



HAL
open science

Web vivant et web archivé, aux sources de l'histoire nativement numérique

Sophie Gebeil

► **To cite this version:**

Sophie Gebeil. Web vivant et web archivé, aux sources de l'histoire nativement numérique. Le web : source et archive, Soutenu par le GIS CollEx-Persée et porté par le Service Commun de Documentation de l'Université de Lille et la Bibliothèque nationale de France, en partenariat avec le GERiCO de l'Université de Lille, Sciences Po et le Campus Condorcet, le réseau ResPaDon se donne pour objectif d'ouvrir un espace de dialogue entre chercheurs et professionnels des bibliothèques et des archives., Apr 2023, Lille, France. hal-04058328

HAL Id: hal-04058328

<https://amu.hal.science/hal-04058328v1>

Submitted on 4 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Web vivant et web archivé, aux sources de l'histoire nativement numérique

Si la consultation d'archives numérisées en ligne est une pratique désormais bien installée, les historien-ne-s qui prennent le web comme source pour une histoire du temps présent demeurent, malgré un intérêt croissant, peu nombreux au regard des autres sciences humaines et sociales. Certes, l'étude de la période très contemporaine ne concerne qu'un segment de la profession. Mais à cette dimension s'ajoutent les difficultés liées à l'appréhension des documents nativement numériques. Depuis la fin du XIX^e siècle, les historien-ne-s n'ont cessé d'intégrer de nouveaux types de sources et se sont émancipés du primat de la source écrite. Dès lors, pourquoi l'intégration du web vivant puis du web archivé comme matériau historiographique n'est-elle pas une simple étape supplémentaire dans ce processus d'élargissement mais invite au contraire à questionner l'ontologie disciplinaire ? Nous verrons dans un premier temps que pour l'historien-ne en terre numérique, les sources en ligne non archivées constituent des documents complexes qui s'accordent mal avec les démarches historiennes. Par conséquent, le recours aux archives du web est incontournable pour une histoire nativement numérique mais, en tant que documents récrés à partir du web, elles imposent aussi la plus grande vigilance et un nouveau rapport à l'archive, ce qui sera abordé dans un second temps. Enfin, les travaux de recherche en histoire basés sur les archives du web depuis le début des années 2000 permettent de mieux comprendre comment l'usage de ce patrimoine nativement numérique recréé participe du renouvellement des méthodes historiques en lien avec les humanités numériques. Étayée par des travaux de recherche développés depuis une dizaine d'années concernant l'histoire des pratiques mémorielles et des mobilisations collectives à partir des archives du web de la BNF et de l'INA, cette communication se propose d'envisager les modalités selon lesquelles le recours aux archives du web comme source en histoire induit un renouvellement des méthodes du fait du caractère computationnel du web et reconfigure le rapport à l'archive, interrogeant ainsi l'ontologie de la discipline.

1. Les spécificités disciplinaires de l'histoire face aux sources en ligne

Dans le premier chapitre de leur *Introduction aux études historiques*, Charles-Victor Langlois et Charles Seignobos fondaient le caractère « scientifique » de l'histoire sur la critique documentaire, affirmant que « rien ne supplée aux documents : pas de documents, pas d'histoire¹ ». Depuis l'École méthodique, l'ouverture de nouveaux champs de recherche s'est traduite par un enrichissement des méthodes d'analyse, au gré des innovations historiographiques : l'École des Annales explora les sources matérielles, l'histoire orale encouragea le recours aux témoignages puis l'histoire culturelle mobilisa les sources audiovisuelles. Ainsi, désormais « tout est source » pour l'historien-ne, « de Goya à Chantal Goya » selon la formule de Pascal Ory². Dans cette perspective, plusieurs chercheur-e-s ont naturellement intégré des documents en ligne dans leurs travaux en tant que sources complémentaires, appréhendant le Web comme « un gisement multiforme de sources

¹ Langlois C.V., Seignobos C., *Introduction aux études historiques*, 2014 (1^e éd.1899), Lyon, ENS éditions, 2014.

² Ory P., *op.cit.*

documentaires³ ». La généralisation d'Internet offre un vaste territoire d'exploration, mais prendre le Web comme source primaire induit une complexification des méthodes. Du fait de leurs spécificités, les sources en ligne remettent en question les notions de matérialité et surtout de stabilité du document historique, aspect déterminant dans l'administration de la preuve et dans l'ontologie de la discipline.

