Revue d'histoire culturelle
XVIIIe-XXIe siècles

Numéros / | 2022 : 5 | Psychanalyse et histoire culturelle / Épistémologie en débats

Préservation et distorsion : l'espace-temps des réseaux socio-numériques et du web archivé

Protection and distortion: the space-time of born-digital heritage

Frédéric Clavert, Sophia Mahroug and Valérie Schafer

DOI : 10.56698/rhc.2791

Abstracts | Index | Outline | Text | Notes | Illustrations | References | Authors

ABSTRACTS

Français | English

Massive data, also known as Big Data - originating from websites and digital social networks, in the form of text, images, videos or metadata -, constitute significant sources for recent and future research in cultural history. These "digital traces", collected by researchers or institutions, require further methodological thoughts - from their archiving to their development, in order to analyse them at different scales (*scalable reading*). Indeed, they allow researchers to identify new spatiotemporal boundaries, but also asymmetries and distortions between the theoretical scope of Big Data (from the millisecond to the long term, from the meter to the globe) and its practical scope (regional inequalities in collection, noise and silences within the archives). Based on several research projects and institutional initiatives, this article aims at thinking about the space-time of born-digital heritage, from the standpoint of data, collections and research, in order to grasp both the consequences of this massive archiving on the shaping of history and the profession of historian, and to identify the ongoing issues of these historical sources for academic research.

INDEX

Mots-clés
patrimoine nativement numérique, réseaux socio-numériques, archive du web, données massives, épistémologie

Keywords
born-digital heritage, social media, web archive, big data, epistemology

OUTLINE

TEXT

1 Des articles, livres ou numéros de revues se sont penchés sur le *Big Data* depuis plus de dix ans, sous un angle parfois très critique[1], souvent interdisciplinaire[2], avec parfois une dimension historique[3]. Certaines publications s'interrogent sur la pertinence de la définition du terme : alors même que la mise à disposition de données de plus en plus abondantes risque d'entraîner de profonds changements disciplinaires et méthodologiques[4], d'autres tentent une vue plus générale sur l'ensemble du champ des sciences sociales[5] et les usages par ces dernières des différents types de données[6].

Les réflexions sur l'espace-temps à l'ère du *Big Data* restent cependant à approfondir. Longtemps figurée sous la forme de deux entités distinctes et immuables, la représentation de l'espace-temps comme un ensemble indissociable au début du XXe siècle a profondément influencé les écoles de pensée en sciences humaines et sociales (SHS). Cette conception bergsonienne d'un espace-temps à vitesse et géométrie variables continue de nourrir les réflexions historiques, comme en attestent les travaux de Reinhart Koselleck[7], François Hartog[8] ou du philosophe Hans Ulrich Gumbrecht[9] et peut être repensée sous l'effet du recours aux données massives et éphémères que sont les données disponibles sur le Web et les réseaux sociaux numériques (RSNs). Ces dernières sont indéniablement des sources pour l'histoire sociale et culturelle, présentant des entrelacs de temps perçus et vécus, à différentes échelles, dans tous les espaces géographiques.

Comme l'archive papier, les traces numériques peuvent permettre l'élaboration d'une chronologie précise à partir d'un événement, à l'instar de la COVID-19 ou des attentats de 2015 en France. Elles peuvent également dessiner des temps longs et alimenter une histoire sérielle à partir de sujets sociaux et économiques du temps présent (précarité, égalité homme/femme…). Elles saisissent *de facto* des « vibrations[10] », des tendances et des inscriptions spatiales[11], comme en témoignent certaines collectes récentes de plusieurs millions
2 de tweets[12], à partir de méthodes d'archivage spécifiques abordées dans cet article.

Celui-ci propose en effet de penser les différents jeux d'échelle et de temporalités qui opèrent au sein de ce patrimoine nativement numérique – en particulier des RSNs et du web archivé, en trois temps : celui de la collecte d'abord, qui pose des enjeux de préservation, sélection, curation, représentativité et accès ; celui des données ensuite, qui présentent des entrelacs spatio-temporels complexes ; celui de la recherche enfin, qui implique une herméneutique numérique appropriée et des outils computationnels comme intellectuels pour les aborder. Cette exploration spatio-temporelle tout au long de la chaîne de préservation et d'analyse des données massives doit permettre de saisir les conséquences de l'archivage massif des données sur la mémoire, l'histoire et le métier d'historien, ainsi que les
3 asymétries et distorsions entre la portée théorique du *Big Data* (de la milliseconde à la longue durée, du mètre au globe) et sa portée pratique (inégalités régionales dans les collectes, silos entre collections, bruits et silences au sein des archives, etc.).

Espace et temps des collectes

Les traces numériques recueillies sur les sites web et les RSNs, qu'elles soient sous la forme d'écrits, d'images, de vidéos ou de
4 métadonnées, constituent des sources de plus en plus convoquées dans les travaux de recherche en SHS. Les historiens s'en sont également emparés dans des études portant sur des dimensions mémorielles en ligne d'événements[13] ou l'histoire du numérique[14], à partir de collectes institutionnelles[15] ou de collectes entreprises individuellement[16].

Étudier des sources « nativement numériques » implique une réflexion sur les paradigmes et cadres sociaux entourant ces données. Si Franck Ghitalla[17] ou Mark Graham[18] ont proposé une cartographie du Web à partir de sa structure technique, d'autres ont utilisé le Web non comme un objet mais comme une variable supplémentaire à leur sujet d'étude. La géopolitique et les études stratégiques repensent ainsi le Web comme un « cyberespace » ou une « datasphère » qui prolonge les territoires physiques afin de comprendre les dynamiques spatiales et les conflits contemporains, de redéfinir des frontières et mettre en lumière des acteurs inconnus jusqu'alors, sous la forme de réseaux culturels, politiques ou encore religieux pluriels et inexplorés. Des études ont par exemple montré la centralité des RSNs dans les mobilisations populaires des années 2010 autour de nouveaux types d'acteurs comme les influenceurs, ou de nouveaux lieux de pouvoir comme les places[19]. Les traces laissées sur *Twitter* et *Facebook* au moment du mouvement vert de 2009 en
5 Iran et des printemps arabes de 2011 en Égypte, en Tunisie et en Syrie, ont permis de démontrer le rôle fondamental (et les limites) de certaines plateformes dans l'organisation des mouvements contestataires[20] et la diffusion d'informations au monde entier, depuis des espaces où la liberté de parole est pourtant limitée.

