Les archives du Web : gouvernance et identités

Francesca Musiani, Valérie Schafer

Les archives du Web : gouvernance et identités

Francesca MUSIANI
Valérie SCHAFER

Introduction

> « D'abord appréhendées comme supports de rechange ou simples outils de codage et de communication de contenus invariants, les technologies digitales réclament d'être désormais pensées comme un véritable écosystème, redéfinissant dans son ensemble la mémoire sociale et culturelle. […] Au lieu de ne voir le numérique que comme un autre moyen de conservation (plus ou moins performant), on verra alors qu'il implique une reconfiguration en profondeur des logiques, des instances et des pratiques mémorielles »[1].

Brewster Kahle et la fondation Internet Archive furent dès 1996 des pionniers de l'archivage du Web. En 2006, l'État français en fait, au titre du dépôt légal, une des missions de la Bibliothèque nationale de France (BnF) et de l'Institut national de l'Audiovisuel (Ina). Ces actions s'inscrivent dans une logique qui ne voit plus le numérique seulement comme un support de conservation patrimoniale mais comme un patrimoine à conserver : dès 2003 une charte de l'Unesco distinguait ainsi patrimoine numérisé et patrimoine nativement numérique (*Born Digital Heritage*)[2], ouvrant la voie à la reconnaissance institutionnelle de la valeur archivistique et patrimoniale de ces traces du passé.

[1] MERZEAU (Louise), « Faire mémoire des traces numériques », *e-dossiers de l'audiovisuel « Sciences humaines et sociales et patrimoine numérique »*, Ina expert, 2012 (http://www.ina-expert.com/e-dossier-de-l-audiovisuel-sciences-humaines-et-sociales-et-patrimoine-numerique/faire-memoire-des-traces-numeriques.html).

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

Or, si en France l'archivage du Web s'inscrit dans un cadre historique long – celui du dépôt légal[1] – il pose aussi, comme le relève Louise Merzeau, des questions spécifiques en termes de sélection, de collecte et d'exploitation. Celles-ci tiennent notamment à l'étroite imbrication entre les « propriétés techniques, symboliques et sociopolitiques »[2] non seulement de la traçabilité des données, mais aussi des processus d'archivage qui leur sont appliqués. Cet article explore donc les réalités de l'archivage du Web, en mêlant ces différentes dimensions, ainsi que l'impact qu'a l'identité des organisations sur les formes de gouvernance de l'archivage du Web et l'identité des fonds et collections. Après une vision d'ensemble, il prendra pour point d'appui l'archivage d'urgence du Web et en particulier de Twitter, au moment des attentats parisiens de janvier et novembre 2015[3].

Les archivages du Web

Plutôt que d'un archivage du Web, on peut parler « des archivages du Web ». De même que les productions sur la Toile mobilisent une diversité d'acteurs, leur conservation met en jeu plusieurs institutions, acteurs humains et techniques. Ceux-ci relèvent à la fois du monde des archives et bibliothèques – qui se sont vu confier la responsabilité d'une partie des traces numériques du passé – et du secteur de l'ingénierie, de la recherche, des entreprises ; ou encore, ils prennent forme à partir d'initiatives de la société civile. Entre ces instances et communautés existent des circulations et concertations, dont témoigne par exemple la large adoption par Internet Archive, puis par le monde des bibliothèques, du robot de collecte Heritrix ou encore, les réunions de l'IIPC (*International Internet Preservation Consortium*). Cependant, à la segmentation des archives – qui conduit aujourd'hui à une mosaïque d'archives européennes, sans passerelle entre les fonds[4] – répond celle des périmètres et cadres juridiques, institutionnels ou encore des modalités d'accès.

---

[1] COHEN (Évelyne) et VERLAINE (Julie), « Le dépôt légal de l'Internet français à la Bibliothèque nationale de France », *Sociétés & Représentations,* 2013, 1(35), p. 209-218.
MUSSOU (Claude), « Et le Web devint archive : enjeux et défis », *Le Temps des médias,* 2012, 2(19), p. 259-266.

