This article appears with permission by the original publisher: Canadian Parliamentary Review, Winter 37(4), 2014.

Open Data's Potential for Political History

Ian Milligan

The recent trend of "open government" initiatives has provided an exciting new source of material for digital humanities researchers. Large datasets allow these scholars to engage in "distant reading" exercises to provide context in ways previously not possible. In this article, the author provides examples of the tools researchers can use to expand their understanding of the country's political history and of the changing nature of parliamentary institutions and debates. He concludes with suggestions for ways to gain the maximum benefit from these data releases.

What could we learn if we read every word of the federal Hansard and explored how the frequency of various 'topics' rose and fell over time? Or, what types of trends might we see if we were able to know the occupation of every candidate for office since 1867? What kind of heretofore unknown value can be discovered in these sorts of extremely large datasets? The answers to all of these questions are promising.

New and newly digitized datasets from parliamentary sources offer considerable potential for historians, political scientists, and other researchers interested in political history. The rise of digital humanities – a hard-to-define and nebulous grouping of humanities scholars who explore the possibilities offered by new media and emerging technologies and present fascinating methods to approach analyzing large quantities of information – as well as exciting releases in the 'open data' sphere, combine to offer new opportunities for understanding the past. In this piece, I highlight some of the possibilities that large datasets present to people interested in parliamentary history, and conclude with suggestions about what governments and funding agencies can do to support this emerging field of research.

Ian Milligan is an assistant professor of Canadian and digital history at the University of Waterloo, leading a Social Sciences and Humanities Research Council-funded exploration of how historians can meaningfully engage with and computationally explore web archives. He is also a founding co-editor of ActiveHistory.ca, a website dedicated to connecting the work of historians with the wider public.

Open Government and the Digital Humanities

'Open data' is the idea that data should be made publicly available for use by anyone for any purpose, including reusing the data, modifying it, and building platforms upon it. 'Open data' is married to the concept of 'open government' – the idea that the people of a country should be able to access, read, and manipulate (in their own applications and on their own terms) the data that a country generates. The current federal government aggressively moved in this direction with the 2011 launch of the Open Government Initiative.[1] When people think of 'open data,' historical research probably does not immediately come to mind. In general, most open data releases tend towards the scientific, the technical, or the immediately applicable: bus route information, for example, or geospatial information about various zoning or infrastructure placements. However, some of these new data releases are increasingly relevant to historians, including the ones alluded to above – all candidates for federal political office, the frequency of words appearing in transcripts from parliamentary debates, etc.

Prior to the advent of these types of initiatives, many humanists would not be able to access these large arrays of information. The dawn of the era of the digital humanities has opened up new exciting possibilities for analysis, however. In English literature, for example, literary scholar Franco Moretti argues for "distant reading" to help understand the rise of the Victorian novel; rather than focusing efforts on a corpus of some two hundred or so books, we can use computational methods to study tens of thousands of novels at once.[2]

While it is still important to read individual books to test theories and explore prose, we cannot read all of them; distant reading lets us further contextualize the ones that we do read.

Using a few parliamentary datasets as examples, let's see some of what a digital humanist can do with access to all of this data.

Topic Modeling and Distantly Reading "Hansard," 1994-2012

The federal government has made its full transcripts of debates since 1994 available online.[3] The transcripts form a relatively large, but not insurmountable, amount of full-text data: 800 megabytes of plain text. Yet it would be nearly impossible to read all of this text, especially if you wanted to be able to do anything else with your time!

We can, of course, query it with full-text searching. Many of us have been doing these types of searches for years, and to good effect in published scholarship on parliamentary history. But meaningful full-text searching is always difficult to carry out; a researcher must know what to look for with a fairly high degree of certainty. Using colloquial keywords, short-hand terms or perhaps being ignorant of a single typographical mistake, can lead to many missed results. Often a researcher would need to know a lot about a topic *before* hitting the search bar. More so, full-text searches in some search engines can skew results, given the algorithms that underlie the search function; results are being ranked in a way that most scholars do not understand.[4] If, however, a scholar is looking for specific discussions, whether it is a particular name of a labour strike or a specific piece of legislation, full-text search can be extremely useful. To try a full text search of Hansard, visit http://www.parl.gc.ca/housechamberbusiness/ChamberHome.aspx and click on "Search and Browse by Subject" in the left-hand column.

Researchers can repurpose the plain text used in subject searches to manipulate and explore these Hansard records themselves. One method that works particularly well with large corpuses is called topic modeling, a textual analysis methodology based on a mathematical concept known as Latent Dirichlet Allocation.[5] As Shawn Graham, Scott Weingart, and I wrote in the *Programming Historian*:

> Topic modeling programs do not know anything about the meaning of the words in a text. Instead, they assume that any piece of text is composed (by an author) by selecting words from possible baskets of words where each basket corresponds to a topic. If that is true, then it becomes possible to mathematically decompose a text into the probable baskets from whence the words first came. The tool goes through this process over and over again until it settles on the most likely distribution of words into baskets, which we call topics.[6]

In other words, imagine that you're writing a brief about the treatment of women workers. When writing sentences and paragraphs about labour unions, you tend to use words like "labour," "agreement," "certified," or "arbitration." When writing about women, you're likely to use words like "differential," "femininity," "inequality," and "maternity." Imagine that all those words are in little buckets sitting on your desk. By the end of your writing, the buckets are empty. Topic modeling tries to reverse that process: putting them back into the buckets from which they most likely came.

