Les Cahiers de Framespa

e-STORIA

42 | 2023
Historien.nes et numérique : pratiques et expériences vécues
Dossier : Historien.nes et numérique : pratiques et expériences vécues

Arpenter et sillonner les archives du Web

Surfing and browsing the web archives

Explorar y navegar por los archivos de la web

Valérie Schafer

Résumés

Français English Español
Cet article analyse en trois temps, correspondant à trois approches méthodologiques des archives du Web, la manière dont la recherche a évolué sous le double effet de nouveaux modes d'archivage et d'accès, développés par les institutions, et des questionnements scientifiques. Il revient sur des projets, de Web90 entre 2014 et 2018, dédié aux cultures numériques des années 1990, à l'actuel projet Hivi (Une histoire de la viralité en ligne), en passant par des recherches consacrées à des évènements spécifiques (attentats de 2015-2016 ; crise de la COVID-19). Ce parcours est mis en relation avec les évolutions de l'état de l'art et des pratiques archivistiques et académiques, face aux masses de données générées, notamment celles des réseaux socio-numériques, et à des thématiques transnationales ou transversales en développement.

This article analyzes in three steps, each one corresponding to a specific methodological approach to Web archives, the way research on this born digital heritage has evolved under the double influence of new modes of archiving and access, developed by institutions, and of research questions. It looks back at research projects, from the Web90 project running between 2014 and 2018, dedicated to the digital cultures of the 1990s, to the current Hivi project (A History of Online Virality), through research about events (terrorist attacks of 2015-2016; COVID-19 crisis). This path is related to the evolution of the state of the art and of archival and academic practices, in the face of the large amounts of data

generated, particularly those of the social networks, and of transnational or transversal research in progress.

Este artículo analiza en tres etapas, correspondientes a tres enfoques metodológicos de los archivos web, el modo en que la investigación ha evolucionado bajo el doble efecto de los nuevos modos de archivo y acceso, desarrollados por las instituciones, y de la evolución de las preguntas de investigación. Repasa proyectos, desde Web90 entre 2014 y 2018, dedicado a las culturas digitales de los años 90, hasta el actual proyecto Hivi (Una historia de la viralidad online), pasando por investigaciones dedicadas a acontecimientos concretos (atentados de 2015-2016 ; crisis COVID-19). Esta trayectoria está vinculada a la evolución del estado de la cuestión y de las prácticas archivísticas y académicas, ante las masas de datos generados, en particular los de las redes socio-numéricas, y a temas transnacionales o transversales en desarrollo.

Entrées d'index

Mots-clés : archives du web, histoire numérique, méthodologies, épistémologie, retour d'expérience
Keywords: web archives, digital history, methodologies, epistemology, reflexive feedback
Palabras claves: archivos web, historia digital, metodologías, epistemología, retroalimentación

Texte intégral

1 Constituées à partir de la seconde moitié des années 1990, notamment au sein de la fondation Internet Archive, née en 1996 à l'initiative de Brewster Kahle[1], ou encore d'initiatives relevant par exemple de la bibliothèque nationale d'Australie[2], les archives du Web mettent plusieurs années avant de sortir de l'ombre. La Wayback Machine[3] permet à partir de 2001 aux internautes d'accéder en ligne aux collections d'Internet Archive, tandis qu'en 2003 l'Unesco définit le patrimoine numérique[4], et notamment né (ou nativement) numérique. Les chercheurs, quant à eux, vont encore mettre presque une décennie avant de réellement commencer à tirer parti de ces nouvelles sources, malgré quelques initiatives pionnières dont celle de Niels Brügger au Danemark, qu'il retrace dans un article du *Temps des médias* en 2012[5].

2 C'est à la faveur de la coordination de ce numéro avec Jérôme Bourdon que les archives du Web attirent notre attention. La fascination qu'elles exercent immédiatement n'a pas cessé, se transformant en parallèle en une expertise technique et scientifique en constante mutation, au fil des changements réguliers qu'ont connu l'archivage du Web, mais aussi les cultures numériques (avec le développement important des réseaux socio-numériques, que ce soit Facebook, Twitter, Instagram, Périscope, TikTok et d'autres) et les questionnements de la recherche.

3 Après une exploration d'abord diachronique et qualitative, les archives du Web sont devenues, face à la masse de données générées, et notamment à la faveur de l'archivage des réseaux socio-numériques, une source également de lecture distante[6], permise par la mise à disposition de données, de métadonnées et d'outils par les institutions. De plus en plus intégrées dans les humanités numériques, les investigations se tournent aujourd'hui vers des méthodes hybrides et interdisciplinaires, entre lecture multi-scalaire[7] (combinant lecture proche et distante) et croisement des données du Web archivé et du Web vivant, comme nous le faisons dans le projet Hivi[8], dédié à l'histoire de la viralité en ligne.

4 C'est cet itinéraire dans les archives du Web, croisant évolutions de l'archivage et de la recherche, que nous souhaitons mettre en perspective, des premiers tâtonnements méthodologiques et épistémologiques et de la lecture proche à la lecture scalable, en passant par des expériences de lecture distante, qui élargissent le macroscope historien, pour reprendre l'expression de Shawn Graham, Ian Milligan et Scott Weingart[9].

