Le potentiel des données ouvertes pour l'histoire politique

Ian Milligan

Les initiatives pour un gouvernement ouvert, une tendance récente, offrent aux chercheurs en sciences humaines numériques une nouvelle source de documentation fort intéressante. Ces chercheurs peuvent obtenir du contexte à même ces vastes ensembles de données grâce à la « lecture à distance ». Dans le présent article, l'auteur fournit des exemples de certains outils à la disposition des chercheurs afin de mieux comprendre le contexte historique politique du pays ainsi que la nature en constante évolution des institutions parlementaires et des débats qui y ont lieu. Pour conclure, il fait des propositions afin de tirer le maximum des données diffusées.

Que pourrait on apprendre si on examinait dans le hansard les variations de la fréquence des divers sujets au fil du temps? Quelles tendances pourrait on observer si on était en mesure de connaître la profession de tous les aspirants candidats au pouvoir depuis 1867? Quel type de valeur inconnue, de cette époque jusqu'à aujourd'hui, recèle cet ensemble de données extrêmement vaste? Les réponses à toutes ces questions sont prometteuses.

Les nouveaux ensembles de données de sources parlementaires, récemment numérisés, offrent tout un potentiel aux historiens, aux politologues et aux autres chercheurs s'intéressant à l'histoire politique. L'essor des sciences humaines numériques – un regroupement difficile à définir de chercheurs en sciences humaines qui explorent les possibilités des nouveaux médias et des nouvelles technologies et présentent des méthodes très intéressantes pour analyser d'énormes quantités d'information – ainsi que la diffusion de données ouvertes fort intéressantes – amènent de nouvelles possibilités pour comprendre le passé. Dans le présent article, je décris certaines des possibilités offertes par les ensembles volumineux de données aux chercheurs qui s'intéressent à l'histoire parlementaire. Je conclus en proposant ce que les gouvernements et les organismes de financement pourraient faire pour appuyer ce nouveau domaine de recherche.

Ian Milligan est professeur adjoint d'histoire canadienne et numérique à l'Université de Waterloo, où il dirige une étude financée par le conseil de recherche en sciences humaines et sociales sur les formes que peuvent prendre les véritables échanges entre historiens et les méthodes de consultation des archives Web. Il est également codirecteur et cofondateur du site Web ActiveHistory.ca, dont le but est de rendre le travail des historiens accessibles au grand public.

Gouvernement ouvert et sciences humaines numériques

Le concept de « données ouvertes » repose sur un principe : rendre les données accessibles au public afin qu'elles puissent être utilisées par tous, que ce soit en vue de les réutiliser, de les modifier ou de s'en servir pour créer des plateformes, notamment. Le concept de « données ouvertes » est rattaché à celui de « gouvernement ouvert », qui lui repose sur le principe selon lequel la population d'un pays doit pouvoir accéder aux données qui y sont produites, les consulter et les manipuler (dans leurs propres applications et de la façon dont elle l'entend). Le gouvernement fédéral actuel est allé sérieusement dans cette direction en lançant en 2011 l'Initiative pour un gouvernement ouvert[1]. Quand on pense aux données ouvertes, ce ne sont probablement pas les travaux de recherche historique qui nous viennent à l'esprit. De façon générale, la plupart des données ouvertes diffusées sont de nature scientifique, technique ou très concrète; il s'agit, par exemple, d'information concernant des itinéraires d'autobus ou de l'information géospatiale sur divers secteurs et infrastructures. Pourtant, certaines de ces nouvelles données diffusées sont de plus en plus pertinentes pour les historiens, y compris celles mentionnées plus haut. Pensons par exemple à tous les candidats aux élections fédérales et à la fréquence des mots figurant dans les transcriptions des débats parlementaires.

Avant l'avènement de ce genre d'initiatives, ces volumineux ensembles de données n'auraient pas été accessibles à bon nombre de chercheurs en sciences

humaines. Or, le début de l'ère des *sciences humaines numériques* a donné lieu à de nouvelles possibilités d'analyse fort intéressantes. Par exemple, selon le professeur de littérature anglaise Franco Moretti, la « lecture à distance » permet de comprendre l'essor du roman victorien. Au lieu de mobiliser ses efforts sur un corpus de quelque 200 livres, des méthodes computationnelles permettent d'étudier des dizaines de milliers de romans à la fois[2]. Si, pour mettre à l'épreuve des théories littéraires et analyser la prose d'un écrivain, il est fondamental de lire des ouvrages, il est impossible de les lire tous. La lecture à distance permet donc de placer les ouvrages lus dans un contexte plus large.

Voyons, au moyen d'exemples tirés des ensembles de données parlementaires, ce qu'un humaniste numérique peut faire pour accéder à toutes ces données.

Modélisation des sujets et lecture à distance du hansard, de 1994 à 2012.

Le gouvernement fédéral a rendu accessible, en format numérique, l'intégralité des transcriptions des débats depuis 1994[3]. Les transcriptions forment un ensemble de données plein texte relativement volumineux, mais non insurmontable : 800 mégabits de texte brut. Pourtant, il serait impossible de les lire intégralement, en particulier si occuper son temps à autre chose!

Évidemment, on peut faire des recherches en texte intégral. Nous sommes nombreux à effectuer ce genre de recherches depuis des années, et à bon escient dans les travaux de recherche sur l'histoire parlementaire qui ont été publiés. Cependant, il est toujours difficile d'effectuer des recherches concrètes en texte intégral puisqu'un chercheur doit savoir assez bien ce qu'il recherche. Le fait d'utiliser des mots clés trop courants ou des termes abrégés ou encore de commettre une seule petite erreur typographique peut éliminer de nombreux résultats. Bien souvent, il faut déjà en savoir beaucoup sur un sujet *avant même* d'effectuer sa recherche. Et la plupart du temps, sur certains moteurs de recherche, les résultats des recherches en texte intégral peuvent être faussés à cause des algorithmes de classement utilisés, ce qui fait en sorte que l'ordre de présentation des résultats peut être incompréhensible à la plupart des chercheurs[4]. Par contre, quand on cherche un débat sur un sujet précis, que ce soit une grève de travail ou un projet de loi en particulier, les recherches en texte intégral peuvent être extrêmement utiles. Pour tenter une recherche en texte intégral dans le hansard, rendez vous à http://www.parl.gc.ca/housechamberbusiness/ChamberHome.aspx?Language=F et cliquez sur Rechercher et explorer par sujet dans la colonne de gauche.

