin and Quan-Haase, "Are E-Books Replacing Print Books?"

26 Maya R. Gupta, Nathaniel P. Jacobson, Eric K. Garcia, "OCR Binarization and Image Pre-Processing for Searching Historical Documents," *Pattern Recognition: The Journal of the Pattern Recognition Society* 40 (2007): 389, RFAI, http://www.rfai.li.univ-tours.fr/fr/ressources/_dh/DOC/DocOCR/OCRbinarisation.pdf.

27 Blanke, Bryant, and Hedges, "Ocropodium," 76.

28 William Noblett, "Digitization: A Cautionary Tale," *New Review of Academic Librarianship* 17 (2011): 3, Taylor-Francis Online, http://www.tandfonline.com/doi/abs/10.1080/13614533.2011.558737#.UiY6grxfRkk.

29 Donald S. Macqueen, "Developing Methods for Very-Large-Scale Searches in ProQuest Historical Newspapers Collections and Infotrac the Times Digital Archive: The Case of Two Million versus Two Millions," *Journal of English Linguistics* 32, no. 2 (June 2004): 127, http://eng.sagepub.com/content/32/2/124.abstract.

30 Blanke, Bryant, and Hedges, "Ocropodium," 83.

31 Simon Tanner, "Deciding Whether Optical Character Recognition Is Feasible," Dec. 2004, King's Digital Consultancy Services, http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf.

32 Dan Cohen et al., "Data Mining with Criminal Intent: Final White Paper," 31 Aug. 2011, With Criminal Intent, http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf.

33 For more, see "About This Project," The Proceedings of the Old Bailey: London's Central Criminal Court, 1674 to 1913, http://www.oldbaileyonline.org/static/Project.jsp.

34 The "artistic" examples come from my own work. See Ian Milligan, "'The Force of All Our Numbers': New Leftists, Labour, and the 1973 Artistic Woodwork Strike," *Labour / Le Travail* 66 (Fall 2010): 37–71, http://www.lltjournal.ca/index.php/llt/article/download/5613/6476.

35 Tim Hitchcock, "Academic History Writing and Its Discontents," *Journal of Digital Humanities* 1, no. 1 (Winter 2011), http://journalofdigitalhumanities.org/1-1/academic-history-writing-and-its-disconnects-by-tim-hitchcock/.

36 "Australian Newspapers Digitisation Program," National Library of Australia, http://www.nla.gov.au/content/newspaper-digitisation-program.