1. Flâner dans le Web du passé

5 Lorsque Niels Brügger évoque son expérience avec les archives du Web, il se remémore une

tentative d'archivage par ses propres moyens. Cette pratique n'a pas disparu et l'on peut par exemple évoquer la collecte de données Twitter de Frédéric Clavert pour sa recherche sur les commémorations de la Première Guerre mondiale[10]. Elle est facilitée dans le cas de Twitter par une API[11] qui offre la possibilité de la collecte, ce qui n'est pas le cas de tous les réseaux socio-numériques (RSN). Toutefois, Niels Brügger évoque également les initiatives de la fondation Internet Archive, tandis que dans la décennie 2000 l'archivage institutionnel du Web se développe considérablement, à l'initiative des bibliothèques nationales. En France, en 2006, il entre ainsi dans les missions confiées à la Bibliothèque nationale de France (BnF) et à l'Institut national de l'audiovisuel (Ina), au titre du dépôt légal (et dans leur périmètre respectif - l'audiovisuel pour l'Ina et le reste de la webosphère française pour la BnF[12]). En parallèle, un même mouvement s'effectue au Danemark, en Grande-Bretagne et dans d'autres pays européens.

6 En 2012, quand nous commençons à nous intéresser aux archives du Web qui concernent la France, la consultation est donc possible via la Wayback Machine, mais aussi dans les fonds de la BnF et à l'Ina. Les années 1996 à 2006 (ou un peu antérieures, car des premières expériences sont menées avant, par exemple dès 2002 à la BnF pour des collectes électorales) sont à peu près concordantes dans les fonds d'Internet Archive et de la BnF : cette dernière récupère les données de la fondation états-unienne pour ce qui est qualifié en interne « d'incunables du Web ». Ensuite l'archivage diverge, chaque institution effectuant le sien propre[13] - et l'Ina avec des « outils maison », adaptés à ses missions liées à l'audiovisuel, qui impliquent de penser le flux, la vidéo, etc. Aussi, ses collectes ne reposent pas comme dans la plupart des bibliothèques européennes sur le robot Heritrix qu'utilise aussi Internet Archive.

1.a. Une approche méthodologique à construire

7 À la faveur du projet ANR Web90 (2014-2018)[14], nous commençons une exploration collective des archives du Web des années 1990. Elle prend deux directions : un aspect méthodologique et épistémologique d'abord, et une approche davantage tournée vers les contenus ensuite. Le premier enjeu rejoint les réflexions de Niels Brügger, ou encore de Peter Webster et Jane Winters en Grande-Bretagne, pour la seconde dans le projet Buddah[15]. Alors que les archives du Web commencent en effet à faire l'objet de projets financés et à intéresser la communauté académique, beaucoup est à faire sur le plan méthodologique pour comprendre et expliquer ces nouvelles sources, qui mettent par exemple à mal la notion d'authenticité. Très rapidement, l'équipe Web90 introduit, et notamment grâce à l'expertise de Francesca Musiani dans le champ des *Science and Technology Studies*, une réflexion sur la gouvernance des archives du Web[16] - réflexion élargie au fil des années et réactualisée sous l'effet de nouveaux défis, que ce soient les collectes d'urgence, la multiplication des acteurs de l'archivage du Web, des enjeux d'inclusion, etc.[17]. Outre la nécessité de mieux comprendre la « fabrique » de l'archive du Web, il s'agit de disséminer les méthodes, ce que l'équipe cherche à faire dans *Qu'est-ce qu'une archive du Web ?*[18]. Bien des tâtonnements sont à l'œuvre et l'exploration est d'abord qualitative. C'est alors une des seules possible, sauf à avoir un accès privilégié aux données, comme c'est le cas de Ian Milligan pour Geocities[19], ce qui lui permet de développer des travaux tournés vers la lecture distante et l'analyse de réseaux.

1.b. Entre bricolage et butinage dans le Web des années 1990

8 De notre côté, c'est davantage le bricolage qui caractérise l'exploration des archives et le croisement des sources permet de se faire une idée de la webosphère française, sans possibilité toutefois d'une vision vraiment globale, à la fois par manque de données sur le Web archivé mais aussi sur la diversité des sites existants dans la seconde moitié des années 1990 en France. De plus, Internet Archive ne permet alors une recherche que par URLs dans ses fonds et les archives du Web de la BnF ne sont pas indexées en plein texte (elles ne le sont toujours pas

intégralement). Elles ne permettent donc pas une recherche dans le texte intégral et en s'appuyant sur tous les mots des sites et pages web. Toutefois, en croisant par exemple le *Guide du Routard de l'Internet* de 1998, qui permet aux néophytes de se repérer dans la Toile française, les données de l'Afnic (Association française pour le nommage internet en coopération) qui enregistre la création de nouveaux noms de domaines (que l'on retrouve grâce aux pages web archivées de l'association), et d'autres méthodes de triangulation des sources, se dessine peu à peu un paysage numérique, retracé dans *En construction*[20]. Les archives du Web y servent à l'appui de la réflexion, mais jouent encore un rôle parfois discret au regard d'autres sources. C'est d'abord par la presse spécialisée, les entretiens oraux, etc., que le paysage du Web des années 1990 est reconstruit. Quelques études de cas sont développées, par exemple sur le « Web des pros[21] » ou des sites web éducatifs ou gouvernementaux. Mais la démarche est celle du butinage au sein d'une typologie des sites qui s'affine à la faveur du parcours Web90, créé en partenariat avec la BnF[22], et qui permet la découverte d'une centaine de sites Web des années 1990. Prédominent alors le choix de contenus qui reflètent les communautés et efforts pionniers, les premiers usages marchands ou commerciaux, des sites remarquables par leur design, ou encore ceux qui proposent un contenu créatif ou/et inattendu. Une forme d'exotisme est à l'œuvre dans cette redécouverte du Web des débuts. Il y a dans cette démarche des lacunes certaines pour reconstruire une vision totale, et le chercheur est davantage *lurker*[23], butineur, profitant de la sérendipité du Web et des liens hypertextes qui pointent vers des contenus associés. *Small is beautiful*, telle est notre approche : tout est à découvrir et à bâtir, dans un espace scientifique qui offre encore peu de modèles. Il faut également noter que les archives du Web des années 1990 sont loin d'être préservées avec la qualité actuelle et que les sauts temporels d'une page à l'autre, les éléments manquants, les images disparues illustrent parfaitement ce que Niels Brügger a qualifié de *reborn-digital heritage*[24], une reconstitution imparfaite de sources nativement numériques, parfois récalcitrantes à la collecte. C'est un parcours assez proche que mène Sophie Gebeil en parallèle, pour sa thèse sur les mémoires maghrébines de l'immigration[25], tout en expérimentant des premières approches de lecture distante, notamment via les hyperliens.

