Compte rendu: **L'histoire à l'épreuve du numérique : un nouveau « goût de l'archive » ?** Lausanne, 4 juillet 2018

Organisé par: Université de Lausanne, Université du Luxembourg, infoclio.ch, Association internationale de l'histoire contemporaine de l'Europe (AIHCE)

Compte rendu par : Jean Rime, Université de Fribourg/Université Paul-Valéry, Montpellier

Fruit d'un partenariat entre les universités de Lausanne et de Luxembourg, le site infoclio.ch et l'Association internationale de l'histoire contemporaine de l'Europe (AIHCE), la journée « L'histoire contemporaine à l'ère du numérique : sources, méthodologies et critiques » (Lausanne, 4 juillet 2018) vise une réflexion collective sur la place des nouvelles technologiques dans la recherche historique, et plus particulièrement sur la manière dont elles affectent le rapport aux sources primaires.

En introduction, les organisateurs, **FRÉDÉRIC CLAVERT** (Luxembourg), **SOLENN HUITRIC** (Lausanne), **ENRICO NATALE** (infoclio.ch) et **RAPHAËLLE RUPPEN COUTAZ** (Lausanne) expliquent l'opportunité d'une telle rencontre par les mutations de la recherche historique tant dans l'accès aux sources (numérisation) que dans leur traitement (mise en données, mise en réseaux) et dans la communication des résultats (outils de visualisation). Les évolutions observées depuis quelques années semblent avoir suffisamment changé les réflexes de la profession pour que ses représentations s'en trouvent ébranlées : le « goût de l'archive » naguère célébré par Arlette Farge[1] semble désormais bien éloigné du quotidien des chercheurs. « Quel est le goût de l'archive à l'âge du numérique ? » Telle serait, en somme, la question au centre de la journée.

Pour y répondre, les interventions se répartissent autour de trois types d'objets : les documents diplomatiques, les archives de presse et les sources nativement numériques. Les rapports d'expérience livrés par les intervenants sont mis en perspective par un discutant de manière à en élargir les enjeux méthodologiques.

La journée s'ouvre sur une session consacrée aux **documents diplomatiques**. **JOËL PRAZ** et **CHRISTIANE SIBILLE** (Berne) présentent DODIS (DOcuments DIplomatiques Suisses,

[1] Farge Arlette : *Le Goût de l'archive* [1989], Paris, Seuil, 1997. Voir notamment son plaidoyer pour « une approche tactile et immédiate du matériau, cette sensation préhensible des traces du passé » (p. 23) face à ce qui était alors le *nec plus ultra* : la reproduction sur microfilm.

https://www.dodis.ch), centre de recherche fondé en 1972 avec la mission de rendre accessibles les sources sur les relations extérieures de la Suisse. Le projet s'est doté dès 1995 d'une base de données numérique, qui compte aujourd'hui quelque 30 000 documents, ainsi que 40 000 individus, 20 000 organisations et 7000 lieux géographiques indexés manuellement. Son activité se concentre dans la sélection et l'édition scientifiques de documents publiés en ligne, dont les plus importants, dotés d'un appareil critique, forment des ouvrages papier de référence. Autour de cet objectif principal gravitent différents supports de valorisation : collection d'études en libre accès (*Quaderni di Dodis*), bibliographie, e-dossiers thématiques. DODIS s'engage en outre dans la fédération de ses ressources avec d'autres instruments de recherche au moyen du portail Metagrid (https://www.metagrid.ch) et de la plateforme HistHub (https://histhub.ch). La longévité du projet permet une réflexion critique sur la pérennité des données produites, sur les normes de transcription pour l'édition en ligne et sur les problèmes que pose la rétro-indexation des volumes antérieurs. Cette expérience conclut notamment à la nécessité d'adopter des standards communs, à l'instar de la TEI (Text Encoding Initiative), pour faciliter l'interopérabilité des descripteurs.