La mainmise de plus en plus affirmée d'acteurs privés (GAFAMI[21] notamment) sur Internet et le Web depuis le début des années 2010
6 a toutefois des conséquences directes sur la pérennité et l'accessibilité des traces numériques. L'idée d'un Internet « souverain » est par ailleurs devenue un moyen de contrôle des idées et des mouvements de populations à des fins politiques, comme en témoignent la mise en place du *Runet* en Russie depuis 2019, la construction de « routes de la soie numériques » par la Chine en 2015 ou encore la validation d'un « plan de protection » du Web iranien en 2021. Les coupures ponctuelles des couches basses d'Internet à des fins répressives, comme au Kazakhstan en janvier 2020 ou en Birmanie en février 2021, tout comme des cas de blocage d'*Internet Archive* et de l'accès à ses archives du Web (en 2017 en Jordanie par exemple) ont des conséquences directes sur la production, la préservation et l'accessibilité des données. Il faut ainsi considérer des asymétries, notamment pour l'étude des régions non-occidentales[22], déjà pénalisées par l'impossibilité d'accéder au terrain[23]. Des initiatives françaises tentent de faire de certaines zones de conflits un terrain de recherche multidisciplinaire, grâce à la disponibilité des traces numériques. L'ANR SHAKK[24] a pour objectif de reconstruire les événements syriens depuis 2011 hors du terrain étudié. Ce projet pluridisciplinaire ambitionne de dégager les nouvelles frontières de la

Syrie, ainsi que l'évolution des pouvoirs et acteurs en place dans le pays. L'équipe convoque des données du Web, vidéos, photographies

ou commentaires publiés par des manifestants, des activistes ou encore des combattants. Dans un même souci de pallier un terrain inaccessible, l'ERC Off-Site[25] propose depuis 2018 une étude ethnographique sur la violence d'État durant les années Khomeini en Iran (1979-1988) à l'aide des sources alternatives (*counter-archives*)[26]. Malgré le caractère novateur de ces projets ou d'une méthodologie de la cartographie appliquée au cyberespace telle que le propose le laboratoire de géopolitique GEODE, à partir de l'exemple russe[27], il reste encore beaucoup à faire et à partager sur la méthodologie de ces terrains numériques.

Le déplacement de la recherche vers des structures dématérialisées n'est par ailleurs pas une réponse absolue : tous les pays ne disposent pas d'archivage institutionnel du Web, tandis qu'*Internet Archive* ne peut couvrir également toutes les régions du globe. De nombreuses données ne sont ainsi pas sécurisées et restent périssables : les suppressions de publications et de commentaires, les fermetures de compte, les changements de logiciel[28] et même les incendies de *datacenters* ou serveurs constituent autant de micro-événements dramatiques pour la préservation des données.

L'oubli, l'irrégularité de collectes ou le manque de pérennité de certaines traces invitent l'historien à se faire l'archiviste de ses propres données, parfois en anticipant des projets futurs. Cette problématique de l'anticipation est bien connue des chercheurs travaillant sur des régions à risque, et qui manipulent des données dispersées sur une multitude de plateformes mouvantes et soumises au risque de potentielles fermetures de domaines internet et aux enjeux de la cybergouvernance[29].

Face à l'urgence, les institutions (bibliothèques en charge du dépôt légal, fondation comme *Internet Archive*, etc.) lancent aussi des collectes spéciales, par exemple lors d'attentats ou de la pandémie de la COVID-19. L'espace-temps des collectes institutionnelles est en effet à la fois régulier et toujours susceptible d'être bousculé par l'actualité[30]. De nombreuses institutions ont mis en place des collectes annuelles ou bi-annuelles pour les sites relevant de leur périmètre, qu'elles ne pourraient pas préserver quotidiennement, alors que l'on parle de millions d'URL et noms de domaine. Des collectes plus régulières, voire quotidiennes, sont prévues sur des sites web spécifiques, par exemple de presse à la Bibliothèque nationale de France (BnF). Certains événements, comme les élections, font l'objet de collectes spéciales anticipées. Enfin des collectes liées à un événement inattendu peuvent bousculer l'archivage du Web, en appelant à une réaction immédiate. Dès 2015, la BnF et l'Ina, confrontés aux multiples expressions numériques que suscitent les attaques terroristes en France, mettent en place des collectes d'urgence[31]. L'événement inattendu implique un archivage qui doit suivre au plus près les traces numériques sans pouvoir totalement les anticiper, que ce soit en termes de tendances ou de temporalités : la collecte crée ainsi une archive vivante (*living archive*[32]), qui s'enrichit au fil des jours de nouveaux sites web (dans le cas de la crise de la COVID-19), et un hashtag peut prendre le pas sur un autre (dans le cas des attentats du Bataclan, le hashtag « tirs » est employé avant que l'on parle d'attentat). Comme l'explique Thomas Drugeon (Ina) au sujet du 13 novembre 2015 : « Les hashtags étaient plus éparpillés en novembre, en janvier presque tout était concentré sur le hashtag #jesuischarlie. En novembre ressortent au moins cinq hashtags et on discerne des mouvements, des cycles également, par exemple jour/nuit en rapport avec le décalage horaire à l'international[33] ».

La collecte d'urgence soulève d'autres problèmes de temporalités : quand commencer l'archivage et quand l'arrêter, lorsque la crise s'installe, par exemple dans le cas de la COVID-19 ? Comment gérer les vagues qui se succèdent et les rendre compatibles avec les autres missions des équipes, le budget et les contraintes techniques et politiques de préservation[34] ? Comme en témoigne Ben Els, archiviste du Web à la Bibliothèque nationale du Luxembourg : « Nous avons commencé le 16 mars 2020, et il est difficile de dire quand nous nous arrêterons. Il s'agit plutôt d'une question budgétaire. [...] J'ai beaucoup de téraoctets à ma disposition, mais nous ne mettrons pas forcément 5 téraoctets de plus[35] ». À des collectes COVID qui peuvent s'arrêter dès 2020 (ce qui ne signifie pas que les archives du Web sont ensuite muettes, puisqu'une collecte régulière comme celle dédiée à l'Actualité à la BnF va inclure de nombreux éléments liés à la crise sanitaire) répondent au contraire des archives qui continuent de s'enrichir, par exemple à l'Ina, au fil des échos que trouvent la référence aux attentats contre *Charlie Hebdo* sur Twitter, que ce soit lors des attentats du Bataclan ou de Nice ou des commémorations et procès qui reconvoquent les hashtags de 2015.

L'exhaustivité est impossible : les archivistes en sont conscients, et essaient de viser une forme de représentativité en documentant parfois les données manquantes, comme à l'Ina pour sa collecte des attentats via l'API[36] de Twitter. La représentativité passe aussi par le soin porté aux traces locales ou régionales, comme dans la collection COVID-19 de la BnF. Une analyse de ses données dérivées permet d'identifier de nombreux sites web régionaux. Comme l'explique Alexandre Faye (BnF), la sollicitation des différents départements de la BnF et des correspondants DLWeb[37] notamment en région, fait que « la moitié des contenus sélectionnés durant le confinement ont un mot clé géographique qui indique que le contenu est en lien avec un territoire précis[38] ».

Certaines régions du monde peuvent connaître une moindre attention : l'épidémie d'Ebola n'a pas d'équivalent d'archivage par rapport à la crise du COVID, ce qui ne s'explique pas uniquement par le caractère plus récent de la seconde. On signalera toutefois les efforts collectifs au sein d'*Archive-It*, lié à *Internet Archive*, pour documenter par exemple le tremblement de terre en Haïti en 2010[39].

