Une histoire par les données ?
Le futur très proche de l'histoire des relations internationales

FRÉDÉRIC CLAVERT

Résumé
Le monde de données dans lequel nous vivons engendre l'émergence de grands ensembles de données dont certains peuvent être utiles à l'histoire des relations internationales, sous condition d'un élargissement du canon méthodologique des historiens par l'appropriation de la notion de lecture distante. Toutefois, l'interrogation majeure de la pratique de l'histoire des relations internationales à l'ère numérique ne concerne pas tant les méthodes que les sources primaires, « numérisées » comme « nées numériques ».
Mots clés : Histoire numérique – Données massives – Lecture distante – Sources primaires – Mise en données.

Abstract
A data driven history? The close future of international history
The data-driven world in which we are living is based on the emergence of large data sets, some of which may be useful for the practice of international history, if historians are able to broaden their methodologies. It notably implies the appropriation of the concept of distant reading. However, the most important question that challenges international history in the digital age is concerning primary sources, whether « digitalized » or « digital born ».
Keywords: *Digital history – Big data – Distant reading – Primary sources – Datafication.*

En 2011, la revue *Science* publie un article retentissant : « Quantitative Analysis of Culture Using Millions of Digitized Books[1] ». Cette étude menée par une équipe en partie issue de la firme Google utilise un corpus de

[1] Frédéric Clavert est maître assistant en section d'histoire à l'Université de Lausanne. Il est membre du bureau du *Laboratoire de cultures et humanités digitales de l'Université de Lausanne* (LaDHUL).

plusieurs millions de livres numérisés, Google Books[2]. Cet article revendique une histoire culturelle sans historiens, laissant les données « brutes » parler d'elles-mêmes. Bien que de sérieux problèmes méthodologiques en limitent largement la portée, ce texte rappelle que l'émergence d'ensembles massifs de données historiques de natures variées appelle à la mise au point de nouvelles méthodologies. La question qui se pose est la suivante : comment « faire » de l'histoire dans un monde de données ? Pour le cas de l'histoire des relations internationales, nous analyserons les possibilités offertes par la « mise en données » du monde, puis nous nous pencherons sur les sources passées et futures de l'histoire des relations internationales à l'ère numérique.

Les possibilités de la mise en données de l'histoire

La mise en données, traduction que nous proposons du terme *datafication*, se définit ainsi : « To datafy a phenomenon is to put it in a quantified format it can be tabulated and analysed[3] ». Appliquée aux usages historiens, la mise en données est un processus contenant toutes les étapes allant de la numérisation d'un artefact physique – les archives numérisées – ou de la captation d'un signal – les sources primaires nées numériques – aux possibilités d'analyser ces artefacts et signaux *via* des outils informatiques, c'est-à-dire une lecture des sources au moyen d'une médiation informatique. La mise en données du monde, telle qu'elle est ici définie, se traduit par l'émergence de grands ensembles de données à disposition des chercheurs, dont Google Books[4] ou Gallica[5]. Comment, en tant qu'historiens, les exploiter au mieux ?

---

2 Jean-Baptiste Michel *et al.*, « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science,* n° 331 (6014), décembre 2010, p. 176-182.

3 « Mettre en données un phénomène revient à le traduire en un format quantifié, qui peut faire l'objet d'un calcul et être analysé », *in* Viktor Mayer-Schönberger et Kenneth Cukier, *Big data: a revolution that will transform how we live, work, and think*, Boston, Houghton Mifflin Harcourt, 2013, p. 72.

4 http://books.google.com [consulté le 2 mars 2016].

L'une des réponses méthodologiques est la notion de « lecture distante » avancée par Franco Moretti : « […] what we really need is a little pact with the devil : we know how to read texts, now let's learn how not to read them[6]. » Cette provocation résulte des travaux de ce dernier sur la littérature européenne des XVIII^e^ et XIX^e^ siècles : doit-on se contenter de faire l'histoire des « grands textes » ou peut-on y intégrer l'histoire de toute la littérature, textes mineurs et oubliés compris ? Dans le second cas, le nombre de sources devient trop important pour se restreindre à une lecture « humaine ». Il faut alors avoir recours à une médiation logicielle, c'est-à-dire à des programmes informatiques permettant à l'ordinateur de lire pour nous[7].

