Culture quantitative et numérique
Les archives à l'ère numérique
Notes d'accompagnement du diaporama

Cours de Caroline Muller

Caroline.muller@univ-rennes2.fr

Introduction

- Le texte d'Arlette Farge, *Le goût de l'archive* (1989) est devenu une sorte de référence de la description d'un « voyage » en archives de l'historienne qui travaille sur les fonds de la Bastille. Sur l'extrait présent dans le diaporama, on lit toute la dimension concrète et sensible du travail en salle de lecture (choix de la place, ambiance, recopiage manuel des documents d'archives)
- Cette description ne correspond plus à la façon dont écrasante majorité des historien(nes) travaille auj. Séjour court en salle de lecture, prise de photographie. Consultation des inventaires à distance. On peut même faire une recherche entière sans aller dans un centre de conservation (fonds numérisés).
- Conséquence aussi de la *« numérisation » du métier d'historien* (voir le texte de Sébastien Poublanc[1]), c'est-à-dire la banalisation du recours à des outils quotidiens : le microordinateur, la tablette, le mini-scanner, le smartphone, le disque dur externe (…) Les articles scientifiques sont aujourd'hui écrit via des logiciels de traitement de texte ; la littérature scientifique disponible en ligne, etc. Peu d'historien-nes « recopient » aujourd'hui le texte d'un doc d'archive comme Arlette Farge le décrit.

Slide 5

Commenter la capture du tweet de Nathalie Raoux. Que dit-elle de son travail quotidien ?

- La place du téléphone
- La découverte régulière d'usages (océriser*, scanner)
- L'énoncé de cette découverte sur un réseau socio numérique, Twitter

[1] Sébastien Poublanc, *Qu'est-ce que le numérique fait aux historiens ?* 2019, https://sms.hypotheses.org/12125 .

I A Qui numérise et pourquoi ?
Slides 6 à 10

Tweet des archives du Canada montre deux choses :

→ Que « numériser » n'est pas un geste simple, intuitif, en un clic : longue préparation nécessaire

→ Que « numériser » relève d'un choix, ici des dossiers militaires, pourquoi ? Cela rejoint en fait les types de fonds numérisés en priorité ; sur AD35 on voit l'état civil, les registres paroissiaux, matricules militaires, actes notariés, presse et iconographie. Chacun de ces choix correspond à une logique

- Logique de préservation / mise à disposition. Un fonds très consulté devient fragile, le numériser, c'est le protéger tout en le rendant plus accessible à tous. Pour la presse, grande fragilité du support qui se désagrège (papier journal XIXe en particulier). Pour les registres, correspond au besoin public et social de généalogie, de faire valoir des droits (cadastres et droits de propriété). Dans les choix de numérisation, on lit les différentes *fonctions* des centres d'archive.
- Penser aussi que les instruments de recherche* sont eux aussi numérisés, pas que les fonds. Exemple du cadre de classement des AD 35 sur diapo : permet d'avoir une idée du contenu d'une boîte.
- La numérisation en elle-même est un geste dont les historien-nes doivent comprendre la logique ; on peut numériser en masse (à la chaîne avec des machines, exemple de la presse) ou document par document (voir le billet de Johanna Daniel) quand c'est fragile ou que ça nécessite des choix scientifiques. Une partition de Rameau, annotée, recoupée, découpée par le compositeur nécessite une *stratégie de numérisation* particulière pour que le ou la lectrice en comprenne le sens une fois derrière son écran. Numériser peut aussi ê « renumériser » à partir de supports déjà secondaires comme les microfilms ; ou reprendre une numérisation ancienne qui ne correspond plus aux standards de qualité. Coûts de tout ceci.

