La notion de source et sa redéfinition à l'ère du numérique

Céline Guyon

RENCONTRE ACADÉMIQUE ÉDUCATION AUX MÉDIAS ET À L'INFORMATION

Organisateur : Délégation académique au numérique pour l'éducation de Lyon

Date et horaire : 25/09/2020

La notion de source et sa redéfinition à l'ère du numérique, **Céline Guyon**, archiviste, maîtresse de conférences associée à l'Enssib, co-responsable du Master 2 ARN - Archives numériques

Il m'a été proposée, par les organisateurs de cette journée, que je remercie, de discuter des sources numériques et de leurs enjeux.

Je vous propose d'en dessiner le panorama, sans prétendre, loin s'en faut, à l'exhaustivité d'autant plus que ce panorama affiche un parti pris, celui du regard d'une archiviste sur ces questions.

La question des sources numériques a été débattu par les SHS, les Sciences de l'information et les professionnels de la gestion de l'information que sont les archivistes, bibliothécaires et documentalistes, tant sur les plans méthodologiques qu'épistémologiques.

S'agissant des archives, les technologies informatiques et le numérique ont transformé les pratiques archivistiques et nous amènent à repenser la définition même des archives.

Le numérique n'est plus seulement un outil mais véritablement un éco système dans lequel nous évoluons et qui reconditionne la manière de produire, d'échanger, de rechercher, de consulter l'information. Cette information numérique est multiforme, protéiforme, massive.

On distingue classiquement deux familles de sources numériques, les premières existent seulement sous leur forme numérique initiale alors que les secondes sont issues d'un dispositif de numérisation de documents existants. Les premières sont qualifiées de sources nativement numériques ou de sources d'origine numérique. Cette distinction a été opérée en 2003 par l'Unesco dans une charte sur la conservation du patrimoine numérique. Cette charte, en précisant que les sources nativement numériques font partie intégrante du patrimoine de l'humanité a ouvert la voie à leur patrimonialisation, c'est-à-dire à la reconnaissance de leur valeur archivistique et mémorielle et donc à la nécessité de leur archivage[1].

Les sources numériques ont, quels que soient leur forme (un fichier informatique, une vidéo YouTube, une base de données, un plan en 3D, une page Web) une caractéristique commune, celle de ne pas être lisibles en dehors des dispositifs techniques dans lesquels elles ont été créées ou sont conservées. Par dispositif technique, il faut entendre aussi bien les infrastructures matérielles de stockage (disque dur ou serveur par ex.) que les dispositifs de lecture.

De ce point de vue, une source numérique peut être vue comme un millefeuille composé de multiples couches qui, assemblées ensemble permettent à l'œil humain de visualiser une information, c'est-à-dire de transformer des 0 et des 1 en un message lisible.

La notion de sources renvoie quant à elle à un ensemble de ressources informationnelles qu'elles soient d'ordre culturel, éducatif ou administratif qui contiennent des informations techniques, juridiques, médicales, personnelles.

---

[1] Voir aussi Camille Paloque -Berges, « Les sources nativement numériques pour les sciences humaines et sociales », Histoire@Politique, [en ligne], n° 30, septembre -décembre 2016, www.histoire -politique.fr

Pour illustrer les enjeux liés aux sources numériques je vous propose de questionner 3 axes ; la fiabilité des sources numériques, les données d'usage et l'archivage du Web.

La question de la fiabilité de l'information numérique renvoie à l'instabilité documentaire des sources numériques qui ne bénéficient plus de la fixation du support papier. Il existe par exemple autant de représentations graphiques différentes d'un Tweet « n'incluant pas forcément les mêmes informations » que de terminaux et logiciels utilisés pour les consulter. « Comment faire confiance à un document qui ne cesse d'être en mouvement » ?[2]

Comment dépasser ce que Bruno Bachimont appelle une « forme de naïveté épistémologique » à savoir qu'on a tendance à croire ce qu'on voit du point de vue de l'utilisateur, du chercheur ? Un Tweet par exemple se compose de beaucoup plus de données que celles visibles. Les interfaces de programmation mises à disposition des développeurs permettent en effet d'accéder, pour chaque tweet, à une trentaine d'attributs ou métadonnées.

