Recherche d'information et humanités numériques : une approche et des outils pour l'historien

Cyrille Suire

Doctorat Université de La Rochelle

THÈSE

pour obtenir le grade de docteur délivré par

École doctorale EUCLIDE

Spécialité doctorale « Informatique et applications »

présentée et soutenue publiquement par

Cyrille Suire

le 13 septembre 2018

Recherche d'information et humanités numériques : une approche et des outils pour l'historien

Directeur de thèse : **Pascal Estraillier**, Professeur, Laboratoire L3i, Université de La Rochelle
Co-directeur de thèse : **Charles Illouz**, Professeur, Laboratoire CRHIA, Université de La Rochelle

Jury

M. Benoist Pierre,	Professeur, Université de Tours	Rapporteur
M. Samuel Nowakowski,	Maître de conférences HDR, Université de Lorraine	Rapporteur
Mme Fatiha Idmhand,	Professeure, Université de Poitiers	Examinatrice
M. Laurent Romary,	Directeur de recherche, INRIA	Examinateur

Université de La Rochelle
Laboratoire Informatique, Image et Interaction (L3i)

Remerciements

Mes remerciements vont tout naturellement d'abord à mes directeurs de thèse, Pascal Estraillier et Charles Illouz. Sans leur confiance, leurs conseils et leur soutien critique, cette aventure interdisciplinaire n'aurait pas pu voir le jour. Je tiens également à remercier les membres du jury, Benoist Pierre et Samuel Nowakowski d'avoir accepté de rapporter cette thèse et Fatiha Idmhand et Laurent Romary de l'avoir examinée.

Je remercie également les membres de l'équipe E-adapt du L3i, pour leur soutien durant ces travaux de thèse et le cadre de travail toujours agréable qu'ils ont su m'offrir. Je remercie tout particulièrement Ronan Champagnat et Mourad Rabah pour leurs conseils et relectures avisés de mes recherches et pour leur accompagnement durant mes enseignements.

À ce titre, je remercie chaleureusement l'équipe du département informatique de l'IUT de La Rochelle, avec qui j'ai appris à enseigner durant mon contrat doctoral et avec qui j'ai eu la chance de vivre une année complète durant mon ATER.

Merci également à tous les membres du laboratoire L3i qui ont participé de près ou de loin à mes travaux. Je tiens à remercier tout particulièrement Mickaël Coustaty et Nicolas Sidère, à qui je dois beaucoup, messieurs, à charge de revanche.

Je ne peux par ailleurs que remercier tous mes collègues et amis doctorants, ceux d'hier et d'aujourd'hui, et leur souhaiter la meilleure réussite possible. Chloé, Damien, Marcela, Guillaume, Christophe, Clément, Élodie, et tout ceux que j'oublie. Quelques mots de remerciements aux amis du bureau 001, Daouda, Nam, Nour, Soraya, Joffrey et Nouredine [1] pour tous ces bons moments. J'adresse une pensée amicale à mes compagnons de galère, Axel, Vincent, Sabrine, finalement nous y sommes arrivés, merci pour tout.

Je salue également le travail de tous les stagiaires qui ont participé aux développements des outils logiciels décrits dans cette thèse, Loïc, Jonathan, Tiago, Hector et tous les autres, à qui j'adresse tous mes vœux de réussite.

Je remercie chaleureusement Jean-Sébastien Noël, pour avoir permis l'organisation des expérimentations de ce travail, pour nos passionnants échanges et toutes les perspectives réjouissantes qui s'ouvrent à nous.

Enfin, et surtout, je ne remercierai jamais assez ma compagne, Annabelle, pour son soutien irremplaçable dans mes hauts et mes bas, et son affection de toujours. Mes derniers mots iront à mes enfants, Gabriel, né au début de cette thèse, qui trouverait sans doute, et j'en suis désolé, que ce manuscrit manque de camions de pompiers, et Élise, dont la naissance est venue émerveiller la fin de ces travaux.

1. Nouredine, à qui je ne peux que dédier une note de bas de page, lui qui a tant œuvré à en limiter le nombre dans le présent manuscrit. Pour tous nos fascinants débats, ton rôle de *bad reviewer* et tes précieux conseils, merci.

Résumé

Les travaux de cette thèse portent sur les conséquences du développement du numérique sur la pratique de recherche en SHS au sens large et en histoire en particulier. L'introduction du numérique bouleverse les pratiques de recherche en histoire en mettant à disposition du chercheur un grand volume de sources numérisées ainsi que de nombreux outils d'analyse et d'écriture. Si ces nouveaux moyens de recherche permettent à la discipline d'adopter de nouvelles approches et de renouveler certains points de vue, ils posent également des questions sur les plans méthodologique et épistémologique.

Devant ce constat, nous avons choisi d'étudier plus en détail l'impact des outils de recherche d'information, bibliothèques numériques et moteurs de recherche de sources sur l'activité de recherche en histoire. Ces systèmes offrent un accès à un grand volume de documents historiques mais leur fonctionnement repose sur des traitements informatiques pour la plupart invisibles aux yeux des utilisateurs, qui peuvent ainsi s'apparenter à des boîtes noires. L'objectif principal de cette thèse est donc de donner les moyens aux utilisateurs d'observer et de comprendre ces processus dans l'optique de leur permettre d'en intégrer les effets de bord à leur méthodologie.

Afin de mieux positionner notre objet d'étude, nous proposons un cadre conceptuel reposant sur la notion de ressource numérique. Ce concept représente les systèmes numériques que nous étudions au sein de leur contexte d'usage, de production et d'exécution, il fait le lien entre des usages attendus par les utilisateurs et des choix méthodologiques ou techniques issus des présupposés de ces concepteurs. Sur la base de ce cadre conceptuel, nous proposons une analyse des bibliothèques numériques et moteurs de recherche de sources en fonction de chacun des contextes.

Ainsi, notre étude propose une analyse des usages de ce type de ressource numérique dans le cadre d'une recherche en histoire en adoptant une démarche expérimentale et en produisant des indicateurs de la pratique. Ces indicateurs sont ensuite croisés avec le fonctionnement du système, dans ces contextes de production et d'exécution, pour en révéler les biais méthodologiques. À l'issue de ces analyses, nous proposons un réinvestissement de ces résultats sous la forme d'un outil logiciel dédié à l'enseignement d'une approche critique de la recherche d'information en ligne pour les apprentis historiens.

