

Bibliothèques numériques du XIX^e siècle : retours sur expériences

Entre jungle et eldorado

Introduction

Présentation

Rapide tour de table

Pour les professionnels de l'information : en quoi leur métier change ? Comment se passe l'interface, l'interaction avec les usagers ?

Pour les doctorants et chercheurs : leur sujet et dans quelles mesures utilisent-ils les « bibliothèques » numériques

Lecture d'un texte en guise de devinette : quelle date ? (à la décennie près...)

« Il y eut un temps où l'on se plaignait en France de la rareté des sources historiques. Peut-être avait-on raison de se plaindre, car beaucoup de découvertes [sic] étaient encore à faire ; et les travaux exécutés par de patientes congrégations ou de doctes académies restaient incomplets ou inaccessibles pour le plus grand nombre.

Quelle différence aujourd'hui ! Les matériaux surabondent, la lumière nous vient de tous côtés, les antiques cités se révèlent, les ruines sortent de dessous terre ; c'est pour le coup que le genre humain a retrouvé ses titres, et la science qui apprend à les reconnaître est devenue une chose facile et familière, un instrument dont chacun peut faire usage pour son compte. [...].

Il faut en convenir aussi, nous sommes de rudes et hardis fossoyeurs ; c'est à qui mettra la main à l'œuvre, à qui découvrira quelque débris inconnu, quelque fait nouveau, quelque renseignement oublié. Encore un peu de temps, et nous serons embarrassés de nos propres richesses, et la vérité s'effacera peut-être obscurcie par la multitude de preuves qui devaient la mettre en lumière, à moins que, par un examen attentif et une exacte appréciation des faits, nous ne fassions régner un peu d'ordre parmi ces éléments épars, quelquefois contradictoires.

Alors, mais seulement alors, l'œuvre de la reconstruction s'accomplira sans danger, et la synthèse historique pourra remplacer cette froide et incomplète analyse dont le moindre inconvénient n'est pas de concentrer l'attention sur les détails en la détournant de l'ensemble. [...]

Nous savons l'effet ordinaire des révolutions physiques et politiques dans les sociétés humaines. Ces révolutions vivifient quelquefois, mais il leur arrive aussi de tuer, et alors elles dispersent les débris à peu près comme les meurtriers qui vont semant çà et là les membres de leur victime. Notre devoir, à nous, est d'exhumer le vieux cadavre, non point par lambeaux (s'il est possible), mais en conservant intactes les parties qui se tiennent encore, ou, si elles sont disjointes, en étudiant les rapports qui serviront à les réunir.

En un mot, la patience des recherches, est chose vertueuse et utile, mais plus utile encore, et, par conséquent, plus honorable est la science qui apprécie sainement les matériaux qui lui sont offerts et qui les dispose de manière à reconstruire les temps passés avec leur vérité de croyances, de mœurs, de langage et de physionomie. »

Réponse

Journal de l'Institut historique, Tome troisième, deuxième année, Paris, 1835, rapport d'Eugène Labat (membre de l'Institut historique) sur la publication des « Archives curieuses de l'histoire de France depuis Louis XI jusqu'à Louis XVIII » par M.L. Cimber et F. Danjou, pp 64-65 ; [<https://books.google.fr/books?id=iqgPAAAAYAAJ>]

Ecrit dans le contexte de la floraison des sociétés savantes, alors très productives tant en terme de fouilles archéologiques que d'écrits, allant parfois jusqu'à un niveau d'érudition qui sera moqué plus tard par Flaubert dans *Bouvard et Pécuchet*. Mais il faut aussi restituer à cette époque son envie de « scientificité » et de rigueur de classement car c'est véritablement à ce moment que les glossaires, thésaurus, tables des matières, sommaires, index, etc., et une véritable attention portée à la conservation des archives (avec des budgets) émergent véritablement ; outils et matériaux dans lesquels les rénovateurs, ou plutôt pères fondateurs de la « science historique », tels François Guizot, Augustin Thierry ou Edmond Michelet vont largement puiser.

Beaucoup plus près de nous, et même de l'ordre du futur, une annonce dans Calenda pour une formation d'une semaine organisée par la MSH Val de Loire, qui se tiendra en octobre prochain, intitulée « Gestion de projets des sources numériques de la recherche en sciences humaines et sociales » qui commence ainsi :

« Cette formation s'adresse aux enseignants-chercheurs, chercheurs, doctorants, ingénieurs et techniciens, bibliothécaires impliqués ou allant être impliqués dans un projet de numérisation des sources de la recherche, **troublés par le domaine des « humanités numériques »**.