Dans une perspective d'histoire des médias, les sources nativement numériques renvoient ainsi à l'ensemble des traces et des documents numériques⁴ mobilisés par l'historien·ne dans l'étude du passé : une page Web, les contenus d'un DVD, des métadonnées, etc. Désignant l'action technique qui consiste en le codage d'une information (son, texte, image) selon le système binaire 0 ou 1 (*Binary Digit* ou *bit*), l'adjectif numérique renvoie aussi aux nouvelles « méthodes technoscientifiques de traitement de l'information par le codage et le calcul à travers les programmes informatiques, et se différencie des techniques dites analogiques⁵ ». Les technologies numériques permettent depuis la fin des années 1980 de créer des contenus multimédias, donnant à voir, sur un même écran, des types de sources diverses, textuelles, sonores ou audiovisuelles, nées numériques ou non. À ce caractère hypermédia s'ajoutent les effets de la dissociation entre l'information et le support qui facilitent la modification, à vitesse rapide, du document numérique, intrinsèquement malléable : « La possibilité de manipulation est co-originale à la nature même du support numérique [...]. Ainsi, l'essence du numérique [...] est-elle : “ça a été manipulé”⁶. »

Par ailleurs une page Web n'existe pas isolément, elle s'inscrit dans un environnement structuré par le système interopérable de l'hypertexte sur lequel repose l'architecture du Web. Dès lors, « le document n'est plus seulement un objet, mais aussi un nœud ou une tête de réseau qui permettent de passer d'un document à un autre⁷. Ainsi Niels Brügger considère que le Web est doublement digital⁸ ayant la particularité d'être « nativement fragmenté », « nativement hyperlié⁹ » et composé de couches de texte qui renvoient à la distinction entre les éléments sémiotiques visibles par l'internaute sur l'écran et le texte écrit en différentes versions de HTML (« code »). À ces caractéristiques techniques sont associées des transformations des pratiques d'élaboration, de circulation et de consultation des documents en réseau qui participent de la complexification documentaire, à l'origine d'un nouvel ordre documentaire caractérisé par la « redocumentarisation¹ ».

Fondée sur la diplomatique, la critique documentaire constitue, depuis la fin du XIX^e siècle, un fondement de la démarche historique déclinée en quatre dimensions : la critique externe (la forme du document, en lien avec les sciences auxiliaires de l'histoire), la critique interne (le discours ou la représentation), la critique de provenance (point de vue et authenticité) et la critique de portée (destinataire)¹⁰. « L'âge des fichiers¹¹ » transforme les

³ Soulet J.-F., *L'histoire immédiate : historiographie, sources et méthodes*, Paris, Armand Colin, 2009, p. 220.

⁴ Camille Paloque_Berges définit le document numérique comme « un document accessible *via* un support numérique et interprétable par un individu ». Paloque-Berges C., « Les sources nativement numériques pour les sciences humaines et sociales », *Histoire@Politique*, 30, 2016, p. 221-244, <<https://www.cairn.info/revuehistoire-politique-2016-3-page-221.htm>>.

⁵ *Ibid.* ; Concernant les technosciences, l'auteure renvoie à Hottot G., *Le signe et la technique*, Paris, Aubier-Flammarion, 1984.

⁶ Bachimont B., *Le sens de la technique: le numérique et le calcul*, Paris, Belles Lettres, 2010 cité par Paloque-Berges C., *op. cit.*

⁷ Salaün J.-M., *op. cit.* p. 36. Ajoutons que ce protocole d'échange d'informations s'appuie sur le format de programmation informatique normalisé HTML ; Brügger N., *op. cit.* p. 23.

⁸ Brügger N., “Where does the born – and reborn – digital material take the Digital Humanities ?”, 22 mai 2017. <<https://fr.slideshare.net/UCLDH/where-does-the-born-and-reborn-digital-material-take-the-digital-humanities>>

⁹ Les éléments visibles sont intrinsèquement « nativement fragmentés », pouvant se trouver hébergés dans des serveurs distincts, et ensuite liés entre eux au sein du fichier HTML. Enfin, l'hyperlien permet de connecter deux éléments sur le même ordinateur ou sur deux objets connectés différents. Étant à l'origine du caractère « nativement hyperlié » du Web, l'hyperlien occupe donc un rôle fondamental et constitue « le fil dont la toile

¹⁰ Langois C. V., Seignobos C., *op. cit.*

modalités de l'enquête et les termes mêmes de l'étude. Par exemple, la critique externe implique de cerner le dispositif de médiation, et donc nécessite d'élargir le spectre des sciences auxiliaires de l'histoire en y intégrant la maîtrise de connaissances et de compétences relevant des cultures numériques. De plus, la critique interne sur un corpus de données massives peut nécessiter le recours à des outils numériques (outils d'analyse lexicologique comme Iramuteq par exemple) au profit d'une « histoire par les données¹² ». L'absence d'outil d'analyse adapté aux spécificités disciplinaires ne facilite pas l'appréhension d'Internet comme source¹³. Une deuxième difficulté est d'ordre ontologique et renvoie à la place du document dans la méthodologie historique. Le caractère « volatile », « fluide » des sources en ligne est unanimement rappelé par les cyber-historien-ne-s au début des années 2000, principalement à l'aune du modèle des sources manuscrites ou imprimées.