La notion de lecture distante et l'appel à ne plus lire les sources, ou plutôt à demander à l'ordinateur de les lire pour nous, interpellent tous les historiens. Qu'en est-il de l'histoire des relations internationales ? Serons-nous amenés à ne plus lire nos sources ? La réponse est complexe. Si Moretti doit être pris au sérieux, sa thèse n'implique pas d'abandonner toute lecture proche et critique des sources, mais bien de varier nos pratiques, c'est-à-dire d'alterner lecture distante et lecture proche et d'intégrer à notre canon méthodologique des éléments numériques.

Dans le cadre, par exemple, d'une recherche sur le groupe Werner sur la base d'archives numérisées et publiées en ligne[8], abstraction faite d'obstacles méthodologiques que nous aborderons plus bas, nous avons utilisé des techniques dites d'analyse de texte, c'est-à-dire un ensemble de méthodes statistiques appliquées au texte qui permet une lecture distante.

---

5 http://gallica.bnf.fr/ [consulté le 2 mars 2016].

6 « Ce dont nous avons besoin est un petit pacte avec le diable : nous savons lire les textes, nous devons maintenant apprendre à ne pas les lire. », in Franco Moretti, *Distant Reading*, London/New York, Verso Books, 2013.

7 Approche détaillée dans Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*, London/New York, Verso, 2007.

8 *Une relecture du rapport Werner du 8 octobre 1970 à la lumière des archives familiales Pierre Werner*, http://www.cvce.eu/recherche/unit-content/-/unit/ba6ac883-7a80-470c-9baa-8f95b8372811 [consulté le 2 mars 2016].

L'usage de ces techniques[9] nous a permis de dégager les grands sujets de discussions abordés par les membres du groupe Werner et de les projeter dans le temps, afin d'établir une chronologie thématique des travaux du groupe Werner sur l'union économique et monétaire au cours de l'année 1970. Pour arriver à ce résultat, la sélection des sources à inclure dans notre corpus et la définition de leurs métadonnées[10] nous a permis d'acquérir une connaissance poussée de cet ensemble d'archives. Ainsi, l'ambition même d'une lecture distante des sources primaires a engendré une lecture proche et critique. De manière plus générale, la lecture proche des sources avec médiation informatique pourrait être nettement améliorée, car la mise en données des sources permettrait d'intégrer par exemple l'historique d'une source – les notes de préparation, les échanges autour du document, ses différentes versions – en une seule interface facilitant une compréhension globale du processus d'écriture de la source, et, au-delà, de la procédure de prise de décision dans les relations internationales.

L'exploitation des archives Werner a soulevé d'autres questions méthodologiques, liées non pas aux techniques de lecture distante, mais à leur application à des sources dont la mise en données n'a pas été des meilleures. L'interrogation sur les méthodes de l'historien à l'ère numérique se déplace alors vers les questionnements liés au processus de mise en données des sources.

---

9 Nous utilisons IRaMuTeQ, http://www.iramuteq.org [consulté le 2 mars 2016], implémentation de la méthode exposée dans Max Reinert, « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars" », *Langage et société*, n° 66 (1), 1993, p. 5-39.

10 Les métadonnées sont les données qui décrivent les autres données. Elles peuvent comprendre l'auteur d'un document, sa date de publication, son destinataire, etc.

Mise en données des sources primaires de l'histoire : biais et interrogations

Notre définition de la « mise en données » des sources historiennes insiste sur la notion de processus, afin de rappeler que la « numérisation » se définit par les choix opérés par les organismes qui y procèdent : les critères d'inclusion de documents dans le corpus à numériser ; la définition des métadonnées ; les éléments du document qui sont numérisés[11] ; le mode d'accès aux sources numérisées ; l'encodage des « entités nommées » – noms propres, lieux, traces de temps, etc. – pour faciliter la recherche ; la mise à disposition d'outils pour l'analyse des documents numérisés ; les technologies assurant la numérisation, l'interopérabilité et la pérennité des données numérisées. Un tel processus de mise en données ne peut qu'interroger la communauté historienne, car il touche en profondeur à notre matière première, les sources primaires. Or, ce processus peut biaiser nos recherches de deux manières : par ce qui n'est pas numérisé ; par ce qui est mal numérisé.

Les absences

Une analyse des mentions des journaux canadiens du XIX^e^ siècle dans les articles de la *Canadian Historical Review* a montré que les grands journaux « centraux » et anglophones, numérisés et disponibles pour les chercheurs, sont de plus en plus utilisés aux dépens des journaux francophones et/ou locaux, engendrant une vision centrale et anglo-saxonne de l'histoire canadienne[12]. Milligan parle ainsi d'un ordre illusoire

[11] Dans William J. Turkel, « Intervention: Hacking history, from analogue to digital and back again », in *Rethinking History*, n° 15 (2), juin 2011, p. 287-296. William Turkel prend l'exemple de lettres du XVIII^e^ siècle, vinaigrées lors des épidémies de choléra : ne pas numériser l'odeur des archives revient ici à perdre des informations.