[2] MERZEAU (Louise), *ibid.*

[3] Nous nous appuierons ici sur les premiers résultats du projet ASAP (Archives Sauvegarde Attentats Paris) financé par le CNRS et mené en partenariat avec l'Ina et la BnF au cours de 2016 (http://asap.hypotheses.org/a-propos).

[4] C'est toutefois l'ambition du groupe européen de recherche RESAW, créé en 2012 à l'initiative de Niels Brügger (Aarhus University, NetLab & the Center for Internet Studies, Danemark) que de réfléchir à la possibilité de développer ces passerelles (http://resaw.eu/about/).

Des pratiques diversifiées

La France n'est pas la seule à être confrontée à la répartition de son archivage entre plusieurs institutions ; c'est notamment le cas au Royaume-Uni entre la UK Web Archive[1] de la British Library et la UK Government Web Archive[2] des Archives nationales. L'exemple français permet toutefois d'illustrer la diversité des archivages du Web, en distinguant des variations entre les collections de la BnF et de l'Ina.

Les différences portent d'abord sur la nature et le périmètre même des collections. L'Ina s'est vu confier, dans le cadre de la loi DADVSI[3], une mission de conservation des contenus qui relèvent de l'audiovisuel, tandis que la BnF prend en charge « le reste » d'un ensemble français. Celui-ci ne se limite pas au .fr, mais prend aussi en considération des extensions territoriales (par exemple le .re) ainsi que des contenus produits par des Français ou des auteurs domiciliés en France, dont les adresses sont en .com, .org., etc. Près de 4,5 millions de sites sont ainsi collectés par la BnF chaque année, en se fondant sur les listes de l'Association française pour le nommage Internet en coopération (Afnic), d'OVH et de plusieurs registres de domaine, ainsi que de l'Office des postes et télécommunications de Nouvelle-Calédonie. La collecte de masse ne permet de garder qu'une fois par an ces sites. Cependant, à cet échantillon large qui se veut représentatif de la diversité des sites, ne discriminant pas entre productions institutionnelles, scientifiques ou personnelles par exemple, s'ajoutent des collectes plus régulières : ciblées sur 20 000 sites, elles sont journalières dans des cas spécifiques, par exemple pour les sites des organes de presse[4]. Si l'ampleur de son périmètre implique de la part de la BnF une collecte à la récurrence variable, l'Ina peut appréhender l'archivage du Web différemment, en vertu de son périmètre plus spécialisé et restreint (environ 14 000 sites). Ce nombre nettement plus réduit lui permet d'effectuer des collectes plus régulières, qui peuvent rejoindre en cela la fréquence de celles dédiées par la BnF aux organes de presse, pour des contenus par ailleurs sujets à de nombreuses modifications, au fil de l'actualité et de la journée. Depuis février 2014 l'Ina archive par ailleurs les comptes de 12 000 utilisateurs liés au monde de l'audiovisuel et 400 mots-dièse. Si elle est loin derrière la *Library of Congress* en termes d'archivage de Twitter – la bibliothèque américaine ayant passé un accord avec Twitter[5] en 2010 – sa collecte des contenus

[1] http://www.Webarchive.org.uk/ukwa/info/about#what_uk_archive
[2] http://www.nationalarchives.gov.uk/Webarchive/
[3] Loi relative au droit d'auteur et et aux droits voisins dans la société de l'information.
[4] http://www.bnf.fr/fr/collections_et_services/livre_presse_medias/a.archives_internet.html
[5] ZIMMER (Michael), « *The Twitter Archive at the Library of Congress* : *Challenges for information*

audiovisuels liés au domaine français va bien au-delà de celle disponible *via* la *Wayback Machine* d'Internet Archive. L'institut a par ailleurs fait le choix d'un autre robot de collecte qu'Heritrix, pourtant largement répandu dans le monde de l'archivage du Web, et a mis au point des technologies spécifiques qui lui ont semblé de nature à mieux capturer ces sites, incluant par leur nature de nombreux contenus audio et vidéos et reposant sur une fluidité et une circulation transmédiatique intense. Le développement d'un outil permettant aujourd'hui de faire défiler en parallèle des archives audiovisuelles et les réactions archivées qui lui sont associées sur Twitter témoigne de la volonté de l'Ina de prendre en compte l'hybridation croissante entre les différents supports et temporalités médiatiques.

