Francesca Musiani, Camille Paloque-Bergès, Valérie Schafer et Benjamin G. Thierry

Qu'est-ce qu'une archive du web ?

OpenEdition Press

Comment naviguer dans l'archive ?

DOI : 10.4000/books.oep.8746
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2019
Collection : Encyclopédie numérique
ISBN électronique : 9791036504709

COMMENT NAVIGUER DANS L'ARCHIVE ?

L'archive du Web cherche à reproduire l'interactivité qui existait au sein du Web vivant en permettant de cliquer sur les liens et de naviguer dans la Toile. Elle présente toutefois des caractéristiques en termes de temporalités, d'interfaces, de granularité, d'accompagnement des données par des métadonnées, qui rendent explicite le fait que l'archive du Web n'est pas une copie à l'identique du Web au moment de son archivage. Naviguer dans la Toile du passé implique donc des défis et des précautions théoriques comme pratiques, qui interrogent au final la possibilité de repenser ce Web du passé en contexte.

Les temporalités de l'archive du Web

La question des temporalités est probablement l'un des enjeux les plus aigus en matière d'exploitation des corpus conservés. L'archive du Web est instable et signe la « fin de la matérialité documentaire » par le rassemblement de documents « modulables et mobiles », en contradiction avec la vision traditionnelle de l'archivage dont la fonction serait de « figer et [de] stabiliser ». Aussi, l'archive en ligne est marquée du sceau d'une « temporalité brève qui s'accorde mal avec le temps de la recherche historique » (Gebeil, 2016).

S'il ne faut pas minimiser les difficultés posées, il s'agit surtout d'acclimater les pratiques de recherche à de nouveaux régimes de temporalité, dont l'archive hérite du Web lui-même[1]. Serge Noiret nous avertissait en 2011 :

1. Voir l'ouvrage collectif *Temps et temporalités du Web*, Presses universitaires de Paris Ouest, Paris, 2018, issu du colloque éponyme organisé à l'Institut des sciences de la communication du CNRS en décembre 2015.

> « le *digital turn* [tournant numérique] a rendu précaire un certain nombre de concepts chers aux historiens comme celui de la pérennité des sources et la capacité de reproduire dans le temps une analyse qui s'y réfère. » (Noiret, 2011)

Comme Joe Chip, le héros plongé en pleine régression temporelle dans *Ubik* de Philip K. Dick (1969), les utilisateurs de l'archive du Web sont soumis à des régimes chronologiques nouveaux. En premier lieu parce que la sauvegarde d'un site aux mises à jour fréquentes se heurte à l'impossibilité d'une captation totale des données qui le composent : toutes les modifications et ajouts ne peuvent pas être archivés (Mussou, 2012). Ainsi, les archives du site tf1.fr entre 1996 et 2000 dans Internet Archive donnent à voir un corpus réalisé au travers de 18 collectes successives. Pour l'année 1997, ce sont trois captations qui permettent de consulter le site de la première chaîne. À la BnF, les collectes portent sur plusieurs millions de sites archivés depuis 2011 à des fréquences variables, d'« une fois par semaine » à « une fois par an », associées à des « collectes projet » autour d'un sujet particulier[2]. Dans ce cadre, aucune garantie n'existe sur la possibilité de retrouver un site dans son état initial à une date donnée (Brügger, 2012a), chaque état étant le patchwork des modifications intervenues depuis la dernière captation.

Dans le cadre d'une navigation entre les sites, l'archive du Web doit être traitée comme un pavage discontinu de couches temporelles différentes : la page du *Monde* dans la Wayback Machine du 21 février 1999 renvoie par le lien « Nouvelles technologies » à celle du 8 février 1999 (Schafer et Thierry, 2015). L'image du réseau donnée par l'archive est temporellement désaccordée.