Ces travaux sont évalués par une démarche expérimentale. Elle est construite sur la base d'un prototype d'observation du comportement des utilisateurs en situation de recherche d'information et des outils de démonstration des biais associés au fonctionnement des processus informatiques impliqués lors des phases de production des contenus et d'exécution du système. Ce prototype a fait l'objet de plusieurs phases d'expérimentation liées à son développement, l'évaluation de ces fonctionnalités et de son impact sur la pratique dans un contexte de formation.

Abstract

The work of this thesis focuses on the consequences of digital technology development on research practice in the humanities in the broad sense and particularly in history. The introduction of digital technology disrupts historical research practices by making available to the researcher a large volume of digitized sources as well as numerous analysis and writing tools. These new capacities of research allow the discipline to adopt new approaches and renew certain points of view, but they also raise methodological and epistemological questions.

Given this observartion, we have chosen to study in more detail the impact of information retrieval tools, digital libraries and search engines on historical research activity. These systems offer access to a large volume of historical documents but they depend on computer processes that are mostly invisible to users and acting as black boxes. The main objective of this work is to give users the means to observe and understand these processes in order to allow them to integrate their side effects in a suitable methodology.

In order to better position our object of study, we propose a conceptual framework based on the notion of digital resource. This concept represents the digital systems that we study within their contexts of use, production and execution. It connects uses expected by users and methodological or technical choices based on the assumptions of system designers. Based on this conceptual framework, we propose an analysis of digital libraries and historical sources search engines according to each context.

Thus, our study proposes an analysis of the uses of this type of digital resource within the framework of a research in history. The study adopts an experimental approach and produces indicators of the practice. These indicators are then crossed with the functioning of the system, in its contexts of production and execution, to reveal the potential methodological biases. Following these analyzes, we propose a reinvestment of these results in the form of a software tool dedicated to teaching a critical approach to online information retrieval for student in history.

This work is evaluated by an experimental approach. It is built on the basis of a prototype of observation of the behavior of the users when they are looking for information. Our experimental approach is also based on demonstration tools of the biases associated with the functioning of the computer processes involved during the contexts of production and execution. This prototype has been the subject of several experimental phases related to its development, the evaluation of these features and its impact on practice in a training context.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

De très nombreux chercheurs utilisent désormais intensément les technologies et médias numériques pour leur recherche. Dans un texte récent, intitulé *Roy's World*, l'historien Dan Cohen nous fait part des sentiments de Roy Rosenzweig, pionnier de la *Digital History*, alors qu'il travaillait ardemment, dans les années 1990, à la création d'un CD-ROM interactif sur l'histoire américaine. Roy Rosenzweig se plaignait alors de la réaction de ses collègues universitaires qui ne voyaient dans cet objet qu'une démarche « superficielle et cartoonesque », bien loin du sérieux d'une monographie académique. Presque trois décennies plus tard, c'est sur son blog [Cohen, 2017] que Dan Cohen nous relate cette courte histoire et nous montre, par l'exemple, que les choses ont bien changé.

L'usage des technologies numériques pour la recherche en Sciences Humaines et Sociales au sens large et en histoire en particulier n'est désormais plus une affaire de spécialiste et concerne tous les chercheurs. Qu'il travaille seul ou en équipe pluridisciplinaire, le chercheur en histoire est nécessairement au prise avec les effets d'une informatique omniprésente de laquelle il n'est plus possible de s'extraire. Le numérique est devenu un aspect important des champs culturels ou académiques, autour desquels gravite l'activité de l'historien (voir figure 1.1, extraite de [Rygiel, 2012]).

Résumé de manière très schématique, et donc largement critiquable, le processus d'investigation de l'historien débute par l'établissement d'un corpus de sources primaires ou secondaires à même de permettre l'étude de sa question de recherche initiale. S'en suit une phase de dépouillement des sources, puis une phase de réflexion durant laquelle se forge le discours historique. Enfin, vient le temps de l'écriture, de la dissémination et de la communication des résultats produits par ce travail. Des allers et retours sont fréquents naturellement, au gré de l'évolution de la réflexion, de la découverte de matériau nouveau, etc. Pour chacune de ces phases entrent désormais en jeux des outils numériques.

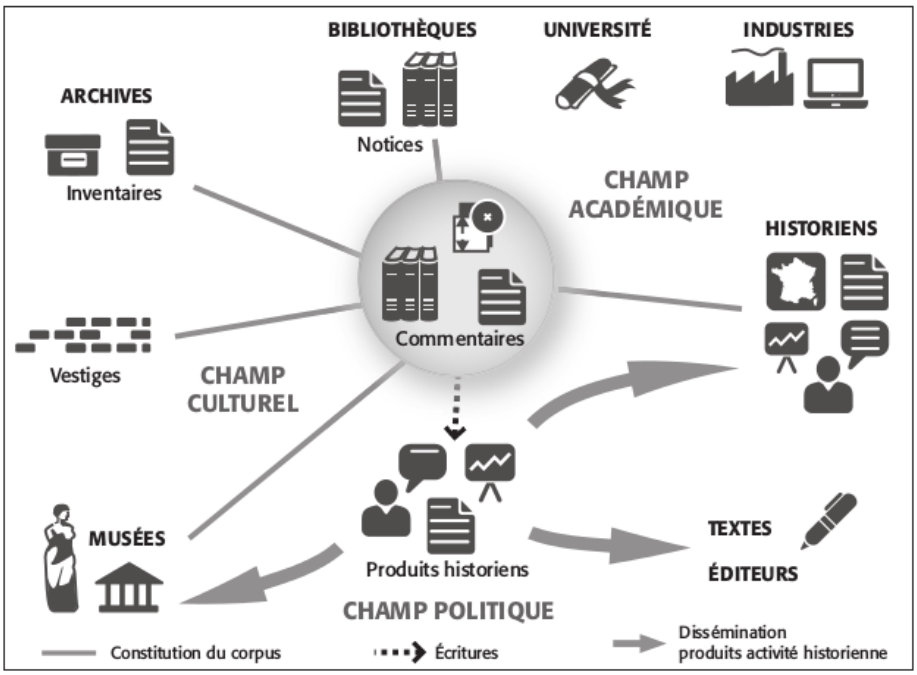
— **Collecte et organisation de la documentation** : De nombreux outils sont impliqués durant cette phase. L'appareil photo numérique permet des prises de vues en archives, les moteurs de recherche de documentation secondaire donnent un accès à la littérature scientifique, les bases de données et bibliothèques numériques offrent quant à eux l'opportunité de consulter de très nombreuses sources primaires, les logiciels de gestion des données permettent d'organiser les contenus recueillis. Ces outils sont soutenus par plusieurs domaines de l'informatique : numérisation et traitement d'images, indexation et gestion des métadonnées, recherche d'information, interfaces homme-machine ;

— **Analyse et réflexion** Durant la phase d'analyse sont employés des outils qui dépendent bien entendu des sources et de l'approche. Il peut s'agir de logiciels de visualisation d'images, d'outils de traitement des données en série, d'analyse de texte et de lecture distante ou de cartographie ;

— **Écriture et production du discours historique** Durant l'écriture peuvent être mis en œuvre des outils d'édition de texte, mais également des outils d'écriture collaborative ou de publication en ligne ainsi que des outils de visualisation des données.