Cette diversité des approches entre BnF et Ina se retrouve à l'échelle européenne : les nuances peuvent porter sur le périmètre (de l'archivage très inclusif des Britanniques au périmètre plus ciblé et sélectif des Suisses[1]). Elles concernent aussi le respect ou non des exclusions signalées par les robots.txt[2]. Les modes d'accès diffèrent également – dans les emprises de la BnF, de l'Ina ou de bibliothèques en région ; en accès libre en ligne pour les archives portugaises sur arquivo.pt[3].

Une influence réciproque entre gouvernance et identité

Si le nombre ou les caractéristiques des documents nativement numériques conservés peuvent inviter à adapter les outils et méthodes employés pour leur collecte, l'influence des périmètres n'est pas seule en cause dans les choix effectués. L'identité de l'institution se révèle notamment dans les interfaces de consultation.

Megan Sapnar Ankerson insiste, par exemple, sur l'influence qu'a eue la Wayback Machine sur la manière dont l'usager appréhende le Web du passé et ses archives[4] (et sur la façon dont les institutions ont également, après elle, développé leurs propres interfaces).

---

practice and information policy », *First Monday*, Volume 20, 7-6 July 2015 (http://firstmonday.org/ojs/index.php/fm/article/view/5619/4653).

[1] SCHAFER (Valérie), MUSIANI (Francesca), BORELLI (Marguerite), « *Negotiating the Web of the Past* », *French Journal for Media Research* [en ligne], 6/2016, La toile négociée (http://frenchjournalformediaresearch.com/lodel/index.php?id=952).

[2] Si le cadre du dépôt légal entraîne des contraintes, notamment en termes d'accès aux archives, la *British Library* depuis l'entrée en vigueur du *Legal Deposit Act*, le 6 avril 2013 n'a plus besoin de demander aux ayants droit l'autorisation de collecter les sites. En vertu du DL Web, la BnF collecte par ailleurs des sites comme lemonde.fr ; ce n'est pas le cas d'Internet Archive qui respecte les robots.txt.

[3] Pour un aperçu des différentes initiatives, depuis celles pionnières de la Croatie ou de la Norvège, voir : http://www.netpreserve.org/legal-deposit

[4] ANKERSON (Megan S.), « *Take me back! Web history as chronotourism of the digital archive* », *Times and Temporalities of the Web International Symposium*, Paris, 2015.

Complémentaire de cette analyse, l'étude « biographique » de l'Internet Movie Database (IMDb) par Fernando van der Vliest et Tessa de Keijser montre que la liste de diffusion de l'IMDb, ainsi qu'elle a été conservée par Internet Archive, révèle nombre de « relocations » et d'évolutions dans sa conception qui « construisent un utilisateur duquel on attend de la production d'ordre dans un ensemble d'éléments *a priori* indifférenciés »[1].

La nouvelle interface de la Wayback Machine, en version bêta début 2017, propose une entrée différente et novatrice. Elle bénéficie à la fois de l'implémentation de la recherche *plein text* dans les pages d'accueil des sites et d'un mode de présentation en rupture, qui met désormais aussi en valeur des métadonnées de nature à documenter l'archive[2].

Le changement de son interface de consultation par l'Ina au premier semestre 2016 ne relève pas non plus de seules questions cosmétiques ou ergonomiques. Comme l'explique Zeynep Pehlivan, ingénieure de recherche au sein du DL Web de l'Ina :

> « Notre choix pour le menu vertical ne se fonde pas seulement sur des préoccupations ergonomiques, mais également sur les demandes de développements futurs. Notre objectif à moyen terme est de proposer une plateforme de recherche uniforme, sans distinguer les différents types d'objets archivés, en utilisant le menu de gauche comme un *dashboard*. Par exemple, si un utilisateur recherche une URL spécifique, on souhaiterait lui fournir non seulement la possibilité de naviguer vers cette URL, mais aussi les métadonnées correspondantes, la liste des vidéos qu'on peut trouver sur cette page et la liste des *tweet* qui pointent vers cette URL »[3].