[12] Ian Milligan, « Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010 », in *Canadian Historical Review*, n° 94 (4), décembre 2013, p. 540-569.

découlant de l'usage de larges bases de données donnant l'impression au chercheur d'avoir procédé à une recherche sur l'intégralité des sources, alors que ces bases sont souvent fortement lacunaires.

Cet « ordre illusoire » vaut pour l'histoire des relations internationales. Dans le cas de l'histoire de la construction européenne, les deux grandes bases de données disponibles sont entretenues par le Centre Virtuel de la Connaissance sur l'Europe (CVCE, aujourd'hui intégré à l'Université du Luxembourg)[13] et *Archive of European Integration* (*AEI*, Université de Pittsburgh)[14]. En utilisant Google Scholar, nous avons constaté que ces deux bases de données ont été utilisées très vite après leur mise en ligne et que le rythme de leur utilisation n'a cessé de croître[15].

Cet usage se fait au risque de la rigueur méthodologique. Les deux sites ne documentent ainsi que très faiblement leurs procédures, notamment car il est très difficile de fixer des critères objectifs d'inclusion d'un document dans un corpus à numériser. Prenons l'exemple de l'*AEI* : les documents numérisés sont des archives officielles des institutions européennes et de la littérature scientifique dite « grise ». La collecte se fait *via* les institutions contributrices et leurs chercheurs. Le matériel soumis doit être d'une utilité potentielle pour la recherche. L'ensemble de ces critères est flou et contraste avec l'édition des *Documents diplomatiques français*, par exemple, où l'on peut clairement identifier les historiens qui ont participé à leur élaboration, qui reposent sur des fonds d'archives cohérents et dont les critères de sélection sont, souvent, rappelés en début de volume. Ainsi, dans leur état actuel, il est important de voir les bibliothèques numériques

---

13 http://www.cvce.eu/ [consulté le 21 janvier 2016]. L'auteur de cet article a été salarié du CVCE.

14 http://aei.pitt.edu/ [consulté le 2 mars 2016].

15 Pour le détail, consulter Clavert Frédéric, « Les biais de la mise en données de l'histoire : lecture d'un article de Ian Milligan », billet de carnet de recherche, <http://histnum.hypotheses.org/2006> [mis en ligne le 14 janvier 2014, consulté le 4 mars 2016].

en ligne comme une première approche à compléter par un séjour en centre d'archives.

Dans le cas des archives diplomatiques, les sources les plus utilisées en histoire des relations internationales, certaines absences sont problématiques. Si les documents diplomatiques suisses – fort partiellement[16] – et les *Foreign Relations of the United States*[17] sont en ligne, ce n'est pas le cas de leurs équivalents français. Or, comme le montre l'article de Milligan pour les chercheurs ou un récent débat publié par l'*International Journal for Higher Education* pour les étudiants[18], ce qui n'est pas en ligne n'existe pas : il y a ici urgence à disposer d'une version en libre accès des documents diplomatiques français. Si, aujourd'hui, les volumes les plus récents sont disponibles en version électronique payante, nous ne pouvons les considérer comme « mis en données » dans la mesure où chaque document ne dispose pas de ses propres métadonnées.

Ce qui est mal numérisé

Un autre biais réside dans ce qui est mal numérisé. Comme l'a rappelé Tim Hitchcock[19], la reconnaissance de texte (OCR), c'est-à-dire le fait de passer d'un texte en image à un texte vu comme texte par l'ordinateur, est un élément fondamental de la mise en données des sources primaires historiennes. Ni le CVCE ni l'*AEI* ne livrent d'indications suffisantes sur le taux de réussite de la reconnaissance de texte. Or, le succès de cette dernière est déterminant : en cas de sélection des archives *via* une recherche « plein texte », les résultats ne seront satisfaisants qu'à la

16 http://www.dodis.ch/ [consulté le 26 février 2016].

17 https://uwdc.library.wisc.edu/collections/FRUS/ [consulté le 26 février 2016].

