Publié sur *URFIST de Paris* (https://urfist.chartes.psl.eu)

atelier [1], Atelier Boudoirs de l'historien(ne) [2], Histoire [3], ressources électroniques [4]

Thématique : Bibliothèques numériques du XIXe siècle, retours sur expériences - Entre jungle et Eldorado

Intervenants

- Claire Lemercier [5], Directrice de recherche CNRS en histoire, Centre de sociologie des organisations ;
- Christophe Meslin [6], Doctorant contractuel, Centre d'histoire culturelle des sociétés contemporaines – UVSQ – Fondation des sciences du patrimoine.

Introduction

En préambule, Christophe Meslin introduit cette séance par une citation de 1835 qui montre bien que l'on n'a pas attendu le XXIe s. pour être « troublé » par la question des sources, trouble renforcé *a fortiori* à l'heure des *digital humanities* comme le porte une annonce récente [7] dans Calenda. Il revient également sur l'usage trompeur du terme de « *bibliothèques* numériques » pour désigner des outils dont l'idée de classement initial est bien souvent absente.

Les deux intervenants reviennent d'abord sur le préalable essentiel qui est de bien choisir les mots recherchés pour toute recherche dans une bibliothèque numérique. A cela deux raisons :

1° *l'évolution des mots-clés*, notamment quand on travaille sur une longue période avec une évolution possible des termes et du vocabulaire. De fait, la notion de « mot-clé » est trompeuse pour ces outils qui ne sont bien souvent pas des outils bibliothéconomiques et ne proposent pas toujours d'accompagnement humain (textes bruts) ; on se reportera alors utilement à Frantext [8], par exemple, outil établi par des linguistes à partir d'un corpus certes restreint et littéraire, mais ouvert pour la période sur la non fiction et replaçant les termes dans leur contexte.Cette question est particulièrement importante pour les étudiants afin de leur faire comprendre qu'il puisse y avoir des recherches autres qu'en texte intégral, mais par exemple par date ou avec un thésaurus…. Comme l'exprime Andrew Abbott dans son récent *Digital Paper. A Manual for Research and Writingwith Library and Internet Materials*, il est essentiel d'enseigner la cohabitation des deux modes d'interrogation ;

2° l'*évolution des outils numériques*. Il est ainsi important de conserver des archives de ses propres recherches et noter les mots-clés et les réponses obtenues, de manière à pouvoir refaire la recherche un an plus tard. De fait, derrière une impression de scientificité, ces outils ne proposent pas toujours des résultats réplicables et les résultats peuvent changer selon le lieu, les changements de la collection ou l'évolution des fonctionnalités avancées des outils. De même, est-il important de garder trace des résultats afin de pouvoir retrouver des documents, *a minima* l'URL (dont on ignore cependant la pérennité), voire le PDF. C. Meslin rappelle que lors des tensions entre la France et Google il y a quelques années, Google livres présentait des résultats vraisemblablement « bridés », mais que l'on pouvait néanmoins toujours accéder aux textes en ligne en passant par l'URL conservée sur leur PDF.

Les incontournables

- **Google livres** [9]

Un constat s'impose dès le départ : Google livres « ne peut pas être ignoré » notamment quand on recherche sur des sujets complémentaires (ex. : aspect culturel + technique), et bien souvent, c'est seulement dans un second temps que l'on utilise d'autres outils, notamment quand ces derniers offrent des corpus constitués. De toutes les bibliothèques numériques existantes, c'est certainement Google livres « la plus grande et la moins rangée ». Il y aurait ainsi 15 millions de documents dans Google livres (mais part de français ?, part de XIXe s. ?). L'intérêt d'une telle masse est de mettre en lumière les écritures « ordinaires », « des anonymes », des écrits « pas très dignes » (la littérature « grise »). Là où le catalogage n'était parfois pas toujours explicite, des choses sont désormais rendues visibles et ce, quel que soit leur auteur, anonyme comme organisation…

Plusieurs points d'attention doivent cependant alerter le chercheur :