La culture des institutions peut également avoir un impact sur les fonds et leur accessibilité : le cas portugais est à cet égard révélateur car l'archivage du Web est lié au réseau de la recherche de ce pays et à une culture plus informatique que patrimoniale. La mise en ligne en libre accès sur arquivo.pt relève d'une culture ouverte proche de celle d'Internet Archive et a notamment permis l'introduction dans sa nouvelle interface de liens et de passerelles vers d'autres archives ouvertes *via* le protocole Memento[4].

---

[1] VAN DER VLIST (Fernando N.), DE KEIJSER (Tessa), « *Reconstructing Web History. Tracing the Conceptual Trajectory of the List on the Internet Movie Database* », Research Report (https://www.fernandovandervlist.nl/papers/reconstructing-Web-history.html).

[2] Voir par exemple la réponse de la Wayback Machine à une requête portant sur CNRS : https://web-beta.archive.org/web/*/cnrs
et le sommaire du site archivé du CNRS: https://web-beta.archive.org/details/http://cnrs.fr

[3] Échange électronique avec Zeynep Pehlivan du 11 février 2016.

[4] http://sobre.arquivo.pt/news/arquivo.pt-supports-memento?set_language=en

Enfin, si l'institution et ses modes de gouvernance internes ont un impact sur les collections, les fonds à collecter peuvent également influencer les manières de procéder. L'organisation Archive-It s'est notamment spécialisée, depuis 2007 et la fusillade sur le campus de Virginia Tech[1], dans la collecte de crise et d'urgence, à la faveur de sauvegardes portant notamment sur le conflit ukrainien, les « printemps arabes » ou encore le tremblement de terre d'Haïti de 2010[2]. Archive-It fait appel à son réseau international d'institutions et de volontaires pour cibler des contenus pertinents. C'est ainsi qu'il a été procédé au moment des attentats contre Charlie Hebdo, la BnF ayant remonté ses URL collectées à Archive-It, lui permettant de repérer les sites intéressants à collecter dans un contexte à la fois d'urgence et de terrain moins familier aux Américains qu'à la BnF. Le cas de l'archivage des attentats parisiens, documenté au sein du projet ASAP, permet par ailleurs, à partir d'une étude de cas concrète, de mesurer l'imbrication entre gouvernance et identité dans le cadre des collections réalisées.

Archivage en tension : le cas des attentats parisiens

Il convient de rappeler que l'archivage d'urgence n'est pas une spécificité de l'archivage du patrimoine nativement numérique[3], que la collecte des éphémères a une longue tradition[4] et a par ailleurs après les attentats porté aussi sur des documents papiers[5]. Il faut également prendre en considération les précédents propres au patrimoine nativement numérique, qu'ils soient exogènes aux institutions que nous étudions (par exemple *The September 11*

Memento introduit la notion de *datetime negotiation* et permet de demander une version d'une ressource telle qu'elle existait à un moment spécifique dans le passé. Voir http://www.mementoweb.org/about/

[1] Voir l'entretien mené avec Jefferson Bailey et Sylvie Rollason-Cass par Valérie Schafer et Marguerite Borelli (11-17 mars 2016) : https://asap.hypotheses.org/125

[2] https://archive-it.org/collections/1784

[3] Agnès Magnien (directrice déléguée aux collections, Ina) se souvient par exemple de la collecte en temps réel des archives d'une politique publique, le « fonds social d'urgence » dans les années 1990 (échange électronique du 12 mars 2016).

[4] HAGE (Julien), « Les éphémères de l'âge de l'imprimé à l'ère numérique. Un champ disciplinaire en révolution », *Fabula*/ Les colloques, Les éphémères, un patrimoine à construire, publié le 8 novembre 2015 (http://www.fabula.org/colloques/document2925.php).

[5] Notamment par les archives de Paris.

Digital Archive[1]) ou qu'ils aient pu être menés au sein de celles-ci. Ainsi Annick Le Follic rappelle-t-elle qu'une réflexion sur la collecte d'urgence a été menée au sein de la BnF en amont de janvier 2015, visant à la fois des événements ponctuels prévus (comme les élections) et inattendus[2]. Les attentats de janvier et novembre 2015 n'en restent pas moins profondément disruptifs et ont provoqué d'intenses « vibrations "en temps réel" »[3]. Les deux institutions ont alors choisi de lancer une collecte d'urgence au sein de laquelle les méthodes et choix effectués révèlent des spécificités.