La matérialité des traces laissées par l'activité humaine est aussi associée à l'élaboration des faits historiques chez Marc Bloch évoquant les « matériaux fournis par les générations passées » sur lesquels les historien-ne-s s'appuient. L'analyse des sources inclut d'ailleurs l'étude du support matériel en relation avec les sciences auxiliaires de l'histoire, comme l'épigraphie par exemple, permettant de cerner les modalités d'élaboration du document. Or, comme l'expliquent Frédéric Clavert et Serge Noiret :

« Les sources primaires ne sont souvent plus reliées au contexte matériel qui leur faisait "prendre sens" et les validait dans leurs contextes : dans le monde numérique, un des grands problèmes est certainement celui de l'individuation des contextes signifiants, ce que les philologues appellent l'histoire de la construction des textes et des documents¹⁴. »

La dissociation entre l'information et son support invite donc à repenser la notion de traces en y intégrant la dimension numérique et interconnectée. Les historien-ne-s des médias audiovisuels et du cinéma savent bien que le caractère « immatériel » des sources ne concerne pas seulement les documents issus du Web puisque la dématérialisation des sources n'est donc pas inédite, mais le numérique induit un changement de paradigme documentaire caractérisé par la « re-construction¹⁵ ». Par rapport aux sources audiovisuelles, les sources en ligne sont l'objet d'une instabilité plus importante car elles s'inscrivent dans un environnement en mouvement qui évolue au gré des interactions émanant des acteur-ric-e-s de la Toile : en consultant une page Web, l'utilisateur traite une actualisation du document à un instant donné, dans un processus de modification constant et de reconstruction à partir de ressources multiples¹⁶. À cela s'ajoute la volatilité des contenus qui demeure un frein fondamental à la constitution d'un corpus nativement numérique dans une perspective historique car elle remet en cause la stabilité et la pérennisation des sources¹⁷. Cette instabilité contraste avec le sentiment d'accessibilité qui prévaut dans l'usage, y compris lorsqu'il s'agit de citer des pages Web dans un travail de recherche. Au-

¹¹ Salaün J.-M., *op. cit.*

¹² Clavert F., « Une histoire par les données ? Le futur très proche de l'histoire des relations internationales », *Bulletin de l'Institut Pierre Renouvin*, 44, 2016, p. 119-130.

¹³ Rygiel P., « L'enquête historique à l'ère numérique », *Revue d'histoire moderne et contemporaine*, 58, 2012, p. 30-40.

¹⁴ Clavert F., Noiret S., *L'histoire contemporaine à l'ère numérique*, *op. cit.*, p. 24.

¹⁵ Bachimont B., « Archivage audiovisuel et numérique: les enjeux de la longue durée », in Leblond C. (dir.), *Archivage et stockage pérennes*, Paris, Hermès, Lavoisier, 2009. <http://cours.ebsi.umontreal.ca/sci6116/Ressources_files/Bachimont-Archivage.pdf>

¹⁶ Tricot A., *Mémoire d'Habilitation à Diriger des Recherches*, p. 13. <http://andre.tricot.pagesperso-orange.fr/Tricot_HDR.pdf> 3 Noiret S., *op. cit.*

¹⁷ La durée de vie moyenne d'une URL étant estimée à 9,3 ans. Musiani F., Paloque-Bergès C., Schafer V., Thierry B. G., *op. cit.*