To demonstrate an example of topic modeling I downloaded all English language Hansard transcripts from 1994 onwards and tried to reconstruct them back into 'topics' within the text using Machine Learning for LanguagE Toolkit, or MALLET. Anyone can try out this tool by following our tutorial at http://programminghistorian.org/lessons/topic-modeling-and-mallet. Once topics in this dataset were established, it was possible to measure how frequently they appeared in Hansard text throughout these years.

A quick note on how the results are displayed: First, the six graphs presented here use a varying y axis interval to show how frequently the topic appears in a given sitting of Parliament. I have elected to change the scale of the y axis for visability purposes, so please note the values being used. Second, the words found in the resulting topics have not been translated. Using the French language plain text Hansards may result in slightly different topic results. Therefore, these graphs solely represent English language topics and the experiment should be conducted separately in French for accurate results.

I think that we can find provocative information with topic modeling. For example, one topic, that we might label "peace and peacekeeping," immediately appeared in MALLET's Hansard analysis (*See Fig. 1*). I was curious to see if establishing the frequency of this topic would allow me to test a hypothesis in the recent book *Warrior Nation: Rebranding Canada in an Age of Anxiety*. Here, Ian McKay and Jamie Swift argue that the Canadian narrative of a peaceful, peacekeeping country is being replaced by the notion that Canada is a warrior nation focussed on military might. They suggest there is evidence of a shift from peace to war in our commemorative strategies, the decisions made

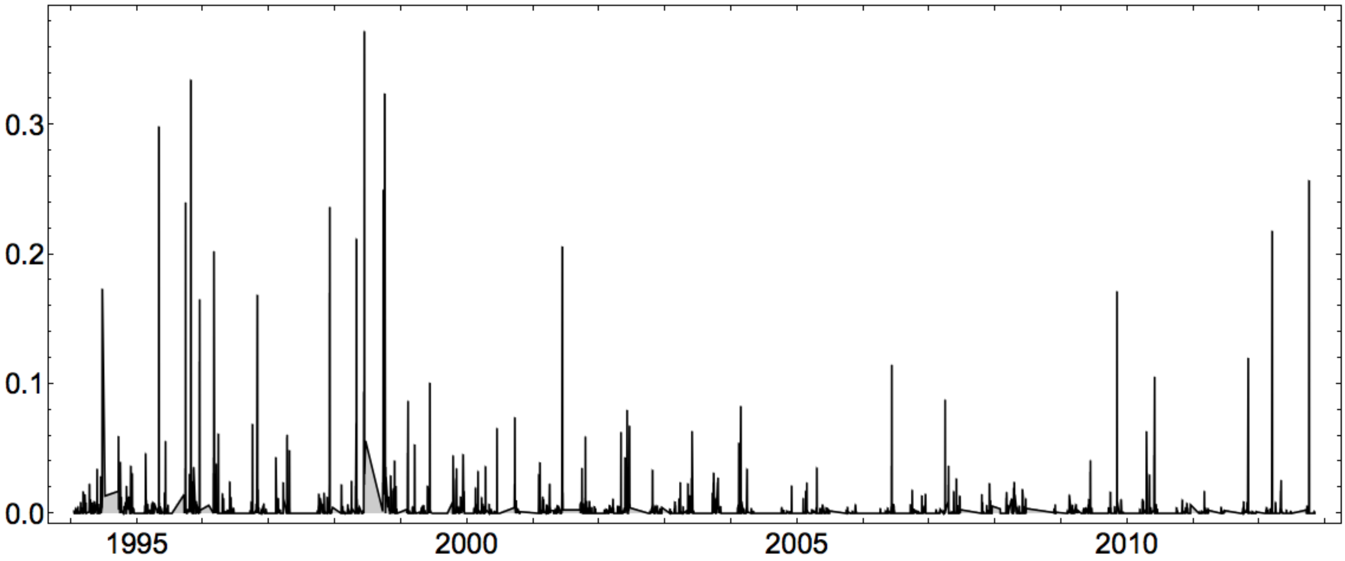

Fig. 1: A visualization of this topic's relative frequency across segments of Hansard. Topic keywords: "international canada peace mr nato war world peacekeeping conflict troops nations united people kosovo situation humanitarian foreign role genocide." Note that it is far more common before 2000 than afterwards (although perhaps we are more recently seeing a resurgence).