De la collecte d'urgence …

Un point notable réside dans l'appréhension différente de la collecte de Twitter. La BnF passe, comme pour ses autres archivages du Web, par le robot Heritrix, tout en notant les limites que peut poser celui-ci pour capturer les RSN. Elle a ainsi

> « testé deux formules : à partir du compte ou du *hashtag* […], le robot s'identifie soit comme navigateur Web, soit comme appareil mobile, puis il archive le premier affichage (pas de déroulé dynamique du "fil"), ce qui représente 20 *tweets*. En pratique, cela donne 4 *screenshots* par jour pour un compte ou un *hashtag* suivi.»[4].

Aux captures qui suivent le déroulé d'une discussion et cherchent à présenter une reproduction fidèle du fil, répondent un mode de collecte et, dès lors, une collection très différents côté Ina : la capture passe en effet par l'API publique de Twitter et archive les tweets plutôt que les pages. « Si on archive les pages twitter, il faut réextraire les posts ensuite, etc. Pour nous cela n'a pas d'intérêt, ce qui nous intéresse ce sont les "données brutes", un tweet c'est plus que 140 caractères, c'est aussi des informations sur la date, l'émetteur, des images, etc. Par contre on n'a pas capté les images de fond. On stocke une structure de données arborescente », expliquait ainsi Thomas Drugeon, responsable du DL Web à l'Ina[5].

---

[1] http://911digitalarchive.org

[2] Entretien avec Annick Le Follic (chargée de collections numériques au département du dépôt légal de la BnF), mené par M. Borelli et V. Schafer le 21 mars 2016.

[3] BOULLIER (Dominique), « Charlie est un phénomène de 3e génération (aussi) », *SHS 3G*, le 1er juin 2015 (http://shs3g.hypotheses.org/114)

[4] Entretien avec Annick Le Follic, *ibid.*

[5] 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).

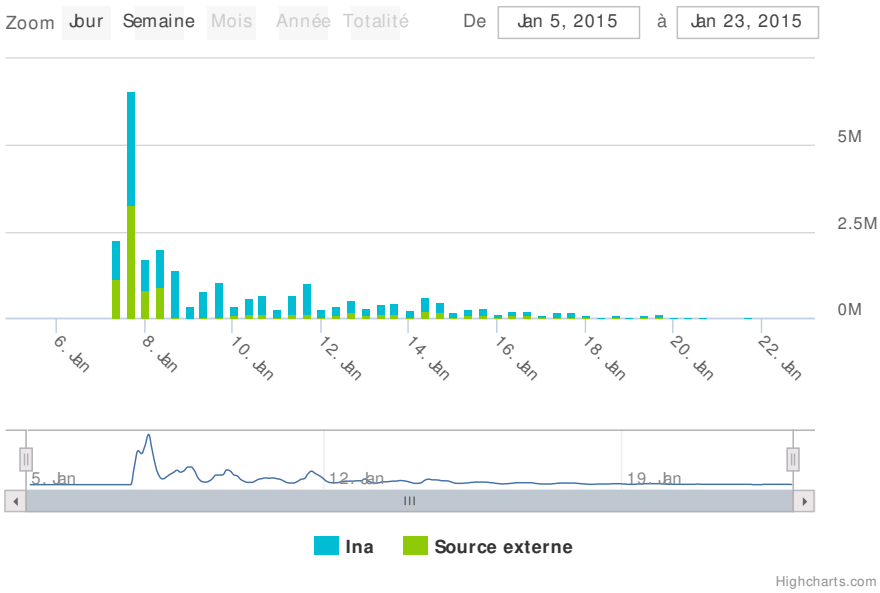
Outre cette différence de collecte, mais aussi de conception de ce que peut être l'archivage de Twitter, des éléments organisationnels et de gouvernance propres aux institutions ont joué au moment de l'archivage : ainsi la BnF a-t-elle choisi de demander à son large réseau de correspondants de remonter des contenus qui pouvaient sembler pertinents au moment des événements survenus dans les locaux de *Charlie Hebdo*, tout en s'interrogeant sur son périmètre et les limites de son archivage. Côté Ina, on reconnaît une moindre réactivité en janvier qu'en novembre 2015 : une opération de maintenance sur le réseau électrique de la salle serveur de l'Ina a entraîné une délocalisation des processus de captation qui a permis à Thomas Drugeon de lancer la collecte Twitter le soir même du 13 novembre – circonstances exceptionnelles doublées de l'expérience, puisqu'en janvier la collecte avait été lancée après 24 heures, passant à côté du pic principal de tweets.