À l'échelle de la page et de ses ressources (images, liens, fichiers embarqués divers…), un temps désarticulé est également à l'œuvre : certains contenus d'une page ne sont pas archivés (les publicités ou les commentaires lorsqu'ils sont

2. Voir http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblees_arch_internet.html.

permis par exemple sur les sites de la presse en ligne) ou recollectés, comme évoqué précédemment. Ce dédoublonnage des ressources conduit par exemple à trouver le logo endeuillé de noir du CNRS sur la page d'accueil du site captée en août 2015 par la BnF alors qu'il a été mis en place suite aux attentats de novembre 2015… Des fonctionnalités récemment introduites dans certaines archives du Web peuvent toutefois permettre d'identifier la date de collecte de chaque élément d'une page web archivée par rapport aux autres qui composent cette même page, rendant désormais visibles et explicites ces patchworks temporels[3].

Enfin, la page web archivée elle-même, en tant qu'espace d'affichage ou contenant informationnel, ne comporte pas forcément de date de création, pas de date de modification, mais seulement une date d'archivage, ce qui rend l'analyse diachronique hasardeuse :

> « *there remains a question of the documents' timestamps: The timestamp of the snapshot of a past version of a URL is that of the date of archiving, not necessarily the last updated date of that URL* [...] *To solve this problem, researchers usually aggregate the archived URLs per year, which results in an approximation of an historical hyperlink network with a large margin of error*[4]. » (Ben-David et Huurdeman, 2014)

À la contrainte des régimes de temporalités désaccordés qu'impose l'archivage s'ajoutent les décalages entre les

3. Voir par exemple dans le cas d'Internet Archive, le billet posté sur leur blog le 5 octobre 2017 par Mark Graham "Wayback Machine Playback... now with Timestamps!": https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps/.

4. « Une question reste en suspens à propos de l'« horodatage » des documents: l'« horodatage » d'une ancienne version d'une URL archivée est l'« horodatage » qui correspond à son moment d'archivage, pas nécessairement celui de la dernière mise à jour de cette URL [...] Pour résoudre ce problème, les chercheurs agrègent les URLs archivées par année ce qui crée un réseau de liens avec une large marge d'erreur dans les dates utilisées. » (Notre traduction.) La question de la datation est également sensible dans la thèse de Quentin Lobbé (2018) sous la direction de Pierre Senellart et Dana Diminescu. Se fondant sur la notion de « fragment Web », il explore la possibilité de retrouver sa date d'édition et non la seule date d'archivage.

temporalités *en ligne* et *hors-ligne*. Comme le rappelle Clément Oury dans le domaine des sites politiques, une fois le scrutin achevé, on observe une rapide disparition des pages utilisées pendant la campagne, notamment sous l'effet des recompositions plus ou moins rapides du paysage politique :

> « On a vu, notamment au lendemain du premier tour des élections régionales de 2010, des candidats fermer définitivement leur blog lorsqu'ils ralliaient une liste d'union. » (Oury, 2012)

Pendant la seule campagne pour les élections législatives de 2007, la moitié des sites créés pour l'occasion avait disparu cinq mois plus tard.

À l'inverse, comme le souligne Claude Mussou (Ina), l'archive du Web se constitue au fil de l'eau, à mesure que le corpus s'alimente par sédimentations successives, les collectes s'ajoutant les unes aux autres (Mussou, 2012).

En outre, le hors-ligne pilote en partie l'archive du Web : face aux attentats qui ont frappé la France et en particulier *Charlie Hebdo* en 2015, la BnF comme l'Ina ont choisi de mener des collectes d'urgence.

Les nouveaux régimes de temporalités de l'archive en ligne nous poussent probablement à rompre avec le confort que comporte l'utilisation d'archives datées et précisément identifiées que l'époque contemporaine nous avait habitués à utiliser. Toutefois les collègues spécialistes de périodes plus reculées et moins prolixes en documentation écrite ont déjà affronté des questions semblables. D'un regard vers le passé peut naître une manière d'envisager l'avenir, fut-il numérique.

Interaction et interactivité avec l'archive du Web

L'exploration des archives du Web implique en outre de se soumettre à un régime d'interactivité porté par les interfaces et services qui mettent à disposition du chercheur les masses de données préservées.