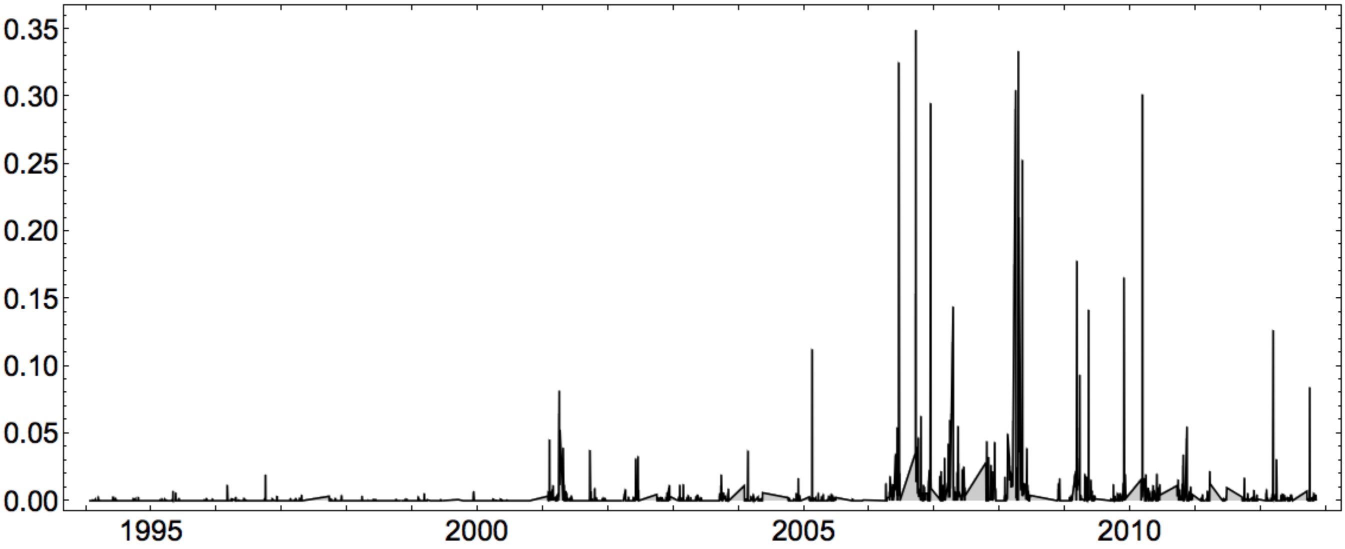

Fig. 2: A visualization of this topic's relative frequency. Topic keywords: "afghanistan mission canada canadian afghan mr minister government troops military security women defence forces international soldiers development motion support." Note again that it is more common after 2001, and notably after 2006. Comparing this to Fig. 1, we can see a transition between the two topics to some degree.

in the new citizenship guide for new Canadians, and several other facets of Canadian society.[7] A constant topic of discussion amongst historians at the Canadian Historical Association and in historical discussion venues such as *ActiveHistory.ca,* could we also see evidence to support this thesis in the Hansard dataset?

Keeping in mind that topic modeling tools *automatically* generate topics from these plain text datasets, and that we must put meaning to the word groups we find, I suggest the change in the topic's frequency from 1994 to present does accord with the *Warrior Nation* thesis. There is a noticeable drop off in this topic after the Conservative election in early

2006; however, the 9/11 attacks could also be seen as a significant fulcrum. We also do continue to see spikes. We don't know what these spikes mean at present, as they may be tied to random mentions of the Afghanistan mission or tied to specific events. More research is warranted. Another topic which appeared could also be relevant to read in tandem with this pattern (*See Fig. 2*).

Here, we see a topic directly related to the war in Afghanistan, albeit defence more generally, as well. The topic first appears briefly in the 1990s, but it accelerates in early 2001 with a news spike about the Taliban and then with the height of Canadian involvement in the Afghanistan war. If we take the two topics together, we can see how the first topic is more dominant near the beginning of the period under study while the second