Les chercheurs peuvent réutiliser le texte intégral dans lequel ils effectuent des recherches par sujet pour manipuler et explorer eux-mêmes le hansard. La « modélisation thématique » est particulièrement efficace sur de volumineux corpus; il s'agit d'une méthode d'analyse textuelle fondée sur le concept mathématique de l'allocation Dirichlet Latent[5]. Voici ce que Shawn Graham, Scott Weingart et moi avons écrit à cet égard dans *Programming Historian* :

> Les programmes de modélisation thématique ne tiennent absolument pas compte du sens des mots en contexte. La composition (par un rédacteur) de chaque fragment de texte est plutôt fondée sur la sélection de mots à partir de paniers probables de mots, dans lesquels chaque panier correspond à un sujet. Si c'est vrai, il devient alors possible de décomposer mathématiquement un texte en paniers d'où ils venaient probablement au départ. L'outil répète systématiquement le processus jusqu'à l'établissement de la distribution la plus probable de mots en paniers, que nous appelons sujets[6].

Autrement dit, imaginez que vous rédigez un mémoire sur les travailleuses. Les passages concernant les syndicats comporteraient des mots tels que « travail », « accord », « accréditée » ou « arbitrage ». Ceux concernant les femmes comporteraient probablement des mots comme « différentiel », « féminité », « inégalité » et « maternité ». Imaginez que tous ces mots se trouvent dans de petits paniers sur votre bureau. Une fois le texte rédigé, les paniers seraient vides. La modélisation thématique veut inverser le processus, c'est à dire remettre les mots dans les paniers d'où il est très probable qu'ils viennent.

Dans l'idée de montrer un exemple de modélisation par sujet, j'ai téléchargé les transcriptions du hansard en anglais de 1994 à ce jour et tenté de les catégoriser par sujets au moyen de l'outil MAchine Learning for LanguagE Toolkit, ou MALLET. Tous peuvent mettre cet outil à l'essai en regardant notre tutoriel à http://programminghistorian.org/lessons/topic-modeling-and-mallet. Une fois que les sujets ont été établis dans l'ensemble de données, il a été possible d'évaluer leurs occurrences dans le hansard au fil des ans.

Petite remarque sur l'affichage des résultats : premièrement, les six graphiques sont présentés ici au moyen d'un axe « y » variable pour montrer la fréquence d'apparition d'un sujet au cours d'une séance au Parlement. Puisque j'ai choisi de modifier l'échelle de l'axe « y » par souci de visibilité, notez les valeurs utilisées; deuxièmement, les mots qui se trouvent dans les résultats de sujets ne sont pas traduits. Les recherches en texte intégral dans le hansard en version française pourraient donner des résultats de sujets légèrement différents. Ces graphiques représentent donc uniquement les sujets dans la version anglaise du hansard, et l'expérience devrait être réalisée séparément dans le hansard en français par souci d'exactitude des résultats.

À mon avis, cet exemple permet de trouver de l'information très intéressante par la modélisation thématique. Un sujet, que l'on peut étiqueter « peace and peacekeeping », apparaît en premier dans l'analyse du hansard réalisée au moyen de l'outil MALLET (*voir fig. 1*). J'ai eu la curiosité de vérifier si, en établissant la fréquence d'apparition de ce sujet, je

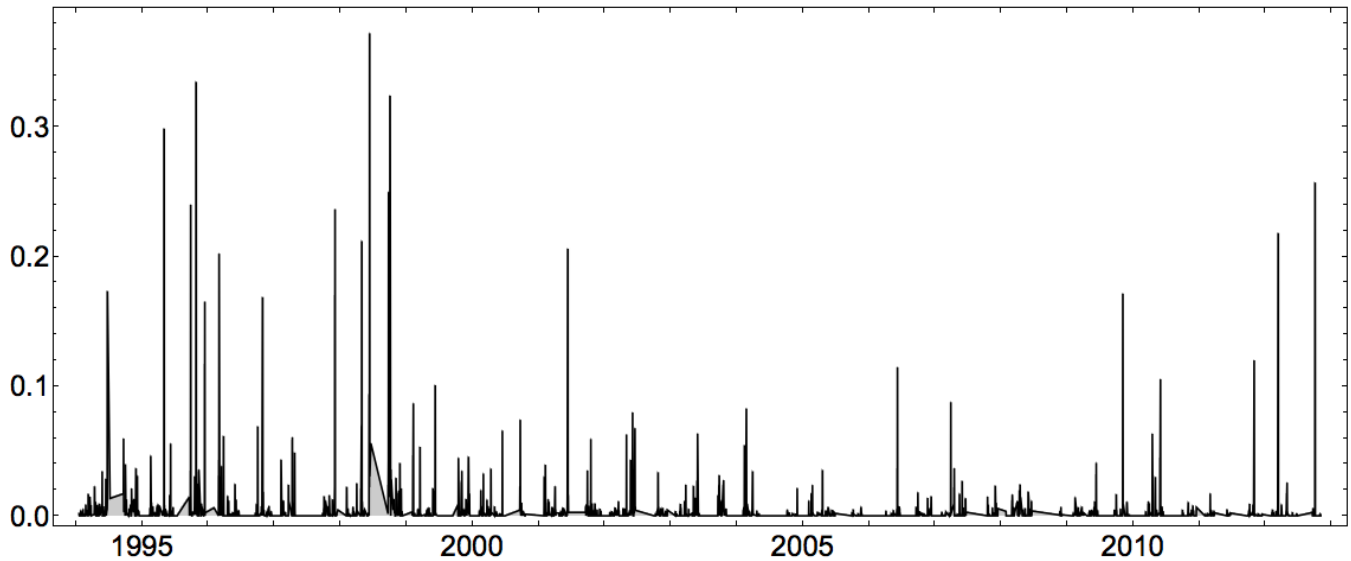

Fig. 1 : Aperçu de la fréquence relative du sujet dans différents segments du hansard : mots clés : « international Canada peace mr nato war world peacekeeping conflict troops nations united people kosovo situation humanitarian foreign role genocide » (international Canada paix monsieur otan guerre monde paix conflit troupes nations unies peuple kosovo situation humanitaire étranger rôle génocide). Il est à noter que ces mots clés sont plus fréquents, et de loin, avant qu'après 2000 (bien qu'on observe peut être plus récemment une recrudescence).