Enfin, des espaces numériques ne sont pas archivés, ou de manière très ponctuelle : outre *Periscope*, utilisé au moment des attentats du Bataclan, *Instagram* ou *TikTok* ne sont guère archivés, tandis que *Vine* a disparu et que *Telegram* révèle son importance dans la guerre en Ukraine, sans bénéficier de réelle politique d'archivage. On notera toutefois dans le cas du conflit ukrainien la vive réactivité à l'égard du patrimoine nativement numérique, avec le lancement de l'initiative participative SUCHO, *Saving Ukrainian Cultural Heritage Online*[40]. Malgré cet exemple extraordinaire, bien des espaces sont et seront oubliés par les archivages d'urgence, alors que les internautes déploient leur inventivité sous forme de mèmes sur les RSNs ou encore de commentaires dans les espaces d'avis des restaurants russes, pour essayer d'alerter la population sur les horreurs de cette guerre[41]. L'archive est toujours lacunaire, mais ici se joue aussi une partie de la guerre : l'oubli est aussi conséquence de rapports de forces, y compris en conflit armé[42].

Malgré des millions de données collectées, de fortes asymétries et des silences traversent les espaces et temps des collectes, qui se manifestent également au sein des collections.

Espace et temps au sein des collections

Les *Big Data* en histoire ne correspondent pas tout à fait à la définition originelle qui fait référence aux « 3V[43] » : Vélocité, Volume et Variété des données[44]. En effet, les données passées font, en toute logique, preuve d'un manque de *vélocité*, c'est-à-dire de renouvellement constant et en temps réel du corpus (à l'exception des études mémorielles, à l'image du corpus déjà évoqué des attentats qui continue d'évoluer sous l'effet de la réutilisation d'hashtags antérieurs). Toutefois, les deux autres V – *volume* et *variety* – concernent directement les historiens et historiennes. Ils permettent de travailler sur l'espace et le temps avec des outils de lecture distante, à des échelles spatiales et temporelles variées. Depuis que Franco Moretti[45] a forgé l'expression de lecture distante, celle-ci a connu un engouement, notamment au sein des humanités numériques, mais aussi de sérieuses critiques[46]. De l'usage et des critiques est ressortie la notion de lecture « multi-échelle » (*scalable reading*)[47] – c'est-à-dire le fait d'imbriquer les différentes échelles de lecture, de ne renoncer ni à la lecture proche, qualitative, ni à la lecture distante –lorsque l'ordinateur et des logiciels « lisent pour nous ». Nous donnerons ici deux types d'exemples de ces dimensions spatio-temporelles des collections : d'abord, les cas de *GeoCities*, de *Mygale.org* et du Web yougoslave – lorsque l'espace-temps est clos et que la collecte des données et leur lecture relèvent de l'urgence ou de l'impossibilité ; ensuite, celui du Centenaire de la Grande Guerre, alors que l'espace-temps de cet événement restait ouvert, rendant la collecte possible voire aisée, mais avec des limites importantes.

Dans ses recherches sur *GeoCities*, en alternant lectures proche et distante, Ian Milligan montre par des visualisations et cartographies de cet espace numérique un type d'agencement des relations sociales[48]. Créée en 1994, *GeoCities* est une plateforme populaire d'hébergement de sites web, fondée sur une métaphore géographique manifeste. Les sites web y sont regroupés en *neighbourhoods* (quartiers) en fonction des sujets traités et des affinités thématiques. Ian Milligan montre comment cet ensemble de sites a permis à de nombreux utilisateurs de tisser des liens au sein d'un Web à l'époque encore balbutiant et parfois intimidant, y compris en construisant des communautés reposant sur l'entraide entre utilisateurs. Mais cette recherche sur *GeoCities* rappelle aussi indirectement la fragilité des sources nativement numériques : *GeoCities* est racheté en 1999 par la société *Yahoo!*, qui décide en 2009 de le fermer puis de l'effacer. S'il est sauvé *in extremis* par l'*Archive team*[49], de nombreuses plateformes et sites web, notamment avant 2000, n'ont jamais été archivés. *Mygale.org* était une forme d'équivalent francophone de *GeoCities*. Ce service, devenu *Multimania* en 1999, n'est archivé par *Internet Archive* qu'à partir de 2000. Estimé à 40 000 sites web en 1998 (soit à peu près la moitié des sites web français d'alors)[50], *Mygale.org* puis *Multimania* est laissé à l'abandon à partir de 2009, lorsque Lycos Europe, devenu son propriétaire, a été liquidé. Au regard des histoires croisées de *GeoCities* et de *Mygale.org*, apparaît un espace-temps du web empli d'absences qu'il faut s'efforcer de comprendre, parfois de combler. C'est ce qu'Anat Ben-David a tenté de faire en reconstituant le domaine yougoslave (.yu), disparu avec la fin de la République fédérale de Yougoslavie en 2003[51].

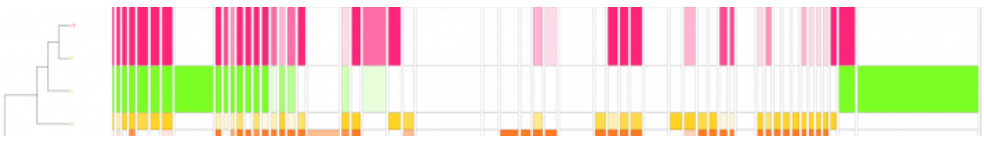
Les RSNs sont également sujets à des failles spatio-temporelles, mais selon des modalités assez différentes. En effet, les médias sociaux tels que *Facebook* ou *Twitter* ont des spécificités fortes en raison des données émises, gigantesques, même en comparaison des archives du web, d'une part ; des modalités de leur archivage, d'autre part[52]. Les archives institutionnelles du Web collectent certains éléments, mais un chercheur ou une chercheuse peut être amené à constituer son corpus lui-même via les API des RSNs, pour des raisons diverses, qui tiennent à un intérêt pour un sujet spécifique ou/et au souhait d'avoir ses propres données accessibles sur son ordinateur et traitables par des outils choisis par exemple. Si la collecte de *Facebook* relève plutôt du parcours du combattant, collecter des données sur *Twitter* est plus aisé, surtout depuis 2021, lorsque la firme californienne a lancé une opération d'ouverture de ses données vis-à-vis des chercheurs et chercheuses[53]. Une fois l'opération reconnue comme telle par *Twitter*, il est possible de collecter 10 millions de tweets par mois dans l'ensemble de l'historique de *Twitter*. Le volume qui peut être acquis nécessite la plupart du temps de faire appel à des méthodologies issues des sciences informatiques (techniques d'apprentissages, *machine* et *deep learning*[54]).