9 Rapidement, toutefois, et au cours même du projet ANR, le souhait des institutions d'archivage de fournir des services supplémentaires et de rendre les archives du Web plus praticables (vers le milieu du projet, Internet Archive introduit par exemple une fonction de recherche par mot-clé dans les pages d'accueil des sites archivés) se matérialise. Web90 et une autre recherche liée aux archives des attentats de 2015 (ASAP[26]) bénéficient du projet d'indexation plein texte d'une partie des fonds (incunables du Web et attentats) par l'équipe du dépôt légal (DL) Web de la BnF et du travail en lien mené sur les interfaces de recherche, dans le cadre d'un projet interne, CORPUS[27]. On est encore assez loin d'un « Web archivé de données », mais de nouvelles approches sont à l'œuvre (c'est le cas également côté Ina, nous y reviendrons, sur les données des RSN).

2. Plonger dans les données

10 Alors que les attentats de 2015 frappent la France, il devient évident que face à l'horreur, l'exceptionnel, le disruptif, les réseaux socio-numériques sont devenus une source importante pour saisir ces évènements. Les millions de réactions presque instantanées au moment des attentats de *Charlie Hebdo* et du Bataclan en témoignent. La BnF et l'Ina lancent des collectes pour capturer les traces de ces évènements inédits, sur le principe d'une *living archive*[28] qui demande une adaptation en temps réel (par exemple pour identifier les mots-dièse pertinents à collecter sur Twitter). Si le principe des collectes spéciales n'est pas inconnu des institutions, par exemple pour des évènements prévus comme des élections, les attentats marquent un tournant, notamment quantitatif, dans la collecte d'évènements imprévus, que prolongent des collections ensuite dédiées par exemple aux mouvements sociaux ou encore à la crise de la COVID-19, qui prend en termes d'archivage une ampleur internationale, à l'instar de la crise elle-même. Les cas des attentats et de la crise de la COVID-19 sont aussi disruptifs pour les chercheurs, par les masses de données collectées ou encore, nous y reviendrons dans le cas de

la pandémie, par une approche par les métadonnées et données dérivées, avant de pouvoir accéder aux contenus même des archives.

2.a. Une approche quantitative des traces numériques des attentats

11 L'archivage des RSN a commencé à l'Ina en amont de 2015, par exemple avec les vidéos qui entrent pleinement dans le périmètre audiovisuel de l'institut (Dailymotion, Vimeo ou YouTube), suivi de Twitter notamment pour des comptes de journalistes ou de chaînes. Lors des attentats de 2015, c'est par l'archivage d'urgence de Twitter que l'Ina, par l'intermédiaire de l'API développeur de ce réseau, constitue une collection sans précédent. Or, l'institut va développer des outils et interfaces de consultation pour aborder cette masse de données, permettant de réaliser des nuages de mots, de trier les résultats par émojis (☐), mots-dièse (hastags, #) ou images, de suivre leur usage dans le temps, etc. Ces fonctionnalités permettent au projet ASAP une première approche de ce fonds, tandis que l'équipe s'attache aussi à documenter la démarche d'archivage d'urgence entreprise autant par l'Ina que par la BnF - avec parfois des limites et lacunes, quand l'API développeur de Twitter par exemple ne permet plus de collecter les tweets car leur nombre dépasse 1 % des tweets totaux émis et atteint un plafond de collecte, ou quand les retweets sont figés au temps t de leur archivage, ne permettant pas de déceler la postérité d'un contenu, et quand certains hashtags sont omis (#jenesuispascharlie), car il faut choisir dans l'urgence ceux à préserver[29]. Il n'en reste pas moins une source inédite et prolifique, suivie d'autres collectes exceptionnelles, par exemple lors des mouvements #metoo ou de la crise de la Covid-19, sur laquelle nous reviendrons.

12 La démarche quantitative est alors nouvelle pour nous, mais d'autres chercheurs l'ont déjà expérimentée, à l'instar de Frédéric Clavert déjà mentionné. C'est aussi dans le cadre d'une recherche sur la Première Guerre Mondiale que Valérie Beaudouin a engagé avec la BnF une analyse de réseaux[30]. Des recherches ont ouvert la voie à des lectures orientées humanités numériques, comme le projet e-diasporas développé par Dana Diminescu[31], qui utilise Navicrawler ou Hyphe[32]. D'autres approches sont à évoquer à l'étranger, notamment celles de Ian Milligan, déjà cité, ou encore d'Anat Ben-David qui s'investit au sein du projet e-diasporas, puis continue ensuite des recherches sur le .yu, disparu avec la Yougoslavie. Inspirée par les *cultural studies* et les études visuelles, elle utilise aussi le logiciel Anaconda pour analyser les couleurs du Web[33].

13 Le potentiel que représentent les données et métadonnées des archives du Web est perçu par les institutions d'archivage. Cet intérêt est partagé par les chercheurs, alors que les contenus ne sont pas exportables ou partageables et souvent seulement consultables dans les salles des bibliothèques. Déjà le projet RESAW[34] qui démarre en 2012 à l'initiative de Niels Brügger et qui peu à peu va constituer un réseau de chercheurs, qui n'obtiendra pas le financement européen pour lequel il s'est créé, mais poursuit des collaborations durables et notamment à la faveur des conférences biennales RESAW, réfléchit à dépasser l'accès fermé aux archives du Web, envisageant notamment de s'appuyer sur un partage de métadonnées.