- le moteur de recherche avancée [10], très utile, est bien caché ;
- l'outil reste discret sur son contenu et son alimentation : impossible de connaître le contenu des derniers versements, ce qui éviterait pourtant de rechercher à nouveau sur des documents déjà

repérés ; impossible de connaître également les fonds numérisés par telle ou telle bibliothèque, ce qui permettrait de repérer des fonds spécialisés ou des collections suivies ;
- à l'instar du moteur de recherche Google, le moteur de recherche de Google livres, bien que très puissant, n'est pas rigoureux : des recherches avec des guillemets, avec la troncature (*) ou encore en respectant la casse n'apportent pas toujours des résultats similaires à la chaîne de caractère saisis et une recherche [miroir + miroirs + miroitiers] sera plus satisfaisante qu'une recherche [miroir*] ; de même, convient-il mieux de forcer la chaîne de caractères même pour un terme unique : la requête [« respectabilité »] évitera ainsi d'obtenir parmi les résultats des termes comme « sort respectable »… ;
- le texte n'est pas systématiquement disponible ; on pourra alors se tourner en priorité vers les bibliothèques numériques Hathi Trus [11]t ou Archive [12], ou chercher la bibliothèque qui a numérisé le texte ;
- enfin, une méthodologie rigoureuse doit être suivie lorsque l'on utilise l'outil Ngram Viewer [13], qui permet de suivre l'évolution de certains termes dans le temps. De fait, s'il s'agit d'un très bel outil exploratoire, pour établir des hypothèses, il n'en est pas un outil scientifique. L'ouvrage de Christophe Charle *La discordance des temps : une brève histoire de la modernité* montrait ainsi, au travers de l'exemple du terme « modernité », les limites du Ngram Viewer : ignorance du corpus interrogé (titres et nombre) et absence de datation de nombre d'ouvrages ; absence de contextualisation du mot ; nivellement des écarts en raison des grandes périodes définies par défaut, etc. Comme le dit Claire Lemercier, « il y a encore une valeur au travail » : à l'historien de revoir le mot dans son contexte écrit, mais aussi historique, en croisant ces premières pistes avec d'autres données et en se méfiant des biais de la numérisation même des ouvrages. C. Meslin rappelle ainsi que, si le développement du terme « trumeau » au cours du XIXe s. peut être lié à une évolution commerciale et à une démocratisation de l'usage des miroirs, les écrits concernés sont encore bien souvent issus d'une littérature de « privilégiés », ce qui pose la question même des lecteurs d'origine des ouvrages numérisés. D'une manière générale, l'outil de Google est plus pertinent pour les noms propres et quasi-noms propres (organisations), sans jamais avoir valeur de preuve. De son côté, C. Lemercier rapporte avoir ainsi repéré un pic de l'expression « Tribunal of commerce » en 1874, avant de se rendre compte que la numérisation concernait essentiellement des journaux australiens, alors qu'elle-même s'intéressait au Royaume-Uni et aux Etats-Unis. L'absence même de corpus véritablement cohérents remet en cause le principe des « *culturomics* », selon lequel les statistiques ont réponse à tout et qu'il n'y a plus besoin d'historiens.

- **Gallica** [14]

La bibliothèque numérique Gallica, c'est actuellement plus de 550 000 livres ou encore plus d'un million de fascicules de presse. De l'avis des intervenants, cette bibliothèque est excellente lorsqu'on sait déjà ce que l'on cherche – le plus simple est d'abord de faire des essais sur Google livres.

Parmi les points forts à souligner :

- les documents sont liés à un catalogue de bibliothèque, avec ses autorités, facilitant ainsi les filtres ou le repérage d'autres volumes ou de collections complètes ;
- certaines parties de Gallica sont pensées comme des corpus (intégral ou cohérent), ce qui permet au chercheur de savoir sur quoi il argumente ;
- il est possible de télécharger le document, voire l'OCR pour faire le copier-coller (visible en passant par affichage > mode texte) ;
- on peut se créer un panier, divisible en dossiers.

Quelques points faibles empêchent néanmoins une utilisation totalement fluide :

- faute de rapprochement entre les différents titres d'une même revue, il est parfois difficile de retrouver l'ensemble des numéros. Le plus simple est de passer par le catalogue général [15], qui fournira la notice mère avec l'historique des titres et les liens vers Gallica ;
- toujours du côté des revues, la recherche doit se faire en deux temps puisqu'il convient de relancer la recherche initiale sur la première liste de résultats obtenus (moteur de recherche en haut à gauche « rechercher dans ce périodique ») ;
- tous les livres ne sont pas encore OCRisés et certains ne disposent encore que du mode image, sans mode texte ; par ailleurs, l'OCR (reconnaissance optique de caractères), surtout sur les quotidiens n'est pas toujours bien satisfaisante : la mauvaise qualité du papier empêche une bonne reconnaissance des mots, tandis que l'étroitesse des colonnes coupent de nombreux noms – ceci explique notamment un projet australien autour de la presse quotidienne [16] avec correction collaborative ;
- le mode d'affichage ne facilite pas toujours la contextualisation (pages séparées éventuellement).