18 Gabriela Ossenbach, « If it's not online, it doesn't exist », *IJHE*, n° 5(1), 2015, p. 80-82.

19 Tim Hitchcock, « Academic History Writing and its Disconnects », in *Journal of Digital Humanities,* n° 1(1), Hiver 2011. En ligne : http://journalofdigitalhumanities.org/1-1/academic-history-writing-and-its-disconnects-by-tim-hitchcock/ [consulté le 27.06.2012].

condition que le texte soit parfaitement reconnu. De plus, non seulement les erreurs liées à l'OCR sont plus nombreuses pour les langues courantes autres que l'anglais, mais elles sont parfois presque insurmontables pour les langues dites « rares »[20]. La mauvaise qualité de cette mise en données a d'autres conséquences. Dans le cas de notre recherche sur le groupe Werner, elle a rendu la lecture distante de ce corpus inexploitable scientifiquement[21] : le chapitre issu de cette recherche a dû être transformé en note méthodologique.

Les sources nées numériques

Enfin, se pose également la question de nos archives futures, qui pour beaucoup seront nées numériques. Les deux biais déjà pointés jusqu'ici s'y appliqueront tout autant et risquent même de s'accentuer. L'une des grandes bases de données, nées numériques, utile à l'histoire récente des relations internationales est la *Global Database of Events, Language, and Tone* (GDELT). Elle rassemble, dans une base de données, « *the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages* ». La GDELT permet d'identifier les personnes, lieux, organisations, thèmes, émotions… afin d'analyser par traitement informatique le monde entier[22]. Fondée sur des sources médiatiques internationales, la GDELT est l'incarnation d'un ordre illusoire né

---

20 Voir pour le cas de l'hébreu : Gerben Zaagsma, « Using Digital Sources in Historical Research. Jewish History on the Internet », in Frédéric Clavert et Serge Noiret (dir.), *L'histoire contemporaine à l'ère numérique*, Bruxelles, PIE – Peter Lang, 2013.

21 Frédéric Clavert, « L'apport du numérique aux sciences historiques : exemple d'une analyse computationnelle des archives Werner », in Elena Danescu et Susana Munoz (dir.), *Pierre Werner et l'Europe : pensée, action, enseignements / Pierre Werner and Europe: His Approach, Action and Legacy*, Bruxelles, PIE - Peter Lang, 2015.

22 http://www.gdeltproject.org/ [consulté le 2 mars 2016].

numérique. Ainsi, en décembre 2013[23], sous prétexte qu'un « demi-milliard de clics ne peuvent mentir », une carte des risques pour l'année 2014 considérait la France comme pays hautement instable et la Syrie comme pays sûr. Les concepteurs de la carte ont ainsi ignoré que la GDELT est une base de données sur la perception du monde engendrée par les médias anglo-saxons et par les pratiques des internautes.

En outre, de nouvelles sources émergent. Les réseaux socionumériques en sont un exemple, notamment utiles pour étudier des mouvements d'opinion touchant à des sujets internationaux. Mais il faut espérer que ces sources soient disponibles : Facebook ne compte pas les mettre facilement à disposition des chercheurs et l'archive de Twitter constitué par la Bibliothèque du Congrès est pour le moment indisponible pour raisons techniques[24]. Au-delà de l'accès, se pose aussi la question de la qualité de la captation des données. Lorsqu'un réseau social n'autorise l'utilisateur qu'à « aimer » la publication d'un autre utilisateur, il est difficile d'interpréter la signification de ce « j'aime ».

Nos sources plus traditionnelles, comme les documents diplomatiques, sont amenées à devenir des archives nées numériques et massives. Si l'administration Johnson a produit environ 40 000 mémos, l'administration Clinton a engendré quatre millions de courriels[25]. La question de l'absence se reposera : l'administration française, dont les archives du Quai d'Orsay,

---

23 Kalev Leetaru, « Half a Billion Clicks Can't Be Wrong » in *Foreign Policy*, <https://foreignpolicy.com/2014/01/03/half-a-billion-clicks-cant-be-wrong/> [mis en ligne le 3 janvier 2014, consulté le 21.01.2016]. Cet article contient de nombreux détails sur la manière dont la GDELT est constituée.

24 Erin Allen, « Update on the Twitter Archive at the Library of Congress », Billet de blog, *Library of Congress Blog*, <http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/>, [mis en ligne le 4 janvier 2013, consulté le 3 mars 2016].

25 Cité par William J. Turkel, K. Kee, S. Roberts, « Navigating the infinite archive », in Toni Weller (*ed.*), *History in the digital age*, London – New York, Routledge, 2013, p. 62.

met au point un système d'archivage de sources nées numériques[26] mais l'*Auswärtiges Amt* archive les courriels en demandant à ses fonctionnaires de les imprimer sur papier[27]. Dans ce second cas, la perte d'information est sensible : l'ensemble des courriels reçus et émis ne peuvent être imprimés, forçant les diplomates à opérer des choix normalement dévolus aux archivistes.