<http://calenda.org/324845>

(clôture des inscriptions le 30 mai)

Un trouble, cela peut être positif ou négatif, mais certainement nous sommes tous, plus ou moins, et de bien des manières différentes, troublés par les « humanités numériques » ou les « sources numériques ». C'est ce trouble que l'on souhaite aborder dans cette séance, dans l'écart, la distance ou le vertige qui peuvent être ressentis entre l'aspect de « jungle » et celui « d'Eldorado », qui tous deux contiennent une dimension d'aventure et d'exploration... mais dans l'une, on s'y perd, dans l'autre, on s'enrichit.

Bref tour d'horizon des « modes de recherche avancée » de 5 principales « bibliothèques » numériques

Google Books (ou Livres) ;
Gallica ;
Internet Archive (Archive.org) ;
Hathi trust ;
GALE.

Rappel de définition :

Bibliothèque (*Le petit Robert*) :

- 1- Meuble ou assemblage de tablettes permettant de ranger et de classer des livres
- 2- Salle, édifice où sont classés des livres pouvant être consultés

Avant-propos : banalité mais souligner l'importance de bien choisir ses mots, son champ lexical, les possibilités de recoupements et d'exclusion (qu'on utilise les opérateurs booléens ou les filtres des moteurs), et si l'on travaille sur une grande période, tenir compte de l'évolution des mots, des glissements ou superpositions de sens ; usage des différents dicos de l'époque mais surtout « sentir » dans la masse des textes, dans le « jus », ces évolutions, les autres mots ou adjectifs qui s'y agrègent, etc.) et ainsi affiner et enrichir progressivement son champ lexical d'interrogation ; et NB, utiliser à la fois singulier & pluriel, mais par contre se méfier des conseils (comme sur le site « Comment ça marche ») préconisant l'astérisque (*) : miroi* (miroir, miroirs, miroitier, miroitiers, miroiter, etc.) Pour Google Book et Gallica, on obtient beaucoup plus de résultats en faisant mot par mot

Insister sur un nécessaire va-et-vient à faire entre le « chaos » qui ouvre des pistes, fait ressortir des documents auxquels on n'aurait pas immédiatement pensé, et des corpus constitué ou à constituer.
Et que ces usages de ressources numérisées sont aussi très relatifs au sujet de recherche que l'on a.

- **Google Livres** (environ 15 millions d'ouvrages numérisées, mais mystère sur le nombre en français, et du XIXe siècle ?)

La recherche avancée « cachée » : https://books.google.fr/advanced_book_search?hl=fr

Google Livres a un double intérêt : la plus grande masse (car ce n'est pas une collection à proprement parler) d'ouvrages numérisés, mais aussi la puissance du moteur de recherche appliqué à des ouvrages qui ne sont pas dans les « collections » de Google mais qui – du coup – nous apparaissent comme existants ailleurs, et avec parfois et par chance, un extrait du texte, quelques lignes, qui peuvent déjà donner une idée de l'intérêt ou non d'une recherche plus poussée à mener sur d'autres moteurs, ou d'une consultation en bibliothèque.

[Soit maintenant, soit après le tour d'horizon, un petit point sur la dimension « gadget » (ou non) de **Google books Ngram Viewer**.

(occurrences d'un mot dans l'ensemble des livres numérisés par Google en pouvant restreindre à une langue) ; <https://books.google.com/ngrams>

Mettre en garde contre les présentations dithyrambiques et très optimistes de la « culturomics » par un de ses fondateurs, Jean-Baptiste Michel. (**pour Claire**)

un exemple problématique de cet usage par un historien patenté :

Christophe Charle et Ngram viewer (Cf. Google) in *La discordance des temps: Une brève histoire de la modernité*, Armand Colin, 2014 ;

<https://books.google.fr/books?id=4gUyZnGHiYsC>

Mais...« modernité »

1800-1870 :

https://books.google.com/ngrams/graph?content=modernit%C3%A9&year_start=1800&year_end=1870&corpus=19&smoothing=3&share=&direct_url=t1%3B%2Cmodernit%C3%A9%3B%2Cco