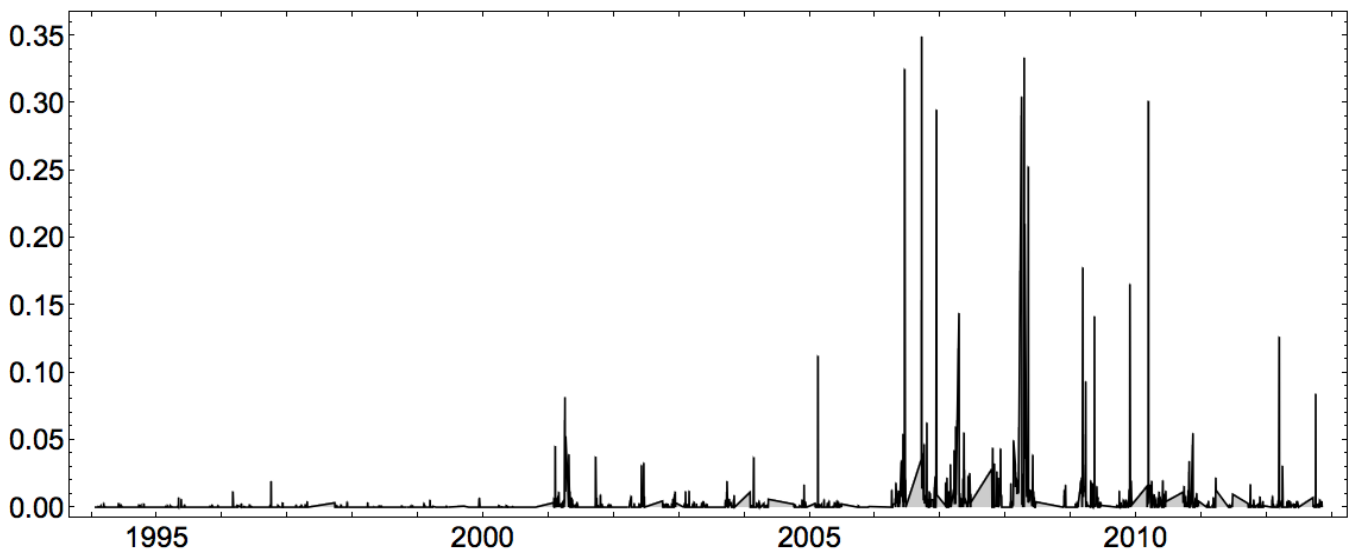

Fig. 2 : Aperçu de la fréquence relative du sujet : mots clés : « afghanistan mission canada canadian afghan mr minister government troops military security women defence forces international soldiers development motion support » (afghanistan mission canada canadienne afghan monsieur ministre gouvernement troupes militaire sécurité femmes défense forces internationales soldats développement motion soutien). Encore une fois, il est à noter que ces mots clés sont plus fréquents après 2001, et surtout après 2006. Par comparaison avec la figure 1, on peut observer une transition entre les deux jusqu'à un certain point.

pourrais vérifier une hypothèse avancée dans l'ouvrage récent *Warrior Nation*. Selon les auteurs Ian McKay et Jamie Swift, contrairement à l'idée reçue, le Canada serait davantage un pays militariste qu'un pays pacifique et gardien de la paix. Selon eux, les données montrent une transition de la paix vers la guerre dans nos stratégies commémoratives, les décisions prises concernant le nouveau guide de la citoyenneté destiné aux Néo Canadiens et plusieurs autres facettes de la société canadienne[7]. Un sujet constant de discussion entre les historiens au sein de la Société historique du Canada

et dans le cadre de tribunes de discussion historique telles que *ActiveHistory.ca* est le suivant : pourrait on également trouver des données à l'appui de cette hypothèse dans l'ensemble de données que constitue le hansard?

Puisque les outils de modélisation thématique génèrent *automatiquement* des sujets à partir de ces ensembles de données en texte intégral et qu'il faut interpréter les groupes de mots trouvés, le changement dans la fréquence d'apparition de 1994 à ce jour confirme à mon avis la thèse avancée dans *Warrior Nation*. On observe en effet une baisse notable de ce thème après l'élection des conservateurs, au début de 2006, mais les attentats du 11 septembre pourraient

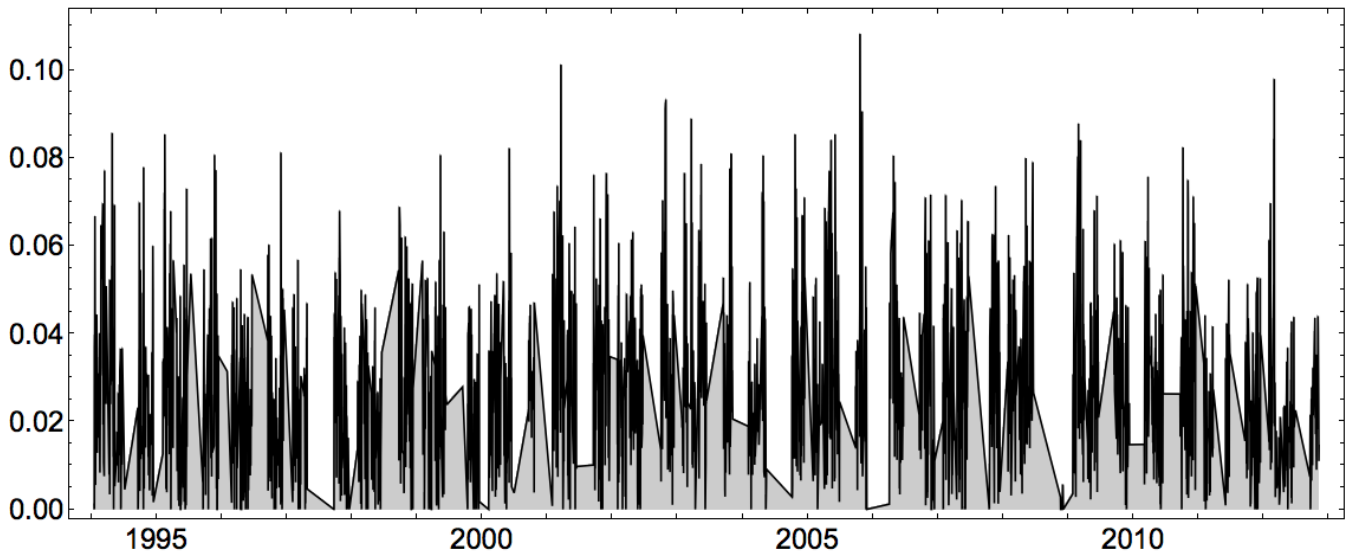

Fig. 3 : Cette figure montre la constitution globale des travaux parlementaires : mots clés : « committee mr report standing important parliamentary speaker work secretary process house issue recommendations review national made ensure information forward » (comité monsieur rapport permanent important parlementaire président travail secrétaire processus Chambre question recommandations examen national effectué assurer information transmise). Comme on peut s'y attendre, la fréquence des mots clés est relativement constante.