I B Consulter : la salle de lecture réinventée
Slides 11-15

Transformation profonde aujourd'hui de la notion de « salle de lecture » qui était jusqu'alors le lieu, dans un centre de conservation, où l'on demandait puis consultait les documents après s'être dûment inscrit, tout ceci étant décrit dans le livre d'Arlette Farge. Une salle de lecture aujourd'hui peut être : la page d'accueil de Gallica, la page d'accueil d'une association qui propose la lecture de fonds d'archives, son salon où l'on consulte ses photos d'archives, bref, toutes sortes de lieux virtuels et concrets.

- Changement majeur de la diffusion de la photographie des fonds d'archives par les chercheurs. Nouveaux objets comme trépied, smartphone, appareils photo et scanners. Photographie en masse : Sean Takats (qui a dirigé l'équipe de développement de Zotero) a mené une enquête montrant que hist. Américains ont en moyenne +5000 photos d'archives dans leurs ordinateurs.
- Conséquences : bouleversement lieu lecture. Chez soi, avec son thé ou son café, derrière écran. Pas de problèmes horaires. Dépendance à une connexion Internet et à son matériel. Posture change aussi.
- Risques : décontextualisation du document. On ne peut plus mener directement la critique externe ; observer la forme, odeur, texture du doc, toutes riches en enseignements. Difficulté d'évaluer sa taille, sa couleur véritable (exemple de l'iconographie numérisée), sa place dans le fonds : comment de boîtes ? de dossiers ? quel est le document juste avant et juste après ?
- Quand travail dans salle de lecture virtuelle (exemple Archives nationales sur diapo), nécessité de *bien connaître le moteur de recherche pour comprendre ce qui sort* : notion de *requête*. Dans quel texte le moteur de recherche va chercher le mot demandé ? Le document lui-même ou ses métadonnées ? Bref, pour chercher efficacement dans une salle de lecture virtuelle, *il faut comprendre comment a été construite l'interface mise à notre disposition*. La recherche avancée peut être salutaire, permet d'affiner la recherche dans des masses de documents. En fait, il faut faire la critique historienne de l'outil qu'on utilise pour découvrir le fonds d'archive, c'est-à-dire *comprendre comment le document arrive sur notre écran, pourquoi l'algorithme a préféré nous montrer celui-ci plutôt qu'un autre* (et il faudrait en parler dans un autre cours, mais on est déjà bien encadré par les intelligences artificielles…)

I C Gérer l'archive-donnée chez soi
Slides 16

Prendre des centaines de photographies ou télécharger des masses de PDF implique d'avoir un plan de gestion de vos données de recherche. Sur deux ans, vous oublierez de façon certaine ce que vous avez pris en photo la 3e semaine de votre M1. Il faut donc trouver une solution pour classer et décrire les données de recherche. Pour les photos d'archives, le logiciel Tropy, gratuit et open source, a été dessiné pour cela. Vous pouvez aussi gérer cela à partir d'une base de données ou d'un simple fichier Excel. L'important est de se doter d'une stratégie de gestion 😊

II A Les archives du Web
Slides 17 à 19

Parmi les nouveautés – plus si nouvelles en réalité mais les étudiant-es y pensent encore peu – nous pouvons travailler sur les archives d'Internet et du Web : archives des sites Web, des réseaux socio numériques, etc. C'est ce qu'on appelle des sources nées numériques ou nativement numériques, i.e. qui n'ont jamais eu d'existence « physique » comme le rapport préfectoral de 1853.

- Archiver Internet et le Web est essentiel : songez à la place que cela occupe dans nos vies quotidiennes aujourd'hui. Cet archivage a déjà une longue histoire (voir la réf biblio sur la diapo) qui même des acteurs privés (les débuts d'Internet Archive), des institutions publiques – bibliothèques-, des États. Là aussi, il faut distinguer : archivage de masse (un robot qui passe sur les pages et les capture – on y reviendra) et stratégie ponctuelle concertée autour d'événements précis. Par exemple, la BNF et l'INA font des collectes à l'occasion d'événements comme les attentats ou les campagnes électorales. Ces choix relèvent de discussions autour des fonctions des archives.
- Archiver le Web n'est pas aisé pour de multiples raisons :
  - → Par définition, une page Web est mouvante (en constante évolution) et relationnelle (elle renvoie à tout un environnement par un jeu de liens) : faire une capture d'écran d'une page le matin ne garantit pas que la page sera identique le soir ! alors, qu'est-ce qu'on garde ?
  - → Cela pose des difficultés techniques, par exemple archiver les images, les vidéos, certains formats de fichier. Comme les archives classiques, elles sont pleines de trous. La différence est que, quand on lit un rapport préfectoral, on a dans les mains le document tel qu'il a été rédigé puis lu ; *pour la page Web, on ne dispose jamais que d'une*