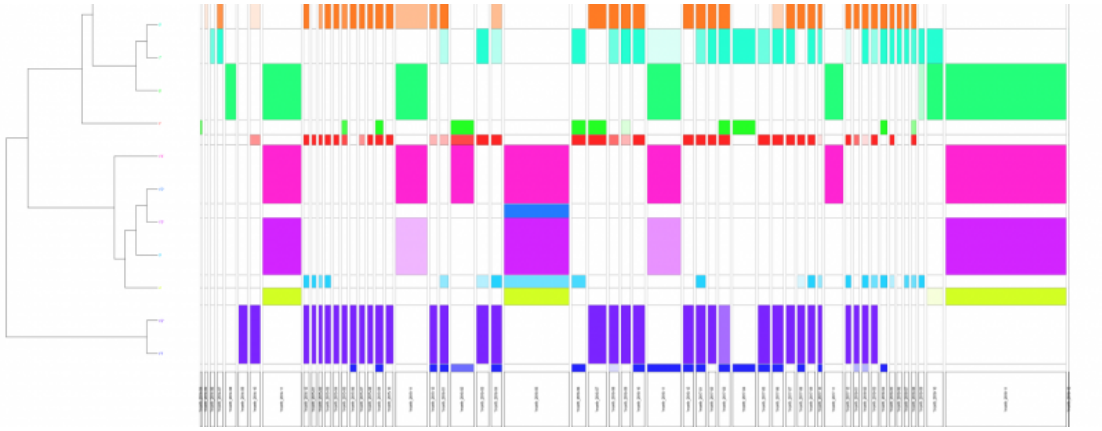
L'une des caractéristiques des tweets moissonnés est de contenir de nombreuses métadonnées, y compris spatiales au sens de géographiques (divers éléments de lieux ou encore la latitude et longitude, si l'utilisateur – cas rare – accepte la géolocalisation), ainsi que des données permettant de reconstituer la position d'un utilisateur de *Twitter* dans un espace défini, une « twittosphère », et ce notamment par des liens. Ces derniers peuvent être de plusieurs sortes : les liens conversationnels proposés par la plateforme comme les retweets, réponses, *likes*, ou les liens hypertexte plus classiques renvoyant vers d'autres espaces du Web. Les métadonnées temporelles ont, elles, la particularité de donner la possibilité (théorique) de reconstituer ce qui se dit, les interactions sociales, à la seconde près. Il devient alors possible de saisir très précisément des « vibrations[55] ». Dans le cas du Centenaire de la Première Guerre mondiale, la conjugaison des possibilités de collecte de données *Twitter* et des outils de lecture distante (IRaMuTeQ par exemple) a permis d'observer et d'analyser les temporalités et espaces de la commémoration en ligne[56]. La temporalité se retrouve dans le nombre de tweets collectés, irrégulier au fil du Centenaire, de 2014 à 2018, et qui se glisse dans le rythme des commémorations officielles. Mais elle se retrouve aussi dans le contenu des tweets, par exemple pour les tweets francophones (Fig. 1 et 2).

Figure 1 – Classification hiérarchique descendante (CHD, méthode Reinert, telle qu'implémentée dans IRaMuTeQ) des tweets francophones du corpus collecté pendant le Centenaire

La CHD (Fig. 1[57]) regroupe des segments de texte – ici, des tweets – dans des classes, selon un raisonnement statistique reposant pour l'essentiel sur la co-occurrence de mots. Les mots affichés (lemmatisés et en bas de casse) sont les lemmes plus pertinents pour chacune des classes. Ainsi, la classe 4, par exemple, regroupe 3,5% des tweets analysés. Les mots les plus pertinents de cette classe (« macron », « président », « hollande », « trump », etc) permettent, avec d'autres fonctionnalités d'IraMuTeQ, d'interpréter le contenu des tweets de cette classe.

Figure 2 - Projection chronologique (par mois) des classes de la CHD

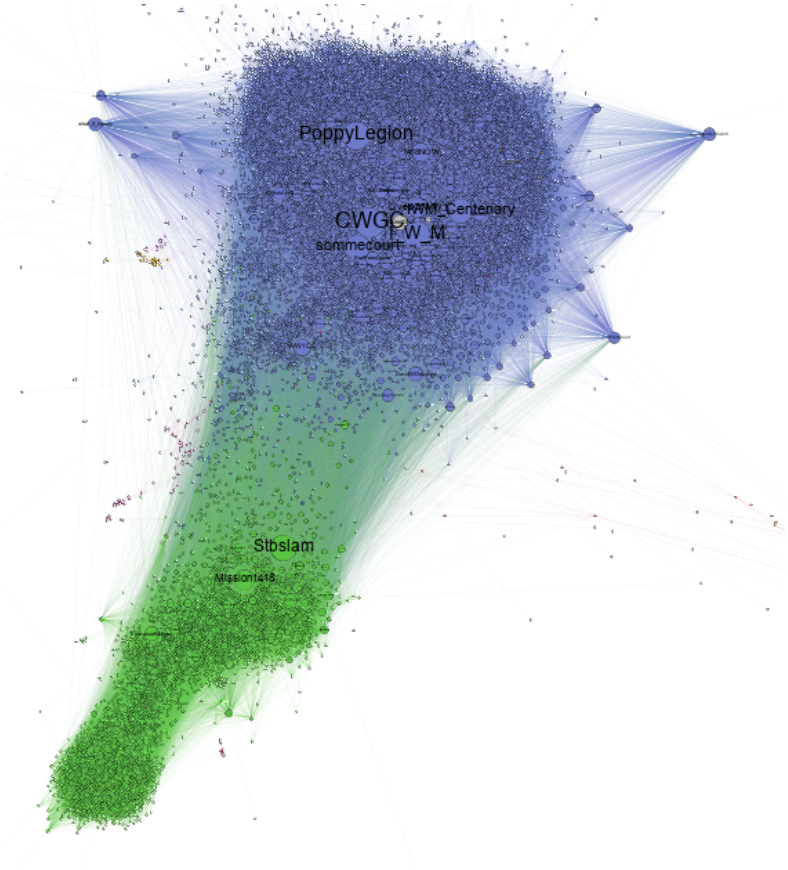




La projection dans le temps des classes de tweets obtenues par la méthode Reinert (Fig. 2) permet par exemple de déduire que l'on ne parle pas du Poilu mort pour la France avec les mêmes mots pendant les grandes commémorations (vocabulaire général de l'hommage) et le reste du temps (hommage à des poilus particuliers), ou que les moments de grandes commémorations (comme les 11 novembre ou, en 2016, la commémoration franco-allemande de Verdun ou franco-britannique de la Somme) apportent toujours leur lot de contestations.

Les relations entre les comptes *Twitter* peuvent également se traduire dans l'espace, compris ici comme l'agencement des comptes *Twitter* les uns par rapport aux autres (Fig. 3).

Figure 3 - Réseau des relations (réponses, citation, retweets) du Centenaire de la Grande Guerre



La figure 3 montre ainsi trois espaces : l'un anglophone et majoritaire (en violet), l'autre francophone et structuré pour l'essentiel autour de la Mission du Centenaire (Mission14-18) et du compte de la journaliste Stéphanie Trouillard (en vert, partie supérieure) et le dernier (en vert, partie inférieure) regroupant des comptes ayant pour beaucoup participé aux controverses de mai 2016 autour de la commémoration de Verdun.

Toutefois, l'abondance de métadonnées spatio-temporelles n'est pas synonyme de reconstitution complète d'un espace-temps, car ces métadonnées ne peuvent que reconstituer un espace-temps géré et défini par la plateforme elle-même. Elles ne sont pas relatives à d'autres temporalités, celles de chaque compte *Twitter* par exemple.