delà de la désagréable sensation générée par l’affichage d’une « erreur 404 » indiquant que la demande ne peut aboutir, cela brise le « contrat de vérité¹⁸ » qui lie l’historien·ne et son destinataire. En effet, l’historien·ne, dans son travail « d’objectivation documentaire » est tenu « de répondre à la confiance que lui accorde son lecteur ». Comme le soulignait March Bloch, la citation des sources utilisées en notes infrapaginales s’inscrit dès lors dans l’établissement de la preuve et offre la possibilité au lecteur·rice de pouvoir consulter le matériau de l’historien·ne¹⁹. Le lien « brisé », à cause d’une adresse URL qui n’est plus valable rend caducs les faits établis à partir des documents cités. La possibilité de vérifier les sources énoncées est un point fondamental de l’administration de la preuve au sein de la discipline. Une première parade réside dans les dispositifs personnels de sauvegarde (captures d’écran, collecte personnelle *via* API, etc.). Cependant, ces procédés n’en garantissent pas la pérennisation et ne permettent pas au futur lecteur de consulter la totalité des sources convoquées. Le caractère volatil des sources n’est pas une nouveauté en soi, une bobine filmique peut par exemple brûler ou se dégrader. En régime numérique, la malléabilité des sources est non seulement généralisée, mais en plus, leur caractère hyperlié en limite la possibilité d’identification dans le temps malgré une apparente traçabilité. En l’absence de « mise en archive » du Web, les sources nativement numériques agissent comme un mirage : elles sont facilement accessibles, consultables en ligne, il semble aisé d’en sauvegarder une trace, mais leur instabilité a pour conséquence une « citabilité » fragile et problématique qui rend leur appréhension comme source historique complexe.

Face à ces défis, les archives du Web sont indispensables pour les historien·ne·s actuels et à venir car elles permettent de stabiliser des corpus et de retrouver des traces qui ne sont plus accessibles sur le Web aujourd’hui. Néanmoins leur exploitation impose de multiples précautions et instaure un nouveau rapport à l’archive.

2. Nouveau rapport à l’archive

Depuis la professionnalisation de l’histoire au XIX^e siècle, l’archive, au sens large, occupe une place centrale dans l’élaboration des faits historiques malgré la diversification des sources. Cette interdépendance est rappelée par Paul Ricœur dans *La Mémoire, l’Histoire, l’Oubli* : « Le moment de l’archive, c’est le moment de l’entrée en écriture de l’opération historiographique²⁰ ». Comme l’illustre l’histoire des médias, de nombreux champs de recherche sont dépendants de l’existence ou non d’un processus d’archivage. Si les documents de l’historien·ne ne se restreignent pas aux seules archives, celles-ci constituent une source privilégiée dans l’établissement de la preuve. Les archives du Web qui permettent aux chercheur·e·s d’accéder à d’anciennes versions de pages Web ou de stabiliser un corpus en ligne sont cependant fondées sur un paradoxe inhérent puisque la pérennisation et la transmission ne sont possibles qu’en figeant un environnement par définition mouvant et instable, à partir duquel les archives sont reconstruites. Ces considérations ne sont pas sans incidence sur l’usage de ces fonds à des fins historiques. Malgré des différences de pratiques, le principe de l’archive du Web par *crawler* est identique. Les contenus sont d’abord collectés puis l’indexation aboutit à la séparation entre les objets

¹⁸ Dosse F., « Le moment Ricœur », *Vingtième Siècle. Revue d’histoire*, 69, janvier-mars 2001, p. 137, www.web/revues/home/prescript/article/xxs_0294-1759_2001_num_69_1_1289.

¹⁹ Bloch M., *Apologie pour l’histoire ou le métier d’historien*, Paris, Armand Colin, 1952.

²⁰ Ricœur P., *La mémoire, l’histoire, l’oubli*, Paris, Seuil, 2000.

archivés (textes, images, etc.) et les informations de présentation permettant d'en structurer l'affichage (ex : feuille de style, métadonnées, etc.). À chaque contenu est associée une clé d'identification. Au moment de la consultation, le contenu et les informations structurantes sont de nouveau assemblés en fonction de la requête de l'utilisateur (page x, à la date t). Il ne s'agit ni de documents à propos du Web vivant, ni de copies intégrales figées et isolées (« *monument* »), ni de simples empreintes. Il en résulte une perte de l'original, considérée comme étant l'une des spécificités de l'archive du Web comme source, n'étant ni un miroir du Web vivant, ni une copie²¹. Ces dernières peuvent être considérées comme des documents *de* l'Internet et du Web. L'ensemble des opérations à l'origine de l'archivage correspond à une re-construction, une re-crétion²² qu'il appelle « *re-mediation* » : « Les contenus déjà médiatisés (sur le Web) forment le matériau parent pour une nouvelle médiation (dans l'archive)²³. » Ces dénominations sont à l'origine de l'expression *reborn digital heritage* fondée sur la notion de patrimoine nativement numérique porté par l'UNESCO en 2003, l'ajout du préfixe visant à en souligner le caractère reconstitué. L'ensemble du processus a pour objectif de faire en sorte que ces « documents du Web » correspondent au maximum aux documents originaux qui étaient initialement présents sur le Web.