version déformée de ce qu'elle était dans son environnement. C'est la question cruciale de l'unité et de l'intégrité du document qui se pose désormais de façon renouvelée.

→ Les réseaux socio numériques posent toute une autre série de problèmes, en particulier la dépendance de l'archivage aux contraintes posées par les plateformes comme Twitter ou Facebook. Par ailleurs, la masse de données collectées pose, en bout de chaîne, une nouvelle difficulté : une fois la collecte faite, il faut développer des outils qui permettent aux historien-nes et chercheur-es de les travailler ; exemple de la Library of Congress et Twitter.

II B Les archives audiovisuelles
Slide 20

D'autres types de sources sont aussi disponibles, celles des média « traditionnels » : la radio ou la télévision.

- Sur la diapo – dont le schéma vient d'une conférence à Reims de Thibault Le Hégarat – on voit une frise de conservation des archives de l'INA – environ 700 000 heures de programmes en 2016. Archiver la télévision pose des problèmes techniques car les supports ont beaucoup évolué dans le temps (kinéscope, magnétoscope, bobines, vinyles, cassettes, CD, dvd, formats nativement numériques, etc.) Cela implique de transférer régulièrement des enregistrements d'un support à l'autre et de conserver des outils de consultation plus disponibles sur le marché courant (essayez d'acheter un magnétoscope aujourd'hui…)
  Ainsi sur n'importe quel corpus d'archives, à l'ère numérique ou non, importance de prendre conscience des logiques et choix de conservation, du poids des supports, qui ont un effet sur la création des vides et des pleins.
- La radio fait elle aussi l'objet de politiques de conservation[2]. Le cas de la radio décrit par Céline Loriou montre qu'il est souvent nécessaire *de se rendre dans l'institution de conservation* pour avoir accès à la source numérisée – ici la BNF relais de l'INA - autrement dit, il ne faut pas assimiler « numérisation des sources » et « accès distant automatique » et mise en ligne.

[2] Pour voir comment on peut travailler sur ce type d'archives, voir Loriou Céline. « Faire de l'histoire, un casque sur les oreilles : le goût de l'archive radiophonique ». In : *La Gazette des archives*, n°253, 2019-1. *Le goût de l'archive à l'ère numérique*. pp. 71-82. DOI : https://doi.org/10.3406/gazar.2019.5687

II C La numérisation de la vie quotidienne : effets
Slide 21

La numérisation du métier d'historien(ne) n'est qu'un aspect de la numérisation générale de la vie quotidienne, avec son lot d'exclus – par exemple devant la dématérialisation des démarches administratives des services publics. Au point de vue archives, les défis sont immenses.

Une vie quotidienne de 1850, 1950 et 2022 ne produit plus du tout la même nature de trace. Exemple du commerce, des registres de compte, des reçus et autres traces de paiements – vs la généralisation de la carte bleue. Nous produisons massivement des traces numériques, du soir au matin et du matin au soir si on utilise un smartphone – traces qui n'ont pas l'inertie du papier qu'on est capable de conserver des milliers d'années. Enjeu archivage aujourd'hui : archiver des flux de données, et pouvoir ensuite fournir des solutions pour les consulter. Autre exemple : les bases de données des entreprises ; si on archive la base de données de toutes les couleurs de brosses à dent produites par l'entreprise d'hygiène Dupont en 2010, il faudrait être certain de pouvoir lire cette base de données sur un ordinateur en 2045. Autrement, *la transmission par inertie* – ce qui reste même sans qu'on y prête forcément attention, les papiers découverts par Nicolas Offenstadt dans l'urbex en Allemagne de l'Est[3] – cette transmission est en train de disparaître.