Ces exemples se rejoignent sur certaines lacunes et biais. La volonté de suivre des phénomènes trans-plateformes, comme dans le cas du

projet HiVi (*A history of online virality*)[58], va cumuler ces difficultés spatio-temporelles : la viralité, en circulant d'une plateforme à une autre, révèle des asymétries en termes d'archivage (YouTube est très bien archivé par l'Ina mais ce n'est pas le cas dans bien des pays européens, *TikTok* et *Instagram* ne sont quasiment pas archivés). Elle pose aussi la question des doublons, nombreux dans le cas des phénomènes viraux, ainsi que des nuances ténues entre les différentes formes que prend un phénomène comme un mème (variation du texte accompagnant une image macro, collage*, mash-up*), ce qui peut poser des problèmes de recherche. Les phénomènes Internet présentent des temporalités parfois très courtes (pic du Harlem Shake[59] en 2013 sur quelques mois), parfois plus longues (stabilité du Rickroll[60]) et invitent à croiser Web vivant et archivé, ainsi que des fonds et interfaces différents, à l'instar de ceux de la BnF et de l'Ina. À l'instantanéité et au caractère volatile, voire éphémère, des données numériques, répondent donc des temporalités de la recherche plus longues[61].

Espace et temps de la recherche

Si l'archivage et les données du patrimoine nativement numérique sont mis au défi de l'espace et du temps, ces derniers sont également à prendre en compte dans le cadre de la recherche scientifique.

Outre les problèmes de maintenance et de durabilité que peuvent poser des collectes effectuées par les chercheurs eux-mêmes (*cf.* les enjeux des FAIR data[62]), la question de la contextualisation des données est centrale. L'entreprise collective au sein du projet WARCnet[63] de documentation des collectes COVID, qui a permis de réaliser de nombreux entretiens oraux avec des institutions d'archivage (au Danemark, au Luxembourg, en Irlande, en France, en Grande-Bretagne, etc.[64]), est une réponse à un besoin de contextualisation et de documentation pérenne.

D'autres enjeux se posent en termes d'accès aux données[65] et d'outils. Les interfaces, les modes de consultation et les outils de recherche sont aussi mouvants que les périmètres d'archivage du Web. Ainsi l'implémentation d'une recherche plein texte dans les pages d'accueil d'*Internet Archive* au cours de la décennie 2010 a-t-elle permis de contourner certaines difficultés antérieures de la recherche par une URL précise dans la *Wayback Machine*. Dans cette dernière, des fonctions de comparaison entre les pages et de visualisation de la structure des sites ont également été ajoutées, tandis que d'autres institutions d'archivage misent de plus en plus sur la fourniture de métadonnées pour une lecture distante des fonds. Des données dérivées telles que l'URL d'un site, son nom de domaine, sa date d'archivage, la nature du fichier (vidéo, texte, audio…) permettent sans entrer dans le contenu d'identifier des tendances (par exemple, une surreprésentation de sites anglophones, gouvernementaux, etc.). Ce mouvement s'amplifie grâce à des initiatives dédiées aux chercheurs comme la récente création du BnF Datalab qui accompagne les chercheurs dans l'exploitation de son patrimoine numérisé ou nativement numérique sous l'angle d'une approche par les données, ou les « datathons » organisés par l'Ina.

La recherche sur des phénomènes transnationaux se heurte quant à elle à la logique spatiale de nombreuses archives du Web, souvent nationales car issues des lois sur le dépôt légal, mais aussi au multilinguisme et aux limitations d'accès aux fonds dans les enceintes des bibliothèques. Seules les métadonnées sont éventuellement exportables, bien qu'elles soient rarement harmonisées d'une archive à l'autre. Le travail sur les métadonnées permet d'amorcer une recherche, mais il est en général insuffisant. Travailler autour de sujets transnationaux implique fréquemment la constitution d'équipes interdisciplinaires multinationales. Le besoin d'approches analytiques complémentaires issues des sciences de l'information, de la sociologie, de l'anthropologie et de l'histoire, par exemple, se conjugue à un besoin de savoir-faire informatique et technique. Intégrer des chercheurs et chercheuses en sciences informatiques – parfois en mathématiques ou physique aussi, comme le rappelle Philippe Rygiel[66] – implique des ajustements d'objectifs de recherche, par exemple pour permettre la publication d'articles plus orientés vers les sciences informatiques. Le projet *Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset* (AWAC2)[67] montre ainsi bien les défis qu'engendre une recherche sur un sujet transnational mené par une équipe européenne sur un financement venant d'un autre continent (le Canada). Ce projet de recherche, financé par le programme canadien *Archive Unleashed* en coopération avec *Archive-It* (une entreprise *non-profit* californienne liée à *Internet Archive*), a pu avoir accès à un corpus d'archives du Web constitué par l'IIPC (*International Internet Preservation Consortium*) en coopération avec de nombreuses institutions nationales d'archivage du web, et ce *via* une interface ARCH et les outils développés par l'équipe d'*Archive Unleashed*. Ces derniers permettent de créer des corpus de données dérivés, plus faciles à utiliser que le corpus d'origine d'un volume de plus de cinq téraoctets. Toutefois, ces jeux de données dérivés restent d'un usage complexe. Ils ont impliqué la mise au point de méthodes de travail issues du monde informatique (utilisation de la plateforme github, écriture de lignes de code au sein de *code notebooks* permettant la reproductibilité et le partage des analyses ainsi développées), ainsi que l'élargissement de l'équipe à un chercheur en sciences informatiques. La fouille de texte étant un outil central, l'arrivée du collègue informaticien a aussi ajouté comme possible but de recherche la comparaison de différents outils de *topic modeling*, l'une des techniques au cœur de la fouille de texte. AWAC2 est ainsi au cœur de plusieurs espaces-temps – celui du corpus, croisant des logiques d'archivage nationales et internationales, dont l'équipe n'a pas encore perçu toutes les subtilités ; ceux de cadres de recherche sur deux

continents ; ceux de disciplines devant trouver des objets communs de recherche. Ce croisement implique une évolution des méthodes et rythmes de recherche, c'est-à-dire le fait de trouver un espace-temps commun à toutes celles et ceux impliqués dans cette recherche.

Il implique également des accords sur l'accès, la sécurisation, le traitement et l'usage des données qui s'inscrivent dans plusieurs cadres, dont le règlement général de la protection des données (RGPD). Tout chercheur de l'Union européenne manipulant des données massives doit se conformer à cette réglementation, et faire une évaluation préalable des dommages potentiels, notamment sur la préservation de l'anonymat. Certains projets se sont également saisis de ces problématiques sous l'angle éthique, comme le projet *Documenting the Now* – fortement actif aux côtés du mouvement *Black Lives Matter*. Bien que le RGPD se soit et ait inspiré de nombreuses législations extra-européennes (le Japon, l'Argentine, la Californie, ou encore l'Uruguay depuis 2008), il ne peut s'appliquer à tous les terrains. En outre, l'encadrement sur les données massives ne peut se contenter d'être juridique, sans infrastructures techniques régulant le stockage et l'exploitation des données. Devant l'influence des GAFAMI, qui repose sur la « prédation de ces traces d'activité, et parfois de données nominatives »[68], les problématiques éthiques doivent aussi constituer l'une des priorités méthodologiques du *Big Data*.

Le travail sur des collectes d'urgence et événements récents met également à jour des temporalités difficilement conciliables. À partir de la mi-mars 2020, le nombre de projets de recherche montés en tant que réponses rapides à la crise sanitaire a été particulièrement important. Cela peut s'expliquer par l'importance de la crise, mais également par son statut : la première crise mondiale de cette ampleur à l'ère des données massives. La temporalité de ces projets de réponse rapide est confrontée à celles de la recherche et de la pérennisation. À ceci s'ajoutent la multiplication des crises (les attentats de 2015 et 2016, la crise sanitaire de 2020, l'agression russe contre l'Ukraine) et la fatigue qui en découle, dont la fatigue informationnelle : le *big data* est aussi l'ère de la circulation de l'information à haute fréquence et la difficulté, y compris pour la recherche, est d'appréhender ce flux constant, dense, connaissant parfois de brèves mais intenses accélérations[69].