Intimement liées à l'esprit du projet initialement conçu par Brewster Kahle à la fin des années 1990, les archives du Web proposent une expérience très proche de celle de la navigation en ligne, progressivement enrichie par de nouvelles fonctions (recherche en plein texte, API diverses, etc.) qui s'adaptent à un enrichissement des corpus, particulièrement avec l'entrée des réseaux socionumériques dans l'orbite de la conservation.

Comme elle avait pesé sur la mise en images et en mots d'Internet, la bibliothèque continue d'être une référence incontournable pour penser l'archive du Web. En 2011 Brewster Kahle rappelait son ambition de faire d'Internet Archive « une bibliothèque numérique » dont la « visée [est] à la fois sociale et technologique » et qui permet un « accès universel à l'ensemble de la connaissance : tous les livres, toute la musique, toutes les vidéos, accessibles partout, par tous » (Kahle, 2014a). C'est cette vision qui l'habite depuis l'origine du projet tel qu'il le décrit en 1997 dans *American Scientific* (Kahle, 1997).

Cette vision explique qu'en 2001, quand naît la Wayback Machine[5] qui permet l'accès aux ressources d'Internet Archive, les sites et leurs pages constituent l'unité de base de la consultation. Internet Archive contient des sites comme une bibliothèque contient des livres.

Encore aujourd'hui, l'entrée principale dans l'archive se fait par l'adresse du site. La navigation dans les versions successivement archivées se fait également à l'échelle du site dans le cadre de ce que Anat Ben-David et Hugo Huurdeman désignent comme une « *single URL approach* [approche par URL unique] » (Ben-David et Huurdeman, 2014).

Bien entendu, une navigation au fil des liens est possible entre les sites archivés, mais sans garantie que les liens aboutissent.

Ce régime d'interaction avec l'archive qui est fondé sur la double métaphore de la bibliothèque et de la toile n'est pas sans poser des problèmes. Le premier d'entre eux, comme le souligne Megan Ankerson, est probablement l'importance disproportionnée donnée au facteur temporel dans une sorte de

5. Pour accéder à la Wayback Machine : https://archive.org/web/.

voyage « chrono-touristique » qui s'impose au chercheur au sein des archives (Ankerson, 2015b). La Wayback Machine ne se prive pas de faire reposer sa communication sur l'invitation à un « voyage dans le passé » mis en avant jusqu'à son interface, surmontée par le slogan « *Explore more than 345 billion web pages saved over time* ».

Ces biais introduits par l'interface et les conditions de collecte des données, les institutions responsables de l'archivage ont tenté de les pallier.

La première étape a consisté à mettre à disposition des chercheurs, souvent après consultation de la communauté des utilisateurs comme à l'Ina ou à la BnF, des outils supplémentaires d'interprétation et d'interrogation des sites archivés.

Le plus attendu a probablement été la possibilité d'une interrogation en plein texte[6] des ressources archivées qui permet d'échapper à une consultation où domine la « *single URL approach* ». Les archives portugaises, françaises (Ina et BnF) ou encore britanniques et japonaises y ont recours.

Cette possibilité enrichit l'expérience de navigation à deux titres au moins. D'abord, la recherche en plein texte permet de thématiser des recherches qui n'auraient pu aboutir par une consultation « à la main » des sites, l'un après l'autre. C'est une étape fondamentale dans la constitution des corpus de recherche et l'émergence de nouveaux objets. Ensuite les résultats obtenus permettent des tris multiples (dates, occurrences d'un terme, d'une expression, présence d'un type de ressources, etc.).

Dernière étape en date de l'évolution des interfaces, la mise en place d'une multitude de « surcouches » de recherche et de manipulation des données qui permettent d'exploiter l'archive et d'en rendre compte sous une forme particulière. En Grande-Bretagne, le moteur Shine[7] permet par exemple de soumettre les résultats d'une recherche à un traitement statistique et de générer une représentation sous une forme proche de Google Ngram. L'archivage

6. La recherche en « plein texte » est ici employée pour traduire l'anglais « *full-text search* ».
7. https://www.webarchive.org.uk/shine.

des réseaux socionumériques par la BnF et l'Ina permet le traitement des métadonnées associées aux messages collectés et donne la possibilité d'interroger les données collectées de manière croisée (par exemple par mot-dièse et langue ou date, etc.) et de représenter les résultats de multiples façons : frises chronologiques, nuage de mots, liste d'emojis les plus utilisés, etc.

