Les sources nativement numériques
Enjeux de documentation et de redocumentarisation

Valérie Schafer, Frédéric Clavert et Caroline Muller

En janvier 2026, comme les années précédentes, se tenait, à l'Unesco à Paris, le symposium annuel de la fondation Software Heritage[1]. Celui-ci accueillait des acteurs et des actrices du patrimoine numérique, des entreprises investies dans le secteur du logiciel et plus largement du numérique, des représentantes et des représentants du monde politique, de la recherche et de l'enseignement intéressés par les données et la science ouvertes, un public universitaire et bien d'autres parties prenantes encore. Toutes et tous étaient sensibles à ce projet, né en 2016 et parmi les derniers venus dans le champ du patrimoine nativement numérique, qui a fait du code source[2], de sa préservation et de son accessibilité sa mission. Défini en 2003 par l'Unesco dans une charte relative à sa conservation[3], le patrimoine dit « nativement numérique » – pour le distinguer du patrimoine

1. Voir, par exemple, la page de l'événement de 2026 : https://www.softwareheritage.org/2025/11/25/software_heritage_2026_symposium_summit/. Pour plus d'informations sur la fondation Software Heritage, voir https://www.softwareheritage.org/?lang=fr.
2. Ensemble d'instructions écrites dans un langage de programmation (comme C, Java, Python, etc.) qui définit les actions que doit effectuer un logiciel.
3. Unesco, Charte sur la conservation du patrimoine numérique, 2003, https://www.unesco.org/fr/legal-affairs/charter-preservation-digital-heritage. Voir également Marta Severo et Séverine Cachat (dir.), *Patrimoine culturel immatériel et numérique. Transmission, participation, enjeux*, Paris, L'Harmattan, 2017 ; Matteo Treleani, *Qu'est-ce que le patrimoine numérique ? Une sémiologie de la circulation des archives*, Lormont, Le Bord de l'eau, 2017.

Annales HSS, 80-4, 2025, p. 791-814, 10.1017/ahss.2025.10139
© Les autrices et l'auteur, 2026. Publié par Cambridge University Press pour le compte des Éditions de l'EHESS. Cet article est en libre accès, distribué sous les termes de la licence Creative Commons Attribution (http://creativecommons.org/licenses/by/4.0), qui autorise la réutilisation, la distribution et la reproduction sans restriction, à condition que l'article original soit correctement cité.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

numérisé – connaît une grande diversité d'approches et d'acteurs. Son histoire commence en amont des années 2000, avec la création, notamment, de la fondation états-unienne Internet Archive en 1996[4]. Ce type de patrimoine englobe des objets aussi variés que des bases de données, des archives du Web, des courriels, des jeux vidéo, des podcasts, des logiciels ou encore leur code source. En outre, les formats, les espaces, les interfaces, les plateformes de collecte comme d'accès et les usages se renouvellent à un rythme soutenu. Les institutions patrimoniales ont peu à peu intégré la préservation des contenus des réseaux sociaux numériques (ou en tout cas de certains d'entre eux), en déployant de nouvelles procédures de collecte, différentes de celles employées pour les sites web. Le développement des systèmes d'intelligence artificielle générative, dont ChatGPT n'est qu'une application parmi bien d'autres, pose aussi aujourd'hui la question de leurs traces : que préserver des *prompts* (les requêtes des utilisateurs), des textes et des images générés par ces systèmes ?

Au sein du patrimoine nativement numérique, ce sont les archives du Web qui ont été les plus explorées par le monde de la recherche[5]. Elles ont bénéficié des efforts des institutions patrimoniales pour les préserver, pour les faire connaître et les rendre accessibles. Elles s'inscrivent en France dans le cadre du dépôt légal du Web, défini par la loi DAVSDI de 2006[6]. Les courriels, objets de collecte dans le champ politique, institutionnel et administratif, sont quant à eux soumis à de forts enjeux légaux et éthiques, renforcés par le Règlement général sur la protection des données (RGPD) et ses différentes implémentations dans les droits nationaux des États membres de l'Union européenne depuis 2016[7].

Les politiques de collecte, de conservation mais aussi de documentation et (re)documentarisation ne sont ainsi pas figées et évoluent au sein d'un écosystème en constante négociation. Il faut ajouter à cela une forte demande sociale sur des contenus et des données qui présentent des enjeux politiques certains, comme ceux, par exemple, qui concernent le mouvement Black Lives Matter[8] ou les archives en ligne des institutions culturelles de l'Ukraine[9], mises en péril par

4. Voir le site d'Internet Archive, https://archive.org.

5. Si elles font davantage l'objet de développements que d'autres sources nativement numériques dans cet article, il n'en reste pas moins nécessaire de ne pas limiter la réflexion à ces sources spécifiques ; le code source, les vidéos en ligne, les réseaux socio-numériques, leurs traces et leurs plateformes, les CD-ROM, les jeux vidéo ou encore les courriels sont également évoqués.

6. Sur cette histoire, voir Évelyne Cohen et Julie Verlaine, « Le dépôt légal de l'internet français à la Bibliothèque nationale de France », nº spécial « Archives et patrimoines visuels et sonores », *Sociétés & Représentations*, 35, 2013, p. 209-218. Les vidéogrammes et documents composites sont soumis au dépôt légal depuis 1975, les objets multimédias, logiciels et bases de données depuis 1992.

7. Outre le volume des données se pose également la question de leur exploitation hors de leur outil de production.

8. Le site web de l'université d'Arkansas fournit des ressources sur l'archivage du mouvement : https://uark.libguides.com/c.php?g=1050881.

9. Saving Ukrainian Cultural Heritage On-Line, SUCHO, http://www.sucho.org/.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

la guerre déclenchée par la Russie[10]. Si ces attentes de communautés, de mouvements activistes, de militants et militantes, ou plus largement de citoyens et citoyennes ne sont pas nouvelles, elles trouvent ici une dimension, une échelle et un écho internationaux inédits.

Il existe à cet égard des asymétries entre pays dotés ou non de politiques de conservation, de moyens techniques et humains plus ou moins conséquents, de périmètres de collecte plus ou moins larges. Les défis et les paradoxes de la conservation et de la documentation sont constants. D'un côté, les limites techniques sont sans cesse repoussées et de nouveaux modes de traitement et moyens sont expérimentés, tandis que des outils d'intelligence artificielle[11], pouvant aider à la collecte, à l'indexation ou à l'analyse, sont introduits. De l'autre, ce champ repose aussi sur des héritages importants, celui des initiatives des grandes bibliothèques nationales, qui relèvent d'une tradition issue du document papier, du dépôt légal et du droit d'auteur. Le mouvement d'incitation à l'ouverture des données dans le monde culturel comme dans celui de la recherche rend cette tension particulièrement visible, car les données ne sont pas toujours aisément accessibles, exportables, exploitables ou partageables, pour des raisons juridiques, mais aussi parfois techniques ou organisationnelles.

Alors que les sciences politiques et celles de l'information et de la communication s'emparent de ces corpus, dans le champ historique, le recours à ces sources concerne aujourd'hui principalement les recherches menées sur la période contemporaine récente et sur le temps présent (après les années 1990), mais il va sans doute se systématiser. Les archives du Web sont, aujourd'hui, encore peu utilisées dans les thèses françaises d'histoire contemporaine[12], mais on imagine difficilement que les historiens et historiennes du XXIe siècle ne mobilisent pas les traces

10. Si notre propos n'est pas ici de démontrer l'enjeu politique du patrimoine nativement numérique, il ne peut être ignoré, à l'heure du sauvetage en urgence de documents nativement numériques liés à l'administration Trump et ses actions antérieures, de la censure qu'Internet Archive a pu connaître dans certains pays au cours de la dernière décennie ou des cyberattaques qui ont affecté aussi bien la fondation en 2024 (voir Mathilde Saliou, « Internet Archive piratée, les données de 31 millions d'utilisateurs sont touchées », *Next*, 10 oct. 2024, next.ink/brief_article/internet-archive-piratee-les-donnees-de-31-millions-dutilisateurs-sont-touchees/), que des bibliothèques nationales (voir le cas de la cyberattaque contre la British Library : Roly Keating, « Learning Lessons from the Cyber-Attack », 3 août 2024, https://blogs.bl.uk/living-knowledge/2024/03/learning-lessons-from-the-cyber-attack.html). Les archives du Web ont parfois contrecarré les ambitions politiques d'effacement des sites web, comme cela a été le cas lorsque le parti conservateur britannique a décidé, en 2013, de « nettoyer » son site web : voir Sebastian Payne, « Why Have the Tories Purged their Website? », *The Spectator*, 13 nov. 2013, https://www.spectator.co.uk/article/why-have-the-tories-purged-their-website/.

11. Voir par exemple Emmanuelle Bermès et Eleonora Moiraghi, « Le patrimoine numérique national à l'heure de l'intelligence artificielle. Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques », *Revue ouverte d'intelligence artificielle*, 1-1, 2020, p. 89-109.

12. Parmi les thèses soutenues en histoire contemporaine en France en 2021 et 2022, 62 auraient pu avoir recours aux archives du Web, 21 mobilisent des contenus en ligne, mais aucune ne mentionne directement un recours aux archives du Web *constituées comme archives*

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

et les archives numériques que les sociétés ont fait naître, et ce sans se limiter aux archives du Web, mais en incluant également les bases de données, les courriels, etc. Or, le développement de telles recherches implique de prêter attention aux logiques de collecte, de documentation et de « mise en données » des sources nativement numériques[13]. Cela renvoie à ce que l'on appelle la « redocumentarisation », en suivant Jean-Michel Salaün :

> *Pour définir la re-documentatisation, il faut commencer par s'entendre sur le terme « documentarisation ». Documentariser, c'est ni plus ni moins traiter un document comme le font, ou le faisaient, traditionnellement les professionnels de la documentation (bibliothécaires, archivistes, documentalistes) : le cataloguer, l'indexer, le résumer, le découper, éventuellement le renforcer, etc. […].*
> *Le numérique, par nature, implique une re-documentarisation. Dans un premier temps, il s'agit de traiter à nouveau des documents traditionnels qui ont été transposés sur un support numérique en utilisant les fonctionnalités de ce dernier. Mais le processus ne se réduit pas à cette simple transposition. En effet, bien des unités documentaires du Web ne ressemblent plus que de très loin aux documents traditionnels. Dans le Web 2.0, dans la construction du Web sémantique ou tout simplement sur les sites dynamiques, la stabilité du document classique s'estompe et la redocumentarisation prend une tout autre dimension. Il s'agit alors d'apporter toutes les métadonnées indispensables à la reconstruction à la volée de documents et toute la traçabilité de son cycle*[14].