- **Archive** [12]

Archive est un bon complément aux deux premiers, notamment quand ils ne proposent qu'un simple aperçu. Si le moteur de recherche avancée [17] est catastrophique - ne remplir que « *anyfield* » -, et si, d'une manière générale, il ne faut pas se fier à l'OCR mais privilégier le PDF, il peut apporter une véritable

plus-value pour certains types de livres numérisés, comme les catalogue de ventes avec des annotations. Il est intéressant notamment en cas de recherche trop fine et qui aurait échoué dans Google livres et Gallica. Enfin, il indique la collection de provenance du document (colonne de droite), ce qui permet de rebondir directement sur les outils spécifiques de l'institution de conservation. Comme pour le lien entre le catalogue général de la BnF et Gallica, on notera l'intérêt de croiser les outils en raison des informations présentées.

Quelques outils complémentaires

A titre de complément et de comparaison, les deux intervenants présentent ensuite quelques autres outils :

- malgré un accès possible via une institution, une grande partie de la base **HathiTrust** [11] est néanmoins accessible librement, au moins pour de simples aperçus, ce qui permet de découvrir des documents à consulter en version papier ;

- établis par des bibliothèques et/ou des chercheurs, les **bases de l'éditeur Gale** [18] présentent les avantages de Google livres par leur taille et la qualité de Gallica par leurs autorités. Outre les bases de presse (*The Times*, *The Economist*) et les collections thématiques par type de documents (ex. : quotidiens, régionaux) ou par sujet (ex. : esclavage), les bases *The Making of the Modern World* sont particulièrement intéressantes pour le dix-neuviémiste, puisque la numérisation systématique de bibliothèques américaines donne ainsi accès à de nombreux livres européens et notamment français. Deux problèmes à signaler : d'une part des difficultés d'accès aux PDF ; d'autre part, des difficultés d'accès aux bases elles-mêmes du fait de leur coût. Rares sont en effet les bibliothèques européennes et *a fortiori* françaises qui donnent accès à quelques titres de Gale (BnF, BIS Sorbonne, ENS).

Parmi les autres ressources signalées lors des échanges avec le public, on notera :

- **Europeana** [19], la bibliothèque numérique européenne, plus basée néanmoins sur des corpus thématiques ;
- **CNUM** [20]du CNAM, pour tout ce qui concerne l'histoire des arts et techniques, intéressant notamment pour des recherches par auteur ;
- **Medic@** [21] de la BIU Santé, pour tout ce qui concerne l'histoire de la santé.

Pour la bibliographie, on se reportera, bien sûr, aux projets numériques suivants :

- **Cairn** [22], **Persée** [23]et **Revues.org** [24] pour la France ;
- le **projet MUSE** [25], plutôt pour le domaine anglo-saxon.

De l'intérêt et des limites des bibliothèques numériques pour l'historien

Pour bien comprendre les bibliothèques numériques, leur contenu et leur fonctionnement, il est indispensable d'avoir, en complément, une bonne représentation mentale des autres outils existants. C'est particulièrement vrai quand il s'agit de former des étudiants en histoire. Sommes-nous en train d'assister à une rupture entre les pratiques anciennes de recherche et les pratiques plus actuelles ? Peut-on craindre que les collègues privilégient les archives papier, considérés comme plus « légitimes » et moins « faciles » qu'internet et que les étudiants, de leur côté, pensent que tout est accessible en ligne ? C'est bien sûr une vue trop schématique de la réalité et qui cache les apports croisés des différentes ressources. Certes, nombre de productions professionnelles restent encore sous format papier et courent ainsi le risque d'être oubliées, mais l'utilisation du numérique permet assurément d'autres modes de publications, de diffusions et donc de travail académique, parfois au-delà du projet initial. On mentionne ainsi le projet d'édition scientifique *L'écho des fabriques* [26] réalisé dans un cadre précis mais qui permettra sans doute d'autres types de recherche. Mais *a contrario* nombre d'historiens soulignent le paradoxe des outils numériques, qui permettent certes d'être moins « prisonnier » du lieu de conservation mais également moins dans l'ambiance, ce qui empêche ainsi de développer d'autres types de contextualisations – on a tous fait l'expérience d'une exploration des rayonnages d'une bibliothèque – et de discussions.