Enfin, des initiatives émergent en périphérie d'Internet Archive et des grandes institutions d'archivage pour donner accès à des outils permettant de nouvelles exploitations des données sauvegardées. Citons par exemple Internet Archive Wayback Machine Link Ripper[8] qui permet de retrouver toutes les URLs archivées dans Internet Archive à partir d'une URL connue ; WebART (pour Web Archives Retrieval Tools) qui est un ensemble d'outils et d'interfaces de recherche proposé par l'équipe Dutch Web Archive de la bibliothèque nationale des Pays-Bas et le Centrum voor Wiskunde en Informatica de l'université d'Amsterdam[9], parmi lesquels on trouve WebARtist, un moteur de recherche en plein texte capable de prendre en compte les temporalités pour retrouver un texte ou une image ; ou encore Wayfinder de Megan Dougherty qui permet de personnaliser son interface de recherche dans les archives du Web en complément de la suite WebArchivist (Dougherty, 2017).

Des outils d'analyse

Si la recherche par mots-clés peut sembler indispensable à des chercheurs, habitués, comme le grand public, aux moteurs de recherche et au plein texte, la fourniture de ces fonctionnalités n'est pourtant pas une évidence. C'est seulement en 2016 que la BnF va implémenter une recherche en plein texte dans ses archives du Web des années 1990, puis dans sa collecte des attentats de 2015, et permettre une recherche avancée par mots-clés, dates, auteurs ou types de formats (.html, .pdf,

8. https://tools.digitalmethods.net/beta/internetArchiveWaybackMachineLinkRipper.
9. http://www.webarchiving.nl/news.

etc.) en adaptant le moteur de recherche utilisé par la British Library, Shine. D'autres indexations sont en cours, mais une partie de la collection de la BnF reste interrogeable seulement en connaissant l'URL du site recherché. La Wayback Machine d'Internet Archive ne fournissait pas non plus de recherche autre qu'une recherche par URL jusqu'à une période récente. Sa recherche par mots-clés comporte par ailleurs le biais de ne fouiller que les pages d'accueil des sites archivés.

Deux remarques s'imposent :

- la première est qu'il faut composer avec des archives en constante évolution, tant par leur mode d'archivage que d'interrogation. Les outils et fonctionnalités offerts par les organisations évoluent au cours même d'un projet et peuvent rendre caduques des méthodologies ou les faire évoluer. Ainsi notre projet Web90 a commencé en 2014 sans autre possibilité de consultation des archives des années 1990 que la recherche par URL (à part pour celles conservées à l'Ina, qui avaient déjà une recherche en plein texte). Quand, en 2016, la recherche en plein texte devient possible, aux heures passées à chercher des sites susceptibles de fournir des informations sur un sujet précis succède une quasi-instantanéité d'accès à des résultats plus variés et détaillés – sans toutefois faire disparaître les biais documentaires, puisque ces résultats comportent des choix introduits dans la conception du moteur de recherche.
- la seconde est le souci des institutions de valoriser ce patrimoine nativement numérique, de le rendre exploitable en fournissant des outils de fouille. Plusieurs éléments expliquent ce choix. Comme le note Thomas Drugeon (DL Web Ina), le chercheur ne peut pas partir avec les données, les sortir des enceintes des bibliothèques en France. Les outils d'analyse se doivent donc aussi d'être disponibles dans l'enceinte de consultation, et ils sont parfois nécessaires pour permettre la lisibilité de plusieurs milliers d'éléments (sites, pages, *hashtags,* etc.) ou les mettre en relation (par exemple au travers d'une recherche linguistique). Si les archives d'Internet Archive

sont en ligne et si on peut avoir le sentiment de pouvoir utiliser plus d'outils ou de les choisir, l'accès aux fichiers WARC[10] n'est pas acquis, et des contraintes techniques (mais aussi économiques) peuvent se poser.