À l'aide d'une série d'exemples tirés de ses propres recherches ou de projets tiers, **FRÉDÉRIC CLAVERT** (Luxembourg) interroge la pertinence des approches spécifiquement numériques, lesquelles ne se contentent pas d'un changement d'échelle dans le traitement des corpus, mais ambitionnent de rendre calculables des phénomènes de masse qui auparavant ne l'étaient pas (« *datafication* ») et de les analyser par l'entraînement des machines (*machine learning*). Problématisant la notion de *distant reading* (F. Moretti), il détaille les apports d'un tel paradigme dans les domaines des statistiques, du *topic modelling* (reconnaissance thématique) ou du *data clustering* (regroupement de données), mais pointe aussi le risque d'une opacification des processus d'analyse, dans la mesure où les visualisations occultent trop souvent la programmation des algorithmes et rendent impossible toute vérification des résultats dans les documents originaux. Or ces résultats pâtissent de certains biais récurrents, que ce soit en raison de la qualité du corpus numérisé et de sa lecture automatique (OCR) ou en raison de ses angles morts : la numérisation croissante mais irrégulière des sources produit un « ordre illusoire » (Ian Milligan) qui cache des secteurs qui n'en bénéficient pas au même degré et crée ainsi des déséquilibres potentiellement trompeurs.[2]

[2] Il en va ainsi de la sur-représentation des journaux numérisés dans les études de presse. Voir Milligan Ian, « Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010 », *The Canadian Historical Review*, 94/4, 2013, p. 540-569, https://muse.jhu.edu/article/527016.

Menée par **EMMANUEL MOURLON-DRUOL** (Glasgow), la discussion se poursuit sur les limites des approches digitales. La masse de documents à numériser prête à s'interroger sur les critères de sélection des archives. Les corpus disponibles ainsi que les outils informatiques utilisés pour les analyser influent de surcroît sur la formulation des questions de recherche, autant de problématiques qui sont toutefois rapportées à des questionnements méthodologiques plus traditionnels.

La deuxième séance est consacrée à la **presse ancienne numérisée**. Représentant d'une tradition de recherche interdisciplinaire (littérature / histoire culturelle), **GUILLAUME PINSON** (Québec), l'un des protagonistes du projet Numapresse (www.numapresse.org), s'intéresse à ces corpus non comme des réservoirs d'informations, mais pour étudier le fonctionnement du médium lui-même et ses imaginaires, objets des projets antérieurs de *La Civilisation du journal*[3] et de *Médias19* (www.medias19.org). La numérisation de presse à l'échelle internationale et la perspective d'une meilleure interopérabilité entre les plateformes élargissent cette recherche, naguère limitée à une classique approche monographique ou nationale, vers une observation du système médiatique dans sa globalité. Le journal est ainsi vu comme un « laboratoire de la circulation des données » marqué par la réimpression et la « viralité » des contenus (cf. Ryan Cordell, www.viraltexts.org). Les nouvelles possibilités offertes par la reconnaissance optique de caractères (OCR) permettent un repérage automatique des circulations, y compris dans des zones du corpus où elles ne sont pas attendues ; elles aident à formaliser et détecter la généricité de certains discours et donc à étudier l'évolution des formes journalistiques. Compte tenu des défis technologiques (interopérabilité), juridiques (journaux en partie sous droits) et scientifiques (statut et complexité du discours journalistique), il semble toutefois plus prudent de militer pour un « numérique léger » qui ne soit pas déconnecté d'une immersion parallèle dans les textes.

À partir d'une étude de cas sur le sentiment anti-européen dans la presse suisse et luxembourgeoise menée dans le cadre du projet Impresso, **ESTELLE BUNOUT** (Luxembourg) montre le caractère complexe des sources numérisées, constituées non seulement de texte scanné, mais aussi d'un ensemble de métadonnées (techniques, bibliographiques, administratives) documentant les processus subis par les sources scannées. Idéalement, l'ensemble de ces couches informatives doit être mobilisé pour la constitution d'un corpus pertinent. Celui-ci ne

[3] Kalifa Dominique, Régnier Philippe, Thérenty Marie-Ève et Vaillant Alain (éds.), *La Civilisation du journal. Histoire culturelle et littéraire de la presse française au XIX^e^ siècle*, Paris, Nouveau Monde, 2011.

procède en effet plus du dépouillement systématique d'une collection, mais d'une interrogation transversale au moyen de mots-clés, de descripteurs jugés pertinents ou de filtrage par types d'article. Le croisement de ces différentes strates du corpus numérisé (qui ne dédaigne pas du reste un recours pragmatique à des méthodes plus classiques en fonction des documents à traiter) est déterminant pour circonscrire, par exemple, l'espace médiatique où s'exprime le sentiment anti-européen, et éviter ainsi de fausser l'analyse par un nivellement illusionniste du médium d'origine.