Pour saisir les pratiques, les réussites et les limites des opérations de (re)documentarisation, il s'agit de penser d'abord la manière dont sont traitées, transformées et « augmentées » les sources nativement numériques – au point que le chercheur Niels Brügger parle pour les archives du Web non plus de « *born digital heritage* » mais de « *reborn digital heritage* »[15]. Ainsi une page web archivée est-elle *redocumentarisée*, et les gestes de collecte et d'archivage, loin d'être des enregistrements d'une information *déjà là*, modifient, nourrissent, augmentent le matériau initial. Il faut alors prendre en compte la diversité et l'évolution des approches de redocumentarisation, mais aussi les cheminements de la recherche, à l'instar

par des institutions. Le répertoire des thèses est disponible sur la liste de l'Association des historiennes et historiens du contemporain (H2C, https://www.asso-h2c.fr/).

13. Valérie Schafer et Sophie Gebeil, « Des archives du Web aux données. Une décennie de nouveaux services et collaborations », *Balisages*, 6, 2023, https://doi.org/10.35562/balisages.1066.

14. Jean-Michel Salaün, « La redocumentarisation, un défi pour les sciences de l'information », A. Béguin, S. Chaudiron et É. Delamotte (dir.), n° spécial « Entre information et communication, les nouveaux espaces du document », *Études de communication*, 30, 2007, p. 13-23.

15. Niels Brügger, « Digital Humanities in the 21st Century: Digital Material as a Driving Force », *Digital Humanities Quaterly*, 10-3, 2016, http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

du développement de méthodes de lecture distante, qui traitent les données au moyen d'outils computationnels[16].

Or, documentation et redocumentarisation conditionnent en grande partie notre capacité à exploiter les archives nativement numériques. Il est dès lors essentiel d'en saisir les caractéristiques, les acteurs et les effets sur la recherche en sciences humaines et sociales, tout en pensant les continuités et les ruptures que cela induit pour les archivistes comme pour la recherche en ce qui concerne, par exemple, les fonds, les collections, les documents[17], l'accès ou la contextualisation.

Un patrimoine redocumentarisé

Les environnements documentaires du patrimoine nativement numérique connaissent de rapides évolutions. Ce n'est certes pas une spécificité de ce patrimoine. Il suffit de penser aux archives audiovisuelles françaises de l'Institut national de l'audiovisuel (INA) : les archives de la radio ou encore de la télévision ont été soumises à des modes de documentation qui ont varié au fil du temps, en termes de traçabilité des choix et des conditions de collecte, de mots-clefs (descripteurs) attachés aux contenus, d'outils de fouille proposés, etc.[18]. Sans décrire en détail, pour chaque type de patrimoine nativement numérique, les caractéristiques de documentation et de redocumentarisation à l'œuvre, il convient d'examiner la question du *reborn digital heritage* et des enjeux que partagent ces patrimoines autour, notamment, de leur accès ou de leur recherchabilité.

Comme précédemment évoqué, le patrimoine nativement numérique est en réalité un « *reborn digital heritage* », soit un patrimoine nativement numérique redocumentarisé. Il a subi au cours de sa collecte et de sa préservation de multiples transformations et augmentations par les professionnels de l'information, que ce soient les archivistes, les bibliothécaires, les conservateurs ou encore les ingénieurs d'études et de recherche. L'ensemble de ces actrices et acteurs est coutumier de la documentarisation. Ce sont elles et eux qui, depuis que des procédés d'archivage

16. Franco Moretti, *Distant Reading*, Londres, Verso, 2013.

17. Si elle s'ancre ici dans l'histoire contemporaine, la réflexion sur les sources et les documents transcende les périodes (voir par exemple Pierre Chastang, « L'archéologie du texte médiéval. Autour de travaux récents sur l'écrit au Moyen Âge », *Annales HSS*, 63-2, 2008, p. 245-269) et la question des sources numériques (voir par exemple La forge numérique, « Cluster 5a : TEI et épigraphie, de l'Antiquité à l'époque moderne », journées Biblissima+, 2022, vidéo sur Canal-U, https://www.canal-u.tv/136258).

18. Voir par exemple Céline Loriou, « Faire de l'histoire, un casque sur les oreilles : le goût de l'archive radiophonique », F. Clavert et C. Muller (dir.), nº spécial « Le goût d l'archive à l'ère numérique », *La Gazette des archives*, 253, 2019, p. 71-82 ; Antonin Segault et Marta Severo, « Les archives audiovisuelles de l'INA, une ressource d'expérimentation entre pédagogie et recherche », T. Ouerfelli et A. Chebbi (dir.), nº spécial « Préservation, usage et ré-usage des archives audiovisuelles numériques », *Les Cahiers du numérique*, 19, 2023, p. 93-118 ; Mileva Stupar, « Décrypter le présent au prisme de la mémoire audiovisuelle et numérique : les collections et les données de l'INA au service de la Recherche », *Arabesques*, 113, 2024, p. 24-25.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

existent, se chargent du traitement des documents[19] : la création d'instruments de recherche, l'indexation, la rédaction de résumés, l'ajout de retranscriptions sont autant de gestes de documentarisation qui visent à rendre le document accessible à son usager. Ces dispositifs remontent au moins au XIXe siècle, si ce n'est plus loin, ainsi que le souligne Olivier Poncet[20].

Aujourd'hui, le processus de redocumentarisation des archives nativement numériques est une condition de leurs collecte, conservation et exploitation, selon des modalités qui diffèrent, pour partie, des premiers gestes de documentarisation des siècles précédents. Le patrimoine nativement numérique oblige à reconsidérer des principes archivistiques anciens et, partant, des perceptions de l'archive par les chercheurs et chercheuses : contenu et support peuvent être séparés (le texte d'un blog peut être copié, collé et modifié à l'infini, sans que cela affecte le site web initial) ; les masses de contenu sont de plus en plus importantes[21] ; la collecte ne répond dès lors plus systématiquement à des logiques de fonds générés par des producteurs stables et identifiés (une administration, un notaire, par exemple), mais peut s'organiser thématiquement et répondre à l'actualité, assemblant des productions de natures hétéroclites.

Les archives du Web offrent un observatoire de ce changement des principes archivistiques : dans leur cas, il est établi que l'on n'a jamais affaire à une copie 1:1 de ce qui a pu être présent en ligne[22]. D'une part, les archives du Web sont collectées non comme des captures d'écran statiques qui feraient de la page des unités, mais par éléments (images, liens hypertextes, etc.), ce qui en fait des composites complexes. D'autre part, la collecte vise à reproduire une forme dynamique dans laquelle les hyperliens renvoient à d'autres pages et contenus[23]. Il ne s'agit donc pas de figer les pages, mais bien de maintenir leur caractère interactif et de reproduire une expérience de navigation au sein du site et entre les pages. Celle-ci reste toutefois imparfaite, pour des raisons variées : un hyperlien peut renvoyer à une page non conservée ; certains éléments résistent à la collecte (comme les nombreuses animations flash et images manquantes dans les archives du Web des années 1990) ; des sauts temporels sont palpables, par exemple quand un hyperlien d'une page web renvoie à une autre page, capturée à une date antérieure ou postérieure. Explorer

19. Voir par exemple, pour les institutions d'archivage, les travaux d'Anne Both et notamment son ouvrage *Le sens du temps. Le quotidien d'un service d'archives départementales*, Toulouse, Anacharsis, 2017.

20. Olivier Poncet, « Du neuf avec de l'ancien : les changements de paradigme des archives », *in* C. Scopsi *et al.* (dir.), *Les nouveaux paradigmes de l'archive*, Pierrefitte-sur-Seine, Publications des Archives nationales, 2024, p. 11-15.

21. 928 milliards de pages web sauvegardées par Internet Archive depuis 1996 au moment où nous écrivons ; la BNF couvre quant à elle actuellement près de 6 millions de noms de domaine, et l'on pourrait multiplier les exemples.

22. Niels Brügger, « L'historiographie de sites Web : quelques enjeux fondamentaux », J. Bourdon et V. Schafer (dir.), nº spécial « Histoire de l'Internet, l'Internet dans l'histoire », *Le Temps des médias*, 18, 2012, p. 159-169.