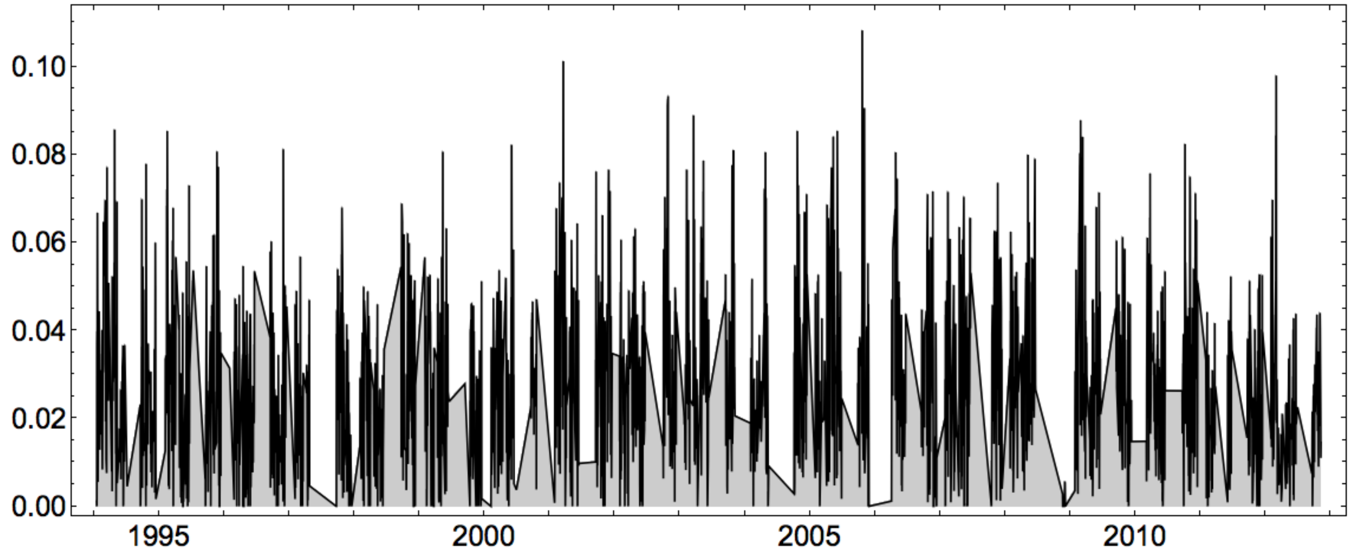

Fig. 3: This figure shows the general scaffolding of parliamentary business. Topic keywords: "committee mr report standing important parliamentary speaker work secretary process house issue recommendations review national made ensure information forward." Note that it is relatively consistent throughout, as should be expected.

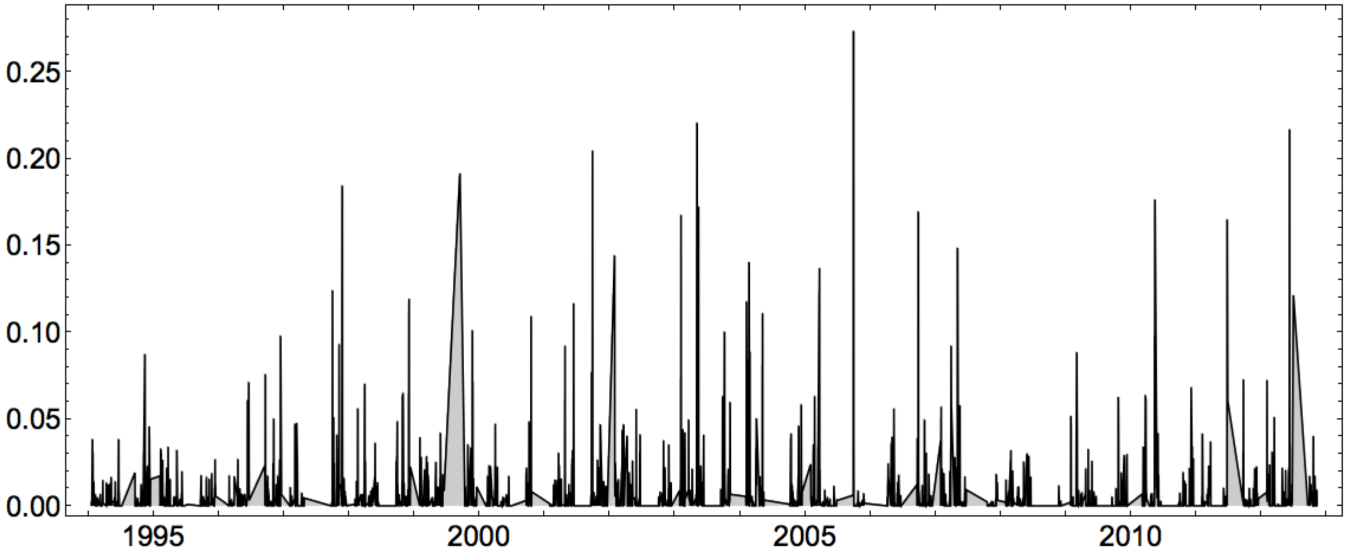

Fig. 4: A visualization of this topic's relative frequency. Topic keywords: "criminal code police sexual children offence mr law child person offences pornography justice dna age defence sex protect arrest." While there are ebbs and flows, it is relatively consistent.

topic is more dominant near the end. We certainly see a transition between the peace and peacekeeping topic and topic related to the military in Afghanistan; once again, this type of trend could potentially support the *Warrior Nation* thesis.

Other topics that appeared in the Hansard plain text modeling are also worth exploring. A topic which includes words likely associated with routine parliamentary business is a constant (*See Fig. 3*). However, two topics (not pictured in graphs) that could be associated with budgets appear to identify shifting rhetoric. Here, a topic with general budgetary language noticeably declines after 2006. Another topic relating to Canada's newer economic action plan appears to replace it, especially by 2009. This topic's