L'organisation différente des modes d'expression en ligne – les mots-dièse étant davantage concentrés en janvier, en particulier sur #jesuischarlie – a par ailleurs un impact sur la collection[1]. C'est en temps réel que l'archiviste doit alors identifier les mots-dièse qui lui semblent représentatifs, quitte à en ajouter au fil des vibrations – ces décisions prises au cœur des événements pouvant amener aussi à négliger des contenus qui pourraient s'avérer ensuite utiles aux chercheurs (voir par exemple l'analyse du #jenesuispasCharlie[2]). Face à des lacunes, même si l'archivage de Twitter ne visait pas à l'exhaustivité mais à la représentativité, l'Ina a par ailleurs fait le choix de compléter sa collecte dédiée aux événements à partir des ID relevés et mis à disposition en ligne par d'autres collecteurs, par exemple Nick Ruest (*Digital Assets Librarian* à la *York University*), tout en effectuant sa propre captation pour garantir l'authenticité des *tweets* et l'homogénéité de la collecte[3].

[1] *Idem*.

[2] BADOUARD (Romain), « "Je ne suis pas Charlie". Pluralité des prises de parole sur le Web et les réseaux sociaux », dans LEFÉBURE (Pierre) et SÉCAIL (Claire), *Le défi Charlie. Les médias à l'épreuve des attentats*, Paris, Lemieux Éditeur, 2016, p. 187-221.

[3] http://ruebot.net/post/exploratory-look-13968293-jesuischarlie-jesuisahmed-jesuisjuif-and-charliehebdo-tweets



Distinction entre sources internes et externes dans l'exploration des *tweets* liés aux événements terroristes de janvier 2015 © Ina

Du côté de la BnF, la conscience de lacunes et des difficultés de gestion de l'urgence est également présente :

> « Pour les attentats de novembre qui ont eu lieu un vendredi soir notamment, nous avons dû attendre le lundi matin avant de lancer la collecte, nos outils étant seulement accessibles sur place pour des raisons de sécurité. Cet impératif de réactivité pose aussi la question suivante : jusqu'où veut-on et peut-on aller dans la disponibilité humaine pour lancer une collecte d'urgence ? »[1].

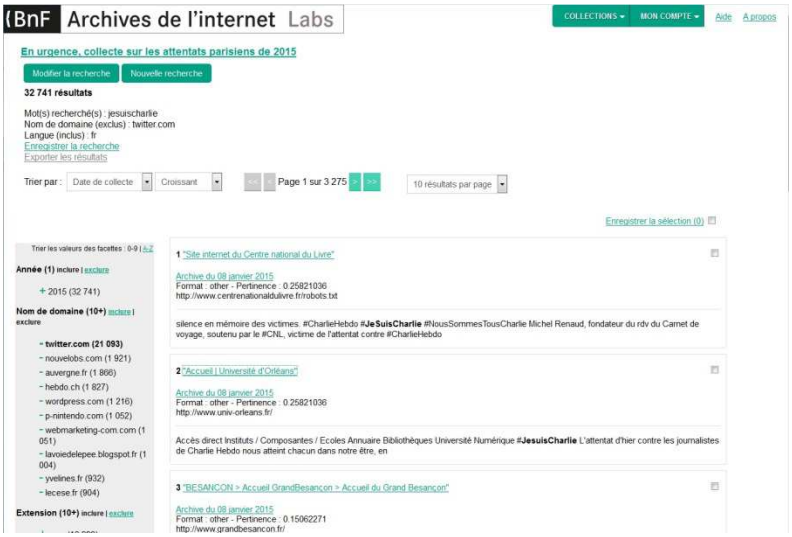
Alors que l'Ina a été plus réactif en novembre qu'en janvier 2015, la BnF a dû, elle, davantage composer avec des contraintes internes :