Synthétisant les questionnements méthodologiques soulevés par les deux interventions précédentes, **ALAIN CLAVIEN** (Fribourg) se demande si la numérisation de la presse rend simplement une source traditionnelle plus facilement consultable, ou si cette opération implicite de « remédiation » crée une source nouvelle, qui ne serait plus assimilable à l'addition de journaux anciens scannés, mais qui pourrait être décrite dans sa généralité comme l'expression collective d'un « discours », aux sens que donnent Foucault ou Angenot à ce terme.

La troisième session se penche sur les **sources nativement numériques**. À partir du cas de son institution, **ALEXANDRE GARCIA** (CICR, Genève) présente le point de vue de l'archiviste. Même si le CICR ouvre ses portes aux chercheurs depuis 1996, ses motivations, en matière de politique d'archivage, sont prioritairement guidées par des objectifs internes : continuité opérationnelle, redevabilité à l'égard des partenaires ou des pays membres, mémoire institutionnelle. L'apparition de documents nativement numériques a modifié le cycle de l'archive. Désormais, la prise en charge des documents se situe bien en amont des procédures de conservation, parfois dès leur création, pour y anticiper l'adjonction de métadonnées propres à faciliter leur classement ultérieur. Si désormais l'archivage des e-mails est bien géré, d'autres types de données sont plus complexes à traiter : données géographiques, *WhatsApp*, médias sociaux, *clouds*. Celles-ci sont caractéristiques d'un mode de vie qui transforme la communication institutionnelle en une information composite et multicanale, évolutive dans son volume, ses formes, son contenu et ses modalités de transmission – autant de mutations qui plaident pour une adaptation du cadre législatif.

Délicate au niveau d'une institution, la collecte des données numériques l'est *a fortiori* à l'échelle du web mondial. **VALÉRIE SCHAFER** (Luxembourg) en présente les différents écueils. Depuis la création d'InternetArchive en 1996 (https://archive.org), 330 milliards de pages sont actuellement conservées. Mais malgré cette explosion quantitative, ce fonds souffre lui aussi de biais structurels qu'il est nécessaire de comprendre pour une utilisation

adéquate. Par exemple, des robots de collecte programmés pour être rapides ne prennent pas en compte certaines images, ou, lorsqu'ils moissonnent à différentes dates des contenus mis à jour, ils ne sauvegardent pas systématiquement des éléments censés demeurer plus stables, comme les bandeaux.[4] Aussi l'archive consultée par l'historien ne restitue-t-elle souvent ni l'aspect ni le fonctionnement du site tel qu'il apparaissait à ses destinataires originels. Cette variabilité de l'archive ne fait que s'accroître aujourd'hui avec la personnalisation des contenus affichés pour chaque utilisateur. Une telle discordance questionne ce que nous pouvons et ce que nous voulons analyser dans les archives du web.

CAROLINE MULLER (Reims) élargit le débat en soulignant la difficulté de suivre ces évolutions rapides dans la formation des historiens, dans leurs pratiques de recherche et même dans les représentations de leur activité. Il est devenu banal d'affirmer que nous formons une « génération de transition », mais la rapidité des mutations technologiques pourrait rendre cet état constant, une perspective à la fois effrayante et stimulante. Si les documents numériques ont d'abord été traités par analogie avec le papier (présentation PowerPoint conservées sous forme de PDF, e-mails imprimés puis scannés, etc.), la nécessité de procédures *ad hoc* se fait de plus en plus sentir pour des sources davantage éloignées des compétences bibliographiques qui nous ont été enseignées : les codes, les algorithmes, les logiciels. Sans une évolution des cadres politiques, légaux et scientifiques des pratiques d'archivage, des pans entiers du patrimoine actuel risquent de disparaître : *quid* des archives de Facebook d'ici quelques dizaines d'années ?

Au terme de cette journée, la diversité des projets présentés aura permis de prendre la mesure des avancées réalisées grâce au numérique dans l'accès aux sources primaires et dans leur traitement, mais aussi des problèmes méthodologiques posés par ce nouvel outillage. Ne cédant ni à la fascination produite par de séduisants outils de visualisation (cartes, nuages, etc.) ni à un scepticisme *a priori*, la réflexion historiographique sur ces objets semble être parvenue à un degré de maturité qu'attestent, au sein d'un auditoire pluridisciplinaire, de nombreuses convergences dans les questionnements comme dans les propositions.