2.b. Une approche par les métadonnées et données dérivées de la crise de la COVID-19

14 Ce qui est pensé comme un possible devient une réalité en 2020 à la faveur de l'archivage du Web lié à la crise de la COVID-19, notamment par les membres de l'IIPC (*International Internet Preservation Consortium*). En effet, dans le cadre du projet WARCnet, porté en partenariat avec Niels Brügger et Jane Winters[35], se développe un programme réunissant archivistes du Web et chercheurs, notamment autour de l'accès transnational aux fonds et du partage des méthodologies. Par exemple, le groupe de travail 1 va explorer des méthodes quantitatives, pour réaliser une cartographie d'une webosphère nationale. Il s'appuie sur les travaux initiés au Danemark[36], qui ont développé une vision de l'évolution des contenus Web

archivés à travers le temps (nombre d'URLs collectés, types de fichiers collectés, etc.), démarche en cours de reproduction au Luxembourg dans le cadre d'une thèse[37] ou encore dans l'essai de Niels Brügger d'étendre la méthodologie danoise à d'autres pays.

15 Or la crise sanitaire éclate au moment du lancement du projet. Le groupe de travail 2 que je coordonne sur l'approche par les évènements décide alors de se pencher sur l'archivage en urgence qui se met en place face à cette crise sanitaire inédite et au caractère global. Une première initiative est de lancer une vaste campagne d'entretiens oraux pour documenter les efforts qui se font jour à travers toute l'Europe pour préserver les traces numériques de la crise[38]. En parallèle de cette collecte orale, le groupe de travail entreprend des discussions, d'abord avec chaque institution d'archivage membre du projet (France, Danemark, Luxembourg, Grande-Bretagne, etc.), pour récupérer des données. Il n'est pas question de pouvoir exporter les contenus archivés, mais les institutions vont faire un pas important en nous confiant leurs données dérivées et métadonnées, qui permettent des comparaisons systématiques des échelles, du périmètre, des temporalités à l'œuvre dans ces collectes d'urgence. Une nouvelle étape est franchie lors de l'appel à projet lancé par l'équipe canadienne d'Archives Unleashed[39] en 2021. Celle-ci développe depuis plusieurs années des outils permettant d'accéder aux archives du Web et elle souhaite développer un système d'accompagnement, de mentorat sur projets, au sein de cohortes aidées sur une base bimensuelle dans leur utilisation de l'interface ARCH[40]. Face à cette opportunité d'aller plus loin dans l'apprentissage des méthodes quantitatives, quelques chercheurs du groupe de travail 2 bénéficient d'un accord avec l'IIPC pour prendre pour cas d'étude cette vaste collecte effectuée au niveau international. C'est alors une exploration collective portée par les outils d'Archives Unleashed qui commence en 2021.

16 Dans ce projet AWAC2, fondé sur les archives de l'IIPC de la crise de la COVID-19, se révèlent des défis importants. À l'enthousiasme d'accéder pour la première fois à une collection internationale aussi étendue, puisque l'IIPC regroupe une trentaine d'institutions d'archivage du Web, auxquelles il a été demandé une sélection de contenus, succède la prise de conscience de l'ampleur de la tâche, et ce dès les tentatives de téléchargement des données, qui échouent régulièrement sur nos ordinateurs aux capacités limitées par rapport à la masse de données à récupérer[41]. C'est un point certes déjà connu depuis plusieurs années d'autres chercheurs, qui analysent le Web vivant ou font leurs propres collectes de tweets par exemple, et ont besoin de capacité de stockage et de calcul importante, ce qui n'avait pas été encore le cas pour notre approche des données du Web archivé, conservées dans les enceintes des institutions ou non accessibles en masse via Internet Archive. Ce problème dépassé avec le soutien de l'équipe du *High Performance Computing* de l'université du Luxembourg, il faut composer avec un double biais des données : elles proviennent d'une première sélection par les institutions d'archivage, puis d'une seconde sélection effectuée pour l'IIPC, qui a notamment proposé d'exclure les RSN. Dans les faits, de nombreux contenus liés au RSN sont conservés, mais là encore avec des biais certains, puisque certains RSN focalisent l'attention, comme Twitter ou YouTube, tandis que d'autres comme Instagram sont très peu préservés à l'heure actuelle. Reste que la collection IIPC est remarquable par son caractère international. Toutefois, qui dit international dit multilingue et les défis à relever sont aussi linguistiques. Ils sont par ailleurs liés aux doublons et aux bruits dans les archives, que l'on trouve notamment dans les sites d'information, préservés régulièrement mais qui donnent des résultats parfois similaires ou/et pas toujours pertinents. Ainsi, la deuxième partie du travail du groupe de recherche au semestre 2022 s'oriente vers l'analyse de la place des femmes dans la crise de la COVID-19. Si un suivi de l'actualité et des tendances permet d'identifier des grandes thématiques comme le *care* et la charge mentale dont peuvent souffrir les femmes (entre travail, école à la maison, tâches domestiques, etc.), d'autres apparaissent à la faveur de la plongée dans les archives. Ainsi, le thème de la grossesse en temps de crise sanitaire, d'abord omis, ressort avec force d'une consultation sur une collection plus petite, celle de l'Ina. Mais si l'on fait par exemple une recherche sur le féminisme dans la collection de l'IIPC, les sites d'information vont remonter de nombreux résultats qui associent féminisme et COVID, alors que les informations les concernant sont en fait différentes et découplées, l'une portant sur la crise sanitaire et l'autre sur le mouvement #metoo. Des va-et-vient entre lecture distante et proche sont donc

indispensables, mais complexes quand cela concerne des millions de résultats. De plus, les limites techniques de l'approche sont vite perçues par l'équipe qui manque au départ de puissance de calcul, de stockage, mais aussi d'outils pour analyser ces résultats.