Notre culture de l'écrit, héritée de l'imprimerie (qui autorise la reproduction à l'identique de multiple copies) s'est construite sur la stabilité de l'écrit assurée par le support. « L'intégrité de l'information que le document porte dépend de celle de son support »[3]: si le support est demeuré intègre dans le temps, on en déduit que l'information qui y est inscrite n'a pas été modifiée et qu'elle est donc complète.

Le numérique (mais déjà avant lui déjà l'audiovisuel) a introduit de ce point de vue un changement de paradigme, en ce sens que l'information et le support matériel sur lequel elle est inscrite sont désormais dissociés. La conservation physique du support ne garantit plus de pouvoir accéder et prendre connaissance de l'information enregistrée sur ce support. Préserver le support physique ne suffit plus pour préserver l'accessibilité aux informations.

Par ailleurs, ce que l'on consulte ne correspond pas à ce qui a été créé ou qui est conservé. Ce que l'on consulte n'est pas ce qui est conservé mais ce qui est reconstruit à partir d'une source préservée. « Le document et son contexte doivent pouvoir être représentés dans un nouvel environnement culturel, social et technologique, un peu comme une composition musicale à partir d'une partition »[4]. Cette recomposition est rendue inexorable par le rythme des évolutions technologiques et son corollaire, l'obsolescence.

Le numérique nous amène, nous dit Bruno Bachimont que je cite, à renouer avec « une économie de la variante où un contenu n'existe qu'à travers des exemplaires qui donnent des versions différentes mais voisines sans qu'aucune ne soit la référence canonique permettant de juger les autres »[5]. Le numérique fait en quelque sorte renaître une certaine culture du manuscrit, avec un contenu toujours en mouvement.

Dans ces conditions, quelle confiance placer, en tant que lecteur, dans l'artefact qui s'affiche à l'écran ? La confiance que nous plaçons dans les sources numériques est, de fait, liée à la fiabilité de leurs conditions de transmission et donc aux conditions de leur conservation.

Les données d'usage

---

[2] Antonin Segault, « Documenter Twitter : défis et méthodes pour la constitution de corpus de tweets », *Balisages* [En ligne], 1 | 2020, mis en ligne le 24 février 2020, consulté le 03 janvier 2024. URL : https://publications-prairial.fr/balisages/index.php?id=280

[3] Robert, P. (2010). *Mnémotechnologies*. Paris, France : Lavoisier

[4] Banat-Berger, F. (2010). Les archives et la révolution numérique. *Le Débat*, 158(1), 70-82.

[5] Bachimont, B. (2017). Patrimoine et numérique : technique et politique de la mémoire. Bry-sur-Marne, France : INA

On use de nombreux qualificatifs pour décrire les données : données publiques, données d'intérêt général, données massives ou big data, données ouvertes, données à caractère personnel, données, lisibles par machine, données primaires, données enrichies, données de référence, données d'usage.

Les données ont souvent été assimilées à une ressource naturelle et qualifiées de nouvel « or noir ». Ce rapprochement entre données et ressources naturelles laisse penser que les données existent à l'état brut. Or, la sociologie, on pensera à l'ethnographie des laboratoires Bruno Latour[6] ou plus récemment à l'ouvrage de Jérôme Denis sur le travail invisible des données[7] ou l'exemple de l'Open data, a rappelé que les données sont le produit d'un travail effectué par des acteurs situés, et non des machines autonomes ; pour reprendre la célèbre formule de Bruno Latour : « il n'y a pas de données mais des obtenues ». C'est-à-dire qu'elles nécessitent un travail souvent invisible.

Les données d'usage sont les données laissées implicitement ou explicitement par les utilisateurs. On les qualifie aussi de traces numériques. Elles sont issues de l'activité de l'utilisateur et, plus spécifiquement, de son interaction avec des contenus.

Il s'agit de données nominatives, personnelles ou de profil (nom, adresse, identifiant, etc.) ; de données de transactions (mode paiement, date, montant, institution financière) ; de données sur les centres d'intérêt (contenus préférés, abonnements) ; de données de comportement (sélection, recherche, consultation, achat, partage, etc.) ; de données collectées par les téléphones mobiles, ordinateurs et objets connectés (appels, activation, localisation) ; de données de navigation.