23. Francesca Musiani *et al.*, *Qu'est-ce qu'une archive du web ?*, Marseille, OpenEdition Press, 2019.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

les interfaces, mais aussi les coulisses et le code source d'une page préservée par Internet Archive permet de se rendre compte des augmentations adjointes aux pages archivées, qui imbriquent des éléments « d'origine » et d'autres ajoutés par la fondation états-unienne (les métadonnées de redocumentarisation). Comme pour d'autres types de patrimoine nativement numérique, les archives du Web incorporent de multiples couches de transformation et d'enrichissement, c'est-à-dire de redocumentarisation : des modifications inhérentes à la collecte, des métadonnées, des permaliens pour permettre des citations pérennes, etc. Ainsi que l'a noté Emmanuelle Bermès[24], cet enrichissement est à mettre en relation avec la tendance qui veut que l'on aille de plus en plus d'un Web archivé vers un « Web des données », c'est-à-dire d'une approche considérant les pages comme des unités documentaires vers une approche plus transversale, morcelée et portant intérêt à certaines données plus spécifiques à l'écosystème du Web (images, nombre de likes, etc.). Cette dynamique invite à réfléchir à la question de l'authenticité documentaire, au cœur des préoccupations des historiens et historiennes depuis longtemps : le patrimoine nativement numérique peut être perçu comme un palimpseste temporel résistant aux outils critiques traditionnels. Le travail sur ce type de sources est alors précisément conditionné par la connaissance que nous avons de leurs contextes et logiques de constitution et de redocumentarisation. E. Bermès note ainsi que « toutes ces procédures aboutissent à la création d'archives numériques dont l'authenticité n'est garantie que par les procédures dont elles font l'objet au sein des institutions patrimoniales, mettant ainsi sur le devant de la scène les pratiques professionnelles des bibliothécaires, archivistes et conservateurs[25] ».

Autre déplacement archivistique, la logique de stock est mise au défi et remplacée par une logique de flux. Comme le notait Louise Merzeau dès 2003 :

> *N'étant plus assigné à un support durable, le document se segmente en éléments plus ou moins autonomes (images, boutons, bandeaux, textes…), que l'on doit traiter isolément. C'est la quantité de ces objets à rapatrier (plusieurs milliards) plus que le nombre d'octets qui entraîne un formidable changement d'échelle de l'archive. […] Chaque parcours de navigation recevra alors une étiquette chronologique, attestant des états dans le temps d'un système toujours en équilibre – le stock dessinant ainsi une sorte de* cartographie horaire *du flux*[26].

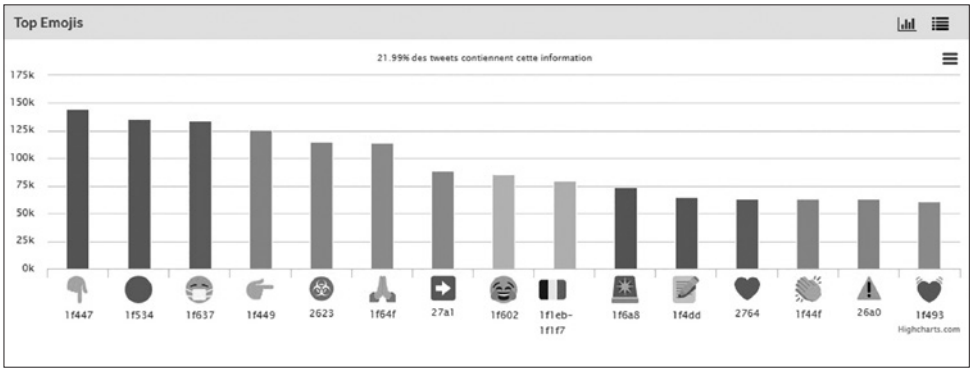
Ces transformations subies au cours des processus de collecte et de traitement ne sont évidemment pas sans conséquence sur le patrimoine nativement numérique tel que mis à disposition de ses utilisateurs et utilisatrices. Dans les archives Twitter de l'INA, il est toujours possible de faire une lecture qualitative des tweets, mais il est aussi possible d'utiliser des outils qui permettent une vue sur des données

24. Emmanuelle Bermès, « Le numérique en bibliothèque : naissance d'un patrimoine. L'exemple de la Bibliothèque nationale de France (1997-2019) », thèse de doctorat, Paris, École nationale des chartes, 2020.
25. *Ead.*, *De l'écran à l'émotion. Quand le numérique devient patrimoine*, Paris, École nationale des chartes-PSL, 2024, p. 83.
26. Louise Merzeau, « Web en stock », *Cahiers de médiologie*, 16-2, 2003, p. 158-167.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

spécifiques. Ainsi peut-on, dans le grand volume de données, ne chercher que les émoticônes : une requête dans les archives Twitter des attentats de 2015 donne la possibilité d'observer, par exemple, les émoticônes associées au mot-dièse (*hashtag*) JeSuisCharlie. Cela permet de distinguer des tendances dans la masse textuelle comme visuelle, à l'instar de l'usage des émojis crayon, drapeau français ou encore mains jointes en signe de prière, associés aux réactions des utilisateurs et utilisatrices. De même, pour la collection liée à la pandémie de COVID-19, il est possible de visualiser les émojis les plus utilisés (fig. 1).

Figure 1 – Émojis les plus utilisés entre février et juillet 2020 pour une requête COVID dans les tweets archivés par l'INA



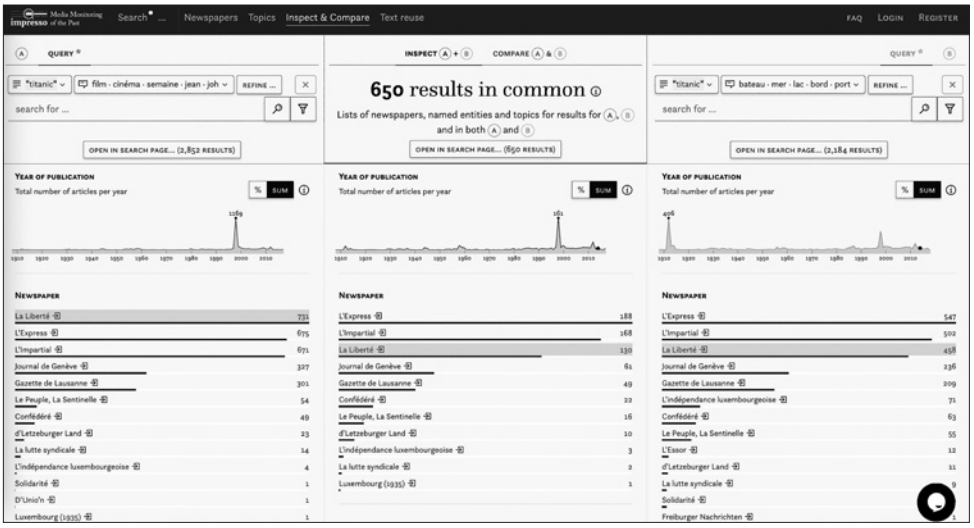
Source : INA.

Cette approche n'est pas spécifique au patrimoine nativement numérique ; il en va de même pour le patrimoine numérisé, par exemple la presse. Là aussi, il est possible d'effectuer des recherches ciblées sur des éléments comme les titres, les mots-clefs ou parfois les images, les encadrés publicitaires, etc., sans forcément passer par la lecture complète de la page qui les accueille. Les modes de lecture se diversifient. C'est le cas avec des plateformes comme Retronews[27], le site presse de la Bibliothèque nationale de France (BNF), ou d'initiatives nées dans le monde académique, à l'instar de Médias 19[28], ou encore avec la plateforme Impresso[29] (fig. 2) qui s'appuient sur des collections de presse numérisée et offrent des interfaces et outils de fouille avancés.

27. https://www.retronews.fr.
28. https://www.medias19.org.
29. https://impresso-project.ch.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

Figure 2 – L'outil « Inspect & Compare » d'Impresso et ses filtres fondés sur la modélisation de sujets



La lecture dépend alors étroitement de l'interface. Or, ces interfaces d'accès, homogénéisées et « naturalisées »[30], peuvent donner l'impression d'un ensemble cohérent, alors qu'elles sont le résultat de différents procédés documentaires et techniques, menés par plusieurs acteurs à des rythmes divers, ainsi que de variations des cadres, périmètres et outils de collecte.

Une variété de protagonistes et d'approches

La redocumentarisation du patrimoine numérique repose en effet sur des protagonistes multiples qui conduisent des stratégies de collecte variées, dépendant à la fois de leurs choix et de contraintes techniques.

Il y a d'abord les institutions patrimoniales nationales et internationales, que l'on songe à la BNF ou à Internet Archive[31]. Il faut ajouter que, sans être des acteurs d'archivage au sens propre, les opérateurs commerciaux conservent nos traces nativement numériques. On peut penser à Twitter, devenu X, et Facebook, mais aussi à des plateformes plus spécialisées comme Know Your Meme[32], qui

30. Mélanie Roustan (dir.), *La recherche dans les institutions patrimoniales. Sources matérielles et ressources numériques*, Villeurbanne, Presses de l'ENSSIB, 2016. La « naturalisation » renvoie ici à la façon dont la construction de l'interface d'accès et de recherche documentaire masque les dispositifs technologiques sous-jacents, rendant la structure informatique invisible et permettant des requêtes en langage dit « naturel ».
31. Il est bien entendu, par ailleurs, que ces deux institutions n'ont ni le même type de statut ni les mêmes modes de gouvernance ou de financement.
32. https://knowyourmeme.com.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

documentent le patrimoine numérique tout en en tirant des bénéfices. En sus des démarches commerciales, on peut aussi penser à ceux et celles qui cherchent à préserver les traces de leurs activités sur le Web, que ce soient les associations ou les initiatives militantes. En effet, la conscience patrimoniale peut être étroitement liée à des enjeux politiques de visibilité et de lutte pour documenter et revendiquer une présence en ligne. C'est le cas par exemple d'initiatives de préservation des mémoires lesbiennes en ligne, qui articulent souvent sensibilité à des archives numérisées et attention au patrimoine nativement numérique, jusqu'à proposer des formations *ad hoc*[33]. Cette conscience patrimoniale pose aussi la question de la relation aux producteurs et aux auteurs. Le patrimoine nativement numérique n'est pas un vaste ensemble désincarné, et la relation au producteur d'archives peut, dans un certain nombre de situations, devenir un atout pour la collecte, la préservation et la contextualisation[34].