> « Pour les attentats contre *Charlie Hebdo*, comme nous étions en période creuse dans nos activités de collecte, nous avons pu mobiliser beaucoup de moyens humains et techniques. […] En novembre, le contexte était différent. Premièrement, nous avions déjà plusieurs chantiers en cours (collecte large annuelle, collectes spécifiques pour les élections régionales, sur la COP21 et sur les réfugiés) et étions donc moins disponibles, en termes de moyens techniques et humains »[2].

[1] Entretien avec Annick Le Follic, *ibid.*
[2] *Idem.*

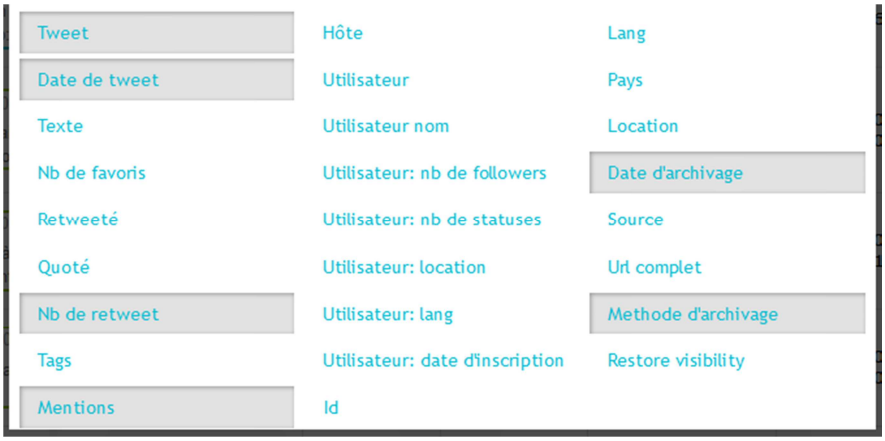
… à son exploitation

Aux moyens techniques et humains à mettre en œuvre lors de la collecte répondent ensuite des besoins, également techniques et humains, pour la mise à disposition du *corpus* et son exploitation. Face à ces collectes de grande ampleur (9 millions d'URL collectés par l'Ina au moment des événements de Charlie Hebdo et 11 millions en novembre 2015 ; du 8 au 16 janvier 2015, 1 581 domaines différents archivés par la BnF et 15,9 millions d'URL) le besoin d'outils d'analyse se pose avec acuité. La BnF a ainsi fait le choix de tester l'implémentation de la recherche *plein text* dans son fonds.



Recherche *plein text* et possibilité d'affiner les résultats à l'aide de facettes © BnF

L'Ina a quant à lui travaillé à fournir des outils, notamment de visualisation, pour exploiter les données et métadonnées du sien. La possibilité de croiser plusieurs éléments tels des mots-dièse, mots-clé, statistiques de langues, nombre de retweets, fait partie des multiples fonctionnalités proposées pour approcher ces masses de données.



Possibilité de croiser les données et métadonnées au cours de l'exploration des tweets et mots-dièse dans l'interface Ina. © Ina

Ces initiatives des institutions sont aussi une réponse au cadre juridique du dépôt légal, comme le relève Thomas Drugeon :

> « L'utilisateur, le chercheur ne peut pas "partir" avec les données du DL Web, les sortir, aussi nos outils doivent répondre à ses besoins, pour qu'avec les outils que nous proposons il puisse faire des analyses pertinentes. Bien sûr il faut comprendre quels sont les vrais besoins des chercheurs et nous devons faire face à un double souci : dans la majorité des cas, les usagers qui viennent consulter un fonds du DL Web le considèrent comme un fonds parmi d'autres au sein de 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... Il y a presque autant de besoins et d'outils que de recherches et de chercheurs »[1].