17 L'arrivée dans l'équipe d'un chercheur en informatique permet de franchir un cap et notamment de travailler sur des algorithmes de *topic modeling*, mais le travail est long, fastidieux, et se réoriente vers des problématiques davantage informatiques, comme la comparaison des résultats avec plusieurs algorithmes (Latent Dirichlet Allocation (LDA), Word2vec, and Doc2vec[42]). Aussi intéressante que soit cette question, elle ne permet pas au terme des six mois d'exploration prévus pour cette étude sur les femmes et la crise sanitaire d'avoir des résultats très élaborés du point de vue des sciences humaines. Le rapport à la recherche est aussi profondément modifié, et là où l'on gagne en questionnements méthodologiques et techniques, il y a une perte certaine de plaisir à découvrir les contenus. Les données fournies permettent un accès au texte, mais ces résultats essentiellement textuels ne reflètent pas la richesse des contenus Web multimédia, tandis que le contexte est difficile à retrouver (or il est important sur cette collection internationale dont les sites web sortent de notre champ familier, quand il s'agit de contenus par exemple d'Amérique du Sud). Ils impliquent de s'intéresser aux effets d'une recherche largement influencée par les données (*data driven*[43]). Il n'en reste pas moins qu'il s'agit là d'une expérience enrichissante, notamment avec l'équipe d'Archives Unleashed, qui permet de travailler en dialogue constant sur des outils développés par des chercheurs pour d'autres chercheurs, en une boucle itérative. Les outils utilisés dans mes recherches étaient en effet jusqu'à présent le plus souvent développés au sein des institutions d'archivage elles-mêmes, certes soucieuses des besoins des chercheurs, mais offerts dans le cadre d'accès au sein des institutions, ne permettant pas l'exportation aisée des corpus et données ou le choix des outils. Toutefois, là aussi, la tendance à la fourniture de réponses au plus proches des projets s'affirme, notamment par la création de *labs* destinés à accompagner les chercheurs et à répondre à des besoins particuliers d'extractions de données, de création de corpus, comme c'est le cas avec le lancement d'un BnF DataLab en 2021 et d'un Lab Ina fin 2022.

3. Naviguer dans le flux

18 C'est à l'occasion d'un appel à projets du BnF DataLab que s'engage une coopération d'un an sur un projet lié à la viralité en ligne, BUZZ-F. Mené en particulier avec Fred Pailler, il s'agit d'un sous-projet au sein d'un projet plus large mené au Luxembourg, Hivi (*A history of online virality*). Ce dernier, qui cherche à retracer depuis le milieu de la décennie 1990 les phénomènes de viralité en ligne et à les saisir sous l'angle notamment des circulations, a placé en son cœur les archives du Web de manière assez naturelle, avant de se rendre compte que les défis méthodologiques et pratiques gagnent en complexité au fil de l'exploration. Alors qu'il s'agit notamment d'avoir une approche de la viralité qui soit contextuelle, qui dépasse l'approche sémiotique, pour penser des modes de partage, de circulation et de réception, le Web archivé présente de sérieuses limites. On pourra notamment noter le fait que l'archivage du Web a évolué et permet un suivi diachronique inégal et incertain des phénomènes viraux, que ceux-ci sont rarement une priorité de l'archivage du Web, que les doublons (dont nous nous plaignions précédemment dans le champ de la crise de la COVID-19) deviennent ici essentiels mais rarement archivés volontairement, que bien des espaces de circulation de la viralité en ligne sont aujourd'hui les plateformes sociales et qu'une toute petite partie en est archivée, que les commentaires ne sont pas forcément préservés (par exemple pour l'archivage de YouTube à l'Ina), etc. Très rapidement, le Web vivant s'impose également comme une source importante, notamment au regard des efforts de patrimonialisation engagés par des plateformes comme Know Your Meme[44] qui, elles, prennent comme cœur de leur approche la viralité[45].

3.a. Retrouver la viralité dans le Web archivé

19 Le projet BUZZ-F lancé à l'automne 2021 va vite réaliser que ses objectifs sont démesurés dans le temps d'un an imparti pour le projet. Comme pour AWAC2, les défis méthodologiques sont tels qu'ils posent la question de la durée nécessaire à ce type de projets et d'accompagnement. Les explorations réalisées en partenariat avec l'équipe du DL Web de la BnF et un ingénieur d'Huma-Num se concentrent alors sur deux cas : le Lip dub et le Harlem Shake[46]. Ces deux phénomènes viraux sont intéressants car ils ne sont pas trop récents (première moitié de la décennie 2010) et permettent de prendre la mesure de la recherchabilité de ces phénomènes (dans des collectes pas toutes indexées en plein texte), de leur archivage à l'époque, et de leur apport au regard du Web vivant. Ainsi une expérience est-elle menée pour comparer les vidéos archivées et celles encore en ligne. Une autre approche méthodologique consiste à s'appuyer sur les noms de domaines, pour essayer de retrouver les contenus spécifiquement dédiés par exemple au Harlem Shake, même s'il faut être conscient de la limite de cette approche puisque de nombreux contenus peuvent se trouver disséminés dans des sites plus généralistes. La presse en ligne archivée (et en plein texte) est également utile. En parallèle de la réalisation d'un corpus Europresse, elle permet de prendre la mesure du rôle performatif de la presse dans la viralité, que confirment également des explorations dans les archives du Web de l'Ina, qui montrent aussi le rôle de l'audiovisuel dans la médiatisation du Harlem Shake. Le travail commun entre chercheurs et archivistes du Web permet de repousser certaines limites. Les équipes de la BnF cherchent des solutions techniques à nos questionnements, proposent des visualisations des contenus et surtout, permettent de créer un corpus que les chercheurs seuls n'auraient pas pu réaliser sans accès aux fichiers WARC. Les lacunes rapidement identifiées dans les archives du Web n'enlèvent rien aux apports de ces contenus, pour beaucoup désormais inaccessibles en ligne.