Si le but de ces projets a très certainement été de *documenter le présent* pour le futur, il n'est pas certain que les moyens de leur pérennisation suivent. Comment s'assurer que, dans vingt ans ou plus, ces amas de données puissent encore être compris ? Et comment pérenniser quand l'on change d'université sans pouvoir garder une main sur les recherches menées dans l'institution quittée ? La préservation du *Big Data* constitue ainsi un enjeu central pour la recherche scientifique, qui doit penser ses infrastructures, ses outils de collecte et de maintenance. En France, les très grandes infrastructures de recherche (TGIR) assurent un cadre depuis une vingtaine d'années en sciences humaines sociales. Pilotées par des comités scientifiques, des TGIR comme Huma-Num[70] ou Progedo[71] mettent en œuvre des infrastructures numériques pour la valorisation des SHS à l'échelle nationale et européenne, et proposent notamment des plateformes de préservation et de diffusion de données (NAKALA), ou encore des moteurs de recherche (Isidore). La valorisation d'une science participative (principe des wikis), l'accès aux données massives à travers de nouveaux outils (git[72]) et la production de nouvelles formes de savoir, grâce à l'interopérabilité des plateformes, dénotent un renouvellement épistémologique, organisationnel et méthodologique nécessaire à l'ère du *Big Data* en SHS. En sociologie notamment, ce renouvellement épistémologique a déjà fait l'objet de publications assez nombreuses[73]. Il reste néanmoins, et plus particulièrement en histoire, à explorer plus avant.

Conclusion

Ce parcours à travers les temporalités et espaces des traces numériques du Web et des RSNs en trois temps (collecte, collection et recherche) permet de démontrer à la fois la densité temporelle et spatiale des données à l'ère du *Big Data*, les possibilités ouvertes par une lecture multi-scalaire (*scalable reading*) pour rendre compte de la granularité spatiale et temporelle au sein de corpus massifs, mais également la profonde illusion de l'exhaustivité.

Que ce soit pour des raisons légales, institutionnelles, scientifiques ou politiques, la préservation et l'exploitation du *Big Data* se heurte à des problématiques déjà connues avant Internet et le Web – tensions, asymétries, silences et bruits, archivage d'urgence – mais implique aussi, face au poids du présentisme en histoire depuis la fin du XXe siècle[74], de repenser nos grilles de lecture pour analyser les sociétés contemporaines, comme l'ont démontré les discussions autour de l'archivage des données à la suite de l'assassinat de George Floyd en 2020[75].

L'imbroglio spatio-temporel évoqué dans cet article montre les faiblesses comme les formidables atouts qu'offrent les sources nativement numériques pour l'histoire culturelle, ce dont témoignent les exemples convoqués, dans le champ des réactions aux attentats, des sensibilités et expériences de la crise sanitaire, des mémoires et commémorations ou encore des cultures numériques. Penser le temps et l'espace du *Big Data* implique l'utilisation de matérialités nouvelles, à partir notamment des serveurs de stockage des infrastructures de recherche.

Les sources nativement numériques invitent à des efforts de renouvellement méthodologique et épistémologique pour comprendre nos sociétés contemporaines. Elles engendrent des glissements de l'archive à la donnée, des enjeux de contextualisation encore insuffisamment théorisés – et complexes en raison de la diversité des plateformes, audiences, communautés numériques, formats, archives, et données –, ainsi qu'une interdisciplinarité certaine pour acquérir de nouvelles « habiletés » cognitives et rhétoriques[76], appréhender de nouvelles « communautés imaginées »[77] et tirer pleinement bénéfice des multiples échelles spatiales et temporelles.

1 danah boyd et Kate Crawford, « CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon », *Information, Communication & Society*, vol. 15 n° 5, 2012, p. 662-679.

2 Mokrane Bouzeghoub et Rémy Mosseri (éds.), *Les Big data à découvert*, Paris, CNRS éditions, 2017.

3 Bruno J. Strasser et Paul N. Edwards, « Big Data Is the Answer … But What Is the Question? », *Osiris*, vol. 32, n° 1, 2017, p. 328-345.

4 Étienne Ollion et Julien Boelaert, « Au delà des big data. Les sciences sociales et la multiplication des données numériques », *Sociologie*, vol. 6, n° 3, 2015, p. 295-310.

5 Gilles Bastin et Paola Tubaro, « Le moment big data des sciences sociales », *Revue française de sociologie*, vol. 59, n° 3, 2018, p. 375 et suivantes.

6 Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, Thousand Oaks, SAGE Publications Ltd, 2014.

7 Reinhart Koselleck, *L'Expérience de l'histoire* (traduit de l'allemand), Paris, Points, 2011 (1ère éd. 1997).

8 François Hartog, *Régimes d'historicité. Présentisme et expériences du temps*, Paris, Points, 2011 (1ère éd. 2003).

9 Hans Ulrich Gumbrecht, *Our Broad Present: Time and Contemporary Culture*, New York, Columbia University Press, 2014.

10 Dominique Boullier, « Les sciences sociales face aux traces du big data : société, opinion ou vibrations ? », *Revue française de science politique*, n° 5, vol. 65, 2015, p. 805-828.

11 Boris Beaude, *Internet, changer l'espace, changer la société : les logiques contemporaines de synchronisation*, Limoges, FYP Éditions, 2012.

12 L'Ina (Institut national de l'audiovisuel) a ainsi recueilli en 2015-2016 plusieurs millions de tweets lors des attentats qui frappent la France. Voir le dossier « Mise en archives des réactions post-attentats : enjeux et perspectives », *La Gazette des archives*, n° 250, vol. 2, 2018. Frédéric Clavert a collecté près de 60 millions de tweets sur la crise sanitaire de la COVID-19 tout en restreignant sa collecte pour des questions techniques. « #covid19 – un pays confiné sur Twitter », C²DH, 22 mars 2020 : https://www.c2dh.uni.lu/data/covid19fr-un-pays-confine-sur-twitter.

13 Frédéric Clavert, « Temporalités du Centenaire de la Grande Guerre sur Twitter », dans Valérie Schafer (dir.), *Temps et temporalités du web*, Nanterre, Presses universitaires de Paris Nanterre, 2018, p. 113-134.

14 Niels Brügger et Ian Milligan (dir.), *The Sage Handbook of Web History*, NewYork, SAGE, 2019.

15 Valérie Schafer, *En construction. La fabrique française d'Internet et du Web dans les années 1990*, Bry-sur-Marne, Ina Éditions, 2018.

16 Frédéric Clavert, *idem*.

17 Franck Ghitalla, *Qu'est-ce que la cartographie du web ? Expéditions scientifiques dans l'univers des données numériques et des réseaux*, Marseille, OpenEdition Press, 2021.

18 Mark Graham, « Geography/internet: ethereal alternate dimensions of cyberspace or grounded augmented realities? », *The Geographical journal*, n° 2, vol. 179, 2013, p. 177-182.