1820-1890 :

https://books.google.com/ngrams/graph?content=modernit%C3%A9&year_start=1820&year_end=1890&corpus=19&smoothing=3&share=&direct_url=t1%3B%2Cmodernit%C3%A9%3B%2Cco

(+ si besoin, exemples sur « armoire à glace » et « trumeau »)

- Gallica

<http://gallica.bnf.fr/advancedsearch?lang=FR>

(Au [12 avril 2014](#), Gallica proposait à la consultation en ligne 3 023 304 documents dont **550 390 livres, 1 347 078 fascicules de presse et revues**, 981 659 images, 44 358 manuscrits, 72 291 cartes, 24 015 partitions, et 3 513 documents sonores – Source Wikipédia)

En caricaturant – un peu – Gallica est excellent quand vous savez déjà ce que vous cherchez, ou que vous pratiquez une recherche spécifique et plus restreinte dans une collection donnée ; sa capacité exploratoire est disons plus faible que Google Book (mieux vaut commencer sur Google, en ayant ainsi affiné un champ lexical par de nouveaux termes environnants le concept ou l'idée initiale de recherche, puis les utiliser sur Gallica) ;

Son mode d'affichage peut parfois – à moins d'une grande patience – décourager certaines requêtes qui se trouvent « cassées » et fait ainsi beaucoup moins ressortir rapidement des résultats plus inattendus, des pistes ou des ouvertures ; l'intérêt majeur de Gallica étant de travailler dans des corpus déjà constitués (tel journal sur telle période, ou une courte période sur un ensemble de titres de presse, etc.).

- **Interne Archive** (a priori, « que » 160 000 livres en français ?)

<https://archive.org/advancedsearch.php>

<https://archive.org/details/texts> ; stats sur les collections (a priori peu de fonds européens, pourtant souvent de bons résultats complémentaires à Google et Gallica)

Préciser les problèmes de la recherche avancée et les moyens de contourner ses défauts

- **Hathi trust**

<http://babel.hathitrust.org/cgi/ls?a=page;page=advanced>

(un « mixte » entre des ouvrages numérisés par Google et donc disponibles sur Google Books, et d'autres plus rares) ; en complément des 3 précédents.

- **GALE (pour Claire)**

<http://access.gale.com/widgets/databases/>

<http://gdc.gale.com/gale-literature-collections/literature-resource-center/>

<http://solutions.cengage.com/Gale/Database-Title-Lists/>

<http://gale.cengage.co.uk/>

Un petit coup de projecteur spécifique sur la presse (et l'avantage de travailler sur des ensembles, des corpus déjà constitués) ?

- les collections de **Gallica**

- **Média 19** (<http://www.medias19.org/>)

Journaux numérisés : <http://www.medias19.org/index.php?id=10576>

Provenant essentiellement de Gallica et de la Bibliothèque municipale de Lyon ;

à l'exception de quelques journaux comme « La Gazette des tribunaux » (<http://www.enap.justice.fr/ressources/index.php?rubrique=108>), mis en ligne par l'ENAP (Ecole nationale de l'administration pénitentiaire) ainsi que le journal lancé par Alexandre Dumas « Le Mousquetaire » (<http://alexandredumas.org/Numeros.php?CF=4>) ; (<http://alexandredumas.org/Corpus/Telechargement>) ; ou encore « Le Petit Comtois » (<http://culture.besancon.fr/ark:/48565/a0112986186328MEpCL>) ; ou encore « Le Rire » (http://collections.citebd.org/lerire/_app/index.php) ; <http://collections.citebd.org/> (Cité internationale de la bande dessinée et de l'Image), surtout pour les dernières décennies du XIXe siècle, puis le XXe. (contient le fonds Cham)

- **L'illustration** : problématique... mais incontournable (1843-1944) pour de nombreuses recherches sur le XIXe, et sur le « background », notamment visuel, diffusé auprès de l'élite et de la classe moyenne (via les cabinets de lecture)

Le site payant <http://www.lillustration.com/> permet de repérer vaguement... PB : ils l'ont modifié récemment : désormais les dates des documents n'apparaissent plus, qui permettaient – avant ! – de repérer les numéros intéressants à consulter en bibliothèque (il y a une formule découverte 24h à 4,90 €) ; en parler à la BNF (pb du possible « deal », collection icono de *l'illustration* en échange de non-numérisation ?)