L'évocation des fichiers WARC renvoie à des pratiques de traitement de données et métadonnées standardisées par le moyen d'outils informatiques (logiciels d'analyse lexicographique par exemple) qui s'apparentent à ce que Franco Moretti a qualifié de *distant reading* (lecture distante), proposant :

> « *What we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them*[11]. » (Moretti, 2013)

Loin de constituer la seule forme de lecture possible des archives du Web, la lecture distante permet toutefois dans le cadre de grands corpus d'avoir un aperçu que les capacités humaines de lecture ne permettent pas.

Les archives du Web passent ainsi sous le « macroscope historien » (Graham *et al.*, 2015). Des outils d'analyse en accès ouvert comme Iramuteq ou Gephi, ou développés par les institutions (pour produire par exemple des *timelines*, des diagrammes représentant les emojis ou images les plus tweetés dans les archives de l'Ina) permettent d'entrer dans les masses documentaires, par le contenu textuel, mais aussi par les images, les émoticônes ou encore les *hashtags* pour Twitter.

La lecture distante a été notamment utilisée pour la reconstruction de Geocities par Ian Milligan (2012-2017). Il a par exemple extrait des images afin de mesurer les promiscuités et récurrences visuelles au sein de ce service de pages personnelles particulièrement populaires dans les années 1990.

10. Le format WARC (Web ARChive), largement adopté depuis le milieu des années 2010, en remplacement de son prédécesseur, le format ARC, permet d'établir des standards en matière de collecte et de stockage des données hétérogènes présentes sur Internet. Pour plus de précisions, voir : http://www.bnf.fr/fr/professionnels/dlweb_boite_outils/a.dlweb_formats_fichiers.html.

11. « Ce qu'il nous faut, c'est un petit pacte avec le diable : nous savons comment lire les textes, apprenons maintenant comment ne pas les lire. » (Notre traduction.)

Les approches inspirées des *cultural* et des *visual studies* d'Anat Ben-David (reconstruction de noms de domaine disparus tel le .yu de l'ex-Yougoslavie, ou analyse de la couleur des domaines nationaux, voir Ben-David 2016 ; Ben-David *et al.*) contribuent également à apporter un nouveau souffle (et de la couleur) dans un paysage académique qui reste par ailleurs toujours très marqué par des approches linguistiques ou politiques, ce que relevaient déjà Dougherty *et al.* il y a quelques années (2010).

Outre le développement d'outils au sein du monde de la recherche, qui doit permettre aux chercheurs d'accéder à de plus en plus de boîtes à outils (voir par exemple The Archives Unleashed Project[12]), les bibliothèques ont également développé des plateformes de consultation, que ce soit la British Library, la BnF, la Bibliothèque royale du Danemark ou l'Ina. Elles sont susceptibles de prendre en charge l'outillage de la recherche à toutes les phases de celle-ci, depuis la recherche dans les fonds (recherche avancée, sélection par facettes de dates, noms de domaine, etc.), puis l'analyse (chronologies, graphiques, statistiques, représentations de tendances linguistiques sur le modèle de Google Ngram) jusqu'à la préservation, voire le partage du corpus.

Élargissant la thématique au-delà des archives du Web, pour considérer les données numériques susceptibles d'être analysées au sein de la bibliothèque de manière plus générale, la BnF a ainsi lancé une enquête prospective en 2017 pour préfigurer un nouveau service de fourniture de données à destination de la recherche, appelé provisoirement Laboratoire d'étude et d'analyse de corpus numériques (Moiraghi, 2018). Un autre exemple récent de ces efforts est fourni par la réalisation à la Bibliothèque royale danoise d'une nouvelle interface (voir « A wayback machine for the UKWA Solr based warc-indexer framework[13] ») incluant, de la recherche à la visualisation des résultats, de multiples fonctionnalités, type cartographie interactive de liens ou encore localisation des images et temporalités des collectes.