Ces outils numériques sont donc désormais bien installés dans le paysage des pratiques de l'historien. Certains d'entre eux sont des outils spécifiques, conçus et développés à

FIGURE 1.1 – « Ceci est un historien », schéma des champs d'intervention de l'activité et des produits de l'historien, extrait de [RYGIEL, 2012].



destination des chercheurs ou des étudiants. Ils ne posent pas de problèmes particuliers en dehors de l'acquisition des compétences nécessaires à leur utilisation. C'est par exemple le cas des gestionnaires de bibliographie tel que Zotero [1] qui sont des moyens pratiques et efficaces pour importer, classer et exporter les références bibliographiques.

D'autres outils largement utilisés par les historiens sont en revanche des outils génériques. Ils n'ont pas été conçus pour un usage de recherche et leurs conséquences sur la pratique restent, pour l'essentiel, à étudier. C'est par exemple le cas des outils ou des sites web permettant de générer, avec très peu de connaissances techniques, des représentations graphiques particulièrement abouties sur le plan visuel. Ce type d'outils, à l'instar du célèbre Google NGram Viewer [2], est très utile pour appuyer un discours ou stimuler l'imagination, mais comme l'expliquait [LAMASSÉ et RYGIEL, 2014], n'est souvent « conforme aux canons d'aucune discipline existante, ni même à certains des principes de la démarche scientifique, du fait en particulier des problèmes tenant à la définition des corpus de références ».

En ce qui concerne le cas particulier des moteurs de recherche, objet central de nos travaux, l'historien Ian Milligan écrivait dans un article de 2014 : « Je demande souvent aux historiens qui affirment ne pas être des historiens du numérique s'ils effectuent leur recherche dans Google et, le cas échéant, s'ils connaissent le fonctionnement de PageRank » [MILLIGAN, 2014]. La question est légitime, si les moteurs de recherche généralistes constituent un point d'entrée naturel vers les contenus du web, leur usage dans un contexte scientifique doit être interrogé. Au centre de la pratique de l'histo-

1. Zotero est un outil de gestion bibliographique *open source* très connu, développé par le Roy Rosenzweig Center for History and New Media (RRCHNM), disponible à l'adresse https://www.zotero.org

2. Qui permet d'observer la fréquence d'usage d'un ou plusieurs termes dans le gigantesque corpus Google Books, accessible à l'adresse : https://books.google.com/ngrams

rien, se posent désormais des questions sur l'environnement, matériel ou logiciel, et le contexte de travail.

1.1 Questionnement scientifique : le numérique, acteur épistémologique, acteur méthodologique

L'introduction du numérique dans les pratiques de recherche des historiens ne s'est pas faite en un jour et n'a naturellement pas manqué d'interroger la discipline. Les historiens réfléchissent depuis longtemps, en témoignent déjà les travaux de Roy Rosenzweig [Rosenzweig, 2003], sur les conséquences du développement du numérique sur leur environnement de travail, leurs sources, leurs méthodes ou leurs écrits. Le numérique est un acteur épistémologique et méthodologique dans la mesure où son introduction dans le processus de recherche a des conséquences, à ces deux niveaux, sur la production de la connaissance historique.

Le numérique modifie en particulier les conditions d'accès aux sources de l'historien. Il faut mesurer ici l'importance particulière de la source pour l'historien. De nombreux auteurs, dont certains des plus grands historiens, ont écrit sur la relation entre l'historien et ses sources, parmi lesquels Paul veyne ou Marc Bloch [Bloch, 1949; Veyne, 1971]. Cette relation est d'abord fondée sur la nature de la démarche scientifique de l'historien, la source est un moyen de transmission des témoignages qui permet de répondre à la question posée par Marc Bloch dans *Apologie pour l'histoire*, « comment puis-je savoir ce que je vais dire ? ». Mais ce rapport au document dépasse la seule information de l'historien. Il est un lien presque physique qui prend corps dans des lieux et dans des temporalités propres à chaque recherche historique, comme l'a magistralement expliqué Arlette Farge dans *Le goût de l'archive* [Farge, 1989]. Dans cette ambiance et dans cette durée, il n'est pas seulement question de rassembler de la documentation et annoter des sources, c'est également là que se forge le discours historique. Modifier les conditions d'accès aux sources, c'est inévitablement bouleverser les conditions du travail de l'historien.

Parmi les études qui montrent les conséquences du numérique sur le plan épistémologique, celle de Lara Putnam, datée de 2016 et intitulée *The Transnational and the Text-Searchable : Digitized Sources and the Shadows They Cast* est particulièrement significative [Putnam, 2016]. L'auteur constate, tout comme nous, l'importance prise par les outils numériques dans l'activité quotidienne de l'historien, pas nécessairement pour des usages nouveaux, mais bien pour effectuer des tâches qui ont toujours été celles de l'historien. Si peu d'historiens utilisent à l'heure actuelle des approches *big data*, une large majorité, en revanche, accèdent aux sources et les consultent en ligne.

Face à ce constat, l'auteur pose la question de la géographie des sources. Là où il fallait régulièrement se contenter de la documentation accessible dans les institutions locales ou nationales de son lieu d'étude, il suffit désormais de se connecter à une interface en ligne pour accéder à des sources issues d'aires géographiques plus éloignées. L'auteur défend l'idée que cet accès rapide aux sources du monde entier favorise l'émergence de problématiques historiques et de points de vue transnationaux. Les conditions matérielles encouragent ainsi l'historien à élargir ses horizons et à penser son objet de recherche au delà des frontières traditionnelles. L'étude est conclue par une série de

questions. Si le numérique semble faire tomber des barrières et permet d'adopter de nouveaux points de vue, rien ne garantit que ces effets stimulants ne soient pas amoindris par une trop grande superficialité de la recherche. Nous pouvons mesurer, à travers une telle réflexion, les questions que soulève le numérique pour l'épistémologie de la discipline.