Ces versions uniques qui composent le Web archivé sont d'autant plus complexes à appréhender que le processus de re-crétion induit des incohérences spatiales et temporelles, visibles en consultation. L'archivage du Web peut en effet créer des originaux n'ayant jamais existé en ligne. Niels Brügger qualifie ce phénomène « d'inconsistance spatio-temporelle » puisqu'à l'agencement d'éléments incohérents s'ajoute la variabilité de l'archivage de la profondeur du site au fil des collectes²⁴. Une page affichée est un patchwork avec des éléments, des *fragments* issus du Web, qui, s'ils sont absents, sont récupérés dans le serveur à la date la plus proche, ce qui amène Niels Brügger à comparer les archives Web à un seau géant de fichiers (*giant bucket of files*)²⁵. Ces spécificités invitent l'historien·ne à repenser son rapport à l'archive et cela commence par questionner la notion d'*archive* pour désigner les différentes modalités visant à collecter et rendre accessible le Web (*via* un *crawler*, une API, une capture d'écran, un film, etc.) et adoptée par les chercheurs et les institutions du domaine. Le patrimoine nativement numérique recréé est en réalité une recréation à partir de documents nativement numériques issus du Web. Les spécificités des archives du Web induisent de plus une grande prudence à l'égard des pages consultées. Face au Web archivé, le croisement des sources intrinsèques à l'enquête historique commence donc par la nécessité d'obtenir des informations sur le processus d'élaboration de l'archive, afin d'évaluer la qualité des pages consultées.

²¹ « Lack of an Original », Brügger N., *op. cit.* p. 104.

²² Bachimont B., « Archivage audiovisuel et numérique : les enjeux de la longue durée », in Leblond C. (dir.), *Archivage et stockage pérennes*, Paris, Hermès/Lavoisier, 2009, p. 195-222 ; Mussou C., « Et le Web devint archive : enjeux et défis », *Le Temps des médias*, 19, 2012, p. 259-266.

²³ Brügger N., *op. cit.*, p. 30.

²⁴ Brügger N., *op. cit.* p. 107-109.

²⁵ *Ibid.* p. 111.

3. Les archives du web comme source : transformation des usages autour des corpus et des méthodes

Les travaux de recherche en histoire basés sur les archives du web depuis le début des années 2000 permettent de mieux comprendre comment l'usage de ce patrimoine nativement numérique recréé participe du renouvellement des méthodes historiques en lien avec les humanités numériques.

Les premiers travaux et initiatives des décennies 1990 et 2000 ont mis en évidence les problématiques spécifiques soulevées par les archives du web, à l'image des collectes du web électoral (BNF) ou du projet e-diaporas (INA)²⁶. Sur ces bases, l'année 2010 inaugure une décennie marquée par l'essor de plusieurs projets de recherche, qui amènent les communautés à appréhender les archives du web comme corpus et à concevoir des méthodologies d'analyse adaptées aux spécificités disciplinaires et aux particularités de ce « patrimoine nativement numérique recréé ». En France, cette dynamique s'effectue autour d'acteurs de l'enseignement supérieur et de la recherche en relation étroite avec les institutions en charge du dépôt légal du web. Par exemple, l'INA inaugure en 2010 une série d'ateliers de recherche méthodologique du dépôt légal du Web média²⁷. En 2014, la BNF organise l'Assemblée Générale de l'IIPC à Paris. Plusieurs projets de recherche en histoire engagent alors des corpus issus du dépôt légal du Web selon des configurations distinctes. Ce mouvement prend la forme de recherches individuelles, à l'image de l'exploration des pages archivées par Internet Archive du site *Geocities* par Ian Milligan²⁸. C'est aussi la période de ma thèse (2010/2015) consacrée aux mémoires de l'immigration maghrébine sur le web français durant les années 2000, fondées sur les archives web de la BNF et de l'INA. Si l'entrée est davantage micro-historique, cela n'empêche pas la masse de données. Les analyses qualitatives des contenus sont associées avec les méthodes issues des humanités numériques. Ainsi Ian Milligan utilise l'analyse des URLs pour mettre en évidence les relations entre les différents groupes du forum. Dans une moindre mesure, l'exploration conduite avec David Rapin en 2014 visait à permettre une lecture distante des sites Oumma.com, et des webdocumentaires de Raspouteam, à travers l'extraction des URLs et des données textuelles.