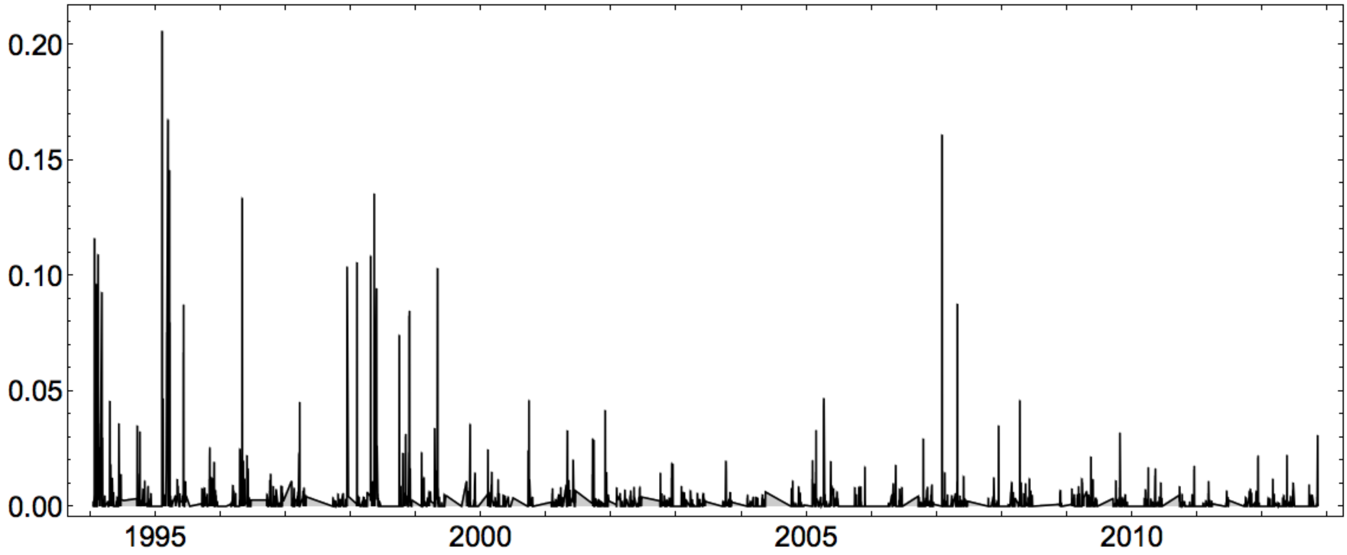

Fig. 5: A visualization of this topic's relative frequency. Topic keywords: "canadian cultural heritage canada culture flag canadians minister industry country mr arts national department world museums film artists quebec."

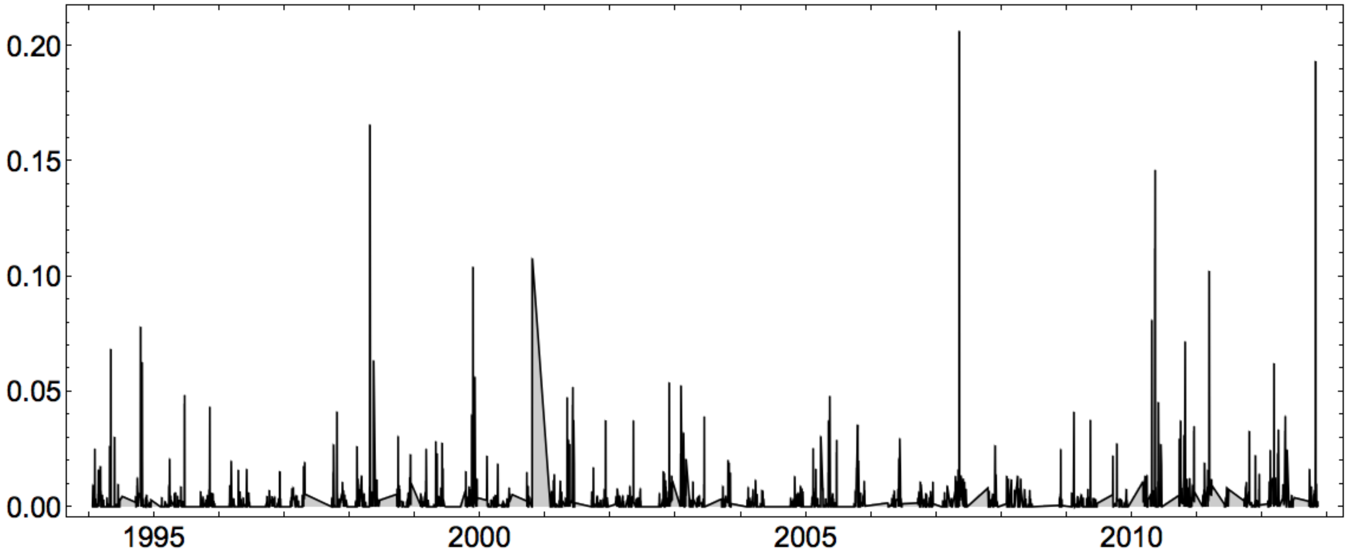

Fig. 6: A visualization of this topic's relative frequency. Topic keywords: "veterans war affairs canadian service mr benefits day world men services support speaker member country forces remembrance committee served." There are spikes around commemorative events, but it has more consistently accelerated since 2010.

keywords include: “economic budget jobs economy canada tax plan mr canadian canadians government measures action businesses support credit world finance crisis.”

A few other topics also appear notable. There is consistent concern within parliamentary debate about the protection of children, as seen in a topic which deals with youth and criminal offences (*See Fig. 4*)

A topic we might label “heritage” (*See Fig. 5)*, seems to be on the decline, though we do see peaks around both the Quebec sovereignty referendum and during the ensuing *Clarity Act* debates. However, a potentially related topic concerning remembrance has seen some spikes in frequency since the beginning of 2010 (*See Fig. 6*).