- La **presse « cachée » de Google** : <https://news.google.com/newspapers>

Très largement anglo-saxon, et pour le francophone, presque uniquement canadien. (exception : « Courrier de Londres » : <https://news.google.com/newspapers?nid=IfoD1wPZYaoC>)

<http://www.theatlantic.com/technology/archive/2011/05/google-shuts-down-newspaper-archive-project/239239/>

New York Times (sur inscription : période d'essai, 4 semaines pour 99 cents !)

<http://timesmachine.nytimes.com/browser>

etc.

Autres bibliothèques (et bases) plus spécialisées
(SOUS RESERVE – l'évoquer mais surtout pour le document de synthèse qui sera envoyé ou mis sur le site de l'URFIST, après la séance)

<https://archive.org/details/sablecentre> ; centre d'études du XIX^e siècle français – Joseph Sablé (Toronto)

<http://www.numerique.culture.fr/pub-fr/index.html> ; patrimoine numérique, catalogue des collections numérisées (920 institutions, 3125 collections)

* <http://www.sitotheque.eu/pages/o-generalites/bibliotheques-numeriques-livres-en-ligne.html>

Sitothèque, la bibliothèque du web ; rubrique « bibliothèques numériques » ; une liste non-exhaustive mais peut donner des idées

* <http://www.wdl.org/fr/>

Bibliothèque numérique mondiale... malgré son titre, de taille très modeste mais ouverture pour des pistes internationales

* <http://www.e-corpus.org/>

<http://www.e-corpus.org/search/index.php>

En particulier, plusieurs journaux régionaux

Bibliothèque numérique de Lyon (numérisation par Google, donc a priori une grande partie se retrouve par Google Book, mais l'intérêt réside dans des possibilités d'interrogation différentes en se limitant à certaines collections, et la presse de la région est un plus)

<http://numelyo.bm-lyon.fr/#>

Pour les thèses anglo-saxonnes depuis milieu XIX^e siècle / sur affiliation

<http://www.proquest.com/search/?searchKeyword=&selectFilter-search=>

Bibliothek für Zeitgeschichte

<http://www.wlb-stuttgart.de/literatursuche/digitale-bibliothek/>

HOPE (Heritage of the people's Europe), histoire sociale et ouvrière depuis la fin du XVIIIe siècle (documents accessibles par Europeana (<http://www.europeana.eu/>) et Social History Portal (<http://www.socialhistoryportal.org/>) / (<http://www.socialhistoryportal.org/browse-collections>)
<http://www.peoplesheritage.eu/fr/index.htm>

Les classiques des sciences sociales (par l'UQAC – Université du Québec à Chicoutimi)

<http://classiques.uqac.ca/>

Archives départementales

CNUM (Conservatoire numérique des Arts et métiers)

<http://cnum.cnam.fr/>

http://cnum.cnam.fr/recherche_avancee/Recherche_avancee.html

Seuls les tables des matières (et index quand il y en a) de nombreuses revues à caractère scientifiques (mais qui peuvent déborder dans les secteurs de l'économie, de l'hygiène au travail ou des arts) sont sous OCR ; en fonction de son sujet de recherche.

The universal digital library (recherche avancée assez limitée... uniquement par titre, auteur ou « sujet »)

<http://www.ulib.org/ULIBAdvSearch.htm>

Blue mountain project (Princeton) / sur les avant-gardes européennes

<http://bluemountain.princeton.edu/index.html>

<http://bluemountain.princeton.edu/bluemtn/cgi-bin/bluemtn?a=cl&cl=CL1&e=-----en-20--1--txt-txIN----->

Digital public library of America (recherche avancée très limitée)

<http://dp.la/>

etc.

En complément bibliographique :

Revue d'histoire moderne et contemporaine supplément 2011-n° 58-4bis.