Dans le même temps se développent des projets associant des équipes de recherche et des archivistes au sein desquels les archives du web sont sollicitées comme espace de dépôt de corpus co-construits ou comme source pour étudier le web passé. Le corpus du projet « Grande Guerre sur le Web » a notamment été constitué lors du projet en croisant des contenus identifiés en ligne et des pages déjà collectées par la BNF²⁹. Cette démarche mixte est également celle qui préside au projet « Web90 – Patrimoine, Mémoires et Histoire du Web » dans les années 1990 porté par Valérie Schafer³⁰.

La place du web archivé dans les corpus diffère en fonction des thématiques et des périodes, allant d'une source principale lorsqu'il s'agit d'étudier les ruines du web, complétées par des

²⁶ « Cliquer, voter : l'internet électoral. Un parcours guidé dans les archives de l'internet », BNF, Mars 2008,

https://www.bnf.fr/sites/default/files/2018-11/cliquer_voter_internet_electoral_parcours.pdf ; e-Diasporas, <http://www.e-diasporas.fr>

²⁷ https://actions-recherche.bnf.fr/bnf/anirw3.nsf/IX01/A2014000123_assemblee-generale-iipc-2014-paris?OpenDocument&onglet=&categorie=&n=&i=E2012000511

²⁸ Ian Milligan (2015). Finding Community in the Ruins of GeoCities: Distantly Reading a Web Archive. UWSpace.

<http://hdl.handle.net/10012/11650>

²⁹ Beaudouin, Valérie, Philippe Chevallier, and Lionel Maurel, eds. *Le Web français de la Grande Guerre: réseaux amateurs et institutionnels*. Presses universitaires de Paris Nanterre, 2018.

³⁰ Schafer, Valérie, ed. *Temps et temporalités du Web*. Nanterre : Presses universitaires de Paris Nanterre, 2018. Web.

<http://books.openedition.org/pupo/5988>.

traces retrouvées sur le web vivant, au dépôt de corpus constitué en ligne comme ce fût le cas dès pour le projet e-diasporas. Ces configurations multiples ont donné lieu à l'essor de méthodologies mixtes, croisant approches qualitatives et quantitatives, élaborées au fil des explorations.

L'approche par cas³¹, centrée sur l'étude d'un nombre réduit de site ou même sur une monographie d'un seul site, offre la possibilité d'historiciser les contenus et de saisir les spécificités et l'évolution des scénographies du passé sur la Toile, identifier les conditions de production et les individus à l'origine des contenus, replacer ces usages très contemporains dans l'enchevêtrement des temporalités en articulant le temps du Web avec la période pré-Web. Pour un-e néophyte en terre numérique, elle constitue d'abord une première expérience de confrontation face aux sources issues du Web et des archives du Web à « taille humaine », permettant ainsi de dédramatiser l'usage de sources nativement numériques contrairement à une collecte de données massives. Dans le même temps, ce premier contact avec un corpus réduit rend aussi compte de la dimension sociotechnique des sources et de la nécessaire acculturation en la matière, qui peut ensuite être mise à profit ultérieurement dans le cadre d'une approche quantitative. Dans le cadre de l'étude des mémoires de l'immigration maghrébine sur le web, quatre d'analyse étaient distingués : l'identification et la fabrique du dispositif, les modes de scénographies numériques et l'expérience usager, l'interprétation et les mises en scène mémorielles ainsi que l'évolution du dispositif dans le temps. Ensuite, elle offre la possibilité d'enrichir le corpus d'étude par des sources complémentaires, notamment diverses productions culturelles contemporaines ou datant de la période pré-Web, tout en replaçant la Toile dans son contexte social³². En outre, cela permet également, dans le cadre d'entretiens, de confronter les créateur-riche-s des contenus aux traces archivées par la BnF ou l'INA lors de l'enquête orale. En somme, la démarche permet d'historiciser les dispositifs en ligne pour leur redonner leur épaisseur humaine et temporelle.

Cependant, l'entrée par des corpus relativement limités, est complémentaire avec le développement de méthodes computationnelles au service d'une lecture distante, incontournable pour explorer des données massives. Les outils et les méthodes développées au service de l'étude des corpus issus du web vivant ne peuvent cependant pas faire l'objet d'une simple reproduction au sein du web archivé qui constitue un environnement spécifique. Plusieurs travaux ont ainsi montré la difficulté de réaliser des graphes de lien dans le web archivé au moyen d'un outil comme Gephi. Ainsi la thèse de Quentin Lobbé soutenue en 2018 apporte un nouvel éclairage³³. En se fondant sur le site Yabiladi archivé par l'INA dans la collection de l'Atlas e-diasporas durant les années 2000, il a conçu un outil « d'exploration désagrégée de corpus d'archives Web » fondé sur la notion de « fragment », défini comme « un sous-ensemble cohérent et auto-suffisant d'une page Web » et constitué en unité d'analyse. Il étudie ainsi les dynamiques propres au collectif, aujourd'hui dissout, d'utilisateurs du forum Yabiladi.fr à partir du corpus archivé.