Parmi tous ces acteurs, les choix, périmètres et stratégies de conservation varient, que l'on considère l'échelle choisie pour les collectes, les logiques qui président à l'assemblage documentaire ou les choix techniques opérés. En effet, préserver un code source, le contenu d'un CD-ROM ou d'une page web requiert des outils différents et seulement partiellement des techniques de documentation communes : certaines pratiques peuvent être partagées comme l'usage de permaliens ou de métadonnées, d'autres non. Les périmètres de conservation et d'approches peuvent aussi évoluer au sein d'une même institution, que ce soit sous l'influence d'initiatives internes, de moyens alloués et de missions supplémentaires ou encore d'incitations nationales et d'émergence de communautés de pratiques, parfois transnationales.

Les ambitions varient également selon l'échelle choisie, allant de collectes ciblées sur un type de contenu, à l'instar de l'initiative Saving Ukrainian Cultural Heritage Online (SUCHO)[35] destinée à sauvegarder le patrimoine culturel en ligne ukrainien, à des opérations étatiques d'archivage du Web menées par les bibliothèques nationales, voire à une visée mondiale avec Internet Archive. Certains projets peuvent englober une diversité de patrimoines nativement numériques – des contenus de CD-ROM, des codes source, des archives du Web, par exemple dans le cas d'Internet Archive, mais également de la BNF –, tandis que de nombreux services d'archives prennent en charge la conservation de bases de données[36].

33. Léna Bouillard, « Du goût à la fièvre : réflexions autour des archives (nativement) numériques lesbiennes », *in* F. Clavert et C. Muller (dir.), *Le goût de l'archive à l'ère numérique*, 2017-2023, https://gout-numerique.net/table-of-contents/archives-nees-numeriques/du-gout-a-la-fievre-reflexions-autour-des-archives-nativement-numeriques-lesbiennes.
34. La question du lien entre l'archiviste et le service producteur des archives nées numériques est abordée notamment dans Céline Guyon, « La fabrique de l'archive : le rituel de la collecte des archives », F. Clavert et C. Muller (dir.), n° spécial « Le goût de l'archive à l'ère numérique », *La Gazette des archives*, 253, 2019, p. 9-16.
35. http://www.sucho.org.
36. É. Cohen et J. Verlaine, « Le dépôt légal de l'internet français… », art. cit.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

À cette diversité des périmètres et des ambitions répond une diversité des héritages et des expériences : alors que les bibliothèques, bénéficiaires de toute une tradition de conservation des publications, notamment papier, et du dépôt légal, inscrivent leur action de collecte du Web dans le cadre de la loi (le dépôt légal de l'Internet) et de missions patrimoniales institutionnelles, certaines initiatives se distinguent comme Arquivo.pt[37] pour le Portugal. Liées au monde des infrastructures numériques et des réseaux pour la recherche, toutes les archives du Web proposées sur la plateforme sont, comme celles d'Internet Archive, en accès ouvert, malgré les enjeux de droits d'auteur. De nouveaux venus s'appuient, quant à eux, sur des initiatives participatives pour faire face à l'urgence. SUCHO est ainsi très rapidement né après l'agression russe contre l'Ukraine en 2022 et a reçu le soutien et les contributions de milliers de bénévoles qu'il a fallu former et auxquels il a fallu donner les moyens de contribuer à la collecte des contenus en ligne[38].

Ces différences ont des conséquences sur les collections. Un exemple est celui de la conservation des traces numériques de la pandémie de COVID-19[39] : nombre d'initiatives indépendantes, plus ou moins pérennes, ont vu le jour, que ce soit de la part des universités, des GLAM (acronyme de Galleries, Libraries, Archives and Museums), des bibliothèques, notamment nationales, etc., donnant lieu à de multiples traces numériques[40]. Un exemple antérieur est celui des attaques terroristes de 2015 en France, notamment contre *Charlie Hebdo* et au Bataclan. La BNF et l'INA ont rapidement lancé des collectes des contenus en ligne[41], mais celles-ci ont aussi émané de chercheurs et de chercheuses ou encore de bibliothécaires : Nick Ruest, depuis le Canada, a par exemple mené des collectes Twitter abondantes[42]. Les contenus conservés par ce dernier et ceux qu'ont recueillis l'INA et la BNF sur Twitter ne sont pas identiques. Il est effectivement difficile de préserver dans son intégralité une telle masse de données : il faut s'adapter quasi instantanément aux tendances et aux changements des mots-clefs pour que les requêtes soient pertinentes, et l'API (interface de programmation d'application) de Twitter limitait,

37. https://arquivo.pt.

38. Sur les initiatives participatives et la formation des volontaires, voir http://www.sucho.org/reports. Voir également Olga Holownia et Kelsey Socha, « Web Archiving the War in Ukraine », International Internet Preservation Consortium, blog, 20 juill. 2022, https://netpreserveblog.wordpress.com/2022/07/20/web-archiving-the-war-in-ukraine/.

39. Amanda Greenwood, « Archiving COVID-19: A Historical Literature Review », *The American Archivist*, 85-1, 2022, p. 288-311.

40. Susan Aasman *et al.*, « Analysing Web Archives of the COVID-19 Crisis Through the IIPC Collaborative Collection: Early Findings and Further Research Questions », International Internet Preservation Consortium, blog, 2 nov. 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/.

41. Une collecte de matériel physique et l'archivage d'objets et messages liés aux hommages est alors également effectuée par des centres d'archives. Voir Maëlle Bazin et Marie Van Eeckenrode (dir.), nº spécial « Mise en archives des réactions post-attentats : enjeux et perspectives », *La Gazette des archives*, 250, 2018.

42. Voir le blog de Nick Ruest, https://ruebot.net/tags/charliehebdo/.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

à l'époque[43], la collecte à un instant T à 1 % du flux mondial de tweets, rendant son exhaustivité impossible[44].

La régularité des collectes au sein d'une même institution relève de choix mais aussi de ressources. Les collectes des archives du Web sont étroitement corrélées aux moyens techniques et humains dont dispose l'institution, ce qui implique des choix de curation et des priorités documentaires. Pour autant, des collections et des thématiques fortes ressortent : dans les bibliothèques européennes, les archivistes du Web sont sensibles aux contenus de la presse en ligne et portent une attention particulière aux institutions culturelles et politiques ; certaines excluent les sites web des internautes ordinaires, tandis que d'autres vont les collecter mais au rythme d'une ou deux collectes annuelles. Par ailleurs, des actions peuvent être engagées plus spécifiquement lors de crises (attentat, mouvement social, etc.), d'événements majeurs (Jeux olympiques) ou lorsque les contenus sont menacés de disparition et demandent une action de préservation immédiate, comme l'ont menée récemment la BNF et l'INA pour les skyblogs[45] à l'annonce de la fermeture de la plateforme[46]. Des sensibilités variées influencent ainsi les approches de ce patrimoine nativement numérique. Les périmètres des collections sont certes liés aux législations et au dépôt légal dans plusieurs pays, mais ils tiennent également aux choix de curation effectués par les équipes. Ces choix ont, parfois, des implications politiques : alors que l'invasion russe de l'Ukraine ou la pandémie de COVID-19 ont fait l'objet de collectes dans de nombreux pays occidentaux, il n'en est pas de même avec les attaques perpétrées par le Hamas contre Israël et la guerre menée depuis par Israël à Gaza[47]. On constate une même indifférence à l'archivage numérique du conflit entre l'Arménie et l'Azerbaïdjan de 2020, alors que celui-ci avait engendré une importante activité en ligne, par exemple sur le site web de discussion Reddit.

Les approches techniques varient aussi : pour collecter le patrimoine nativement numérique, il faut trouver des solutions techniques permettant de préserver des contenus composites, mêlant texte, image, son, code, le tout articulé en fils, discussions, nouvelles itérations et versions, *remixes*, etc. Ainsi, la BNF et l'INA ont adopté une approche différente des réseaux sociaux numériques et en particulier de Twitter. La BNF capture les informations en conservant la forme des fils de message, tandis que l'INA a choisi de passer par l'API de Twitter et de davantage

43. Ce type de collecte n'est aujourd'hui plus possible. Voir *infra*.

44. Valérie Schafer *et al.*, « Paris and Nice Terrorist Attacks: Exploring Twitter and Web Archives », K. Niemeyer et S. Ericson (dir.), nº spécial « Media and Terrorism in France », *Media, War & Conflict*, 12-2, 2019, p. 153-170.

45. Blogs hébergés par la plateforme Skyrock, qui a décidé de leur fermeture en 2024, après que Skyblog a connu une grande popularité dans les années 2000.

46. Sur les initiatives menées à la BNF, voir « La Bibliothèque nationale archive les Skyblogs », https://www.bnf.fr/fr/la-bibliotheque-nationale-archive-les-skyblogs. Et sur celles menées à l'INA, voir Bruno Texier, « Les archives de Skyblog ont été sauvegardées par l'Ina », *Archimag*, 27 juin 2023.

47. À notre connaissance, seules des institutions libanaises ont opéré un archivage du Web palestinien pendant cette crise.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

traiter les contenus comme des données, sans préserver l'interface originelle, dans l'objectif de favoriser les lectures distantes par mots-dièse, images, etc.[48], en prêtant donc une attention importante à la redocumentarisation. De plus, toute la masse de pages archivées n'est pas forcément traitée en plein texte (*plain text*)[49]. Seules les pages d'accueil des sites web conservés par Internet Archive sont en effet actuellement recherchables en plein texte : une recherche par mot-clef ne donne ainsi que des résultats partiels parmi les contenus archivés.