Si le numérique peut se traduire par des conséquences sur le plan épistémologique alors même qu'il n'est utilisé, dans la grande majorité des cas, que pour des tâches classiques et quotidiennes de l'activité de l'historien, nous pouvons nous demander s'il n'intervient pas également sur le plan de la méthode. Lorsque l'historien prend des photos en archive, cherche de l'information via les moteurs de recherche ou consulte des sources numérisées, cela a-t-il un impact sur ses productions ?

S'il ne fallait retenir qu'une étude qui montre des impacts de cette nature, il faudrait probablement choisir celle de Ian Milligan, intitulée *Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997–2010*, publiée en 2013 [Milligan, 2013]. Dans ce texte, l'auteur recense le nombre de citations de certains journaux dans les thèses d'histoire du Canada contemporain entre 1997 et 2010. Ce décompte précis donne des résultats frappants, le *Toronto Star* est cité 74 fois dans les 67 thèses de 1998, il est cité 753 fois dans les 69 thèses de 2010. Dans l'intervalle, explique Ian Milligan, s'opère un changement majeur. En 2010, l'historien n'allume plus le lecteur de microfilm pour parcourir les journaux, page après page, il saisit quelques mots clés dans un moteur de recherche et obtient immédiatement une liste de résultat qu'il peut parcourir sur son écran.

Si l'étude s'était arrêtée là, elle aurait seulement montré à quel point les moteurs de recherche de journaux étaient devenus des ressources centrales, mais elle a été plus loin. Elle a montré que si certains titres étaient nettement plus cités qu'auparavant, d'autres l'étaient beaucoup moins, parce qu'ils avaient été numérisés dans une moindre qualité où parce que l'interface de consultation était moins aisément accessible [3] Lorsque ces titres de presse reflètent des positionnements politiques différents [4], l'impact sur la production du discours historique est majeur. L'usage des moyens numériques de recherche et de consultation des sources a donc des conséquences qui semblent parfois devenir des biais. Ces derniers requièrent donc un positionnement méthodologique adapté à la fois aux problématiques historiques et aux outils disponibles.

Les deux études que nous venons de citer ne sont pas les seules à s'intéresser à ces questions. Qu'il s'agisse de retour d'expériences impliquant le numérique pour les SHS [Grandi et Ruiz, 2012], de montrer que celles-ci seraient en « retard » sur l'étude des problématiques du numérique [Wieviorka, 2013], pour mettre avant un nécessaire point de vue critique [Bourdeloie, 2014] ou pour étudier le renouvellement méthodologique qui semble s'imposer [Mounier, 2015], toutes mesurent un bouleversement profond. Ces études témoignent d'une grande diversité de points de vues et sont riches de questionnements. Elles nous enseignent finalement que les conséquences du développement du numérique dépassent de loin la question strictement instrumentale des nouveaux outils de recherche permis par l'informatique. Ces nouveaux moyens font

3. Une étude similaire a récemment été menée en Grande-Bretagne et montre des phénomènes similaires [Gooding, 2017].

4. En l'occurrence des positionnements se situant plus ou moins en faveur ou défaveur des gouvernements fédéral ou provincial du Canada.

certes gagner du temps et ouvrent de nouvelles perspectives, mais ils ont en parallèle des conséquences, voire des biais, qui ne peuvent être négligés. L'usage des outils numériques pour la recherche historique n'est ainsi pas si neutre qu'il y parait.

Il serait toutefois illusoire de penser qu'il puisse exister une réponse unique à ces questionnements compte tenu de la grande diversité des problématiques, approches, sources et outils de l'historien. Il serait également vain de chercher des réponses strictement informatiques, la communauté des historiens est la seule à même d'intégrer à sa démarche et à sa méthodologie les biais dont nous venons de parler. Cependant, il faut qu'elle en ait les moyens, les moyens de comprendre, de modeler et de configurer les outils pour, si l'on reprend les mots de [Lamassé et Rygiel, 2014], inscrire en leur cœur ses propres normes.

1.2 Problématique : impact des outils numériques de recherche d'information sur la pratique

Sur la base de ce constat, nous avons choisi d'orienter nos travaux vers le cas particulier des bibliothèques numériques et des moteurs de recherche de sources primaire ou secondaire. Ces outils sont intéressants parce qu'ils sont centraux pour l'historien en tant que passerelle principale vers les sources numérisées. Leur usage est donc très régulier, à tous les stades de la recherche [Audenaert et Furuta, 2010; Kemman et collab., 2013].

Même si elle peut s'appliquer à d'autres contextes, notre proposition de recherche est focalisée sur un cadre précis. L'utilisateur, historien ou étudiant en histoire dans notre cas, qui utilise un moteur de recherche pour accéder à des sources. Le processus classique, représenté en noir sur la figure 1.2 est le suivant. L'utilisateur formalise son besoin d'information par une requête composée de mots-clés ou de filtres qui est transmise au moteur de recherche qui peut ainsi produire des résultats de recherche. L'utilisateur est libre, sur la base de ces résultats, de faire évoluer le besoin d'information et donc de produire une nouvelle requête.

Le problème est clair, le moteur de recherche est une boîte noire. L'utilisateur n'a, en dehors des résultats de recherche, aucun moyen d'évaluer la pertinence de sa stratégie de recherche, il ne lui est par exemple pas possible de savoir comment sa requête a été traitée ni comment ont été classés les résultats. La progression de son processus de recherche d'information ne peut se faire qu'à l'aveugle, sur la seule base des résultats produits dont il ignore aussi bien la pertinence que l'exhaustivité. Ces effets de boîtes noires sont causés par des choix de conception qui trouvent leurs origines dans différentes étapes des processus qui conduisent les sources physiques vers l'environnement numérique :

— La dématérialisation : qui vise à numériser les contenus lorsqu'ils ne sont pas nativement numériques et qui impliquent des phases d'extraction de l'information, par exemple par reconnaissance optique de caractères ;
— L'enrichissement de l'information : qui vise à documenter, par l'entremise de métadonnées, les contenus numérisés et qui suppose des enjeux de modélisation et de gestion de ces métadonnées ;

— La recherche d'information : dont l'objectif est de rendre accessible les contenus numériques et qui emploie une algorithmique complexe pour traiter les requêtes des utilisateurs et calculer des résultats de recherche.

Avant d'être disponibles dans une bibliothèque numérique, les contenus que les historiens sont habitués à manipuler, sources primaires comme secondaires, subissent donc de nombreux traitements. Les contenus initiaux peuvent rapidement s'effacer au profit de représentations intermédiaires, éventuellement surchargées d'informations complémentaires, sous forme de métadonnées ou de notes. Le travail de l'utilisateur, lorsqu'il cherche de l'information ou des sources s'opère le plus souvent sur de telles représentations. Les algorithmes de recherche et de classement de l'information travaillent par exemple sur des métadonnées et des extractions des sources, sous la forme de texte la plupart du temps, qui peuvent s'avérer incomplètes, dégradées ou décontextualisées.