Ces données ont la caractéristique d'être massives et d'échapper à l'usager. Elles sont aussi la capacité à créer de la valeur ajoutée, de nouveaux services et d'être convoitées. Les GAFAM tirent un profit maximum de la captation de ces données. Leur collecte massive par les plateformes soulève des questions de souveraineté et le RGPD a pour ambition de (re)donner aux utilisateurs une forme de maîtrise de leurs données par un encadrement de leur usage.

Ces données d'usage n'en demeurent pas moins de nouvelles sources pour les SHS. A ce titre, je voudrai évoquer un projet de recherche sur les logs de connexion à Gallica conduit en partenariat entre la BNF et Télécom ParisTech.[8]

Traditionnellement, on analyse les usages des lecteurs au travers d'enquêtes sous la forme d'entretiens collectifs, individuels ou de questionnaires en ligne. Là, il s'agissait d'étudier les usages des lecteurs de Gallica par une analyse des logs de connexion en leur appliquant des méthodes de fouille de données. Initialement, les logs étaient conservés pour des raisons de sécurité et d'évaluation de la qualité du service.

Le but de l'étude était de catégoriser les usages et donc de permettre le regroupement d'utilisateurs de Gallica en fonction de similitudes dans leurs comportements sur Gallica. Similitudes qui ont été repérées grâce à des méthodes de fouille de données.

L'un des enseignements les plus surprenant de l'étude est que les consultations dans Gallica restent largement monotypes alors qu'on présuppose qu'une bibliothèque numérique est de nature à favoriser l'exploration des fonds numérisés dans toute leur diversité documentaire et profondeur

[6] Latour, B et Woolgar (S). (1979). *La Vie de laboratoire : la Production des faits scientifiques*. Paris, France : La Découverte

[7] Denis, J. (2018). *Le travail invisible des données. Éléments pour une sociologie des infrastructures scripturales*, Paris : Presses des Mines

[8] Adrien Nouvellet, Valérie Beaudouin, Florence d'Alché-Buc, Christophe Prieur, François Roueff. Analyse des traces d'usage de Gallica : Une étude à partir des logs de connexions au site Gallica. [Rapport de recherche] Télécom ParisTech; Bibliothèque nationale de France. 2017. ffhal-01709264

historique. L'importance de la sérendipité si souvent mise en avant avec le numérique, n'a donc pas été confirmée par l'étude.

Archiver le Web ou vouloir fixer ce qui bouge tout le temps

Ce troisième axe me permet de revenir sur l'une des caractéristiques des sources numériques, c'est-à-dire leur instabilité et interroger ce qu'on archive : un contenu ou une expérience utilisateur ?

L'archive du Web, nous disent les auteurs de *Qu'est-ce qu'une archive du Web* cherche à reproduire l'interactivité qui existe au sein du Web en permettant de cliquer sur les liens et de naviguer dans la Toile ; Naviguer dans l'archive du Web est possible par l'intermédiaire de la Way Back Machine d'Internet Archive.

Si l'archive du Web est dynamique elle n'est pas pour autant une copie à l'identique du Web. En premier lieu parce qu'il est impossible d'archiver toutes les modifications d'un site toute simplement pour des questions techniques. En effet, « la durée de captation est supérieure au rythme de mise à jour du site ». En second lieu, parce que l'archive du Web est une recomposition, elle rassemble en fait des parties de site renvoyant à des temps ou époques différents du site ; le site archivé n'a jamais existé comme tel. Certains contenus d'une page ne sont pas archivés (ex. la publicité ou les commentaires sur les sites de presse en ligne), d'autres sont récupérés, comme les logos (les logos ne sont pas collectés à chaque fois pour éviter les doublons) ce qui peut conduire à des situations anachroniques comme par exemple retrouver le logo endeuillé de noir du CNRS sur la page du site captée en aout 2015 par la BNF alors qu'il a été mis en place suite aux attentats, en novembre 2015. Enfin, tous les liens ne fonctionnent pas puisqu'on moment de la collecte, on paramètre la profondeur des liens collectés. Les contenus d'une page Web archivée peuvent donc avoir été collectés à des moments différents et séparément et le lecteur de se retrouver confronter à des sauts temporels en cliquant sur un lien hypertexte.