L'émergence de grands ensembles de données touchant l'histoire des relations internationales place cette dernière à un tournant dont la nature est profondément méthodologique, riche en possibilités mais également en incertitudes. Comme l'a argumenté Matthew Connelly[28], l'avenir de l'histoire des relations internationales est susceptible d'être radieux. Connelly se projette dans les prochaines trente années de l'histoire des relations internationales et appelle de ses vœux la mise en place d'un agenda commun, notamment pour la création d'un dépôt centralisé d'archives nées numériques ou numérisées accessibles aux journalistes et chercheurs, utilisant des techniques informatiques avancées pour traiter de très grandes masses de sources. Nous adhérons à cette vision de l'histoire des relations internationales de l'avenir, malgré un désaccord ponctuel sur le rôle de la lecture proche dans nos recherches, qui nous semble sous-estimé par Matthew Connelly. Aux questions soulevées par la mise en données de nos sources primaires futures (« nées numériques ») et passées (« numérisées »), une réponse émerge par la notion de lecture distante qui est un enrichissement fondamental et nécessaire de nos

[26] « Vitam : vers un socle d'archivage électronique commun à toute l'administration », <http://www.modernisation.gouv.fr/ladministration-change-avec-le-numerique/par-son-systeme-dinformation/vitam-vers-un-socle-d-archivage-electronique-commun-toute-l-administration>, [mis en ligne le 18 mars 2015, consulté le 3 mars 2016].

[27] Cette information est issue d'une conversation de l'auteur avec un diplomate allemand.

[28] Matthew Connelly, « The Next Thirty Years of International Relations Research », *Les cahiers Irice,* n° 14 (2), juillet 2015, p. 85-97.

méthodes. Toutefois, nous aimerions rappeler que, si les techniques de lecture distante sont déjà aujourd'hui exploitables, il reste des interrogations majeures du côté des sources.

En conséquence, deux questions déterminantes sont désormais ouvertes et nécessitent une réponse qui nous semble urgente. La première interrogation est vieille de 25 ans[29] : quelle formation pour les historiens à l'ère numérique ? Si les opinions diffèrent[30], nous plaidons pour notre part pour une approche par la notion d'« alphabétisation numérique » passant par l'enseignement d'une culture numérique – plus large qu'une approche strictement informatique – qui permettrait aux historiens d'aborder non seulement avec aisance des logiciels de collecte et d'analyse de données mais aussi plus largement des pratiques numériques d'écriture et de lecture. Plus déterminante sera la capacité à travailler de manière interdisciplinaire avec des mathématiciens ou des informaticiens et à collaborer plus souvent et plus densément avec des corps de métiers liés à la recherche, bibliothécaires et archivistes notamment[31]. Ces derniers ont d'ailleurs souvent bénéficié d'une très bonne formation numérique[32].

---

29 Jean-Philippe Genet, « La formation informatique des historiens en France : une urgence », *Mémoire vive*, n° 9, juin 1993 ; Émilien Ruiz et Franziska Heimburger, « Faire de l'histoire à l'ère numérique : retours d'expériences », *Revue d'histoire moderne et contemporaine,* n° 58 (4bis-5), 2011, p. 70-89 ; Frédéric Clavert et Serge Noiret, *L'histoire contemporaine à l'ère numerique / Contemporary History in the Digital Age*, Bruxelles, P.I.E.-Peter Lang S.A, 2013.

30 Deux approches différentes se font face dans Olivier Le Deuff, *Le temps des humanités digitales : la mutation des sciences humaines et sociales*, Limoges, Fyp éditions, 2014, dans les chapitres de Frédéric Clavert, « Vers de nouveaux modes de lecture des sources » d'un côté et Frédéric Kaplan, Mélanie Fournier, et Marc-Antoine Nuessli, « L'historien et l'algorithme » de l'autre.

31 Stéphane Lamassé et Philippe Rygiel, « Nouvelles frontières de l'historien », *Revue Sciences/Lettres* (2), octobre 2013.

32 On pourra citer en exemple le Master Technologies numériques appliquées à l'histoire de l'École des Chartes.

La seconde question se pose à toutes les disciplines historiques, mais doit appeler une réponse propre à l'histoire des relations internationales : quels sont nos besoins informatiques et numériques ? Si cet article, nous l'espérons, donne un début de réponse, cette interrogation doit également nous pousser à poursuivre une réflexion commune sur le futur de nos sources passées (leur mise en données) et sur le présent de nos sources futures (la conservation des sources nées numériques).