En fournissant à la fois les données et les outils pour les exploiter, les institutions d'archivage assument dès lors un rôle central qui implique de la part du chercheur une vigilance et un effort pour comprendre à la fois les apports et biais des *corpus*, mais aussi ceux des outils fournis. Si les bibliothèques et le monde de la documentation ont su s'emparer de manière pionnière d'un certain nombre d'outils des humanités numériques[2], il est du

[1] Entretien avec Thomas Drugeon, *ibid.*

[2] CARACO (Benjamin), « Les digital humanities et les bibliothèques », *Bulletin des bibliothèques de France (BBF)*, n° 2, 2012, p. 69-73 (http://bbf.enssib.fr/consulter/bbf-2012-02-0069-002).

devoir des chercheurs de se saisir de ceux-ci[1] avec la conscience que la neutralité des données comme celle des outils est illusoire[2]. Dans le même temps, la mise en place de projets de recherche qui permettent aux chercheurs de remonter des URL au moyen de l'outil BnF Collecte du Web, ou les ateliers du DL Web Ina (dont un Lab a été consacré au premier semestre 2016 à l'archivage de Twitter au cours des attentats de 2015[3]) montrent 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, ce qui à terme influence également l'identité des fonds et les modes de gouvernance.

Conclusion

Cet article a montré comment la conception des infrastructures d'archivage du Web, ainsi que les choix des contenus archivés, contribuent à définir le périmètre et la nature même des archives Web comme patrimoine numérique. Il s'est particulièrement concentré sur la relation entre l'identité des institutions et organisations qui s'occupent de la collecte et de l'archive, et les formes de gouvernance de l'archivage du Web ; pour ce faire, il s'est appuyé sur l'étude en cours de l'archivage d'urgence du Web et des réseaux socio-numériques, en particulier Twitter, au moment des attentats parisiens de janvier et novembre 2015.

La gouvernance des archives Web, véritable microcosme de questions plus larges de gouvernance de l'Internet, s'explicite en effet à la fois dans les infrastructures et les artefacts techniques (les interfaces, les bases de données, les méthodes de collecte) et dans les identités des institutions, les missions dont elles sont investies et les traditions culturelles dont elles proviennent, ainsi que leurs « relations de pouvoir ». Ces deux niveaux sont d'ailleurs très étroitement liés, comme les approches inspirées des *Science and Technology Studies* (STS) contribuent à l'éclairer. L'organisation humaine et technique des institutions n'est pas dissociable ; l'ouverture plus ou moins importante des modèles

[1] PLANTIN (Jean-Christophe), MONNOYER-SMITH (Laurence), « Ouvrir la boîte à outils de la recherche numérique. Trois cas de redistribution de méthodes », *tic&société*, vol. 7, n° 2 (https://ticetsociete.revues.org/1527).

[2] GITELMAN (Lisa, dir.), *Raw Data Is An Oxymoron*, Cambridge, MA, The MIT Press, 2013.

[3] MUSSOU (Claude), « Retour sur la séance du Lab Ina DL Web #0 », Blog Ateliers - Dépôt légal du Web, 8 avril 2016 (http://atelier-dlWeb.fr/blog/?p=1677).

d'archivage répond à des besoins techniques mais également à une vision d'ensemble de ce qu'être un acteur de l'archivage veut dire ; les interfaces et les choix de conception qui les sous-tendent produisent des publics et en sont informées en retour.

Les moments de tension et d'épreuve, comme ont pu l'être les « collectes d'urgence » lors des attentats parisiens, contribuent à « ouvrir les boîtes noires » de l'archivage : ils permettent d'éclairer les manières dont les visions que sous-tendent les institutions d'archivage, celles de leur gouvernance, de leurs missions, du Web lui-même – passé et présent – s'entremêlent et ne peuvent qu'être étudiées ensemble en un écosystème complexe mais cohérent.

Francesca MUSIANI
Chargée de recherche ISCC (CNRS/Paris-Sorbonne/UPMC)
francesca.musiani@cnrs.fr

Valérie SCHAFER
Chargée de recherche ISCC (CNRS/Paris-Sorbonne/UPMC)
valerie.schafer@cnrs.fr