12. https://archivesunleashed.org.
13. Pour un descriptif et des captures d'écran de l'interface du projet, voir https://github.com/netarchivesuite/solrwayback.

La situation du chercheur en 2018 face aux archives du Web n'a ainsi plus rien à voir avec celle du début de la décennie. Reste que ces outils, s'ils peuvent simplifier la recherche, impliquent aussi de penser les biais et la couche de médiation supplémentaire qu'ils introduisent. Les travaux de Noortje Marres (2012, 2015), de Bernhard Rieder et Theo Röhle (2012) notamment, ont montré que le chercheur en sciences sociales doit conserver une distance critique face aux « présupposés épistémologiques contenus dans les outils » (Mabi, Plantin et Monnoyer-Smith, 2014).

On rappellera à cet égard les réflexions stimulantes d'Anat Ben-David et Hugo Huuderman sur les moteurs de recherche dédiés aux archives du Web (2014), ou de Megan Ankerson (2015a) sur les interfaces de consultation des archives du Web. Les outils d'analyse donnent également matière à réflexions méthodologiques, par exemple dans les travaux liés à la reconstruction de Geocities (Milligan, 2017) ou de domaines nationaux (Brügger, 2017a ; Brügger, Laursen et Nielsen, 2017).

Dans le panorama des différents outils d'exploitation des archives du Web, une situation particulière s'est présentée lors des collectes « d'urgence » qui ont suivi les attentats parisiens autour de *Charlie Hebdo* et ceux du 13 novembre 2015 : elle a amené la BnF et l'Ina à questionner leurs outils. En effet, si d'importants moyens techniques et humains ont été mis en œuvre lors de la collecte, la nécessité d'outils d'analyse performants s'est posée clairement face à ces collectes de grande ampleur.

La BnF a ainsi fait le choix de tester l'implémentation de la recherche en plein texte dans son corpus ; l'Ina a, de son côté, travaillé à fournir des outils, notamment de visualisation, pour exploiter les données et métadonnées du sien[14]. Les entretiens menés avec les porteurs de ces initiatives institutionnelles révèlent que celles-ci se trouvent souvent face à une tension :

> « dans la majorité des cas, les usagers qui viennent consulter un fonds du dépôt légal du Web le considèrent comme un fonds parmi d'autres au sein de

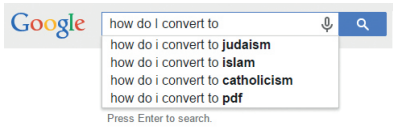
14. Parmi les fonctionnalités proposées : la possibilité de croiser plusieurs éléments tels des mots-dièses, mots-clés, statistiques de langues ou encore nombre de retweets.

> leurs recherches, ils ne vont pas dépenser une énergie énorme pour comprendre les limites. Mais certains vont chercher à aller plus loin. Nous sommes tiraillés entre ces besoins pointus et ceux de la majorité des usagers, pour lesquels il ne faut pas trop spécialiser l'outil, sinon il devient incompréhensible[15]. »

À n'en pas douter, en fournissant à la fois les données et les outils pour les exploiter, les institutions d'archivage assument un rôle central. Le chercheur se doit donc de déployer une vigilance et un effort pour comprendre à la fois les apports et biais des corpus, mais aussi ceux des outils fournis, en gardant à l'esprit que la neutralité des données comme celle des outils est illusoire (Plantin et Monnoyer-Smith, 2013). Dans le même temps, la mise en place de projets de recherche qui permettent aux chercheurs de signaler des URLs à archiver au moyen de l'outil BnF Collecte du Web, ou des ateliers du DL Web Ina, montre une attention aux besoins des chercheurs et aux contributions qu'ils peuvent apporter dans le cadre de l'exploitation du patrimoine numérique ; les institutions cherchent à penser leurs publics et saisir leurs demandes parfois très différentes.