3.b. Une lecture scalable des phénomènes Internet

20 Les contenus de la presse en ligne explorés dans BUZZ-F et le corpus Europresse invitent à une lecture multi-scalaire, car le phénomène du Harlem Shake connaît des développements internationaux en 2013, mais avec des inscriptions et circulations spatiales précises et bien documentées par la presse régionale. En effet, le phénomène se déploie aussi dans l'espace physique et est associé à un établissement universitaire, une association, ou encore une ville. Déjà, Louise Merzeau dans ses premières explorations des archives du Web avait insisté sur la notion de flux[47], qui prend tout son sens quand on souhaite étudier des phénomènes viraux, marqués par une grande intensité de circulation. Saisir les espaces de déploiement d'une culture présentée comme mondialisée, mais qui connaît des déclinaisons nationales voire locales, ou encore communautaires, ainsi qu'une circulation au sein de plusieurs espaces numériques, est essentiel à la compréhension de la viralité. Cela implique un travail minutieux de suivi des phénomènes, souvent pensés en nombre de visites sur YouTube, de vidéos téléchargées, de tweets contenant un hashtag, mais rarement spatialisés. Aussi, dans le projet Hivi, est rapidement ressenti le besoin de saisir des déclinaisons plus locales, et le Harlem Shake est à l'évidence un exemple plus simple à appréhender que des phénomènes Internet plus durables, mais peu ancrés dans l'espace, comme le Rickroll.

21 C'est donc le cas d'étude choisi pour essayer de développer une vision scalable et, comme nous le défendons dans un chapitre dédié à ces approches, une *medium reading*, attentive non seulement au contenu mais au contenant et à la circulation transmédiatique[48]. Cette recherche croise analyse de corpus presse, corpus audiovisuels, archives du Web et enfin un vaste corpus de tweets, et permet de dégager certaines tendances, sur les échos médiatiques du phénomène et leur temporalité, l'ancrage spatial du Harlem Shake ou la participation des internautes. Dans la collecte Actualités des archives du Web de la BnF, une recherche de janvier à mai 2013 par fréquence d'apparition du terme Harlem Shake dans les noms de domaine fait clairement ressortir la presse régionale. Or, selon la base Europresse, certains titres régionaux ont publié sur le Harlem Shake au moins autant que la presse nationale, voire deux fois plus souvent, comme c'est le cas pour *La Montagne, Le Berry Républicain, Le Dauphiné Libéré,* etc.

22 Cet exemple permet de retrouver pleinement les approches historiennes et notamment le

rapport au croisement des sources. Déjà présent dans les approches de Web90 lors de la lecture qualitative, il devient aussi nécessaire pour une lecture distante, en un jeu d'échelles constant. Bien évidemment, la volonté de dépasser la césure entre approche qualitative et quantitative n'est pas nouvelle, l'approche transmédiatique non plus, et l'approche transnationale a fait l'objet de réflexions importantes. Mais la rencontre de ces enjeux au sein des analyses portant sur les archives du Web témoigne sans doute de la fin des tâtonnements méthodologiques originels, pour inviter à de nouveaux bricolages méthodologiques certes, mais qui de plus en plus cherchent à repousser les limites de la compréhension des cultures numériques.

23 **Conclusion**

24 Ce parcours à travers dix ans de recherche dans les archives du Web permet de prendre la mesure d'une évolution accélérée des pratiques, usages, méthodes, et celle-ci est loin d'être achevée. Les recherches sur les archives du Web ont largement traité des aspects méthodologiques. Beaucoup reste à faire, notamment dans le domaine visuel, multimédia, dans l'interopérabilité des données entre le Web archivé et d'autres archives (presse, audiovisuel, etc.) dont nous avons souligné l'importance, par exemple pour l'étude des phénomènes de viralité. De même, à l'heure du *FAIR data* et de démarches de plus en plus collectives d'investigation des données, se pose avec acuité la question du partage des données, des corpus et de leur ouverture. L'archivage du Web et les méthodes d'analyse de ces sources s'adaptent en permanence à des nouveaux défis, mais doivent aussi composer avec les évolutions très rapides des cutures numériques. En outre, comme l'a fait remarquer un des coordinateurs de ce dossier, Sébastien Poublanc, l'irruption d'évènements (attentats, crise sanitaire) entraîne des changements d'échelle et de pratiques, des recompositions dans les collectes, dans l'accès, dans les méthodes, de la part des acteurs, qu'ils soient liés aux institutions patrimoniales ou au monde académique. Les affres que connait Twitter depuis l'arrivée d'Elon Musk et le passage d'une partie des usagers sur Mastodon montrent bien les défis de l'archivage, de même que la guerre en Ukraine et l'investissement en urgence de bénévoles au sein du projet SUCHO[49] pour sauvegarder le patrimoine culturel ukrainien en ligne. L'entrée de plus en plus transversale et transnationale dans les archives du Web pose aussi de nouveaux défis méthodologiques, à résoudre en outre de manière interdisciplinaire. La lecture scalable invite quant à elle à réexplorer des questions déjà rapidement abordées, mais seulement à l'échelle nationale ou encore par une approche très qualitative, par exemple pour les mouvements sociaux en ligne[50]. Plus la recherche sur les archives du Web avance, plus elle découvre de nouvelles frontières à dépasser, de nouveaux terrains d'investigation, sans oublier la question essentielle de la participation des publics, des associations, et son potentiel de contribution à l'histoire publique. L'archive du Web est ainsi un terrain permanent de renouvellement qui met au défi le chercheur, en apprentissage constant, afin de saisir des sources qui sont loin d'être figées et qu'il voit se constituer et évoluer au fil de ses recherches, le faisant à la fois progresser et redébuter en permanence. Toutefois, les tâtonnements méthodologiques doivent aussi être dépassés pour révéler plus pleinement les apports des archives du Web dans les sujets de recherche. Il ne s'agit plus comme il y a dix ans d'ouvrir des pistes d'usage, de s'interroger sur le statut de l'archive du Web, de réévaluer les spécificités de ces archives (en termes d'authenticité, d'auctorialité, etc.). Une littérature maintenant conséquente l'a fait. Il convient de démontrer leurs apports certains pour l'écriture de l'histoire, certes sans taire les difficultés à concilier les compétences en sciences humaines et celles que requièrent les archives du Web, notamment en termes techniques, sans cacher non plus les bricolages qu'imposent des démarches souvent empiriques, mais en allant dans le sens d'une pleine intégration des archives du Web à des études en histoire du temps présent, où elles joueraient de manière fluide aux côtés d'autres archives leur rôle de sources historiques.