Au final, deux risques existent face aux textes numérisés :

- le **syndrome du lampadaire** : on ne consulte que ce qui est facilement accessible – des études au Canada ont ainsi montré que les seuls quotidiens cités dans les travaux étaient des quotidiens numérisés ;
- l'**illusion de l'exhaustivité** : les outils numériques favorisent explicitement de nombreux usages exploratoires qui n'étaient pas possibles avec les catalogues. Mais, si disposer d'une masse de documents est intéressant pour le chercheur, il est essentiel de savoir cerner les travaux et ce qui a été fait. Le caractère évolutif même de ces outils impose d'encore plus réfléchir aux usages que l'on a à chaque recherche et de documenter sa propre pratique : vocabulaire utilisé, méthodes (type de recherche, recherche par corpus cohérents…) et résultats. Face à l'impression qui peut le saisir, le (jeune) chercheur doit comprendre que l'exhaustivité, qui n'a jamais existé, n'existera pas davantage

à l’ère numérique. Il s’agit bien plutôt de réfléchir en termes de corpus prioritaires. C’est particulièrement important dans le cadre d’un mémoire ou d’une thèse : il est crucial de ne pas rester au premier état (rassembler les infos, malgré les outils d’alertes), mais d’avancer. Si la méthode et les interrogations sont bonnes, ce n’est pas grave de penser qu’une recherche de documents dans 3 mois peut avoir des résultats différents d’une recherche aujourd’hui. Il y aurait en revanche un véritable problème si, dans 3 ans, la même recherche entraînait une thèse totalement différente, avec d’autres conclusions…

Pour aller plus loin :

- document de C. Lemercier [27]
- document de C. Meslin [28].

Mots-clefs

Types de publics concernés par cette page
Doctorant [29], Enseignant du supérieur, chercheur [30], Professionnels de l’information [31]
Equipe
Aline Bouchard [32]

Liens
[1] https://urfist.chartes.psl.eu/tags/atelier [2] https://urfist.chartes.psl.eu/tags/atelier-boudoirs-de-l-historienne [3] https://urfist.chartes.psl.eu/tags/histoire [4] https://urfist.chartes.psl.eu/tags/ressources-%C3%A9lectroniques [5] http://www.cso.edu/cv_equipe.asp?per_id=168 [6] http://www.chcsc.uvsq.fr/centre-d-histoire-culturelle-des-societes-contemporaines/langue-fr/l-equipe/doctorants-et-post-doctorants/contrats-doctoraux/m-meslin-christophe-280474.kjsp [7] http://calenda.org/324845 [8] http://www.frantext.fr/ [9] https://books.google.com/?hl=fr [10] https://books.google.fr/advanced_book_search?num=20&hl=fr [11] https://www.hathitrust.org [12] https://archive.org/ [13] https://books.google.com/ngrams [14] http://gallica.bnf.fr/ [15] http://catalogue.bnf.fr [16] https://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf [17] https://archive.org/advancedsearch.php [18] http://www.cengage.com/search/showresults.do?N=197 [19] http://www.europeana.eu/portal/ [20] http://cnum.cnam.fr/ [21] http://www.biusante.parisdescartes.fr/histmed/medica.htm [22] http://www.cairn.info/ [23] http://www.persee.fr [24] http://www.revues.org/ [25] https://muse.jhu.edu/ [26] http://echo-fabrique.ens-lyon.fr/ [27] https://urfist.chartes.psl.eu/sites/default/files/ab/boudoirs/URFIST_Boudoirs_2015_05_18_Lemercier.pdf [28] https://urfist.chartes.psl.eu/sites/default/files/ab/boudoirs/URFIST_Boudoirs_2015_05_18_Meslin.pdf [29] https://urfist.chartes.psl.eu/types-de-public/doctorant [30] https://urfist.chartes.psl.eu/types-de-public/enseignant-du-sup%C3%A9rieur-chercheur [31] https://urfist.chartes.psl.eu/types-de-public/professionnels-de-l%E2%80%99information [32] https://urfist.chartes.psl.eu/urfist-de-paris/aline-bouchard