Penser l'archive du Web en contexte

Figure 1- Mème circulant largement sur la Toile



15. Entretien avec Thomas Drugeon (responsable du DL Web à l'Ina), mené par V. Schafer et M. Borelli le 21 mars 2016 (https://asap.hypotheses.org/tag/ina).

Un mème[16] valant parfois mieux qu'un long discours, celui qui illustre ce début de section rappelle combien la prise en compte du contexte se révèle indispensable pour prétendre à une réelle compréhension de l'archive numérique.

Bien que vrai en soi puisque ce mème reproduit le résultat d'une recherche réellement effectuée et devenue virale, son propos ne l'est qu'à l'aune du rapprochement des différentes requêtes des utilisateurs fait par le moteur de recherche de Google.

L'archive du Web est issue d'un contexte global de production. La grande simplicité de la structure des sites des premières années du Web (des années 1990 au début des années 2000 dans la majorité des pays occidentaux) nous rappelle par exemple de quel poids pesaient encore les offres d'abonnement à la minute sur la consultation et par conséquent sur l'offre informationnelle proposée à l'internaute. Associée aux débits offerts par les modems de l'époque, cette structure des coûts de consultation explique en partie la faible profondeur des sites et la place marginale des images qui ne peuvent en conséquence être analysées hors de ces contraintes externes au Web lui-même.

Dans le contexte actuel, la production des contenus « générés par les utilisateurs » (*user-generated content*) est aussi influencée, dans une large mesure, par les dispositifs eux-mêmes qui récoltent, traitent et analysent ces données, invitant en outre à aimer, retweeter, etc. L'activité « dynamique » et automatique de nombre d'outils web, notamment de robots, doit également être prise en compte pour cerner la complexité du Web contemporain.

Si l'on pousse plus loin la prise en compte des agents techniques, ce qui apparaît à la surface de la page n'est que le rendu visuel d'un ensemble de codes informatiques. Ces derniers, à commencer par le .html, contiennent non seulement la trace des opérations de formatage des données et des logiciels, mais aussi des informations qui peuvent aller au-delà des paramètres techniques et relèvent des contextes de production.

16. Un mème internet est un élément de contenu (sous la forme de texte, image fixe ou animée, ou encore son, et selon des formats très divers) repris et décliné massivement sur la Toile (parfois transformé d'un format à un autre).

Il faut également faire une place aux contextes de réception des sources archivées. L'analyse quantitative de Twitter nous en donne un exemple saisissant. Comment juger de l'importance d'un tweet ou d'une série de tweets ? Faut-il l'analyser à l'aune de sa place dans l'espace de communication du réseau (ses retweets, ses likes, etc.) ? Selon quelle métrique ? Faut-il éventuellement s'ouvrir à une dimension plurimédiatique en soulignant que certains messages, du fait de la notoriété de leur auteur ou de son ancrage dans une communauté spécifique, connaissent un écho important hors du réseau lui-même (on pense en particulier aux relais que les journalistes offrent à certains messages dans un article, un journal télévisé, une émission de radio dont les printemps arabes ont été un exemple poussé jusqu'à l'absurde[17]) ? Impossible de décontextualiser totalement l'analyse pour faire du tweet un élément parmi d'autres. Bien entendu, la lecture distante propose une autre approche des corpus en faisant émerger des relations entre entités et groupes. Mais elle ne peut faire l'économie de la lecture attentive (close reading), sous peine de décontextualisation. Rendre compte d'un contexte global, ce n'est pas se tenir à distance, c'est rendre compte d'un va-et-vient entre les échelles de lecture et de compréhension d'un corpus.