Les conséquences de cette informatique et des choix de conception sur le travail de l'historien sont loin d'être anecdotiques. Quelques études ont permis de faire progresser notre connaissance de leurs impacts. Outre l'étude déjà citée de Milligan [2013], des recherches ont permis d'identifier des sources de biais majeures. A titre d'exemple, l'étude de Traub et collab. [2015] s'intéresse aux effets des erreurs de reconnaissance optique de caractères dans la presse ancienne. Elle nous rapporte que des historiens interrogés dans le cadre de cette recherche ont rapporté avoir, par le passé, publié des travaux fondés sur des analyses quantitatives de données issues de texte, mais qu'ils ne le feraient sans doute plus aujourd'hui, ayant pris conscience des biais induits par les technologies nécessaires à de telles analyses.

Pour les utilisateurs, les conséquences des biais dont nous parlons sont complexes à appréhender parce qu'ils sont pour la plupart invisibles, alors qu'ils impactent les sources essentielles du travail de production de la connaissance historique. Notre proposition vise donc à limiter les effets de ces biais. Pour ce faire, nous proposons de fournir à l'historien les moyens de mieux s'approprier le système qu'il utilise. Ces moyens se matérialisent sous deux formes, représentées en rouge dans la figure 1.2 :

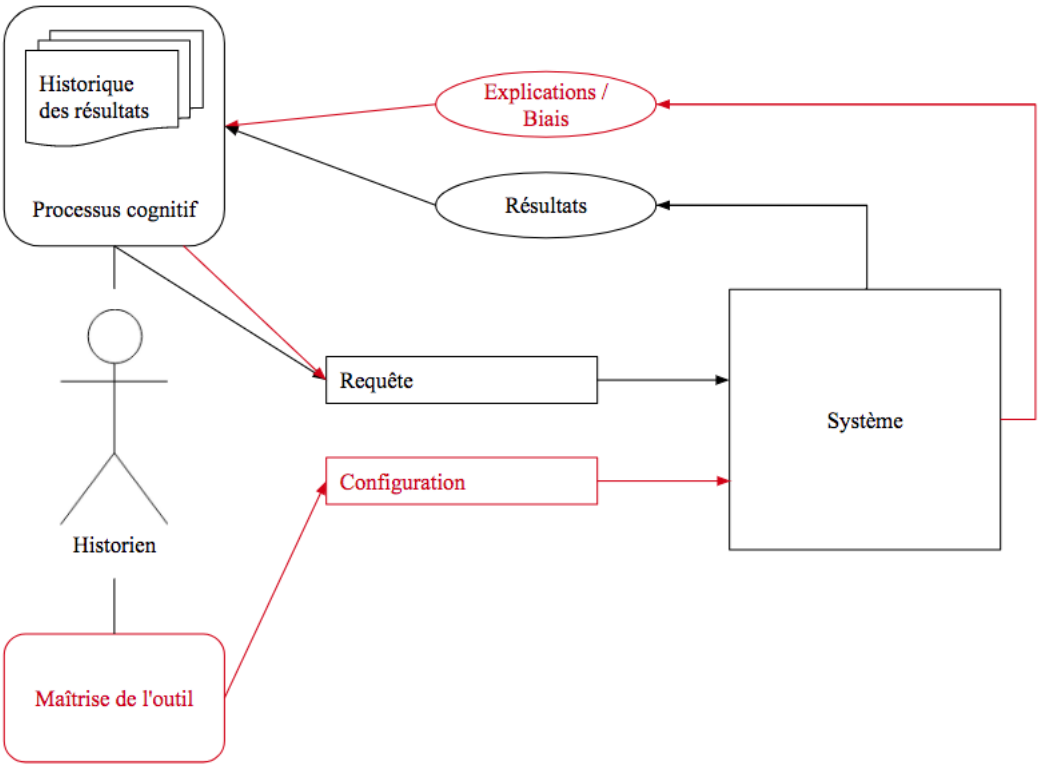
— De l'information complémentaire, accompagnant les résultats de recherche et visant à permettre à l'utilisateur de comprendre les éventuels biais produits par le système qu'il utilise.

— Des moyens d'action et de configuration qui visent, quant à eux, à permettre à l'utilisateur de développer sa maîtrise du système et sa stratégie de recherche en associant requêtes et éléments de configuration.

Notre proposition de recherche s'inscrit dans différentes thématiques de recherche. Du côté du système, et dans le domaine de recherche de l'informatique, il s'agit essentiellement des thématiques de recherche d'information (RI) et d'analyse de documents. Au sein de la RI, notre démarche emploie des résultats scientifiques de l'*information retrieval*, thématique focalisée sur les moyens théorique et technologique de la recherche dans des corpus, mais également de l'*information seeking* qui associe les aspects humains à l'environnement technologique. L'analyse de documents est par ailleurs une thématique de recherche importante dans notre contexte, dans la mesure où elle est à la base de nombreux traitements liés à la dématérialisation.

Du côté de l'utilisateur, notre démarche est bien entendu proche des développements de l'histoire numérique, en anglais *digital history* qui fait référence, dans son acception la

FIGURE 1.2 – Représentation de notre proposition de recherche



Le processus classique d'utilisation d'un moteur de recherche de source est représenté en noir, les éléments de notre proposition de recherche le sont en rouge.

plus large, aux formes de recherche historique qui emploient des sources primaires numérisées ou nativement numériques. Cet espace de réflexion propose aussi bien des travaux de recherche académique que des contributions méthodologiques et pédagogiques. Ces dernières, qui font partie des spécificités de l'histoire numérique dans l'ensemble des humanités numériques [ANDERSEN, 2002; ROBERTSON, 2016], sont fondamentales pour notre approche, dans la mesure où nous traitons largement du problème de la compétence et de la maîtrise technique des systèmes. Dans le même ordre d'idée, la thématique de l'*information literacy*, développée aussi bien en informatique et sciences de l'information, qu'en sciences humaines éclaire également les capacités des utilisateurs à comprendre et utiliser l'information.