19 David M. Faris, « La révolte en réseau : le « printemps arabe » et les médias sociaux », *Politique étrangère*, n° 1, 2012, p. 99-109.

20 Tufekci, Zeynep, *Twitter et les gaz lacrymogènes: Forces et fragilités de la contestation connectée*, Caen, C&F Editions, 2019.

21 Acronyme faisant référence aux plus grosses firmes du monde numérique que sont Google, Apple, Facebook, Amazon, Microsoft et IBM.

22 Voir, par exemple, Anat Ben-David et Adam Amram, « The Internet Archive and the Socio-Technical Construction of Historical Facts », *Internet Histories*, vol. 2, n°1-2, avril 2018, p. 179-201. Les auteurs montrent comment *Internet Archive* a pu archiver le web nord-coréen, inaccessible hors du pays.

23 Le Livre Blanc du GIS MOMM de 2014 souligne l'inaccessibilité au terrain de plus en plus généralisée sur la région du Moyen-Orient. Le Livre Blanc de septembre 2020 soulignait l'importance des outils et des méthodes numériques. Voir « Vers la science ouverte : la transition numérique et la recherche sur le Moyen-Orient et mondes musulmans en France : état des lieux et perspectives », *Rapport du GIS MOMM*, janvier 2020. http://majlis-remomm.fr/livre-blanc/livre-blanc-sur-les-humanites-numeriques

24 https://shakk.hypotheses.org/

25 https://offsite.hypotheses.org/

26 Sur l'importance du *counter-archiving*, voir Anat Ben-David : https://medium.com/copenhagen-institute-for-futures-studies/counter-archiving-combating-data-colonialism-be17ffead4

27 Voir Frédérick Douzet, Kévin Limonier, Selma Mihoubi et Élodie René, « Cartographier la propagation des contenus russes et chinois sur le Web africain francophone », *Hérodote*, 2020/2-3, n° 177-178, p. 77-99.

28 La fin *d'Adobe Flash* en janvier 2021 par exemple.

29 Voir les travaux du laboratoire GEODE et du Centre Internet et Société du CNRS sur les réseaux ukrainiens et russes.

30 Niels Brügger, *The Archived Web. Doing History in the Digital Age*. Cambridge, MA, The MIT Press, 2018 ; Francesca Musiani et al., *Qu'est-ce qu'une archive du Web?*, Marseille, OpenEdition Press, 2019.

31 Valérie Schafer *et al.*, « Paris and Nice Terrorist Attacks: Exploring Twitter and Web Archives », *Media, War & Conflict*, avril 2019, p. 153-170.

32 Tamara Rhodes, « A Living, Breathing Revolution: How Libraries Can Use “Living Archives” to Support, Engage, and Document Social Movements », Singapour, IFLA WLIC ; Sylvie Rollason-Cass et Scott Reed, « Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest », *New Review of Information Networking*, n° 2, vol. 2, 2015, p. 241-247.

33 Entretien entre Marguerite Borelli et Thomas Drugeon, mars 2016, https://asap.hypotheses.org/173.

34 Bruno Bachimont, *Patrimoine et numérique. Technique et politique de la mémoire*, Bry-sur-Marne, Ina Éditions, 2017.

35 Ben Els et Valérie Schafer, « Exploring special web archive collections related to COVID-19: The case of the BnL », *WARCnet Paper*, Aarhus, Université d’Aarhus, 2020.

36 Une API (*Application Programming Interface*) est un programme permettant la communication de données et de services entre deux applications.

37 Un réseau de correspondants, notamment de bibliothécaires, qui contribuent à aider aux missions du dépôt légal que relève la BnF.

38 Véronique Tranchant *et al*., « Dans les coulisses de la collecte Covid-19. Entretien sur les pratiques des correspondants du DLWeb », *Web Corpora*, 16 novembre 2020, https://webcorpora.hypotheses.org/953

39 https://archive-it.org/collections/1784

40 https://www.sucho.org

41 « Google désactive les avis sur des lieux en Russie, inondés de commentaires sur l’Ukraine », *Radio Canada*, 3 mars 2022, https://ici.radio-canada.ca/nouvelle/1866348/google-maps-russie-belarus-ukraine-internaute-desinformation

42 Dietmar Schenk, « Pourvoir de l’archive et vérité historique », *Écrire l’histoire. Histoire, Littérature, Esthétique*, n° 13-14, 2014, p. 34.

43 Doug Laney, « 3D data management: Controlling data volume, velocity and variety », *META Group Research Note* 6, 2001, p. 1-4.

44 Nous gardons « Variété » pour son ‘V’. Il s’agit ici plutôt de « diversité », c’est-à-dire de données peu structurées ou des amas de données n’ayant pas les mêmes structures.

45 Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*, Londres et New York, Verso, 2007.

46 Frédéric Clavert, « History in the Era of Massive Data », *Geschichte Und Gesellschaft*, 46.1, 2021, p. 175-194.

47 Andreas Fickers et Frédéric Clavert, « On pyramids, prisms, and scalable reading », *Journal of Digital History*, n° 1, 2020. https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb.

48 Ian Milligan, « Welcome to the web: the online community of GeoCities during the early years of the World Wide Web », dans Niels Brügger et Ralph Schroeder (dir.), *The Web as History: Using Web Archives to Understand the Past and the Present*, Londres, UCL Press, p. 137-158.

49 https://archive.org/details/archiveteam-geocities

50 Stéphanie Goutte, « Que le grand clic les croque », *Libération*, 14 février 2009, https://www.liberation.fr/medias/2009/02/14/que-le-grand-clic-les-croque_310119/

51 Anat Ben-David, « What Does the Web Remember of Its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain », *New Media & Society*, vol. 18.7, 2016, p. 1103-1119.

52 D’autres médias sociaux (*Whatsapp*, *Snapchat*, etc.), plus orientés vers des systèmes de messagerie, ne sont pas du tout archivés.

53 « Twitter API for Academic Research », Twitter Developer Platform, https://developer.twitter.com/en/products/twitter-api/academic-research

54 Le *machine learning* est une technique d’intelligence artificielle consistant en l’apprentissage automatique d’une machine pour se perfectionner à partir de données. Le *deep learning*, ou apprentissage profond, est issu des techniques de *machine learning* qui s’appuient sur un modèle de réseaux de neurones, inspiré du fonctionnement du cerveau humain.

55 Dominique Boullier, *op. cit*.

56 Frédéric Clavert, *op. cit*.

57 Sur la méthode, voir Max Reinert, « Les “mondes lexicaux” et leur “logique” à travers l’analyse statistique d’un corpus de récits de cauchemars », *Langage et société*, vol. 66, n°1, 1993, p. 5-39. Sur IRaMuTeQ : https://iramuteq.org

58 Projet (C20/SC/14758148) soutenu par FNR de 2021 à 2024 : hivi.uni.lu

59 Même internet devenu viral en 2013, notamment sur YouTube, où un individu puis un groupe de personnes dansent costumés sur la musique de DJ Baauer.

60 Phénomène Internet né en 2007 et consistant à renvoyer un internaute, attiré souvent par la promesse d’un contenu attrayant, vers la chanson et/ou le clip de Rick Astley « *Never Gonna Give You Up* » de 1987, souvent au moyen d’un lien hypertexte.