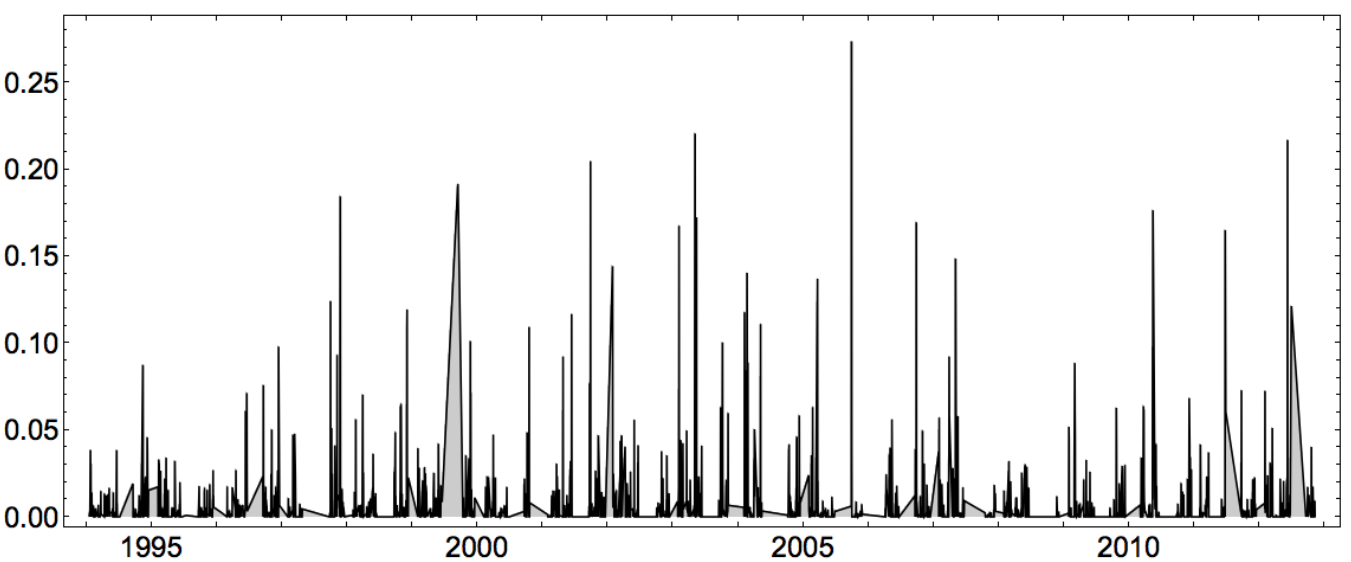

Fig. 4 : Aperçu de la fréquence relative du sujet : mots clés : « criminal code police sexual children offence mr law child person offences pornography justice dna age defence sex protect arrest » (code criminel police infraction sexuelle enfants monsieur loi enfants personnes infractions pornographie justice adn âge défense sexe protection arrestation). Malgré des fluctuations, la fréquence est relativement constante.

également être un tournant. On observe encore également des hausses. On ignore ce qu'elles signifient pour le moment, car elles pourraient être attribuables à des mentions arbitraires à la mission en Afghanistan ou à des événements précis. Voilà qui soulève d'autres questions pour la recherche. Un autre thème apparu pourrait également être pertinent à interpréter parallèlement à cette tendance : (*voir fig. 2*).

Un thème directement lié à la guerre en Afghanistan, mais plus globalement à la défense, apparaît ici. On l'observe d'abord brièvement au cours des années 1990, mais plus fréquemment au début de l'année 2001 à la suite d'une augmentation des nouvelles sur les talibans et de la présence canadienne à la guerre en Afghanistan. En guise de comparaison, le premier thème est prédominant au début de la période à l'étude, tandis que le second thème l'est davantage vers la fin de la période. On peut certainement observer une transition entre le thème de la paix et du maintien de la paix et le thème associé à la présence militaire

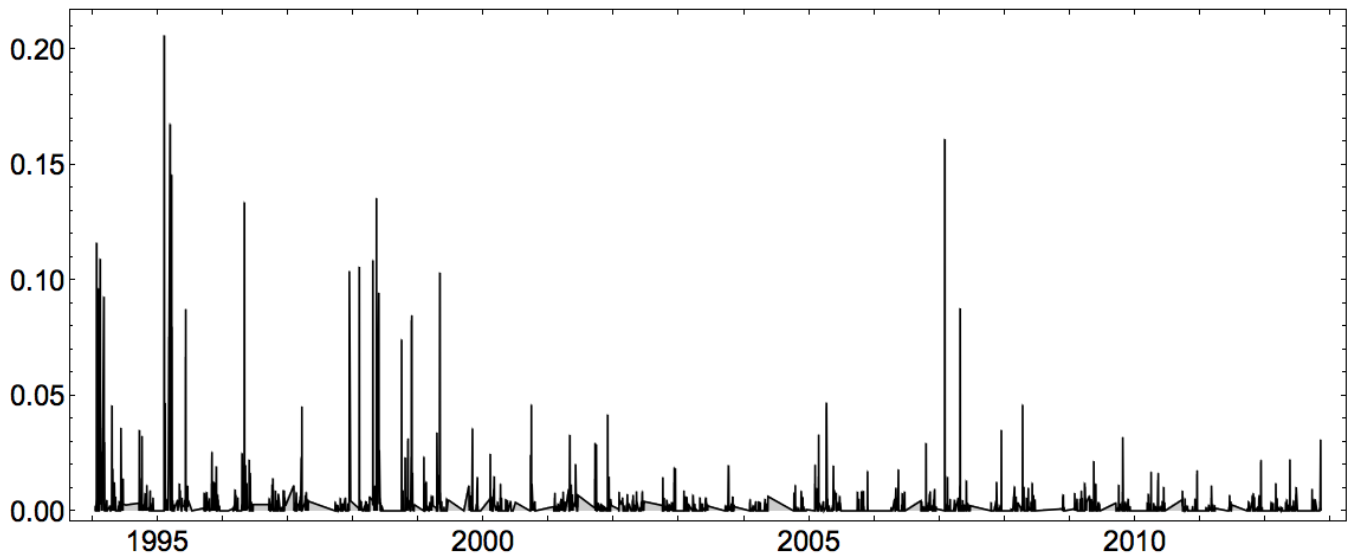

Fig. 5 : Aperçu de la fréquence relative du sujet : mots clés : « canadian cultural heritage canada culture flag canadians minister industry country mr arts national department world museums film artists quebec » (patrimoine culturel canadien canada culture drapeau canadiens ministre industrie pays monsieur arts national ministère monde musées film artistes québec).