Although these examples offer only a brief exploration of some possibilities, by employing these types of tools we can pull our gaze back from individual debates to consider overall debates patterns.

Open Data and Parliamentary Candidate Occupations

Let’s examine another file: “History of the Federal Electoral Ridings, 1867-2010.” Available in both English and French at http://data.gc.ca/data/en/dataset/ea8f2c37-90b6-4fee-857e-984d3060184e, this large file contains information on 38,778 candidates for federal office in Canada. It comes in a 13-column comma-separated value (CSV) file with the following fields:

- Election Date, Election Type, Parliament, Province, Riding, Last Name, First Name, Gender, Occupation, Party, Votes, Votes (%), Elected.

The data in each field is then just a series of lines in text format; for example:

- 2008-10-14, Gen, 40, Quebec, PAPINEAU, Trudeau, Justin, M, teacher, Liberal, 17724, 41.47, 1.

We can move from left to right and gather the data: here we see current Liberal leader Justin Trudeau’s first election, in the 40th Parliament, a general election, with 17,724 votes (41.47 per cent of the total vote count), and who was successfully elected (indicated by the value of ‘1’ in the elected column). CSV files are very useful to researchers because they can be read by multiple types of software: Microsoft Excel, a programming language, or Google Docs.

Using a programming language I was able to control for one or more of these data fields. One value in the occupation field that appeared to be that of ‘lawyer.’ When I pulled the most frequent occupations, here is what appeared:

Table 1: Candidate Occupations

lawyer	3730
farmer	2587
Null	2308
teacher	1415
merchant	1194
businessman	1125
physician	999
barrister	981
parliamentarian	816
student	795
journalist	497
retired	476
manufacturer	425
manager	355
Member of Parliament	351
administrator	298
accountant	271
consultant	267
contractor	267
notary	224
engineer	223
housewife	196
salesman	195
agent insurance	190
professor	184
secretary	179
editor	164
-at+barrister-law	163
educator	145
broker insurance	144

Note that the data is not perfect (it *never* is). 2,308 occupations were listed as 'Null,' which means there was nothing entered in the field. This deficiency mainly results from inconsistent or absent data entry about defeated candidates prior to the 14th Parliament. Nevertheless, we see some occupations we would expect to see: lawyers, farmers, teachers, merchants, businessmen, doctors, etc.

At a glance, we see another problem with this data: "merchant" and "businessman" might be considered part of the same category. Similarly, lawyers appear variously as "lawyers," "solicitors," "barristers," and even "-at+barrister-law." This lack of uniformity in data isn't abnormal, and decisions must be made at all stages about how to interpret it. People create the data, and people – historians or political scientists, for example – must then interpret it. We have to be very careful before taking such data at face value, especially as some re-elected MPs apparently just wrote 'Member of Parliament' or 'parliamentarian' whenever they were re-elected. All of these provisos help point us towards the importance of actually looking at our data, rather than just trusting portals to do the work for us. We can use a program called Google Refine to clarify the data if we want to, or we can manually explore it. Data is not neutral, it's created by humans under subjective conditions.

Returning to "lawyers," how common is this occupation within the candidate pool? More so, do they have a disproportionate level of success at being elected? We know they were common as candidates in the 19th century and continue to be so today.

I generated two graphs, drawing on the 14th sitting of Parliament onwards (the point when data collection improved). Note that I did not control for by-elections within parliaments. Consider *Fig. 7* and *Fig. 8* (the x-axis refers to sittings of Parliament):

From this, we see that in the 14th Parliament nearly 11 per cent of all candidates for seats, whose occupations were listed, gave their occupation as lawyer (there were some solicitors too, but lawyer was overwhelmingly

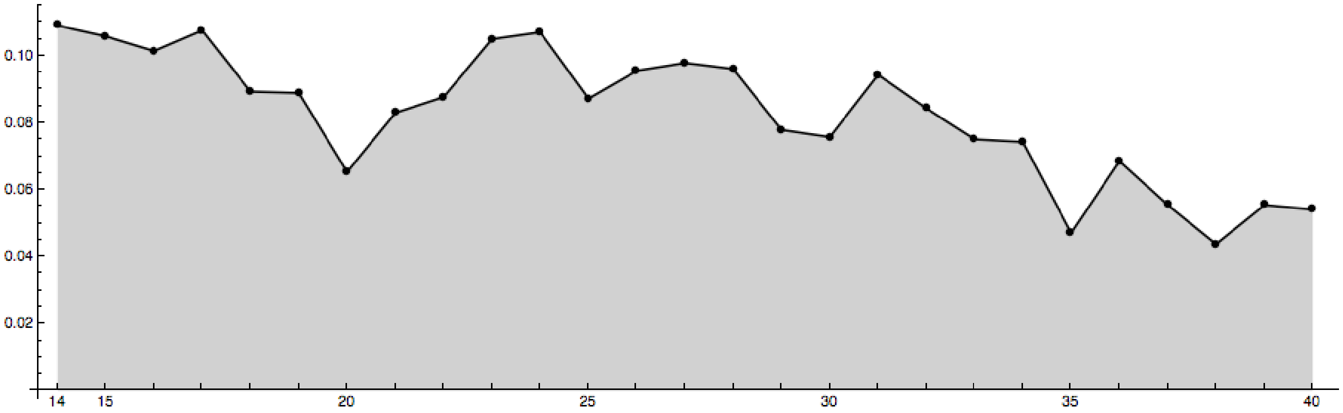

Fig. 7: Frequency of 'lawyer' occupation appearing in all candidates occupation listing, 14th-40th Parliaments