61 Christine Barats, Valérie Schafer et Andreas Fickers, « Fading Away… The challenge of sustainability in digital studies », *Digital Humanities Quaterly*, vol. 14 (3), 2020.

62 Barend Mons, « *Data Stewardship for Open Science. Implementing FAIR* Principles », Boca Rota, CRC Press, 2018.

63 *Web ARChive studies network researching web domains and events*. https://cc.au.dk/en/warcnet/

64 Ces entretiens sont disponibles à https://cc.au.dk/en/warcnet/warcnet-papers

65 Jane Winters et Andrew Prescott, « Negotiating the born-digital: a problem of search », *Archives and Manuscripts*, n° 47, vol. 3, 2019, p. 391-403.

66 Philippe Rygiel, *Historien à l'âge numérique*, Villeurbanne, Presses de l'ENSSIB, 2017, p. 15 et p. 24.

67 Valérie Schafer et Frédéric Clavert, « From WG2 Datathon to AWAC2. Exploring IIPC special COVID collection thanks to the Archive Unleashed Project », WARCnet Aarhus Conference, 3 novembre 2021. https://www.slideshare.net/WARCnetWebArchiveStu/from-wg2-datathon-to-awac2-exploring-iipc-special-covid-collection-thanks-to-the-archive-unleashed-project

68 Dominique Boullier, *Sociologie du numérique*, Paris, Armand Colin, 2019, p. 528.

69 La surcharge informationnelle n'est évidemment pas propre aux chercheurs. Voir Francis Jauréguiberry, « Désir et pratiques de déconnexion », *Hermès, La Revue*, vol. 84, n°2, 2019, p. 98-103.

70 https://www.huma-num.fr/

71 https://www.progedo.fr/

72 Git est un projet porté notamment par Linus Torvalds (l'initiateur de la plateforme d'opération libre concurrente de Windows, Linux) permettant la gestion des projets de développement informatique, qui promet notamment un suivi très fin des modifications et des versions des logiciels (https://git-scm.com/). Il fait l'objet d'expérimentations d'écriture en histoire, par exemple par le *Journal of Digital History* (https://journalofdigitalhistory.org), lancé par le C²DH auquel l'auteur et les autrices de cet article sont affiliés.

73 Gilles Bastin et Paola Tubaro, *op. cit.*, mais aussi Dominique Boullier, « Vie et mort des sciences sociales avec le big data », *Socio. La nouvelle revue des sciences sociales* (4), 2015, p. 19-37.

74 François Hartog, *op. cit.*

75 Valérie Schafer et Jane Winters, « The values of web archives », *International Journal of Digital Humanities*, 2 (1-3), 2021, p. 129-144.

76 Imad Saleh et Hakim Hachour, « Le numérique comme catalyseur épistémologique », *Revue française des sciences de l'information et de la communication*, n°1, juillet 2012.

77 Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, Londres, Rev. Ed, 1991 (1ère éd. 1983).

ILLUSTRATIONS

Figure 1 – Classification hiérarchique descendante (CHD, méthode Reinert, telle qu'implémentée dans IRaMuTeQ) des tweets francophones du corpus collecté pendant le Centenaire
docannexe/image/2791/img-1.png

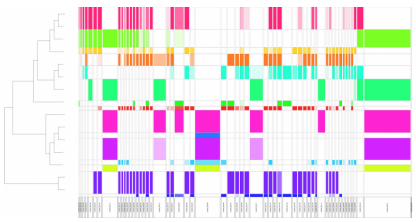


Figure 2 - Projection chronologique (par mois) des classes de la CHD
docannexe/image/2791/img-2.png

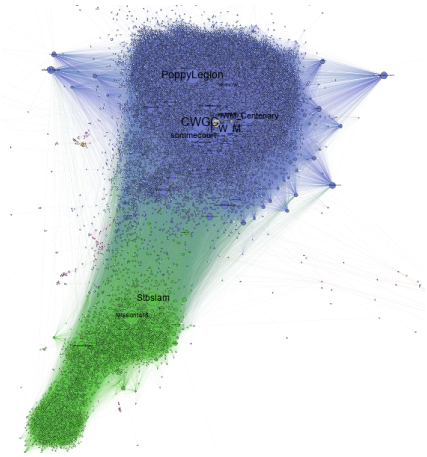


Figure 3 - Réseau des relations (réponses, citation, retweets) du Centenaire de la Grande Guerre
docannexe/image/2791/img-3.png

REFERENCES

Electronic reference

Frédéric Clavert, Sophia Mahroug and Valérie Schafer, « Préservation et distorsion : l'espace-temps des réseaux socio-numériques et du web archivé », *Revue d'histoire culturelle* [Online], | 2022, Online since 15 octobre 2022, connection on 20 décembre 2022. URL : http://revues.mshparisnord.fr /rhc/index.php?id=2791

AUTHORS

Frédéric Clavert
Frédéric Clavert (C²DH, Université du Luxembourg) est professeur assistant en histoire contemporaine au C²DH (Centre d'histoire contemporaine et numérique) de l'Université du Luxembourg depuis septembre 2017. Il s'intéresse aux traces du passé en ligne, notamment sur les réseaux sociaux numériques et plus particulièrement sur Twitter d'une part, sur la relation renouvelée des historiens et historiennes à leurs archives à l'ère numérique d'autre part. Il co-dirige le livre en ligne *Le Goût de l'archive à l'ère numérique* avec Caroline Muller (Rennes 2) et est *managing editor* du *Journal of Digital History.* frederic.clavert@uni.lu

Sophia Mahroug
Sophia Mahroug (CRHXIX, Sorbonne-Université/ C²DH, Université du Luxembourg) est doctorante en histoire contemporaine à Sorbonne Université (CRHXIX) et à l'Université du Luxembourg (C²DH). Dans le cadre de sa thèse de doctorat, elle étudie la culture de guerre en République islamique d'Iran à travers les archives du Web et les réseaux sociaux numériques de la fin des années 1990 à nos jours. Ses recherches ont pour objectif de reformuler une histoire politique de l'Iran du XXIe siècle à partir des sources nativement numériques et d'une réflexion épistémologique sur l'apport des humanités numériques dans l'étude du Moyen-Orient contemporain. Sa dernière intervention au congrès du GIS Moyen- Orient en juin 2021 s'est focalisée sur le rôle de l'armée et des militaires dans la culture nationale en Iran. mahroug.sophiac@gmail.com

Valérie Schafer
Valérie Schafer (C²DH, Université du Luxembourg) est professeure d'histoire européenne contemporaine au C²DH (Centre d'histoire contemporaine et numérique) de l'Université du Luxembourg depuis février 2018. Elle a auparavant travaillé au CNRS et est chercheuse associée au Centre Internet et Société (CIS - CNRS UPR 2000). Elle est spécialisée dans l'histoire de l'informatique, des télécommunications et des réseaux de données. Ses principaux centres de recherche sont l'histoire de l'Internet et du Web, l'histoire des cultures numériques européennes, notamment les phénomènes de viralité, et le patrimoine nativement numérique (en particulier les archives Web). valerie.schafer@uni.lu

By this author

« Never gonna give you up »
Published in *Revue d'histoire culturelle*, | 2022

Electronic ISSN 2780-4143