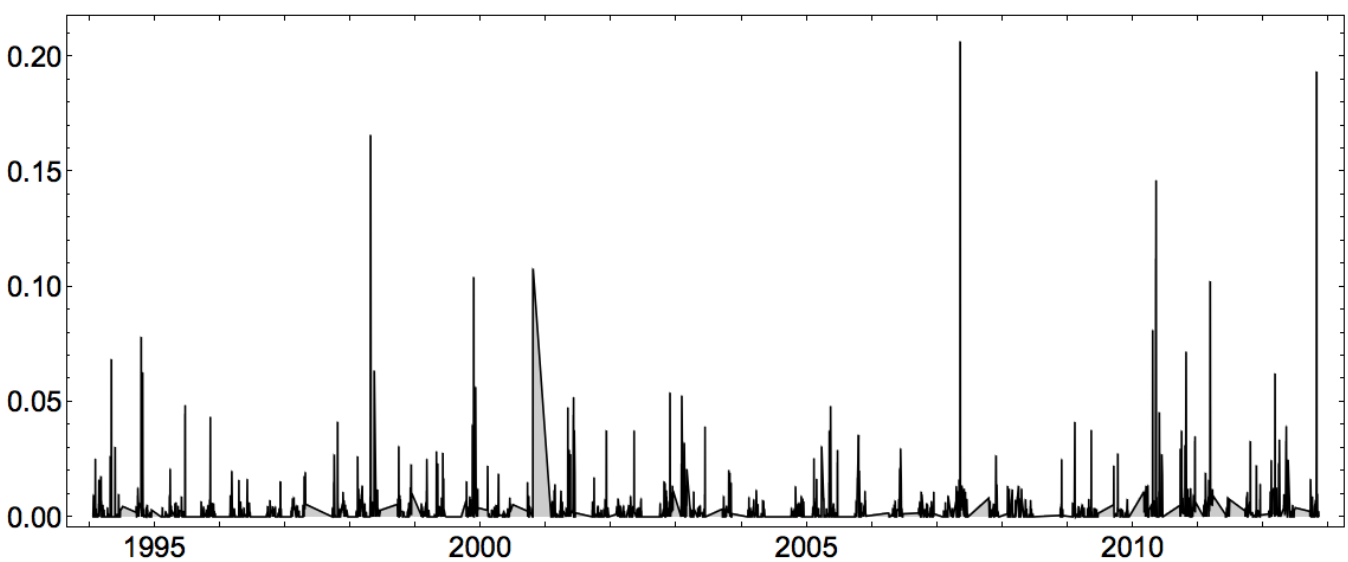

Fig. 6 : Aperçu de la fréquence relative : mots clés : « veterans war affairs canadian service mr benefits day world men services support speaker member country forces remembrance committee served » (anciens combattants affaires canadiens services monsieur bienfaits jour mondial hommes services soutien président député pays forces commémoration comité service). On observe des hausses pendant les événements commémoratifs, mais le sujet a pris de l'ampleur depuis 2010.

en Afghanistan, mais il faudrait pousser les recherches pour établir une corrélation ou voir si l'on peut attribuer cela à la thèse avancée dans Warrior Nation.

D'autres thèmes apparus dans l'analyse de modélisation en texte intégral du hansard sont également dignes d'être explorés. Un thème qui regroupe des termes probablement rattachés aux affaires parlementaires courantes est une constante (*voir fig. 3*). Cependant, deux thèmes susceptibles d'être rattachés aux budgets semblent dénoter un changement dans le discours. Ici, la fréquence d'un thème auquel sont rattachés des termes de nature générale du domaine budgétaire diminue de façon notable après 2006. Ce thème semble être remplacé par un autre thème lié au récent Plan d'action économique du Canada, surtout à partir de 2009. Les mots clés rattachés à ce thème comprennent les suivants : « budget économique emplois économie canada plan fiscal monsieur canadien canadiens gouvernement mesures action entreprises soutien crédit crise mondiale finance ».

D'autres thèmes semblent également notables. La protection des enfants dans le contexte des infractions criminelles visant les jeunes apparaît comme une préoccupation constante dans les débats au Parlement (*voir fig. 4*).

Un thème que l'on pourrait étiqueter « patrimoine » (*voir fig. 5*) semble être moins présent, bien que l'on observe des hausses au cours de la période entourant le référendum sur la souveraineté du Québec et dans le cadre des débats concernant la Loi sur la clarté qui en ont découlé. Toutefois, la fréquence d'apparition d'un thème éventuellement connexe et rattaché à la commémoration est en hausse depuis le début de 2010 (*voir fig. 6*).

Bien que ces exemples n'offrent qu'un survol de certaines possibilités, ce type d'outils permet d'envisager les débats dans leur intégralité plutôt qu'en pièces détachées. Cette perspective nous amènera à concevoir un peu différemment l'histoire du Parlement.

Données ouvertes et professions des candidats comme députés fédéraux

Examinons un autre fichier : « Historique des circonscriptions, 1867-2010 ». Accessible en anglais et en français à http://ouvert.canada.ca/data/fr/dataset/ea8f2c37-90b6-4fee-857e-984d3060184e, ce volumineux fichier contient de l'information sur les 38 778 candidats aux élections fédérales canadiennes. Il se présente sous la forme d'un fichier de 13 colonnes dont les valeurs sont séparées par des virgules, comportant les champs suivants :

- Date d'élection, Type d'élection, Parlement, Province, Circonscription, Nom, Prénom, Sexe, Occupation, Parti, Votes, Votes (en %), Élu(e).

Les données sont ensuite simplement présentées ligne par ligne en format texte, par exemple :

- 2008 10 14, Gen, 40, Québec, PAPINEAU, Trudeau, Justin, M, enseignant, Libéral, 17724, 41.47, 1

Interprétation des données de gauche à droite : première élection de Justin Trudeau, au cours de la 40e législature, à une élection générale, 17 724 votes (41,47 % du compte de vote total), et élu (indiqué par la valeur « 1 » dans la colonne Élu(e)). Les fichiers en format CSV sont d'une grande utilité pour les chercheurs, car ils sont lisibles dans tous les logiciels, dont Microsoft Excel, au moyen de tous les langages de programmation, ou dans Google Docs. Je reviens sur leur importance dans la conclusion du présent article.