Si la diversité est de mise, les institutions partagent cependant des points communs. La fondation Software Heritage a adopté des identifiants uniques pour les codes source préservés, à l'instar de ce que font les bibliothèques pour les archives du Web, avec la création de permaliens qui permettent de citer les sources de manière pérenne. Il existe des espaces de partage, de mise en commun et de circulation des expériences et des savoir-faire (qui reposent d'ailleurs sur une tradition archivistique plus longue) : dans le cas des archives du Web, on peut citer l'International Internet Preservation Consortium (IIPC)[50] ou encore les conférences qui mêlent le milieu des archives et celui de la recherche, comme les conférences RESAW (A Research Infrastructure for the Study of Archived Web Materials) organisées depuis 2012[51]. En France, la BNF et l'INA ont créé des DataLabs[52] pour développer les littératies numériques, diffuser les compétences en analyse de données et aider les chercheurs et chercheuses. Ces laboratoires peuvent s'appuyer sur des infrastructures de recherche, comme la très grande infrastructure de recherche (TGIR*) Huma-Num, pour mener des collaborations.

Il n'en reste pas moins que le patrimoine nativement numérique est soumis à des évolutions constantes, liées à la fois à des changements sociotechniques et légaux, et à des enjeux éthiques. La question des mèmes[53] – ces textes, images ou vidéos qui sont rapidement diffusés par réplication ou reproduction dans les courriels, les billets de blog, sur les médias sociaux, etc. – en fournit un bon exemple, puisqu'ils posent des défis d'archivage. Ils sont disséminés dans les captations des réseaux sociaux numériques ou encore des sites web. La bibliothèque du Congrès aux États-Unis a néanmoins lancé des initiatives dédiées pour documenter et archiver les sites qui recensent les mèmes ou permettent de les générer, dans une

48. Alexandre Faye, Jérôme Thièvre et Valérie Schafer, « Le temps des plateformes : enjeux, différences et complémentarité de l'archivage des médias sociaux numériques à la Bibliothèque nationale de France et à l'Institut national de l'audiovisuel », *in* C. Scopsi *et al.* (dir), *Les nouveaux paradigmes de l'archive*, *op. cit.*, https://books.openedition.org/pan/7344.

49. La recherche en plein texte examine tous les mots d'un document et ne se limite pas à des métadonnées ou à des champs prédéfinis.

50. Voir le site de l'IIPC : https://netpreserve.org.

51. Voir le site de RESAW : https://cc.au.dk/en/resaw.

52. Voir le BnF DataLab, https://www.bnf.fr/fr/bnf-datalab, et INA le lab, https://inalelab.hypotheses.org.

53. Sur les mèmes et leur archivage, voir Valérie Schafer et Fred Pailler, « 'All Your Image Are Belong to Us' : Heritagization, Archiving and Historicization of Memes », E. D'Armenio et M. G. Dondero (dir.), n° spécial « Hyper-Visuality », *Visual Communication*, 23-3, 2024, p. 527-543.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

approche liée à la préservation du folklore, ici numérique[54]. SUCHO a créé un mur de mèmes[55] dédiés à la guerre en Ukraine et aux contenus échangés sur des plateformes comme Telegram[56]. Ces contenus éphémères se transforment et circulent rapidement[57], ce qui les rend difficiles à documenter et pose des défis de redocumentarisation, essentiels pour les saisir et les contextualiser pleinement comme des formes de cultures numériques mais aussi de communication[58]. Un exemple très concret est celui du mème « Disaster Girl », une image macro très populaire représentant une fillette au sourire ambigu devant une maison en feu. On peut en retrouver des traces dans les archives du Web sans pour autant pouvoir suivre précisément sa circulation, ses usages et *remixes*, en particulier parce que le mème résiste aux outils classiques de découvrabilité, comme la recherche par image ou par mot-clef (la photographie est rarement associée à la mention « Disaster Girl »). Si l'on revient au mur de mèmes de SUCHO, on trouve des images complexes à recontextualiser : ce mur ne dit rien de la plateforme sur laquelle le mème a été trouvé, de l'ampleur de sa circulation, d'éventuelles modifications, de doublons, et ses codes culturels et ses références peuvent nous échapper[59].

Cette recontextualisation des sources, si elle reste au cœur de tout travail historique, dépend aussi étroitement des métadonnées et des politiques menées par les acteurs des collectes en matière d'accessibilité, d'ouverture des données et de partage.

54. https://blogs.loc.gov/loc/2017/06/remix-slang-and-memes-a-new-collection-documents-web-culture/.

55. « SUCHO Meme Wall », https://memes.sucho.org/.

56. Application de messagerie instantanée et service de partage de fichiers qui s'appuie sur le cloud lancés en 2013.

57. À cet égard nous tenons à souligner que l'intérêt pour la circulation des images ou des signes n'est pas propre à l'étude des cultures virales ou mémétiques en ligne telle que menée par exemple au sein du projet Hivi (C20/SC/14758148) dédié à la viralité en ligne. En effet, dans le cadre du projet ANR CROBORA « Crossing Borders Archives: The Circulation of Images of Europe », Matteo Treleani et son équipe ont suivi et reconstitué la circulation d'images et de symboles européens dans la sphère médiatique, notamment audiovisuelle, avec des défis qui partagent des liens avec ceux du projet Hivi. Voir notamment Matteo Treleani et Dario Compagno, « A Premediation of Brexit: Genesis, Circulation and Naturalization of an Emblem », E. Reyes *et al.* (dir.), n° spécial « Web Studies », *IJDST. International Journal of Design Sciences & Technology*, 25-2, 2023, http://ijdst.europia.org/index.php/ijdst/article/view/98 ; Shiming Shen *et al.*, « From Stock Shots to Ghost Data: Tracking Audiovisual Archives About the European Union », *VIEW: Journal of European Television History & Culture*, 12-23, 2023, p. 4-23.

58. V. Schafer et F. Pailler, « 'All Your Image Are Belong to Us' », art. cit.

59. Voir par exemple, sur l'usage de mèmes mêlant le Shiba inu, symbole de résistance, et la pastèque, emblème de la ville de Kherson, Anna Rakityanskaya, « The SUCHO Ukrainian War Memes Collection », *Slavic & East European Information Resources*, 24-1, 2023, p. 53-70.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

Des enjeux d'accessibilité, d'ouverture et de partage

L'accessibilité du patrimoine nativement numérique et la disponibilité des données soulèvent en effet de nombreuses questions qui concernent leur valeur culturelle et scientifique, mais aussi les valeurs éthiques qui les sous-tendent[60]. En particulier, l'application des principes FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*)[61] à ce patrimoine interroge. Les acteurs se heurtent en effet à plusieurs obstacles en termes de recherchabilité, d'accessibilité, d'interopérabilité et de réemploi des contenus nativement numériques préservés. Le premier concerne les droits d'auteur : souvent, pour cette raison, les résultats des collectes menées par les institutions dans le cadre du dépôt légal ne sont accessibles qu'au sein de leurs établissements. Autre difficulté, la collecte comme la mise à disposition des données n'échappent pas à des questions industrielles et commerciales : Software Heritage s'est chargé de la mission « de collecter, préserver et partager tous les logiciels disponibles publiquement sous forme de code source[62] ». En d'autres termes, Software Heritage ne collecte pas (et ne peut pas collecter) les logiciels d'usage très courants dont le code source est propriétaire. De même, les réseaux socio-numériques témoignent de ces enjeux, puisque la dynamique récente de fermeture des interfaces de programmation (API)[63] de X et de Reddit est révélatrice d'une stratégie de monétisation de l'accès à la collecte des contenus, ce qui met en péril des pratiques qui bénéficiaient auparavant d'une politique plutôt ouverte, comme c'était le cas à l'INA. On a même pu parler d'« APIcalypse »[64] pour désigner ce mouvement de fermeture. Des sites d'organes de presse interdisent de leur côté leur archivage par Internet Archive, afin de garder le contrôle sur la monétisation de leurs contenus[65]. Ces logiques économiques contraignantes ne sont certes pas nouvelles et ne concernent pas seulement le patrimoine nativement numérique ou numérisé : en 2008, un groupe d'historiens

60. Valérie Schafer et Jane Winters, « The Values of Web Archives », *International Journal of Digital Humanities*, 2, 2021, p. 129-144.
61. Mark Wilkinson *et al.*, « The FAIR Guiding Principles for Scientific Data Management and Stewardship », *Scientific Data*, 3, 2016, nº 160018, https://doi.org/10.1038/sdata.2016.18.
62. https://www.softwareheritage.org/mission/?lang=fr.
63. Nous renvoyons également à une stimulante étude sur l'évolution de l'API de Facebook : Fernando N. van der Vlist *et al.*, « API Governance: The Case of Facebook's Evolution », *Social Media + Society*, 8-2, 2022, https://doi.org/10.1177/205630512210862.
64. Axel Bruns, « After the 'APIcalypse': Social Media Platforms and Their Fight Against Critical Scholarly Research », *Information, Communication & Society*, 22-11, 2019, p. 1544-1566.
65. Ils sont toutefois collectés par la BNF, mais consultables sur site uniquement. *Le Monde* et d'autres organes de presse ou des sites web de différentes natures utilisent le fichier robots.txt qui donne des instructions aux logiciels d'indexation (ceux des moteurs de recherche comme ceux des archives du Web et, aujourd'hui, ceux des grandes plateformes d'intelligence artificielle générative). L'analyse de ces fichiers peut fournir d'importantes informations. Voir Katherine Mackinnon et Emily Maemura, « From IA_Archiver to OpenAI: The Pasts and Futures of Automated Data Scrapers », *AoIR Selected Papers of Internet Research*, 2025, https://doi.org/10.5210/spir.v2024i0.13995.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

économistes avait par exemple tiré la sonnette d'alarme quand BNP-Paribas avait envisagé d'envoyer au pilon une partie de ses archives historiques[66].