Le métier d'historien à l'ère numérique : nouveaux outils, nouvelle épistémologie

<https://www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2011-5.htm>

en particulier :

- « Portrait de l'historien-ne en cyborg », par Nicolas Delalande et Julien Vincent

- « L'enquête historique à l'ère numérique », par Philippe Rygiel

-« L'horizon nouveau de l'historiographie expérimentale », par Éric Brian

- « Institutions et pratiques d'archives face à la « numérisation ». Expériences et malentendus », par Yann Potin

- « Faire de l'histoire à l'ère numérique : retours d'expériences », par Franziska Heimbürger et Émilien Ruiz

Pour une veille sur les débats, séminaires, etc. :

<http://calenda.org/search.html?primary=fsubject&fsubject=298&start=20>

cf. la rubrique « Digital humanities », mais aussi « Approches de corpus, enquêtes, archives » et « Méthodes de traitement et de représentation »

Pour info, et plus pour les documentalistes de musées qui seront présentes (mais connaissent sûrement) :

<http://observatoire-critique.hypotheses.org/1952>

ressources numériques et histoire de l'art

(un blog qui n'est plus alimenté)

ANNEXE

Eloge du chaos ?

Google Books au service de la recherche historique, ou éloge du chaos (mai 2012 –in BBF)

Par Valérie Neveu (ancienne chartiste, maître de conférence en bibliothéconomie, et Responsable de la formation Master métiers des bibliothèques, Université d'Angers.

<http://bbf.enssib.fr/consulter/bbf-2012-03-0060-010>

« ... **la masse des données disponibles, abondance inégalée à ce jour** [...]. Dans cet amas de données non hiérarchisées, la confusion peut donner le tournis. Mais, en contrepartie, **des perles** attendent le chercheur en quête de sources nouvelles ou inédites.

Avant tout, il convient de prendre Google Books pour ce qu'il est, un corpus de textes intégraux, que l'on ne pourra exploiter pleinement que si l'on rentre dans le jeu des outils Google. Google Books n'est pas vraiment une bibliothèque numérique, mais un service, autrement dit une application construite autour du moteur de recherche, le cœur de métier de Google qui continue à faire sa force. Le modèle Google reposant sur l'alliance de l'indexation plein texte, du moteur de recherche et de l'affichage ultrarapide des données trouvées reste d'une efficacité redoutable, même si les progrès constants des produits concurrents tendent, jour après jour, à effriter la position dominante du produit américain.

Le chercheur avisé commencera par poser un signet sur la page « Recherche avancée », étape indispensable pour sélectionner les ouvrages numérisés en plein texte (page que Google tend maintenant à cacher au grand public, pour des raisons sans doute liées à sa stratégie commerciale). Puis il fera une utilisation judicieuse des cases « pages contenant ». C'est là que le moteur de recherche donnera sa pleine mesure, pourvu que l'on soit plus rusé que la machine. [...]

Grâce à son océrisation intégrale, et malgré ses erreurs souvent grossières, Google Books reste aujourd'hui supérieur à toute autre bibliothèque numérique partiellement océrisée ou convertie en mode texte, pour la recherche de mots dans les livres anciens, tous pays et tous siècles confondus : rien ne vaut Google pour un balayage efficace et rapide de l'ensemble de la production, au moins dans une première phase de recherche. [...]

Au sein de ces réponses, ce sont surtout les auteurs « petits, obscurs et sans grade » que l'on pourra redécouvrir, car ceux-ci sont numérisés et exposés dans la liste des résultats de recherche au même titre que les grands noms, sans aucune discrimination.[...]. Or les seconds couteaux, que la postérité n'a pas inscrits au catalogue des classiques, ne sont-ils pas d'aussi bons témoins, et parfois de meilleurs, des mentalités de leur temps ? Et la numérisation de leurs œuvres n'offre-t-elle pas une excellente occasion de renouveler le corpus de sources à étudier ?

Je vois un troisième avantage à l'aimable désordre cultivé par Google Books. Dans sa collecte tous azimuts, Google ne recule pas devant des pièces ne relevant pas de la « littérature » dans l'acception commune ; il fait ainsi une large place aux publications officielles, aux brochures techniques, aux publications éphémères publicitaires ou de circonstance, autant de documents précieux pour l'historien. [...]

À mettre enfin au crédit de Google Books : la diversité linguistique. Si le projet initial a suscité des craintes quant à l'hégémonie possible de l'anglais, force est de constater que ce phénomène ne s'observe pas du tout pour les livres anciens. Grâce au non-choix fait par Google, la composition des bibliothèques partenaires, situées dans plusieurs pays, s'en trouve fidèlement reflétée, et la palette de langues représentée permet au chercheur de trouver des sources issues de toutes les cultures européennes... »