Si l'état de l'art de ces différentes thématiques scientifiques apporte bien entendu des résultats intéressants pour aborder notre problématique, il souffre également de certaines limites. A titre d'exemple, les processus informatiques à l'œuvre dans les moteurs de recherche ou les bibliothèques numériques, comme la reconnaissance optique de caractère, sont largement étudiés [HOLLEY, 2009; TANNER et collab., 2009]. Ils le sont toutefois rarement dans l'objectif de confronter pratique et choix de conception mais presque exclusivement dans une perspective strictement informatique, visant à améliorer les performances techniques. Les études qui adoptent une perspective utilisateur sont quant à elles souvent focalisées sur des problématiques d'interface, dont l'objectif est souvent limité à l'amélioration de l'expérience utilisateur [XIE, 2006, 2008]. Cet enjeu de simplification et de facilitation de l'accès au système conduit souvent à en dis-

simuler les détails de fonctionnement, limitant la capacité du chercheur à se l'approprier et renforçant ses effets de boîte noire.

De la même manière, les développement de l'histoire numérique ou de l'*information literacy* ont permis l'émergence d'initiative pédagogique, visant le transfert de compétences entre les différents membres de la communauté, en témoigne par exemple l'excellent site *The programming Historian*, qui propose un grand nombre de très bons articles expliquant le fonctionnement et encourageant l'usage d'outils numériques de recherche [Afanador-Llach et collab., 2017]. Il n'existe toutefois pas d'étude, à l'exception de celle de [Jackson et collab., 2016] dans un contexte de *Big Data* qui investiguent le cas particulier des moteurs de recherche dans le cadre de l'histoire numérique.

Notre travail a donc pour objectif de contribuer à l'étude de ce cadre particulier, dans le contexte de l'histoire numérique. Il vise à adopter un point de vue global entremêlant autant que possible les versants utilisateur et système de la problématique dans une démarche conceptuelle globale et à répondre aux questions suivantes :

— Sur le plan conceptuel, comment croiser les exigences de la démarche scientifique de l'historien avec les effets des choix de conception et des compromis techniques opérés par l'informatique des moteurs de recherche de sources ?

— Du côté du système, bibliothèque numérique et moteur de recherche, par quels moyens mettre au jour et expliquer à l'historien les biais induits par les processus informatiques impliqués depuis la numérisation jusqu'à la recherche d'information ?

— Sur la base de ces moyens, par quelle démarche favoriser une meilleure maîtrise du système et une meilleure appropriation méthodologique des outils de recherche de sources en ligne ?

1.3 Approche de recherche : observer, éclairer, réinvestir

Comme tout travail académique, notre approche requiert, dans un premier temps, un effort de positionnement. Nous proposons une lecture des bibliothèques numériques à travers un cadre conceptuel qui nous permet de positionner les processus informatiques qui nous intéressent entre les contenus et les utilisateurs, autrement dit, dans notre contexte, entre les sources et l'historien.

Cadre conceptuel

Ce cadre conceptuel repose sur la notion de **ressource numérique**. Elle se définit comme un système informatique fondé sur des données, ici les sources numérisées ou nativement numériques, qui fournit un service par l'encapsulation d'outils de traitement, un moteur de recherche ou un outil de visualisation de données par exemple et qui est fortement dépendant du contexte dans lequel il est utilisé [Picard, 2011; Sawadogo, 2016]. Le concept de ressource numérique représente donc un système, au sein de son environnement, qui fait le lien entre des usages attendus par les utilisateurs

et des choix méthodologiques ou techniques issus des présupposés de ces concepteurs. Il est donc un cadre d'analyse pertinent pour l'évaluation globale des systèmes numériques dans leur écosystème, des logiques de conception initiales, jusqu'aux usages finaux.

Les bibliothèques numériques peuvent être situées dans ce cadre conceptuel. Elles fournissent un service dont les résultats dépendent de nombreux facteurs contextuels. Le schéma en figure 1.3 représente un système de recherche d'information sous la forme de ressource numérique, environné des contextes suivants :

— **Le contexte d'usage** représente la stratégie de recherche des utilisateurs, leurs modes de navigation et la pertinence de leurs requêtes. Il est déterminant pour la qualité des résultats ;

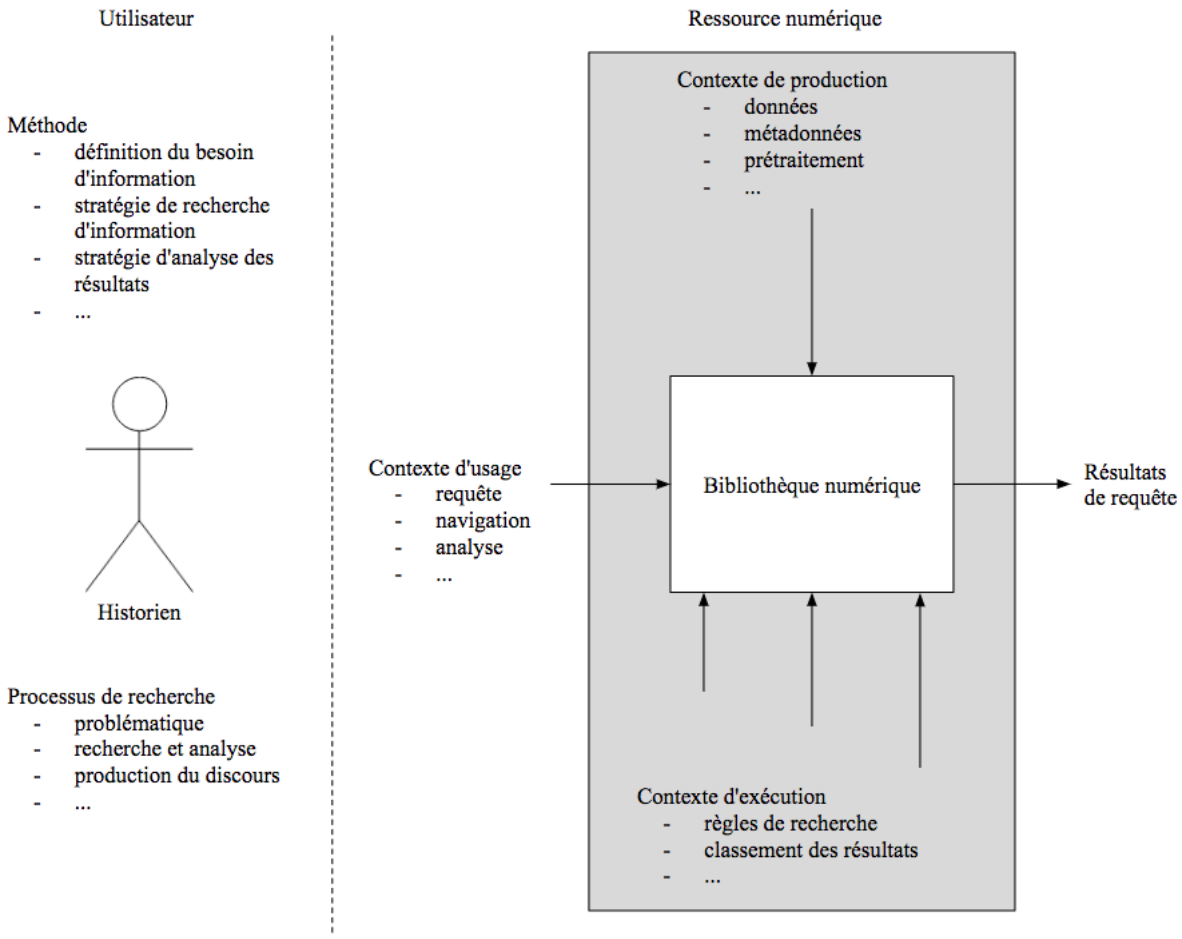
— **Le contexte de production** dépasse les seules qualité et typologie des contenus initiaux, il est constitué du résultat des nombreux choix de conception opérés lors de la phase de production des contenus. Il a naturellement un impact majeur sur les résultats de la recherche, qui varient en fonction de la qualité et de la pertinence des données et des métadonnées ;