À ces contraintes externes s'ajoutent des défis techniques : les contenus doivent pouvoir être retrouvés (c'est la « découvrabilité ») dans un paysage foisonnant dans lequel il est parfois difficile de comprendre qui collecte quoi et comment y accéder. La performance virale « Harlem Shake » fournit un bon exemple des difficultés posées. Cette performance en ligne, musicale et dansée, gagne une popularité internationale en 2013[67] et entre dans les contenus archivés par la BNF *via* différents canaux, sans pour autant faire l'objet d'une collecte spéciale. C'est à la faveur de collectes des sites web, de presse ou encore de contenus vidéos (Dailymotion) que des traces en sont gardées. Or, seule une partie de la collecte du début des années 2010 est requêtable en plein texte à la BNF, notamment celle qui concerne la presse en ligne. Ainsi faut-il, pour retrouver les contenus liés au « Harlem Shake », développer des stratégies et passer par une extraction des adresses URL[68] afin d'isoler celles qui contiennent le terme, et ce avec l'appui des bibliothécaires[69]. Cette méthode de recherche signifie que l'on passe à côté de contenus « Harlem Shake » dans des sites non dédiés, plus généralistes, dont les URLs ne comportent pas ces mots.

Le cas du « Harlem Shake » témoigne également de la fragmentation de l'archivage qui, sans être nouvelle, oblige à réfléchir aux périmètres d'archivage des institutions patrimoniales. On trouve ainsi la trace du « Harlem Shake » dans les archives de l'INA. Cependant, si l'on peut dénicher un tweet mentionnant le « Harlem Shake » en lien avec une émission télévisée, on ne peut pas retrouver l'ensemble des tweets qui lui sont dédiés, ni l'ensemble des vidéos. Une bonne connaissance de l'histoire des pratiques numériques est par ailleurs souvent requise : dans le cas du « Harlem Shake », les traces retrouvées doivent être replacées dans le contexte des usages de Twitter ou encore de YouTube en 2013[70].

En définitive, les contenus sont dispersés, les lacunes sont nombreuses et ne sont perceptibles qu'à la condition d'avoir une maîtrise suffisante des questions d'archivage. Retrouver les contenus du patrimoine nativement numérique implique

66. Marc Flandreau, « Lettre ouverte de la communauté internationale d'histoire économique à Michel Pébereau », 8 déc. 2008, https://groupes.renater.fr/sympa/arc/histoire_eco/2008-12/msg00011.html.

67. Phénomène viral Internet, il s'agit d'une série de vidéos mettant en scène un individu puis un groupe de personnes qui dansent de manière exubérante et souvent déguisés sur un extrait de la musique du DJ Baauer.

68. L'URL (*Uniform Resource Locator*) est une adresse web qui spécifie l'emplacement d'une ressource, dont les pages web et le protocole pour y accéder. Extraire les URLs, ici, est une opération qui consiste à trouver des URLs provenant des sites web archivés par la BNF, contenant le terme « Harlem Shake ».

69. Dans le cadre du projet BUZZ-F, mené en 2021/2022 au sein du BnF DataLab (https://www.bnf.fr/fr/les-projets-de-recherche-bnf-datalab#bnf-ann-es-pr-c-dentes).

70. Fred Pailler et Valérie Schafer, « Keep Calm and Stay Focused: Historicising and Intertwining Scales and Temporalities of Online Virality », *in* F. Armaselu et A. Fickers (dir.), *Zoomland: Exploring Scale in Digital History and Humanities*, Berlin, De Gruyter, 2023, p. 119-150.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

donc d'avoir en tête cette cartographie des initiatives, des acteurs, des modes de collecte, de leurs contraintes, et de comprendre le fonctionnement des interfaces de recherche mises à disposition[71]. Ces dernières donnent accès au patrimoine nativement numérique en lissant un grand nombre des logiques de la redocumentarisation. La même interface de la BNF qui propose l'accès aux contenus web de la fin des années 1990 et à ceux produits au cours des années les plus récentes ne permet pas aux chercheurs et chercheuses de distinguer de manière immédiate les changements profonds qui ont affecté les choix de collecte et les méthodes documentaires au cours de plus de vingt années de dépôt légal. L'archivage a pourtant beaucoup varié[72]. Ces interfaces sont aussi hétérogènes et peu interopérables, en particulier quand il s'agit de sortir des périmètres nationaux. Il n'est dès lors pas étonnant que les recherches aient d'abord porté sur des « Webs nationaux »[73], quand des recherches transnationales tardaient à voir le jour. Des expériences ont cependant montré l'importance des liens entre les collections à l'échelle mondiale : pour reconstruire l'histoire du site web de l'université de Bologne, Federico Nanni s'est appuyé sur des archives du Web du Danemark, qui, par le jeu des hyperliens, avaient conservé des captures du site web italien[74]. Récemment, l'étude des contenus liés à la pandémie de COVID-19 a été l'occasion de mesurer et de discuter les écarts de pratiques de collecte qui peuvent exister ne serait-ce qu'en Europe ; une campagne d'entretiens oraux avec des personnels de plusieurs grandes bibliothèques européennes les a documentées[75], ce qui devrait permettre de mieux comprendre comment croiser et comparer les collections en tenant compte de ces variations[76].

En effet, la compréhension du patrimoine nativement numérique est étroitement liée à la capacité à documenter collectivement les gestes de collecte et de redocumentarisation, dans un dialogue entre bibliothèques et institutions patrimoniales

71. Voir à ce sujet Caroline Muller, avec Frédéric Clavert, *Écrire l'histoire. Gestes et expériences à l'ère numérique*, Malakoff, Armand Colin, 2025.
72. Louise Merzeau, « Vers un Web temporel ? Constituer des corpus pour la recherche contemporaine : de l'archivage du web à son analyse », Conférence du Consortium international pour la préservation de l'internet (IIPC), BNF, Paris, 2014.
73. Voir par exemple les recherches menées au Danemark autour de Niels Brügger pour étudier la « web sphere » danoise et la qualifier : Niels Brügger, Jane Nielsen et Ditte Laursen, « Big Data Experiments with the Archived Web: Methodological Reflections on Studying the Development of a Nation's Web », *First Monday*, 25-3, 2020, https://doi.org/10.5210/fm.v25i3.10384.
74. Federico Nanni, « Reconstructing a Website's Lost Past: Methodological Issues Concerning the History of www.unibo.it », *Digital Humanities Quarterly (DHQ)*, 11-2, 2017, p. 1-37.
75. « Web ARChive Studies Network Researching Web Domains and Events » (WARCnet), projet international financé de 2020 à 2023 par l'Independent Research Fund Denmark/Humanities (9055-00005B). Voir les entretiens oraux conservés dans les WARCnet Papers : https://web.archive.org/web/20230902141816/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports.
76. Ces entretiens ont été analysés dans Friedel Geeraert *et al.*, « Oral Histories and Scalable Reading: Analysing Born-Digital Collecting Practices During the COVID-19 Pandemic », *in* S. Aasman, A. Ben-David et N. Brügger (dir.), *The Routledge Companion to Transnational Web Archive Studies*, Londres, Routledge, 2025, p. 121-141.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

et de recherche, qui construisent des espaces d'échanges autour de ces questions[77]. Les expertises sont plus que jamais complémentaires : les chercheurs et chercheuses ont rapidement perçu la nécessité de prendre en compte dans leur analyse les conditions de création des collectes, les atouts, les biais et les limites des données et des outils mis à disposition. Des problématiques liées à l'authenticité du patrimoine nativement numérique[78], la perspective d'appliquer à ces contenus des méthodes issues de la philologie[79] et le besoin d'une critique et d'une herméneutique numériques[80] se sont en effet posés dès que les chercheurs et chercheuses ont commencé à se saisir du patrimoine nativement numérique.

Les outils de la recherche

Les recherches sur les archives du Web[81] et leurs usages historiques ont pris de l'ampleur dans les années 2010, ouvrant des perspectives à la fois dans le champ de l'histoire numérique[82] et, plus récemment, dans celui des *memory studies*[83]. Le patrimoine nativement numérique fait désormais partie des sources qu'il n'est plus besoin de chercher à légitimer. Pour autant, d'autres pans de ce patrimoine sont nettement moins investis, bien que des travaux en cours de jeunes chercheurs et chercheuses sur le code source et l'histoire des langages informatiques laissent présager des développements intéressants[84]. Des espaces de réflexion collective ouvrent aussi la voie, comme le séminaire « Codes sources »[85] organisé depuis 2015, qui prolonge des travaux internationaux menés au sein des *Critical Code Studies*[86].

77. Sam Fritz *et al.*, « Fostering Community Engagement Through Datathon Events: The Archives Unleashed Experience », *Digital Humanities Quaterly (DHQ)*, 15-1, 2021, http://www.digitalhumanities.org/dhq/vol/15/1/000536/000536.html.
78. Mais aussi à leur conservation, par exemple entre migration des contenus, virtualisation et émulation dans le cas du jeu vidéo ou des CD-ROM.
79. Niels Brügger, « The Archived Website and Website Philology: A New Type of Historical Document », *Nordicom Review*, 29-2, 2008, p. 155-175 ; Frédéric Clavert et Caroline Muller, « Les archives du web : du 'goût de l'archive' à des cultures historiennes renouvelées ? », *Les Cahiers du numérique*, 20-3/4, 2024, p. 163-179.
80. Andreas Fickers et Juliane Tatarinov, *Digital History and Hermeneutics: Between Theory and Practice*, Berlin/Boston, De Gruyter, 2022.
81. Daniel Gomes *et al.* (dir.), *The Past Web: Exploring Web Archives*, Cham, Springer, 2021.
82. Par exemple Frédéric Clavert et Serge Noiret (dir.), *L'histoire contemporaine à l'ère numérique/Contemporary History in the Digital Age*, Bruxelles, Peter Lang, 2013.
83. Andrew Hoskins (dir.), *Digital Memory Studies: Media Pasts in Transition*, Londres, Routledge, 2017. Plus récemment, Silvana Mandolessi, « The Digital Turn in Memory Studies », *Memory Studies*, 16-6, 2023, p. 1513-1528. Voir également Sophie Gebeil, *Website Story. Histoire, mémoires et archives du web*, Bry-sur-Marne, INA, 2021.
84. Voir par exemple la thèse de Titaÿna Kauffmann Will sur l'usage du code source comme source historique, en préparation à l'université du Luxembourg.
85. Séminaire « Codes sources », https://codesource.hypotheses.org/.
86. Voir Mark C. Marino, *Critical Code Studies*, Cambridge, The MIT Press, 2020. Ce domaine de recherche à la croisée des études sur les logiciels, des humanités numériques et de l'informatique vise à analyser la signification culturelle du code informatique.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