Le contexte de réception est aussi fortement influencé par la structure en réseau du Web et de ses archives. La viralité des informations, leur reprise et leur modification entre sites et même entre pages est un élément d'appréciation de contexte important, comme le souligne Clément Oury (2012). Leur instabilité et leur volatilité en sont un autre, non négligeable. Une consultation des archives gagne à inclure une réflexion sur ce qui n'est pas archivé, ou ce qui risque de ne pas l'être. Une page web avec une série de liens vers des documents non archivés travaille la suggestion, l'évocation – voire la frustration du lecteur. Le chercheur doit travailler « en creux », multiplier les

17. Quand l'Occident relaie les contestations qui émergent à partir de fin 2010 en Tunisie, le rôle d'Internet et des réseaux sociaux fait l'objet d'analyses enthousiastes qui relèvent souvent du *solutionnisme* technologique (Morozov 2014), c'est-à-dire d'une pensée qui prête aux nouvelles technologies la capacité à résoudre tous les grands problèmes, de la faim dans le monde à la maladie. Les espoirs sont rapidement déçus questionnant l'impact réel des mobilisations en ligne (Bortzmeyer, 2016).

sources et ne pas s'en tenir à l'illusion d'une archive universelle et exhaustive. Au-delà des archives web, la presse spécialisée, des entretiens oraux ou les archives audiovisuelles livrent ainsi de multiples pistes pour reconstituer l'histoire du Web (Schafer, 2015).

Enfin, le contexte d'archivage informe sur le traitement donné à l'archive du Web. Une collecte n'est jamais une sauvegarde neutre des données : c'est une construction d'événements préjugés. Lorsqu'une collecte est décidée pour documenter un événement, une période ou un sujet d'intérêt, un ensemble de critères est mis en place pour sélectionner ce qui sera conservé. Comme l'archiviste le fait avec les masses de papier qui lui parviennent, sans en prendre exhaustivement connaissance, un tri est effectué a priori. Ainsi, il a été choisi par les institutions françaises d'archivage de poursuivre un objectif de représentativité et non d'exhaustivité en matière de conservation des sites des partis durant les campagnes électorales : les sites des petits partis aux extrémités du spectre politique sont conservés pour que l'ensemble soit représentatif des équilibres du spectre et non du poids respectif des formations en ligne.

Outre-Atlantique, d'autres considérations commencent à entrer en ligne de compte. Notamment sur le plan politique, certains acteurs entendent créer dans les corpus conservés une dimension « non oppressive », c'est-à-dire faire une place clairement identifiée et assumée à des groupes et des individus minoritaires au sein de la société et des flux de données en ligne. Le projet Documenting the Now[18] organise ainsi depuis 2016 une collecte des archives de Twitter selon des thématiques choisies en matière de genre, de critères « ethno-raciaux » anglo-saxons (dans le cadre entre autres du mouvement Black Lives Matter) ou de diversité culturelle.

Cette question des équilibres et de la représentativité est bien entendu critique pour les institutions en charge de la conservation, et pose des questions de fracture numérique, comme on a pu le montrer précédemment. Les faiblesses de la représentation en ligne des Suds (Gomes *et al.*, 2011) préoccupent l'IIPC. Les archivages des domaines .ao et .cv (angolais et cap-verdien)

18. https://www.docnow.io.

par les institutions portugaises en vertu de l'histoire coloniale du pays questionnent quant à eux les logiques éventuelles d'appropriation culturelle que risquent de faire émerger ces pratiques.

Si cette question de la bonne pratique en matière de construction des collections n'est pas tranchée et ne le sera probablement jamais de manière totalement satisfaisante, la participation des usagers des archives semble constituer une voie féconde d'amélioration. Cette association des chercheurs et des usagers au processus de collecte et aux règles qui la gouvernent se multiplie, à l'image des pratiques de la BnF et de l'Ina. Bien entendu, cette inclusion n'est pas nouvelle : de Michelet, chef de la section historique aux Archives nationales, à Jean-Noël Jeannenet, président de la Bibliothèque nationale de France, l'historien en particulier a toujours eu à cœur de participer aux politiques de conservation de son temps par la coconstruction des contextes d'archivage.

Les divers enjeux posés par les contextes de production, de réception et d'archivage illustrent la multiplicité des problématiques qu'il s'agit d'entrelacer au cœur des analyses qui prennent l'archive du Web comme support. Loin de bouleverser les règles traditionnelles de l'analyse, le contexte continue d'enrichir la compréhension du contemporain et, demain, d'un passé dont les traces sont d'ores et déjà lisibles en ligne.