— **Le contexte d'exécution** inclue tous les traitements qui se situent entre les deux contextes précédents et qui manipulent les données. Ces traitements sont eux aussi fortement dépendants de choix de conception et ont un impact important sur les résultats. Les règles et les algorithmes de recherche et de classement des résultats font partie des principaux traitements de ce contexte.

Cet ensemble conceptuel est important car il nous permet de représenter et positionner les différents acteurs de notre problématique et la questionner. Les éléments représentés au milieu du schéma de la figure sont ainsi ceux qui sont généralement invisibles pour l'utilisateur. De son point de vue, il n'existe que très peu de moyen de prendre connaissance des processus situés entre ce qu'il fournit au système, le contexte d'usage, et les résultats. Ce manque d'information se traduit par une incertitude du fonctionnement du système. Pour éclairer ces effets de boîte noire et contribuer à lever cette incertitude, il convient de s'intéresser à tous les facteurs qui influencent le système et mesurer leur impact respectif. Cependant, pour permettre au public de s'approprier ces effets et apporter une réelle plus-value, les mesurer indépendamment ne suffit pas. La démarche a mettre en place doit permettre de les donner à voir et les expliquer.

La démarche de recherche décrite dans ce manuscrit vise donc à proposer et mettre en place un cadre méthodologique et un outil permettant de confronter différentes instances des contextes et des traitements qui tiennent une place importante dans les bibliothèques numériques et les moteurs de recherche. Il ne parait en effet pas possible, pour des raisons évidentes, d'expérimenter et étudier tous les biais et effets potentiels des systèmes de recherche pour tous les types de sources et tous les champs de recherche de l'histoire, tant ceux-ci sont diversifiés. Ce qui nous semble possible, en revanche, c'est de mettre en place une démarche générique, capable de révéler les effets dont nous parlons pour une large variété de contexte d'usage, de production et d'exécution. Cette démarche est fondée sur une approche globale visant à étudier les effets des différents contextes sur le fonctionnement de la ressource et à trouver les moyens de réinvestir les résultats produits pour en limiter les impacts.

FIGURE 1.3 – Représentation du système de recherche d'information associé à ces contextes d'usage et de production



Positionnement de l'utilisateur et de la ressource numérique. Celle-ci est composée du système environné de ces différents contextes, les contextes producteurs d'effets de boites noires sont représentés sur fond gris.

Observer et décrire l'usage du système de recherche d'information

L'objectif de cette facette du travail est de se doter d'une méthodologie d'observation et d'explication du contexte d'usage. Elle doit permettre de comprendre comment les utilisateurs d'un système de recherche d'information, historiens ou étudiants en histoire dans notre cas, recherchent l'information qui est nécessaire à l'accomplissement de leur tâche. L'état de l'art fournit déjà des modèles de comportement de recherche d'information adaptés aux publics que nous visons, obtenus par des moyens d'enquêtes quantitative ou qualitative [MEHO et TIBBO, 2003; RHEE, 2012]. Ces modèles ne proposent toutefois pas de méthode pour mesurer directement la pratique dans les systèmes de recherche d'information manipulés par notre public. Ils ne permettent donc pas de décrire suffisamment finement les usages pour qu'ils puissent être confrontés avec le fonctionnement du système et en faire émerger les biais. Les objectifs de cette partie de notre travail requièrent donc la construction d'un système d'observation et une approche de validation expérimentale à même de fournir des indicateurs pertinents de la

pratique.

Éclairer les contextes de production et d'exécution

Si les pratiques des utilisateurs sont nombreuses, les contextes de production et d'exécution qui sont appliqués aux données le sont tout autant. Pour être parfaitement exhaustif, il faudrait réfléchir et proposer des solutions pour mettre en lumière tous les traitements effectués. Cette tâche dépasse naturellement notre travail et requiert de limiter l'étude. Nous avons ainsi choisi d'étudier plus en détail l'un de ces traitements, en l'occurrence le problème des erreurs de reconnaissance de caractères sur les entités nommées [5]. Il s'agit en effet d'un des processus informatiques produisant sans doute le plus de biais et qu'il est par ailleurs très difficile de révéler par des moyens automatiques, sans recourir à une vérité terrain [Chiron et collab., 2017; Traub et collab., 2015]. Ce cas d'étude a vocation à montrer le plein potentiel de la démarche et l'importance des contributions de l'informaticien pour la création de ressources numériques pertinentes pour l'histoire.

Réinvestir les résultats de l'informatique pour la pratique et la méthodologie de recherche

Si l'informatique peut sans doute trouver des moyens de montrer et d'expliquer les effets de bord des systèmes de recherche d'information, seul l'historien est à même de les intégrer dans sa méthodologie et d'adopter, vis à vis d'eux, le positionnement critique qui fait toute l'essence de sa recherche. Analyser les contextes environnant la ressource n'a ainsi que peu d'intérêt si aucune solution n'est proposée pour permettre au public de les visualiser et de les comprendre. Les résultats des parties précédentes de notre approche se présentent sous forme d'indicateurs chiffrés ou de représentations visuelles. À elles seules, elles ne suffisent pas à permettre d'éclairer significativement les boites noires de la recherche d'information dans le cadre d'une démarche historique. Autrement dit, elles ne permettent pas d'être réinjectées à l'état brut dans le contexte d'usage, pour fournir une aide méthodologique à la recherche.