Notes

1 Organisme à but non lucratif, Internet Archive est une fondation et bibliothèque numérique états-unienne. https://archive.org

2 Kieran Hegarty, « The invention of the archived web: tracing the influence of library frameworks on web archiving infrastructure», *Internet Histories*, 2022, vol. 6, p. 432-451.

3 https://web.archive.org.

4 Unesco, *Charte sur la conservation du patrimoine numérique,* 2003. http://portal.unesco.org/fr/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html.

5 Niels Brügger, « L'historiographie de sites Web : quelques enjeux fondamentaux », *Le Temps des médias*, vol. 18, no. 1, 2012, p. 159-169.

6 La lecture distante (*distant reading*) est une notion créée par Franco Moretti, spécialiste d'histoire littéraire, pour désigner une approche qui aborde les textes non par une lecture détaillée (lecture proche, *close reading*) mais via des corpus massifs à l'aide d'outils computationnels. Franco Moretti, *Distant Reading*, Londres et Brooklin, Verso, 2013.

7 Julia Flanders, Matthew L. Jockers, « A Matter of Scale », Keynote lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA, 2013.

http://digitalcommons.unl.edu/englishfacpubs/106.

8 Hivi (A history of online virality) est soutenu par le Fonds national de la recherche luxembourgeois de 2021 à 2024 (C20/SC/14758148), https://hivi.uni.lu.

9 Shawn Graham, Ian Milligan, Scott Weingart, « Exploring Big Historical Data. The Historian's Macroscope », London, Imperial College Press, 2015, https://themacroscope.org.

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

11 *Application Programming Interface*. Il est possible via cette interface de programmation d'obtenir des données de la part du logiciel.

12 Dans le cadre de la loi DADVSI sur les Droits d'Auteur et Droits Voisins dans la Société de l'Information. Voir notamment Emmanuelle Bermès, « Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial », *Quaderni : communication, technologies, pouvoir*, n°98, 2018-2019, p. 73 et Claude Mussou, « Et le Web devint archive : enjeux et défis », *Le Temps des médias*, 2012, n° 19, p. 259-266.

13 Internet Archive et la BnF ou l'Ina peuvent archiver le même site, mais ce ne sera pas forcément à la même fréquence, ni avec la même profondeur, au même moment, etc. Le périmètre plus restreint de l'Ina lui permet par exemple des captures très fréquentes des sites liés à l'audiovisuel, sans comparaison avec ce que peut faire Internet Archive sur des sites comme tf1.fr.

14 https://web90.hypotheses.org. Membres de l'équipe : https://web90.hypotheses.org/equipe.

15 Sur le projet *Big UK Domain Data for the Arts and Humanities*, voir https://buddah.projects.history.ac.uk/.

16 Valérie Schafer, Francesca Musiani, Marguerite Borelli, « Negotiating the Web of the Past », *French Journal for Media Research*, nº 6, 2016, http://frenchjournalformediaresearch.com/lodel/index.php?id=952.

17 Valérie Schafer, Jane Winters, « The values of web archives », *International Journal of Digital Humanities*, 2021, p. 129-144.

18 Francesca Musiani, Camille Paloque-Bergès, Valérie Schafer, Benjamin Thierry, *Qu'est-ce qu'une archive du Web?*, Marseille, OpenEdition Press, 2019.

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

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

21 Valérie Schafer, Benjamin Thierry, « The 'Web of pros' in the 1990s: The professional acclimation of the World Wide Web in France », *New Media & Society*, vol. 18 (7), p. 1143-1158.

22 https://www.bnf.fr/sites/default/files/2018-11/parcours_web90%20final.pdf.

23 Valérie Schafer, Francesca Musiani, « The Historian of the Web: Crawler, Browser or Lurker ? », webarchivehistorians.org, 2015, https://webarchivehistorians.org/?s=schafer.

24 Niels Brügger, « Digital Humanities in the 21st Century : Digital Material as a Driving Force », *Digital humanities quaterly*, vol. 10, n° 2, 2016, http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html.

25 Sophie Gebeil, La fabrique numérique des mémoires de l'immigration maghrébine sur le web français (1999-2014), Thèse, Université Aix-Marseille, 2015.

26 Archives Sauvegarde Attentats Paris, https://asap.hypotheses.org, est un projet soutenu par le CNRS en 2016.

27 Programme de recherche de 4 ans, de 2016 à 2019, inscrit dans le plan quadriennal de la recherche

de la BnF et visant notamment à « créer un nouveau service de fourniture de données à destination de la recherche ». Eleonora Moiraghi, *Le projet CORPUS et ses publics potentiels*, Paris, BnF, 2018. https://hal-bnf.archives-ouvertes.fr/hal-01739730

28 Tamara Rhodes, « A Living, Breathing Revolution: How Libraries Can Use “Living Archives” to Support, Engage, and Document Social Movements », Singapour, IFLA WLIC, 2013.