III A L'essentiel est invisible pour les yeux ? Invisibilité et visibilité des archives
Slide 22

- Le risque du syndrome du lampadaire. Ne travailler que sur les archives visibles, référencées, mises en ligne et décrites, ou même numérisées. Effet très important du choix du corpus sur les résultats d'une recherche. Effet de surutilisation d'archives plutôt que d'autres (exemple de la presse numérisée). Contexte social et politique de cela, manque de financement de la recherche, qui peut conduire à faire des choix adaptés aux contraintes économiques et moins coûteux en temps (consulter le fonds numérisé plutôt que faire un voyage dans un lieu où se trouve un fonds non numérisé)
- La recherche par mots clefs, nouveau réflexe devant n'importe quel moteur de recherche, a tendance à faire oublier les modes de constitution des fonds d'archives, par producteur. Revenir toujours à : qui produit quel type de document ? Se forcer à intégrer cette interrogation à son travail sous peine de ne voir « sortir » du moteur de recherche qu'un part infime de ce qui pourrait être intéressant

[3] Offenstadt, Nicolas, *Urbex RDA : L'Allemagne de l'Est racontée par ses lieux abandonnés*, Paris, Albin Michel, 2019, 255 pages.

- Pour autant, ne pas basculer dans l'angoisse inverse de vouloir disposer de l'ensemble des données, ou d'un volume impossible à traiter dans les temps d'une recherche de master. Problème : quand on travaille sur archives photographiées, pas de « boucle de rétroaction » au sens ou pas de familiarisation préalable avec le fonds. Tendance à tout prendre en photo, et vaille que pourra. Alterner lecture proche et lecture distante est essentiel quand on découvre un champ de recherche.

B Nouvelles questions, nouveaux modes de lecture des archives
Slide 23

- La lecture distante sur la presse : exemple des apports et difficultés. Le projet *Numapresse* et la question de la viralité, qu'on peut poser désormais grâce aux grands volumes de données. La lecture très très proche : exemple du travail extrêmement pointu qu'on peut mener sur les couches de manuscrit et rouleaux grâce à la numérisation 3D[4].
- L'historien(ne) devient gestionnaire de ses données, ce qui a un effet jusque dans son écriture. Exemple du *Journal of Digital History* et des façons de donner avoir les données exploitées pour une recherche (règle qui veut qu'en histoire, les résultats doivent offrir au lectorat la possibilité de refaire l'enquête…donc citer les sources.)
- Sans aller forcément jusque-là, prendre conscience de ses pratiques numériques discrètes. Photographier et stocker des images d'archive, utiliser la recherche plein texte dans une transcription, annoter numériquement une photo de manuscrit, télécharger des notices de documents numérisés, zoomer/dézoomer sur un détail, etc.

III C Du trop-plein au manque : que restera-t-il de nos archives en 2900 ?
Slide 24

Fragilité de nos données malgré leur massivité. Exercice d'imagination de ce qui se produirait sans réflexion concertée sur l'archivage : les blancs de l'histoire ? Un siècle où l'on n'aura jamais autant produit d'information paradoxalement plongé dans l'inconnu pour nos futurs camarades. Difficulté de s'adapter à la vitesse d'évolution des technologies (question de la télévision plus haut, aujourd'hui des applications). Plus généralement, contexte sociopolitique à prendre en compte à nouveaux ; poids des grandes plateformes et question de la souveraineté des États.

[4] Giovacchini Julie. De la source à l'image : y a-t-il une philologie numérique ? In : *La Gazette des archives*, n°253, 2019-1. Le goût de l'archive à l'ère numérique. pp. 53-70.
DOI : https://doi.org/10.3406/gazar.2019.5686