J'ai réussi, au moyen d'un langage de programmation, à manipuler ces valeurs. L'une des occupations m'a paru d'un intérêt particulier, celle d'« avocat ». En extrayant les professions les plus fréquentes, j'ai obtenu ce qui suit :

Tableau 1 : Occupations des candidats

lawyer (avocat)	3730
farmer (agriculteur)	2587
Null	2308
teacher (enseignant)	1415
merchant (marchand)	1194
businessman (homme d'affaires)	1125
physician (médecin)	999
barrister (avocat)	981
parliamentarian (parlementaire)	816
student (étudiant)	795
journalist (journaliste)	497
retired (retraité)	476
manufacturer (industriel)	425
manager (gérant)	355
Member of Parliament (député)	351
administrator (administrateur)	298
accountant (comptable)	271
consultant (consultant)	267
contractor (entrepreneur)	267
notary (notaire)	224
engineer (ingénieur)	223
housewife (femme au foyer)	196
salesman (vendeur)	195
agent insurance (agent d'assurance)	190
professor (professeur)	184
secretary (secrétaire)	179
editor (rédacteur)	164
-at+barrister-law (avocat)	163
educator (éducateur)	145
broker insurance (courtier d'assurance)	144

Il est à noter que les données ne sont pas parfaites (elles ne le sont *jamais*) : il y a 2 308 occurrences de « Null » pour l'occupation, ce qui veut dire que rien n'a été entré dans le champ. Cette lacune est principalement attribuable au manque d'uniformité des données ou à l'absence de données sur les candidats défaits avant la 14e législature. Néanmoins, certaines des professions qu'on s'attend à voir apparaissent effectivement : avocat, agriculteur, enseignant, marchand, homme d'affaires, médecin, etc.

En survolant les données, cependant, on constate un autre problème : les occupations de « marchand » et d'« homme d'affaires » pourraient être considérées comme faisant partie de la même catégorie. De même, la profession d'avocat apparaît sous diverses appellations en anglais : « lawyers », « solicitors », « barristers », et même « -at+barrister-law ». Le manque d'uniformité dans les données n'est pas anormal, et il faut décider à toutes les étapes comment les interpréter. Des gens créent des données, et d'autres – des historiens et des politologues, par exemple – les interprètent. Il faut être très vigilant avant de prendre ces données au pied de la lettre, d'autant plus que certains députés réélus semblent avoir simplement écrit « député » ou « parlementaire » après leur réélection. Toutes ces conditions montrent à quel point il est important d'analyser les données au lieu de compter sur les portails. Le programme Google Refine permet de préciser les données si on le souhaite ou de les explorer manuellement. Les données ne sont pas neutres; elles sont créées par des humains dans des conditions subjectives.

Pour revenir aux avocats, quelle est la fréquence de cette occupation parmi les candidats? Plus particulièrement, réussissent ils plus que les autres candidats à se faire élire, et dans une plus grande proportion? Ils étaient nombreux à se porter candidats au XIXe siècle, et ils le sont encore aujourd'hui.

J'ai généré deux graphes à partir de la 14e législature (le moment où les données se sont améliorées) jusqu'à aujourd'hui. Il est à noter que je n'ai pas manipulé les données sur les élections partielles des différentes législatures. Examinez les figures 7 et 8 (l'axe « x » est celui des législatures).

Dans ce graphe, on constate que, au cours de la 14e législature, près de 11 % de tous les candidats, dont l'occupation figurait dans la liste, ont inscrit être avocat (en anglais, certains ont inscrit être « solicitor », mais la très

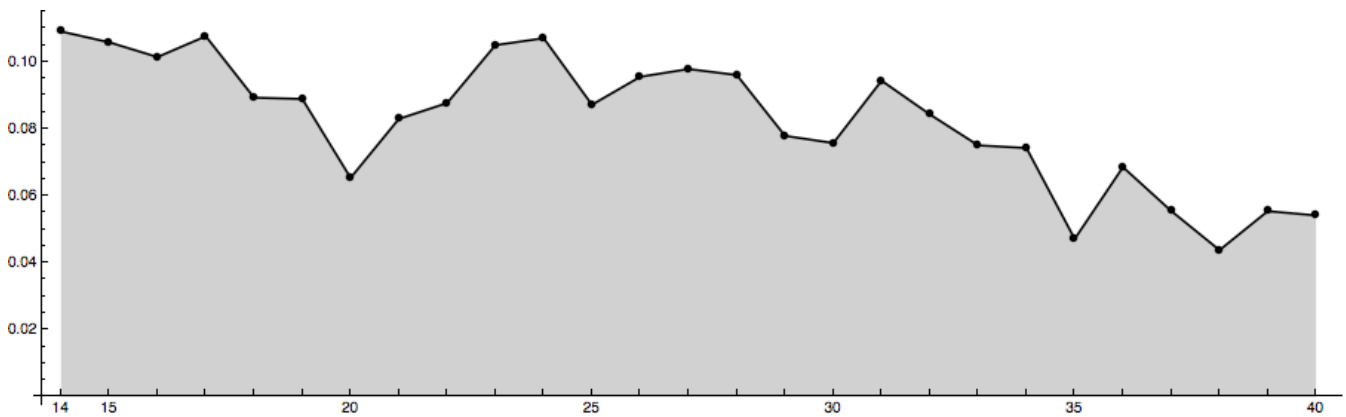

Fig. 7 : Occurrences de la profession d'« avocat » dans toute la liste des professions des candidats, de la 14e à la 40e législature.

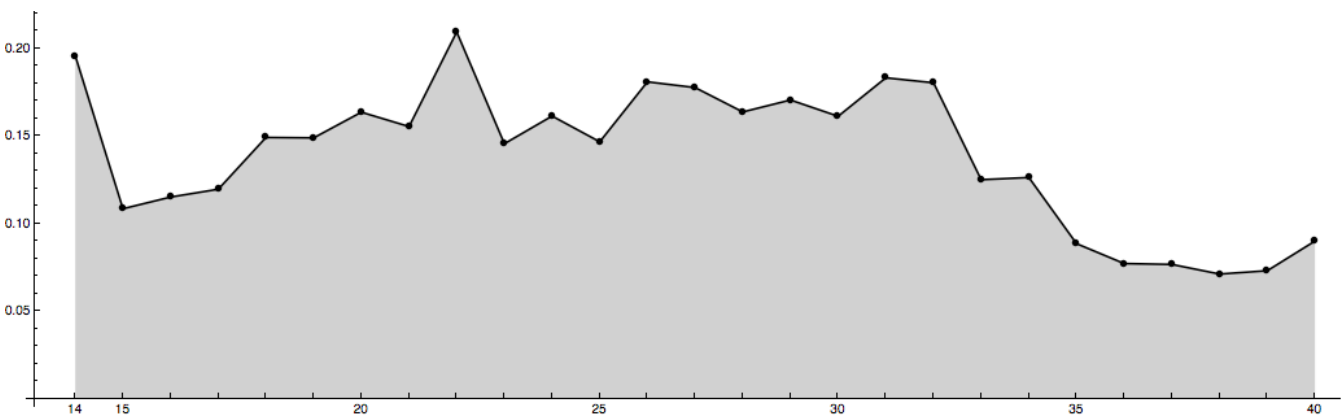

Fig. 8 : Occurrences de la profession d'« avocat » parmi les candidats élus, de la 14e à la 40e législature.