Toutefois, un problème double persiste souvent : accéder aux sources primaires et pouvoir ensuite les « lire »[87]. En amont, les chercheuses et chercheurs peuvent être directement impliqués dans le processus de redocumentarisation, en particulier quand elles et ils se penchent sur les médias sociaux. Si des tweets sont archivés par les institutions patrimoniales, nombre de spécialistes doivent collecter les données de leur côté pour d'autres sujets. Avant son changement de nom (et de nature) en X, Twitter permettait de collecter des données, parfois même avec des accès spécifiquement mis en place pour les chercheurs et chercheuses, qui se faisaient alors aussi archivistes. Si ce phénomène n'est pas nouveau, il s'agit ici de l'appliquer à des masses de données bien plus étendues pour lesquelles les chercheurs et chercheuses doivent redocumentariser les données obtenues. Ils et elles peuvent alors trouver de l'aide dans les ressources internes des universités, notamment infrastructurelles, et dans des initiatives de répertoires plus larges (HAL, Zenodo, etc.), afin de déposer et documenter leurs jeux de données et d'en assurer une forme de maintenance. Cependant, l'expérience des années 1990 a montré que la concurrence de nombreux systèmes de gestion de bases de données (SGBD) de l'époque a abouti à la disparition de la plupart d'entre eux. Certaines bases de données sont ainsi perdues et, même lorsqu'elles pourraient être retrouvées, sont stockées sur des supports difficilement lisibles aujourd'hui, notamment par manque de matériel et/ou de logiciel adaptés pour les lire (disquettes 5 pouces ¼ par exemple). Reconstituer ces environnements informatiques obsolètes reste possible, mais le chemin est complexe. Les politiques infrastructurelles des universités ou de certaines structures interuniversitaires – les TGIR Huma-Num et Progedo en France, par exemple, mais de manière plus générale la mise en place de dépôts de données pérennes – ont tiré des leçons, dans toutes les disciplines, des décennies passées. Toutefois, la vulnérabilité de ces infrastructures (en termes de financement, de maintenance, de pérennité, de sécurité, etc.) doit être envisagée.

Une fois le corpus collecté, constitué et redocumentarisé, chercheurs et chercheuses peuvent alors « lire » ces sources. Cette lecture dépend toujours de la qualité des métadonnées et de la capacité à décomposer et recomposer leurs sources, à les qualifier et à les contextualiser, donc de la qualité de la redocumentarisation. Ainsi, pour assurer la faisabilité d'une démarche diachronique, des métadonnées d'horodatage s'avèrent nécessaires, alors qu'elles ne sont pas toujours présentes : si, dans les archives du Web, connaître la date et l'heure d'une collecte de données est possible, il n'est que rarement possible de déterminer la date précise de publication d'un contenu en ligne. En outre, on a vu qu'une page web, souvent dynamique, peut être (re)constituée d'éléments dont la date de publication diverge.

Les historiens et historiennes peuvent aussi utiliser des outils permettant une lecture distante[88] – l'une des échelles de la lecture, lorsque l'ordinateur « lit pour nous » – de leur corpus. Pour la fouille de données, l'un des logiciels les plus

87. Ian Milligan, « Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives », n° spécial « The Future of Digital Methods for Complex Datasets », *International Journal of Humanities and Arts Computing*, 10-1, 2016, p. 78-94.
88. Voir F. Moretti, *Distant Reading*, *op. cit.*

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

utilisés aujourd'hui en France est IRaMuTeQ. Issu des mêmes méthodologies que le logiciel commercial Alceste – la théorie des mondes lexicaux[89] –, il est, au contraire de ce dernier, libre, c'est-à-dire que son code source est téléchargeable et modifiable[90]. On peut néanmoins s'interroger sur la pérennité à terme du logiciel : si sa communauté d'utilisateurs et d'utilisatrices est grande, celle des développeurs et développeuses semble l'être beaucoup moins. Toutefois, IRaMuTeQ est un logiciel qui documente lui-même les opérations qu'il mène sur un corpus : en ce sens, chercheurs et chercheuses peuvent revenir sur leurs analyses et les reprendre de manière plutôt aisée et rigoureuse.

Au-delà de la question de la pérennisation d'un logiciel en particulier se pose aussi celle des méthodes informatiques permettant la fouille de données. Par exemple, le *Latent Dirichlet Allocation*[91] (LDA), un type de *topic modelling* (modélisation de sujets), a été « standard » dans les humanités numériques, notamment *via* le logiciel MALLET ; il est aujourd'hui souvent délaissé pour les vecteurs de mots ou de documents, eux-mêmes perfectionnés par les *transformers*, mis au point par Google[92], *transformers* qui sont à la base des intelligences artificielles génératives de texte. Ces dernières posent des problèmes de compréhension : elles sont opaques et, bien qu'elles dépendent d'un entraînement sur de grands ensembles de données, elles restent, malgré des améliorations récentes, peu capables de citer leurs sources et ne sont pas compatibles avec le standard FAIR[93].

En somme, pour comprendre le patrimoine nativement numérique et le mobiliser comme source primaire, il est nécessaire de s'intéresser aux bruits et aux silences des données, à leur constitution, à la représentativité des données archivées, aux atouts et aux limites des méthodes et des outils computationnels employés. Cependant, pour pouvoir remettre ces corpus en perspective, en saisir l'usage, l'audience, la place dans l'écosystème médiatique, il est crucial de les

89. Voir Max Reinert, « Les 'mondes lexicaux' et leur 'logique' à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, 66, 1993, p. 5-39. IRaMuTeQ permet, comme Alceste, de visualiser de différentes manières les grands thèmes traversant un corpus de texte, grâce à différentes méthodes statistiques fondées notamment sur la cooccurrence des mots au sein d'un même segment de texte (phrase ou partie de phrase).

90. https://pratinaud.gitpages.huma-num.fr/iramuteq-website/. Développé notamment par Pierre Ratinaud et Pascal Marchand, Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales (LERASS), université Toulouse Jean Jaurès.

91. David M. Blei, Andrew Y. Ng et Michael I. Jordan, « Latent Dirichlet Allocation », *Journal of Machine Learning Research*, 3, 2003, p. 993-1022. Le LDA est un type de modélisation de sujets reposant sur de l'apprentissage automatique, une branche de l'intelligence artificielle, pour créer des ensembles de mots (*bag of words*) à partir de leur cooccurrence dans le corpus de texte analysé. Ces ensembles de mots fournissent une base d'interprétation pour comprendre les grandes thématiques abordées dans le corpus étudié.

92. Ashish Vaswani *et al.*, « Attention Is All You Need », 2 août 2023, https://doi.org/10.48550/arXiv.1706.03762.

93. Pour un aperçu de ces méthodes et de leurs usages, on pourra se référer à Ian Milligan, Scott B. Weingart et Shawn Graham, *Exploring Big Historical Data: The Historian's Macroscope*, Singapour, World Scientific Publishing, [2015] 2022.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

mobiliser comme des sources parmi d'autres, et donc de les croiser avec des corpus de différentes natures.

Ainsi, pour l'histoire des développements numériques en France[94] ou au Luxembourg[95], pour celle des mémoires en ligne sur le Web[96] et des forums de discussion[97], la recherche exploite des archives du Web et d'autres sources nativement numériques comme les Google Groups qui ont gardé trace des échanges Usenet[98], mais aussi la presse ou des entretiens oraux. Un même projet peut à la fois s'appuyer sur des sources « analogiques », papier et nativement numériques. Le projet CD-Hist sur l'histoire du CD-ROM, mené à l'université du Luxembourg[99], s'appuie par exemple sur des corpus de presse (généraliste et spécialisée), des entretiens oraux, des publicités (audiovisuelles et de presse), une multitude de manuels techniques et multimédia d'époque, des archives d'entreprises, des collections privées (par exemple, la collection numérisée d'Alain Letenneur de kits de connexion Internet[100]) ainsi que sur tout un patrimoine nativement numérique disponible sur Internet Archive de contenus de CD-ROM émulés[101]. S'y ajoutent d'autres collections, disponibles par exemple à la BNF[102], tels environ 50 000 CD-ROM préservés dans le cadre du dépôt légal. Insistons aussi sur le rôle extrêmement actif des communautés amateures (voire pro-ams[103]), notamment dans le champ de la préservation vidéo-ludique et du retrocomputing, soit l'utilisation de matériel et logiciels obsolètes, et même sur l'activité engagée notamment dans la sphère des usagers, parfois pirates, qui ont filmé et mis en ligne des navigations, des contenus, etc. On peut citer