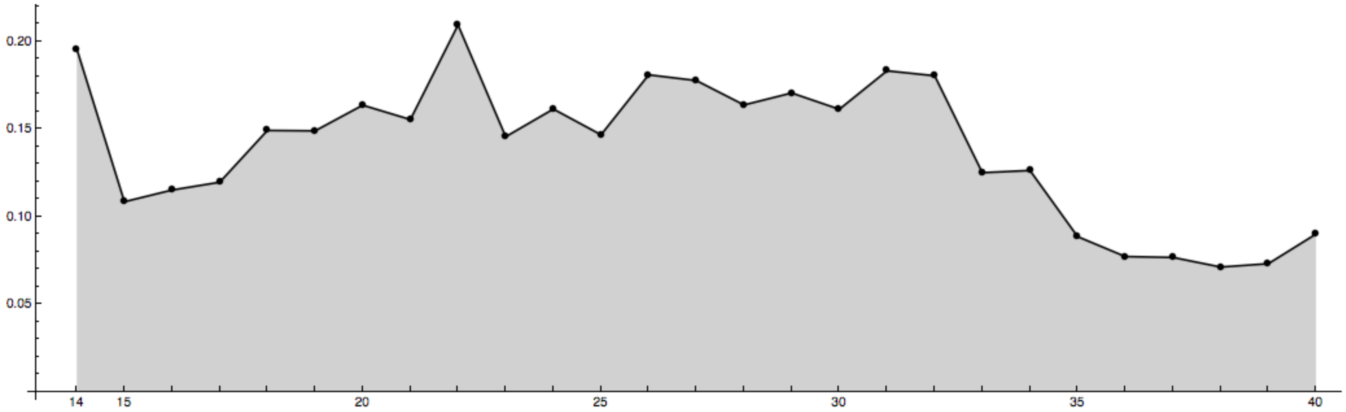

Fig. 8: Frequency of 'lawyer' occupation appearing as an elected candidate, 14th-40th Parliaments.

the way they recorded their occupation). Yet if we drop all the defeated candidates, we see that almost 20 per cent of the successful candidates during that Parliament were lawyers

There appears to have been a dramatic decline in the number of parliamentarians who are lawyers since that time – around nine per cent of our elected candidates in the 40th Parliament listed lawyer as their occupation. Though, of note, as discussed earlier – more lawyers may have listed their occupation as businessman, perhaps, or simply parliamentarian if they were seeking re-election.

Nevertheless, as imperfect as the data can be for exact statistics, it can be used to paint a general picture of candidate pools and the types of people who tended to run for various parties. For example, let's find the top 50 Liberal Party candidate occupations from 1962 onwards and compare to the New Democratic Party's candidates during the same period. I've chosen to use the Liberal and New Democratic parties due to their relatively consistent constitutions as the contemporary Conservative party has undergone several permutations during the same period of time. The resulting data speaks volumes about the make-up of the two parties:

Table 2: Top 50 Occupations for Liberal Party Candidates from 1962 Onwards

lawyer	737
parliamentarian	412
businessman	251
farmer	212
Member of Parliament	142
teacher	138
administrator	82
consultant	71
politician	68
physician	56
barrister	56
merchant	54
manager	53
economist	52
accountant chartered	49
accountant	44
journalist	43
professor	41
retired	38
engineer	37
manufacturer	36
businesswoman	31
broker insurance	31
educator	30
barrister and solicitor	29
business person	27
broadcaster	26
NULL	25
principal school	25
public servant	24
agent insurance	22
director executive	21
cabinet minister	21
publisher	20
notary	19
contractor	19
consultant management	18
housewife	17
engineer professional	16
-at+barrister-law	16
mayor	16
executive	15
business executive	14
doctor medical	13
student	13
social worker	12
clergyman	12
veterinarian	11
realtor	11
manager sales	11

Table 3: Top 50 Occupations for New Democratic Party Candidates From 1962 Onwards

teacher	484
student	192
lawyer	179
farmer	150
professor	71
retired	70
representative union	69
social worker	52
parliamentarian	51
Member of Parliament	48
journalist	43
businessman	43
administrator	38
consultant	37
professor university	37
housewife	36
electrician	34
economist	33
NULL	32
secretary	31
educator	31
representative	31
physician	29
clergyman	29
high school teacher	27
salesman	27
researcher	25
school teacher	23
writer	22
manager	22
-employed+self	20
minister	19
organizer	18
steelworker	18
machinist	17
business manager	17
agent business	16
trade unionist	16
engineer	16
clerk	16
accountant	14
contractor	14
college instructor	13
assistant executive	13
instructor	13
director executive	12
unemployed	12
nurse	12
driver truck	12
sociologist	12

Although I am not a scholar of parliamentary politics, in just a few minutes of tinkering I have already begun to generate good, meaningful data about the composition of our federal parliaments and the candidates who stand for election within them. I present this data warts and all because it shows, once again, that data should be taken with a grain of salt: this data, for example, treats “high school teachers” and “school teachers” differently. That might help one researcher, but might hinder many others.