vaste majorité a inscrit « lawyer »). Pourtant, si on élimine tous les candidats défaits, on constate que près de 20 % des candidats élus au cours de cette législature étaient avocats.

Depuis cette époque, le nombre de parlementaires qui sont avocats de profession semble avoir chuté de façon considérable : environ 9 % des candidats élus au cours de la 40e législature ont inscrit « avocat » comme profession. Fait à noter, cependant, on se fie aux données des tableaux examinés plus tôt. Il se peut donc qu'un bien plus grand nombre d'avocats aient inscrit être « hommes » ou « femmes d'affaires », ou encore simplement « parlementaires » s'ils aspiraient à se faire réélire.

Néanmoins, les données, aussi imparfaites soient elles pour obtenir des statistiques parfaites, peuvent servir à brosser un portrait global des candidats et du type de personnes enclines à se porter candidates pour les divers partis. Voyons, par exemple, quelles sont les 50 occupations les plus fréquentes parmi toutes les professions occupées par les candidats du Parti libéral à partir de 1962, comparativement à celles des candidats du Nouveau Parti démocratique au cours de la période. J'ai choisi le Parti libéral et le Nouveau Parti démocratique depuis cette date en raison de la constance relative de leur composition, le Parti conservateur actuel ayant subi plusieurs réaménagements au cours de la même période. Les données qui en résultent sont très révélatrices de la formation des deux partis.

Tableau 2 : 50 Principales occupations des candidats du parti Libéral de 1962 à aujourd'hui

lawyer (avocat)	737
parliamentarian (parlementaire)	412
businessman (homme d'affaires)	251
farmer (agriculteur)	212
Member of Parliament (député)	142
teacher (enseignant)	138
administrator (administrateur)	82
consultant (consultant)	71
politician (politicien)	68
physician (médecin)	56
barrister (avocat)	56
merchant (marchand)	54
manager (gérant)	53
economist (économiste)	52
accountant chartered (comptable agréé)	49
accountant (comptable)	44
journalist (journaliste)	43
professor (professeur)	41
retired (retraité)	38
engineer (ingénieur)	37
manufacturer (industriel)	36
businesswoman (femme d'affaires)	31
broker insurance (courtier d'assurance)	31
educator (éducateur)	30
barrister and solicitor (avocat)	29
business person (homme d'affaires)	27
broadcaster (annonceur)	26
NULL	25
principal school (directeur d'école)	25
public servant (fonctionnaire)	24
agent insurance (agent d'assurance)	22
director executive (directeur exécutif)	21
cabinet minister (ministre)	21
publisher (éditeur)	20
notary (notaire)	19
contractor (entrepreneur)	19
consultant management (conseiller en administration)	18
housewife (femme au foyer)	17
engineer professional (ingénieur)	16
-at+barrister-law (avocat)	16
mayor (maire)	16
executive (gérant)	15
business executive (chef d'entreprise)	14
doctor medical (médecin)	13
student (étudiant)	13
social worker (travailleur social)	12
clergyman (membre du clergé)	12
veterinarian (vétérinaire)	11
realtor (courtier immobilier)	11
manager sales (gérant des ventes)	11

Tableau 3 : 50 principales occupations des candidats Nouveau Parti démocratique de 1962 à aujourd'hui

teacher (enseignant)	484
student (étudiant)	192
lawyer (avocat)	179
farmer (agriculteur)	150
professor (professeur)	71
retired (retraité)	70
representative union (représentant syndical)	69
social worker (travailleur social)	52
parliamentarian (parlementaire)	51
Member of Parliament (député)	48
journalist (journaliste)	43
businessman (homme d'affaires)	43
administrator (administrateur)	38
consultant (consultant)	37
professor university (professeur d'université)	37
housewife (femme au foyer)	36
electrician (électricien)	34
economist (économiste)	33
NULL	32
secretary (secrétaire)	31
educator (éducateur)	31
representative (représentant)	31
physician (physicien)	29
clergyman (membre du clergé)	29
high school teacher (enseignant au niveau secondaire)	27
salesman (vendeur)	27
researcher (chercheur)	25
school teacher (instituteur)	23
writer (écrivain)	22
manager (gérant)	22
-employed+self (travailleur autonome)	20
minister (ministre)	19
organizer (organisateur)	18
steelworker (ouvrier de l'acier)	18
machinist (machiniste)	17
business manager (gérant)	17
agent business (agent d'affaires)	16
trade unionist (syndicaliste)	16
engineer (ingénieur)	16
clerk (commis)	16
accountant (comptable)	14
contractor (entrepreneur)	14
college instructor (professeur)	13
assistant executive (adjoint exécutif)	13
instructor (instructeur)	13
director executive (directeur général)	12
unemployed (chômeur)	12
nurse (infirmier)	12
driver truck (camionneur)	12
sociologist (sociologue)	12

Bien que je ne sois pas politologue, il m'a suffi de quelques minutes de remaniements pour générer des données de qualité et utiles sur les débats et la composition du Parlement fédéral ainsi que les candidats aux élections fédérales. Je présente ces données malgré leurs lacunes parce qu'elles montrent, encore une fois, que les données doivent être prises avec parcimonie : selon ces données, par exemple, « enseignants au secondaire » et « enseignants » sont considérés comme deux professions distinctes. Cela pourrait aider un chercheur, mais en importuner bien d'autres.

Outre les comptes rendus du Parlement, il existe bien d'autres ensembles de données qui pourraient être dignes d'intérêt à divers chercheurs, notamment les certificats de naissance, pour les noms les plus populaires de bébés, les certificats de mariage, pour les noms de villes et de villages et le nom de tous les soldats qui se sont enrôlés dans le Corps expéditionnaire canadien. Les possibilités d'études sont presque illimitées.