Dernièrement, plusieurs projets d'envergure se sont engagés dans des approches mixtes, alliant étude qualitative et lecture distante, au profit de démarches agiles en capacité de

³¹ Passeron J.-C., Revel J., *Penser par cas*, Paris, l'École des Hautes Études en Sciences Sociales, 2005.

³² Barats C., *Manuel d'analyse du Web en sciences humaines et sociales*, Paris, Armand Colin, 2013.

³³ Quentin Lobbé, *Archives, fragments Web et diasporas. Pour une exploration désagrégée de corpus d'archives Web liées aux représentations en ligne des diasporas*. Université Paris-Saclay, 2018. <<https://hal.inria.fr/tel-01963548>>, Thèse récompensée par le Prix de la recherche de l'INA thèque.

s'adapter au gré des découvertes et des perspectives d'étude, notamment grâce à l'émergence d'environnement facilitant l'échange des données archivées et le déploiement d'outils d'exploration spécifiques. C'est notamment le cas du projet AWAC2³⁴, qui à partir de la collecte internationale « Novel Coronavirus » de l'IIPC, a développé une démarche adaptative (*scalable reading*), alliant lecture distante grâce à la suite mise à disposition dans le cadre de *The Archive Unleashed Project*³⁵ et une étude plus fine des questions de genre dans les publications en ligne liée à la pandémie de Covid-19. De même, le projet *Polyvocal Interpretation Of Contested Colonial Heritage* se propose d'aborder l'évolution de la médiatisation en ligne de la Marche pour l'égalité et contre le racisme de 1983, en mettant l'accent sur les périodes commémoratives³⁶. En partenariat avec l'INA, l'étude mêle une analyse distante des données issues du dépôt légal du web média remontant jusqu'en 2003 à travers le développement d'outils d'extraction et d'analyse sémantique en sciences de l'informatique, et l'examen historique d'études de cas telles que les controverses en ligne qui entourent la sortie en salle du film *La Marche* en 2013, la dimension marseillaise de l'histoire de l'événement, l'évolution des représentations télévisuelles et l'étude des reconfigurations mémorielles du fait de l'émergence de demandes de réexamen de la place du passé colonial dans les politiques culturelles³⁷. Cet aperçu de la place du web archivé dans les recherches historiques et du développement d'outils et de méthodes interdisciplinaires, illustre la vitalité d'un champ et la capacité des archives du web à constituer une *trading zone* entre les historiens et d'autres domaines disciplinaires.

Conclusion

Nouveau matériau « doublement digital » et recréé, le recours au web archivé comme source pour les historien.ne.s impose finalement de reconsidérer le rapport à l'archive, à la critique documentaire et invite à transformer les méthodes d'analyse historiques au profit d'approches interdisciplinaires et multiscalaires.

Les récentes mises à disposition auprès des publics de services fondés sur l'intelligence artificielle permettant à tout internaute de produire plus rapidement encore des textes automatisés mais aussi des images photographiques artificielles constituent de nouveaux défis pour écrire une histoire nativement numérique. Ces outils déjà exploités dans le cadre de projets d'analyse visuelle ou sémantique, renforcent cependant la complexité à établir des faits historiques en contexte numérique mais nous rappellent aussi combien une histoire culturelle du web, attachée à contextualiser et à décortiquer la fabrique des contenus en ligne, les représentations et les formes de réappropriation pour cerner les transformations à l'échelle du temps du web mais aussi au regard des pratiques culturelles préexistantes, est plus que jamais d'actualité.

³⁴ Aasman, S., Brügger, N., Clavert, F., de Wild, K., Gebeil, S. & Schafer, V., "Analysing Web Archives of the Covid-19 Crisis through the IIPC collaborative collection: early findings and further research questions", IIPC netpreserve.org, 2021, <https://netpreserveblog.wordpress.com/2021/11/02/analysing-web-archives-of-the-covid-19-crisis-through-the-iipc-collaborative-collection-early-findings-and-further-research-questions/>

³⁵ The Archives Unleashed Project, <https://archivesunleashed.org>

³⁶ Site du projet : <https://picch-project.org> ; présentation du volet français : <https://madi.hypotheses.org/883>

³⁷ Gallot Fanny, « Entretien avec Sophie Gebeil », *Les cahiers de la lutte contre les discriminations*, 2023.