94. Valérie Schafer, *En construction. La fabrique française d'Internet et du Web dans les années 1990*, Bry-sur-Marne, INA, 2018.
95. Carmen Noguera, « Retrieving Traces of the Luxembourg WebSphere Through Its First Websites », conférence de clôture, WARCnet, Aarhus, 18 oct. 2022.
96. Sophie Gebeil, « Les mémoires de l'immigration maghrébine sur le web français de 1999 à 2014 », R. Besson et C. Scopsi (dir.), n° spécial « La médiation des mémoires en ligne », *Les Cahiers du numérique*, 12-3, 2016, p. 115-138.
97. Camille Paloque-Berges, *Qu'est-ce qu'un forum internet ? Une généalogie historique au prisme des cultures savantes numériques*, Marseille, OpenEdition Press, 2018.
98. Usenet est un réseau comprenant des forums populaires notamment dans les années 1980 et 1990 et dont le développement est donc antérieur au Web. Camille Paloque-Berges, « Usenet as a Web Archive: Multi-Layered Archives of Computer-Mediated-Communication », *in* N. Brügger (dir.), *Web 25: Histories from the First 25 Years of the World Wide Web*, Londres, Peter Lang Publishing, 2017, p. 227-250.
99. Projet démarré au 1er juin 2024 avec le soutien du FNR (C23/SC/18097856/CD-Hist): https://www.uni.lu/c2dh-en/research-projects/cd-hist/.
100. http://www.letenneur.com.
101. L'émulation simule l'environnement informatique initial du CD-ROM, ce qui permet de lire sur un ordinateur du présent des contenus conçus pour un environnement matériel et logiciel antérieur. Beaucoup de contenus de ce type sont disponibles sur Internet Archive, allant des *sharewares* qui permettaient d'installer, de tester et de découvrir des logiciels, à des CD-ROM culturels, des jeux vidéo ou encore des jeux érotiques.
102. Près de 50 000 CD-ROM ont été préservés dans le cadre du dépôt légal et font l'objet d'une politique de virtualisation veillant aussi à préserver les contenus matériels – boîtiers, enveloppes, livres et magazines associés.
103. Expression désignant des collaborations entre professionnels et amateurs.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

enfin d'autres sources au périmètre spécifique, à l'instar des dossiers de demandes de subventions pour les réalisations multimédias, déposées au Centre national du cinéma et de l'image animée (CNC) dans le cas de la France, ou encore des collections d'archives institutionnelles, comme celles de l'Office des publications de l'Union européenne. Cet organisme a conservé dans ses archives à Luxembourg-ville tous les CD-ROM qu'il a publiés depuis 1997, contenant par exemple le Journal officiel jusqu'en 2010, quand le DVD a remplacé le support CD-ROM tandis que la pratique du Web généralisait aussi l'accès en ligne. Cet exemple montre la complémentarité des fonds mais aussi des logiques à l'œuvre et les moyens très différents dont disposent les divers acteurs des collectes et de la conservation.

Enfin, si chercheuses et chercheurs doivent documenter leurs méthodes et leurs recherches, et croiser les sources, se pose aussi pour eux la problématique plus large de se (re)documentariser soi-même. Une recherche est rarement le résultat de démarches strictement individuelles. La question de l'auctorialité émerge, tout comme celle des évolutions des acteurs de la conservation et de la recherche : « […] un troisième personnage apparaît entre le conservateur et le chercheur : l'ingénieur[104] ». De plus, les rencontres lors de colloques ou au sein d'un centre de recherche, les discussions en ligne sur les médias sociaux participent toutes à l'élaboration d'une recherche. Outre des notes personnelles, qu'en est-il aujourd'hui des logiciels utilisés par nos universités et qui documentent nos discussions, y compris scientifiques, au jour le jour ? Si la question des traces logicielles des discussions scientifiques n'est pas nouvelle[105], l'usage de logiciels de grandes sociétés appartenant aux GAFAM (Google, Apple, Facebook, Amazon, Microsoft) ou entreprises proches comme Slack (Salesforce), difficilement pérennisables, archivables, difficilement même redocumentarisables, pose un problème pour la redocumentarisation de la recherche elle-même.

Le patrimoine nativement numérique présente des défis nouveaux pour la recherche en termes de préservation, de documentation et d'exploitation, dans la mesure où il implique de sortir du modèle de conservation du papier pour aller vers un paradigme de la redocumentarisation. Cette redocumentarisation est menée par différents types d'institutions patrimoniales et, si nous avons insisté sur Internet Archive, les grandes bibliothèques nationales ou encore Software Heritage, ces enjeux sont présents dans tous les centres d'archives à toutes les échelles : tous sont aujourd'hui confrontés à des sources nativement numériques, des archives départementales aux archives universitaires ou d'entreprise. Les historiens et historiennes sont aussi parties prenantes de cette redocumentarisation. La diversité des protagonistes ainsi impliqués dans la collecte et la conservation des archives nativement numériques, allant des institutions

104. E. Bermès, *De l'écran à l'émotion*, *op. cit.*, p. 202.
105. Voir, par exemple, le cas du laboratoire d'intelligence artificielle de Stanford : Alina Volynskaya, « Querying the Digital Archive of Science: Distant Reading, Semantic Modelling and Representation of Knowledge », thèse de doctorat, EPFL, 2024, https://doi.org/10.5075/epfl-thesis-10732.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

patrimoniales aux initiatives de recherche et participatives, témoigne de la complexité de cette tâche. Les processus de redocumentarisation transforment ce patrimoine en un *reborn digital heritage*, nécessitant une approche critique et réflexive.

Pour les historiens et historiennes, l'utilisation de ces sources implique une acculturation à des compétences et des méthodologies numériques parfois inédites[106]. La constitution de corpus, leur documentation, la lecture distante et l'analyse de grandes masses de données nécessitent des outils spécifiques dont la pérennité n'est pas toujours assurée. La qualité des métadonnées et la compréhension des contextes de collecte et de préservation sont cruciales pour une exploitation rigoureuse de ces sources. En outre, la rapidité des évolutions technologiques pose la question de la durabilité des méthodes et des outils de recherche, ainsi que de la préservation des traces numériques que produisent les chercheurs et chercheuses.

Toutefois, et si nous avons jusqu'ici insisté sur les différences entre le patrimoine nativement numérique et le patrimoine « papier » plus traditionnel, nous pouvons aussi rappeler qu'une source reste une source : les archives nativement numériques ont des absences comme les archives papier ; la nécessité d'avoir une connaissance fine des archives, de leur collecte et de leur organisation, quelles qu'elles soient, se vérifie plus que jamais. Si savoir lire une cote permet de situer précisément un carton d'archives dans un contexte particulier et d'en retirer des informations riches pour une recherche historique, il en va de même de la capacité à savoir lire l'URL d'une page web archivée. La connaissance de traditions archivistiques qui diffèrent – entre les *series* au Royaume-Uni et aux États-Unis ou le principe du respect des fonds en Europe continentale (pour la période contemporaine) – est souvent nécessaire pour orienter au mieux ses recherches. Dans le cas des archives du Web, il ne s'agit pas de cote ou de carton, mais d'URL ou de fichier WARC (Web ARChive). Si l'on ne parle plus de fonds, il demeure des distinctions, par exemple à la BNF, entre les collections : l'une, « Actualités », se réfère à la presse en ligne, dont les contenus sont archivés quotidiennement, alors que d'autres collectes sont effectuées de manière plus ponctuelle, comme les collectes annuelles larges de tous les noms de domaine français identifiés. E. Bermès le rappelle : « cet effet 'déconstructeur' du numérique sur l'unité documentaire » est pensé dès les années 1990, notamment pour le patrimoine numérisé, avec la TEI (*Text Encoding Initiative*) qui cherche à y répondre par un encodage des textes mais aussi de leur contexte[107].

Penser aujourd'hui la masse des données, soit ce qui a été appelé le *big data*, en histoire implique aussi de penser ce qui manque, pour les archives nativement numériques comme pour les archives numérisées. Outre les plateformes de moins en moins archivables – Facebook depuis 2016, X depuis 2023 –, il y a aussi tout ce que ces plateformes prétendent pouvoir saisir mais ne saisissent que de manière imparfaite. Comment comprendre un *like* même décliné en émotions différentes ?

106. Ian Milligan, « You shouldn't Need to Be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure », WARCNet papers, Aarhus, 2020, https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf.

107. E. Bermès, *De l'écran à l'émotion*, *op. cit.*, p. 33.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press

Si des approches différentes sont explorées pour mieux archiver ces plateformes dont la conception rend (souvent volontairement) difficile l'archivage[108], ces méthodes ne sont pas encore en place et ne sont pas toujours compatibles avec les pratiques des institutions patrimoniales. Celles-ci, comme le souligne Lise Jaillant, sont prises en tenaille entre la rigueur de l'archivage, qui implique souvent une certaine lenteur – due autant à des contraintes juridiques, à des limites de ressources, techniques, financières ou humaines, qu'au temps de traitement –, et la demande des autorités publiques, des chercheurs et des chercheuses de fournir un accès rapide aux données qu'elles collectent. *In fine*, nous partageons ce constat de L. Jaillant :

> *Les archives sont faites pour être utilisées et non pas verrouillées. Pour déverrouiller leur potentiel culturel, nous devons travailler de façon interdisciplinaire et exploiter les technologies les plus récentes. L'accès aux archives numériques est essentiel, mais il faut aussi anticiper le moment où les documents [*records*]*[109] *nativement numériques seront plus aisément accessibles. Pour donner un sens à cette masse de données, il est urgent de mettre au point de nouvelles méthodologies, combinant les méthodes traditionnelles des sciences humaines avec des approches fondées sur la richesse en données*[110].

Valérie Schafer
C²DH, Université du Luxembourg
valerie.schafer@uni.lu

Frédéric Clavert
C²DH, Université du Luxembourg
frederic.clavert@uni.lu

Caroline Muller
Université Rennes 2 – Institut Universitaire de France
caroline.muller@univ-rennes2.fr

108. Anat Ben-David, « Counter-Archiving Facebook », *European Journal of Communication*, 35-3, 2020, p. 249-264.
109. Comme l'a noté l'un des relecteurs de cet article, que nous remercions chaleureusement ainsi que les autres expertes et experts anonymes pour leurs remarques et retours, le terme de « document » ne rend malheureusement que partiellement compte de l'enjeu sous-jacent au choix du mot anglais « *records* » par Lise Jaillant.
110. Lise Jaillant, « More Data, Less Process: A User-Centered Approach to Email and Born-Digital Archives », *The American Archivist*, 85-2, 2022, p. 533-555.

https://doi.org/10.1017/ahss.2025.10139 Published online by Cambridge University Press