À quelles fins ces données devraient-elles servir?

Les ensembles de données comportent un potentiel énorme pour transformer les pratiques de recherche, mais on n'a pas encore pris conscience de toute la valeur que sont ces riches sources d'information. Les universitaires et les législateurs doivent prendre en considération les aspects qui suivent avant d'entreprendre des études sur les bases de données.

Premièrement, il peut être difficile de réaliser des études interdisciplinaires au Canada. Cette année, le Conseil de recherches en sciences humaines a décidé d'abandonner les « domaines prioritaires ». Les demandes de subvention concernant les applications numériques auraient auparavant été soumises à un comité d'« économie numérique », alors qu'elles sont maintenant examinées par des pairs de la discipline. Le jury ignore si un tel changement sera positif ou négatif, mais il me semble que l'utilisation, qui est en pleine transformation, des nouveaux médias et des nouvelles

technologies devrait être examinée par des comités qui y sont étroitement liés. Des chercheurs traditionnels adhèrent aux technologies, alors que d'autres les ont ouvertement rejetées. Le principal problème tient au fait que les projets numériques mobilisent souvent des équipes interdisciplinaires, qu'il s'agisse de chercheurs anglais qui ont adopté la lecture à distance, des informaticiens qui connaissent les tenants et aboutissants des algorithmes beaucoup plus que ne le peuvent les chercheurs en sciences humaines. Les historiens travaillent généralement seuls, ce qui fait en sorte qu'il peut être difficile d'évaluer le travail dans le cadre des projets faisant appel à des équipes nombreuses et multiples. Il faut se méfier des obstacles institutionnels à l'adoption du numérique, particulièrement en raison de leurs implications pour l'embauche, le maintien en poste et la promotion dans le milieu de la recherche.

Par l'entremise des organismes subventionnaires, les gouvernements peuvent contribuer à façonner la recherche future. Les chercheurs devraient être des chefs de file en matière de recherche, tout en respectant les principes de la liberté universitaire et de l'exploration thématique, mais ils fonctionnent au sein des structures établies par les gouvernements.

Nous devrions également favoriser la publication d'un plus grand nombre de données et prendre conscience que, lorsqu'elles deviennent accessibles, elles doivent être lisibles par machine (par exemple, dans des fichiers de texte brut ou des tableaux dans lesquels les valeurs sont séparées par des virgules. On peut créer des interfaces de programmation d'applications (API) complexes, qui sont en quelque sorte des couches d'accès aux données d'un ensemble de données, mais l'idéal est souvent de laisser simplement les chercheurs *télécharger* eux mêmes les données (en respectant les contraintes de sécurité, évidemment). Je me réjouirais si, lorsqu'on créé des ensembles de données, on se demandait s'il était possible de laisser quiconque télécharger les données. Le cas échéant, pourquoi ne pas mettre un gros bouton rouge tout en haut à partir duquel on exporterait les données? Un chercheur peut toujours rêver.

En conclusion, je crois qu'il est important de souligner que les travaux de recherche de ce genre s'accéléreront. Dans le cadre de mon projet de recherche actuel, je me penche sur l'utilisation que les historiens pourront faire des archives Web. Je suis fermement convaincu qu'on ne peut mener de recherche sur l'histoire des années 1990 ou 2000 ni écrire l'histoire de cette période *sans* avoir recours aux archives Web. Tous n'écriront pas l'histoire du Web, bien entendu, mais le contenu du Web est un pan inestimable des archives historiques. Les chercheurs qui étudient une élection récente doivent s'intéresser aux billets sur les babillards électroniques, aux sites Web électoraux, aux gazouillis et aux vidéos, notamment. Ils font tous partie intégrante des archives.

Les années 1990 font maintenant partie de l'histoire passée. Les étudiants qui écriront l'histoire de cette période viennent probablement à peine d'entamer leurs études postsecondaires. Pourront ils avoir recours aux archives Web? Surtout, pourront ils les utiliser par des méthodes computationnelles? Il est, après tout, impossible de lire tous les sites Web – si on croyait qu'il y avait trop de romans victoriens, on peut imaginer le nombre de gazouillis mis en ligne chaque jour. Il faut poser les fondements de la littérature numérique pour la prochaine génération.

Les données existent. On a maintenant besoin d'une génération qualifiée de chercheurs en sciences humaines qui se posent des questions intéressantes, qui sont capables de manipuler des données et qui peuvent participer à l'entrée de la recherche en sciences humaines menée au Canada dans le XIX^{e} siècle. Alors que les historiens se tournent de plus en plus vers les sources en lignes tels que *Programming Historian*, qu'ils bloguent et qu'ils s'intéressent aux données, la nature même de leur profession sera amenée à changer en conséquence. Il est à espérer que les gouvernements continueront eux aussi d'appuyer la recherche en sciences humaines numériques en rendant accessibles des ensembles de données qui soient intéressants de façon à en maximiser l'utilité pour les chercheurs, présents et futurs.

Notes

1 http://ouvert.canada.ca/fr/mise-en-oeuvre-du-plan-daction-du-canada-pour-un-gouvernement-ouvert-annee-1-rapport-dautoevaluation

2 Franco Moretti, « Graphs, Maps, Trees: Abstract Models for Literary History », Verso, 2007.

3. http://www.parl.gc.ca/housechamberbusiness/ChamberSittings.aspx?View=H&Mode=1&Parl=41&Ses=1&Language=F

4 Je demande souvent aux historiens qui affirment ne pas être des historiens du numérique s'ils effectuent leur recherche dans Google et, le cas échéant, s'ils connaissent le fonctionnement de PageRank. Pour en connaître davantage sur cet outil, consulter : Ted Underwood, « Theorizing Research Practices We Forgot to Theorize Twenty Years Ago », Representations, vol. 127, n° 1, août 2014, p. 64–72, doi:10.1525/rep.2014.127.1.64.

5 Le concept est décrit dans David M. Blei, Andrew Y. Ng, et Michael I. Jordan, « Latent Dirichlet Allocation », Journal of Machine Learning Research, vol. 3, 2003, p. 993–1022. On peut également trouver une excellente explication dans Matthew L. Jockers, « The LDA Buffet Is Now Open; Or, Latent Dirichlet Allocation for English Majors », Matthew L. Jockers Blog, 29 septembre 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 et Ian Milligan, « Getting Started with Topic Modeling and MALLET », *Programming Historian*, 2 Septembre 2012, http://programminghistorian.org/lessons/topic-modeling-and-mallet.

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