Naviguer dans les archives du Web implique des précautions théoriques face à ce que V. Schafer et B. G. Thierry qui sont les auteurs d'un livre collectif sur les archives du Web[9] appellent des « régimes de temporalité désaccordés » qui poussent à rompre avec « le confort que comporte l'utilisation d'archives datées et précisément identifiées ». Par ailleurs, La collecte n'est pas une sauvegarde neutre des données, c'est un choix raisonné qui s'appuie sur des critères pour sélectionner ce qui sera conserver, critères qui peuvent varier dans le temps. En France, la collecte des archives du Web se fait dans le cadre du dépôt légal, par la BNF et l'INA. Il existe une autre initiative, celle de la fondation Internet Archive (Brewster Kahle).

On estime à 400 milliards le nombre de pages web archivées. Les archives du Web sont donc massives sans être pour autant exhaustives. Comment exploiter cette masse d'archives quand on est historien ?

L'exploration des archives du web impliquent donc des outils performants qui sont eux aussi loin d'être neutres et qui placent les utilisateurs dans une situation de dépendance vis-à-vis des institutions qui collectent et conservent les archives du Web.

Longtemps, la navigation dans les archives du Web n'a été possible qu'en saisissant l'adresse URL d'un site ; la recherche plein texte est très récente. Jusqu'en 2016 il n'était pas possible de composer des corpus thématiques. Ces dispositifs de médiation présentent eux aussi des biais, par exemple la recherche par mots clefs de la Wayback Machine d'Internet Archive ne permet de fouiller que les

[9] MUSIANI, Francesca ; et al. *Qu'est-ce qu'une archive du web ?* Nouvelle édition [en ligne]. Marseille : OpenEdition Press, 2019 (généré le 03 janvier 2024). DOI : https://doi.org/10.4000/books.oep.8713

pages d'accueil des sites. L'historien doit donc composer avec des archives en constante évolution et est dépendant des outils proposés par les institutions qui assurent leur collecte.

Avec Valérie Schafer et Francesca Musiani, on peut conclure que « 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 ».

On l'a vu les sources numériques sont intrinsèquement instables. L'instabilité documentaire n'est pas une nouveauté en ce sens que le numérique renoue avec le temps d'avant l'imprimerie et la critique de la fiabilité des sources numériques renoue avec des méthodes bien connues des archivistes, la diplomatique, pour gérer la prolifération des variantes.

Les questions liées à la fiabilité et à la fidélité des sources ne sont pas non plus spécifiques au numérique, on les retrouve à chaque innovation technologique. Dans son ouvrage *Ecrire, calculer, classer, comment une révolution de papier à transformer les sociétés contemporaines (1800-1940)*, D. Gardey évoque par exemple « les discussions sur la valeur juridique d'une lettre dactylographiée ou sur la fidélité des écritures comptables tenues sur des fiches, plutôt que dans un registre cousu »

On dit aussi du numérique qu'il est le régime de la surabondance. Là encore ce sentiment de surinformation, n'est en rien spécifique à notre époque. Les historiens de l'Europe moderne ont montré « que la Renaissance éprouvait un sentiment de surinformation à une échelle jusque-là inconnue », comme l'analyse Ann Blair dans *Tant de choses à savoir, comment maîtriser l'information à l'époque moderne*.

Si l'instabilité, l'abondance et la redondance de l'information ne sont pas *in fine* des caractéristiques singulières du numérique, les sources numériques soulèvent par contre des enjeux d'ordre éthique et déontologique liés à l'oubli et au respect de la vie privée. L'équilibre précaire entre mémoire et oubli est en effet régulièrement rediscuté. En 2013, par exemple, l'association des archivistes français a dû mener bataille contre la commission européenne qui souhaitait interdire la conservation des sources numériques contenant des données à caractère personnel.