Un premier axe s'esquisse sur la portée des innovations technologiques. Se limitent-elles à faciliter le travail de l'historien ou opèrent-t-elle un renversement épistémologique ? Si l'échelle des corpus s'en trouve indéniablement modifiée, plusieurs participants soulignent

[4] Un autre exemple emblématique : des pages censées avoir été enregistrées en été présentent des cartes météorologiques hivernales…

que les résultats produits à grands frais ne sont souvent pas radicalement nouveaux, et que plutôt rares sont même les cas où les hypothèses émergent véritablement du travail de l'ordinateur. Du reste, pointer et résorber les différents biais des corpus d'archives numériques ou numérisées (trous, disproportions, etc.) revient finalement à leur appliquer de classiques méthodes de critique des sources. Au fantasme d'une archive totale se substituent ainsi des procédures raisonnées pour assurer aux corpus une cohérence et une représentativité en fonction de problématiques clairement prédéfinies et spécifiques à chaque recherche. Si saut paradigmatique il y a, il n'est pas dû à la seule utilisation des machines, mais d'abord à la conceptualisation d'objets nouveaux, imputable à une conjonction de facteurs plus complexe, comme ce fut le cas à l'avènement de l'histoire sociale.

Symptomatique de cette voie médiane entre le tout-numérique aveugle et les pratiques traditionnelles, la négociation entre « lecture distante » et « lecture rapprochée », ou « profonde », semble faire consensus. Contrastant avec le « petit pacte avec le diable » conclu par Franco Moretti,[5] le choix d'une méthode hybride semble s'imposer, non seulement pour pallier les insuffisances de la numérisation (sources uniquement physiques ou illisibles), mais aussi pour affiner les résultats quantitatifs au moyen d'analyses qualitatives, quitte à se priver par toutes ces précautions de certaines des possibilités les plus innovantes – mais aussi les plus spéculatives et les moins contrôlables – des technologies digitales.

Si la méthode historique, notamment la critique externe, sort donc renforcée de son contact avec le numérique, les imaginaires professionnels, eux, s'en trouvent inquiétés. La perspective d'un accès immédiat à des sources auparavant difficilement accessibles ou d'une structuration de la recherche inspirée par les sciences « dures » attire autant qu'elle peut rebuter, voire entraîner une forme de nostalgie. Le rapport au numérique, avant d'être conditionné par des données techniques, l'est par un ensemble de représentations ou d'affects. Il convient d'en faire également l'histoire pour en inventer l'avenir, car c'est d'eux, avant peut-être que des innovations matérielles, que dépendra le « goût de l'archive » de demain.

Jean Rime
jean.rime@unifr.ch

[5] Moretti Franco, « Conjectures on World Literature » [1999], *Distant Reading*, London / New York, Verso, 2013, p. 48. Version française : *Études de lettres*, 2, 2001.

Programme du colloque

Module 1 - Les documents diplomatiques

Christiane Sibille & Joël Praz, collaborateurs scientifiques, Dodis : « Les Documents Diplomatiques Suisses à l'ère numérique : pratiques et perspectives »

Frédéric Clavert, senior research scientist, C2DH, AIHCE : « Du document diplomatique à l'histoire des relations internationales par les données massives »

Discussion par Emmanuel Mourlon-Druol, senior lecturer, Université de Glasgow, AIHCE

Module 2 – La presse numérisée

Guillaume Pinson, professeur titulaire, Université de Laval, membre du projet Numapresse : « La presse ancienne à l'ère numérique : enjeux scientifiques d'une remédiation »

Estelle Bunout, research associate, C2DH : « Une recherche plus fouillée dans un corpus imparfait ? L'étude de la question européenne dans la presse numérisée suisse et luxembourgeoise (1848-1945) »

Discussion par Alain Clavien, professeur ordinaire, Université de Fribourg

Module 3 – Les sources nées numériques

Alexandre Garcia, archiviste, CICR : « Comment se constitue un fonds d'archives aujourd'hui ? L'exemple du CICR »

Valérie Schafer, professeure d'histoire contemporaine, C2DH : « Les archives du Web, une lecture à plusieurs niveaux »

Discussion par Caroline Muller, PRAG, Université de Reims