Beyond parliamentary records, many other datasets may be of interest to various researchers, including birth registrations, most popular baby names, marriage registrations in various cities and towns, names of soldiers who enlisted in the Canadian Expeditionary Force, and so on. The opportunities for study are nearly limitless.

What Should We Do With This Data?

Datasets hold great potential for transforming research practices, but the full value of these rich information sources has not yet been realized. Academics should consider the following points before engaging in work with datasets.

First, it can be difficult to do interdisciplinary work in Canada. The Social Sciences and Humanities Research Council of Canada decided this year to discontinue the use of ‘priority areas’. Grant applications dealing with digital applications would previously have gone to a specific 'digital economy' committee, whereas now disciplinary peers review them. The jury is out on whether this change will be positive or negative, but the transformative use of new media and emerging technologies strikes me as something that should be reviewed by committees closely related to the subjects. Some traditional academics embrace technology while others quite openly shun it. More problematically, digital projects tend to involve interdisciplinary teams: from English scholars who have embraced distant reading, to computer scientists who understand the nuts and bolts of algorithms far better than humanists can. Historians generally operate on a sole-author, lone practitioner model, which means that we sometimes have trouble evaluating the work of large team-based projects. We need to keep an eye on institutional barriers to digital adoption, particularly as they have implications for hiring, tenure and promotion within the academy.

Our granting councils are one area where governments can support and help to shape the form of research to come. Academics should take the lead on research, in keeping with dictates of academic freedom and abstract exploration, but we operate within structures set up by governments.

We should also encourage the release of more data, and realize that when data is being made available it needs to be machine-readable (for example, as plain text files, or formatted comma-separated value sheets). We can create complicated Application Programming Interfaces (APIs), which are layers to put atop of a dataset to let computers talk to each other, but often just letting scholars *download* the data themselves is ideal (privacy concerns being respected, of course). If datasets are created, I'd love it if people always thought "could we let anybody download this?" And if so, why not put a big red button at the top saying "export data"? A scholar can dream.

Finally, I think it's important to note that that this type of work is going to accelerate in the future. My current primary research project examines how historians will be able to use web archives, and I firmly believe that a history of the 1990s or 2000s cannot be researched and written *without* using web archives. Not everyone will write histories of the web, but what happens on the web is an invaluable part of the historical record. Scholars studying a more recent election, must concern themselves with posts on message boards, electoral websites, tweets, videos, and so forth. These are all part of the record.

The 1990s are now distant history; students who will begin to write our histories of that period are probably just now entering the post-secondary sector. Will they be able to use web archives? More importantly, will they be able to use web archives through computational methods? We cannot read every website, after all – if we thought there were too many Victorian novels, just imagine how many tweets there are on a single day. We need to lay the groundwork of digital literacy for our next generation.

The data is there. We now need a trained generation of humanists who ask interesting questions and can manipulate data to help bring Canada's humanities scholarship into the 21st century. As historians increasingly turn to online sources like the *Programming Historian*, begin to blog and engage with data, the shape of our profession will begin to shift accordingly. Hopefully, governments will continue to support digital humanties research by making datasets available in a way that will maximize their utility to present and future scholars.

Notes

[1] See: *Implementation of Canada's Action Plan on Open Government (Year-1) Self-Assessment Report.* http://open.canada.ca/en/implementation-canadas-action-plan-open-government-year-1-self-assessment-report.

[2] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*,Verso, New York, 2007).

[3] See: http://www.parl.gc.ca/housechamberbusiness/ChamberSittings.aspx.

[4] When other historians tell me that they are not digital historians, I often ask them if they use Google to help with their research - and if so, if they know how PageRank works. For more on this, see Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations* 127, no. 1 (August 2014): 64–72, doi:10.1525/rep.2014.127.1.64.

[5] The concept is described in David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022. An extremely good explanation can also be found in Matthew L. Jockers, "The LDA Buffet Is Now Open; Or, Latent Dirichlet Allocation for English Majors," *Matthew L. Jockers Blog*, September 29, 2011, http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/.

[6] Shawn Graham, Scott Weingart, and Ian Milligan, "Getting Started with Topic Modeling and MALLET," *Programming Historian*, September 2, 2012, http://programminghistorian.org/lessons/topic-modeling-and-mallet.

[7] Ian McKay and Jamie Swift, *Warrior Nation: Rebranding Canada in an Age of Anxiety* (Toronto: Between the Lines, 2012).