Bibliographie

- Aasman, S., Brügger, N., Clavert, F., de Wild, K., Gebeil, S. & Schafer, V., 2021, "Analysing Web Archives of the Covid-19 Crisis through the IIPC collaborative collection: early findings and further research questions", IIPC netpreserve.org, <https://netpreserveblog.wordpress.com/2021/11/02/analysing-web-archives-of-the-covid-19-crisis-through-the-iipc-collaborative-collection-early-findings-and-further-research-questions/>
- Beaudouin Valérie, Chevallier Philippe, Maurel Lionel, 2018, *Le web français de la Grande Guerre : réseaux amateurs et institutionnels*, Presses Universitaires de Paris Nanterre.
- Bloch March, 1952, *Apologie pour l'histoire ou le métier d'historien*, Paris, Armand Colin,
- Brügger Niels, Milligan Ian (dir.), 2019, *The SAGE handbook of web history*, London, Thousand Oaks, SAGE Publications.
- Brügger Niels, Milligan Ian, Ben-David Anat, Gebeil Sophie, Nanni Federico, Rogers Richard, Turkel William J., Weber Matthew S., Webster Peter, 2019, « Internet histories and computational methods: a "round-doc" discussion », *Internet Histories*, vol. 3, n° 3-4, p. 202-222.
- Brügger Niels, Schroeder Ralph, 2017, *The Web as History: Using Web Archives to Understand the Past and the Present*, UCL Press.
- Clavert Frédéric, « Une histoire par les données ? Le futur très proche de l'histoire des relations internationales », *Bulletin de l'Institut Pierre Renouvin*, 44, 2016, p. 119-130.
- Clavert Frédéric, Noiret Serge, 2013, *L'histoire contemporaine à l'ère numérique*, Bruxelles, Peter Lang.
- Dosse François, « Le moment Ricœur », *Vingtième Siècle. Revue d'histoire*, 69, janvier-mars 2001.
- Clavert Frédéric, Mahroug Sophia, Schafer Valérie, 2022, « Préservation et distorsion : l'espace-temps des réseaux socio-numériques et du web archivé », *Revue d'histoire culturelle*, <http://revues.mshparisnord.fr/rhc/index.php?id=2791>
- Gebeil Sophie, 2021, *Website Story. Histoire, mémoires et archives du web*, INA.
- Gebeil Sophie, *La fabrique numérique des mémoires de l'immigration maghrébine sur le web français (1999-2014)*, Aix-Marseille Université (à paraître aux PUP)
- Merzeau Louise, 2006, « Mémoire », *Médium*, vol. 9, n° 4, p.152-163.
- Milligan Ian, 2019, *History in the age of abundance : how the Web is transforming historical research*, Montréal, McGill-Queen's University Press.
- Milligan Ian, 2015, *Finding Community in the Ruins of GeoCities: Distantly Reading a Web Archive*. UWSpace. <http://hdl.handle.net/10012/11650>
- Milligan, Ian, 2022, *The transformation of historical research in the digital age*. Cambridge University Press.
- Musiani Francesca, Paloque-Bergès Camille, Schafer Valérie, Thierry Benjamin G., 2019, *Qu'est-ce qu'une archive du Web ?*, Marseille, OpenEdition Press.
- Paillet Fred, Schafer Valérie, 2022, « Never gonna give you up », *Revue d'histoire culturelle*, <http://revues.mshparisnord.fr/rhc/index.php?id=3314>
- Paloque-Bergès Camille, « Les sources nativement numériques pour les sciences humaines et sociales », *Histoire@Politique*, 30, 2016, p. 221-244, <<https://www.cairn.info/revuehistoire-politique-2016-3-page-221.htm>>.
- Ricœur Paul, 2000, *La mémoire, l'histoire, l'oubli*, Paris, Seuil.
- Rogers Richard, 2015, *Digital methods*, Cambridge, First paperback edition, The MIT

Press.

Ruest, N., Fritz, S., Deschamps, R., Lin, J. & Milligan, I., "From archive to analysis: accessing web archives at scale through a cloud-based interface", *International Journal of Digital Humanities*, 2021.

Rygiel Philippe, 2017, *Historien à l'âge numérique*, Villeurbanne, Presses de l'ENSSIB.

Schafer Valérie, Gebeil Sophie, 2023, « Des archives du Web aux données, nouveaux services et collaborations », *Balisages*.

Soulet J.-F., *L'histoire immédiate : historiographie, sources et méthodes*, Paris, Armand Colin, 2009, p. 220.