29 Entretien avec Thomas Drugeon, responsable du DL Web à l’Ina (21 mars 2016), https://asap.hypotheses.org/173#more-173. Voir aussi Valérie Schafer, Gérome Truc, Romain Badouard, Lucien Castex, Francesca Musiani, « Paris and Nice terrorist attacks: Exploring Twitter and web archives », *Media, War & Conflict*, 2019, vol.12, no. 2, https://journals.sagepub.com/doi/full/10.1177/1750635219839382.

30 Valérie Beaudouin, Zeynep Pehlivan, « Cartographie de la Grande Guerre sur le Web : Rapport final de la phase 2 du projet ‘Le devenir en ligne du patrimoine numérisé : l’exemple de la Grande Guerre’ », 2016 https://hal.archives-ouvertes.fr/hal-01425600.

31 Sur cet ambitieux projet, ses méthodes et visualisations pionnières, voir http://www.e-diasporas.fr.

32 https://medialab.sciencespo.fr/outils/navicrawler/.

33 Anat Ben-David, Adam Amram, Ron Bekkerman, « The colors of the national Web: visual data analysis of the historical Yugoslav Web domain », *International Journal on Digital Libraries,* 2018, p. 95-106.

34 A research infrastructure for the study of archived web materials, http://resaw.eu.

35 Web Archive studies network researching web domains and events,

https://cc.au.dk/en/warcnet.

36 Niels Brügger, Ditte Laursen, Jane Nielsen, « Exploring the domain names of the Danish web » dans *The Web as History. Using Web Archives to Understand the Past and the Present,* Niels Brügger, Ralph Schroeder (dir.), Londres, UCL Presss, 2017, p. 238-248.

37 Thèse en cours de Carmen Noguera au C²DH sur l’informatisation et la numérisation du Luxembourg.

38 Les entretiens oraux réalisés sont tous disponibles en accès ouvert à https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports.

39 Sur ce projet canadien et son équipe, voir : https://archivesunleashed.org. 3. Grâce à une bourse de la *Andrew Mellow Foundation* le projet a développé de 2017 à 2020 des outils de recherche et d’analyse de données pour le Web archivé avant de rentrer dans sa seconde phase.

40 ARCH, pour *Archives Research Compute Hub* est une interface développée par l’équipe d’Archives Unleashed avec Internet Archive pour permettre l’analyse du Web archivé. https://archivesunleashed.org/arch/

41 Susan Aasman, Niels Brügger, Frédéric Clavert, Karin de Wild, Sophie Gebeil, Valérie Schafer, « Analysing Web Archives of the COVID-19 crisis through the IIPC collaborative collection: early findings and further research questions », netpreserveblog, 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/.

42 Susan Aasman, Niels Brügger, Frédéric Clavert, Karin de Wild, Sophie Gebeil, Valérie Schafer, Joshgun Sirajzade, « Studying Women and the COVID-19 Crisis through the IIPC Coronavirus Collection », netpreserveblog, 2022, https://netpreserveblog.wordpress.com/2022/12/20/studying-women-and-the-covid-19-crisis-through-the-iipc-coronavirus-collection/.

43 La *data-driven research* est souvent opposée à la *theory-driven research*. Pour schématiser grossièrement, dans le premier cas les données guident/définissent la question de recherche, dans le second cas les données doivent répondre à une question de recherche préalablement définie. Wolfgang Maas et al., « Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research », *Journal of the Association for Information Systems*, 2018, 19 (12), p. 1253-1273.

https://core.ac.uk/download/pdf/301378837.pdf

44 https://knowyourmeme.com

45 Mais avec d’autres limites qu’il ne convient pas ici de développer, voir Fred Pailler, Valérie Schafer, « ‘Never gonna give you up’. Historiciser la viralité numérique », *Revue d’histoire culturelle*, 2022, http://revues.mshparisnord.fr/rhc/index.php?id=3314.

46 Le lip dub consiste à faire du playback sur une bande sonore préexistante. Un exemple en est en France le lip dub des jeunes de l’UMP en 2010 : https://www.youtube.com/watch?v=VyLOy7l_jP4&t=196s. Autre phénomène viral, le Harlem Shake est une dance sur la musique de DJ Bauuer qui se déploie mondialement en 2013 et notamment via YouTube. https://knowyourmeme.com/memes/harlem-shake

47 Voir notamment la séance de 2015 des ateliers du dépôt légal du Web à l’Ina, co-organisée avec Claude Mussou, sur le thème « Du site aux applications, des applis au streaming… »,

https://merzeau.net/des-sites-aux-applications/.

48 Fred Pailler, Valérie Schafer, « Keep calm and stay focused. Historicising and intertwining scales and temporalities of online virality », dans *Zoomland. Exploring Scale in Digital History and Humanities,* Florentina Armaselu, Andreas Fickers (dir.), Berlin, De Gruyter, 2022, en cours de publication.

49 *Saving Cultural Ukrainian Heritage Online*. Sur cette initiative portée par de nombreux bénévoles, voir https://www.sucho.org.

50 Valérie Schafer, « Sous les pavés l'archive ! Luttes sociales et archives du Web », *Le Temps des Médias*, 2020, vol. 35, p. 121-138.

Pour citer cet article

Référence électronique
Valérie Schafer, « Arpenter et sillonner les archives du Web », *Les Cahiers de Framespa* [En ligne], 42 | 2023, mis en ligne le 04 juillet 2023, consulté le 05 juillet 2023. URL : http://journals.openedition.org/framespa/13971

toire européenne contemporaine à l'université du Luxembourg, au *digital history*). Elle est également chercheuse associée au centre pécialisée dans l'histoire du Web, d'Internet et des cultures