Il est donc fondamental de travailler à la production d'outils, sous forme de fonctionnalités intégrables à un système de bibliothèque numérique, ou sous forme autonome, qui permettent ce retour réflexif. Les biais que nos résultats informatiques permettent de révéler doivent pouvoir être expérimentés par les chercheurs et étudiants pour qu'ils puissent être en capacité de se forger une approche critique de l'usage de la recherche d'information en ligne. Il est donc nécessaire de proposer une démarche capable de permettre un réinvestissement des résultats précédents sur le terrain méthodologique. Cette démarche a vocation à répondre aux besoins suivants :

— montrer les biais créés par les processus à l'œuvre dans les différents contextes de la ressource aux utilisateurs ;

— expliquer l'origine technique des biais et leur manifestation aux utilisateurs ;

5. Les entités nommées correspondent, dans le jargon informatique, à des noms propres désignant des personnes, des lieux ou des organisations.

— permettre aux utilisateurs de développer des stratégies de recherche d'information adaptées à leur contexte d'usage ;

— donner les moyens aux utilisateurs de rendre compte du positionnement méthodologique adopté vis à vis des biais identifiés dans leur contexte de recherche.

Plus-value de la proposition de recherche

Cette proposition de recherche permet d'abord d'apporter une vision globale, utilisateur et système, de la problématique. Ce cadre, conjugué avec les résultats produits pour chaque contexte de la ressource numérique, nous permet d'envisager la construction d'une solution à même de répondre à l'expression des besoins schématisés dans la figure 1.2. Ces résultats, matérialisés dans notre cadre conceptuel à la figure 1.4 sont les suivants :

— **contexte d'usage :** mise en place d'une mécanique d'observation, production et validation d'indicateurs de la pratique des utilisateurs dans le cadre de tâche de recherche en histoire ;

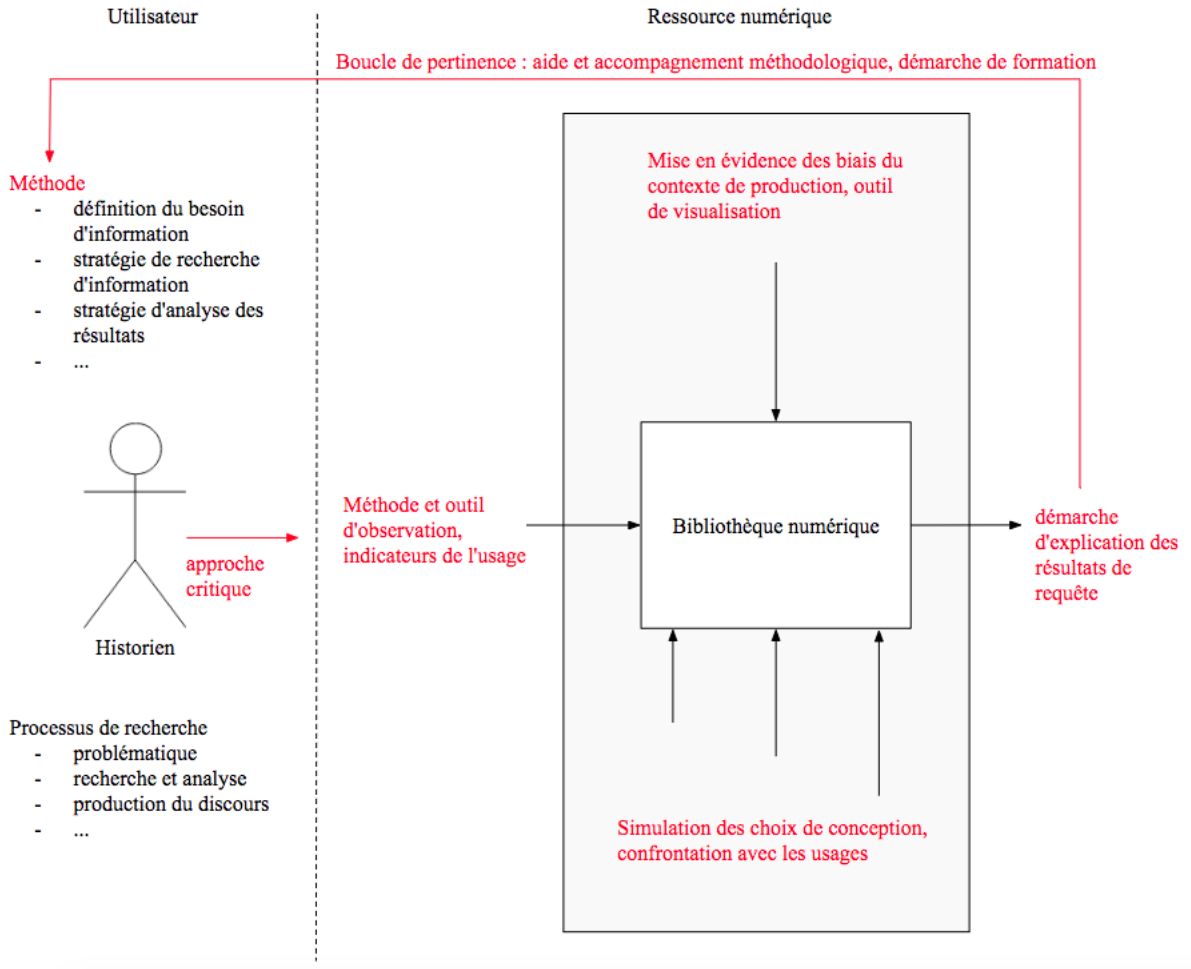
— **contexte de production :** étude des enjeux et de l'impact des choix de conception pour la recherche historique, en particulier pour le cas, très important pour l'historien, de la reconnaissance optique de caractères ;

— **contexte d'exécution :** mise en place d'outil de simulation des contextes et développement de fonctionnalités de visualisation de leurs effets ;

— **méthodologie :** proposition et validation d'une démarche pédagogique associée aux résultats précédents pour favoriser l'adoption par les étudiants en histoire d'une approche critique de la recherche de sources en ligne.

Ces éléments de contributions permettent d'éclairer le fonctionnement global du système et d'expliquer les résultats qu'il produit. Ils permettent par ailleurs de confronter l'usage avec les choix de conception opérés pour la production des contenus et la mécanique de recherche de l'information. De cette manière, la proposition de recherche offre une plus-value globale, sous la forme d'une boucle de pertinence, également représentée à la figure 1.4.

FIGURE 1.4 – Représentation des plus-values de la proposition de recherche



Les principaux éléments de notre contribution sont ici représentés en rouge.

1.4 Contexte académique de la thèse et organisation du manuscrit

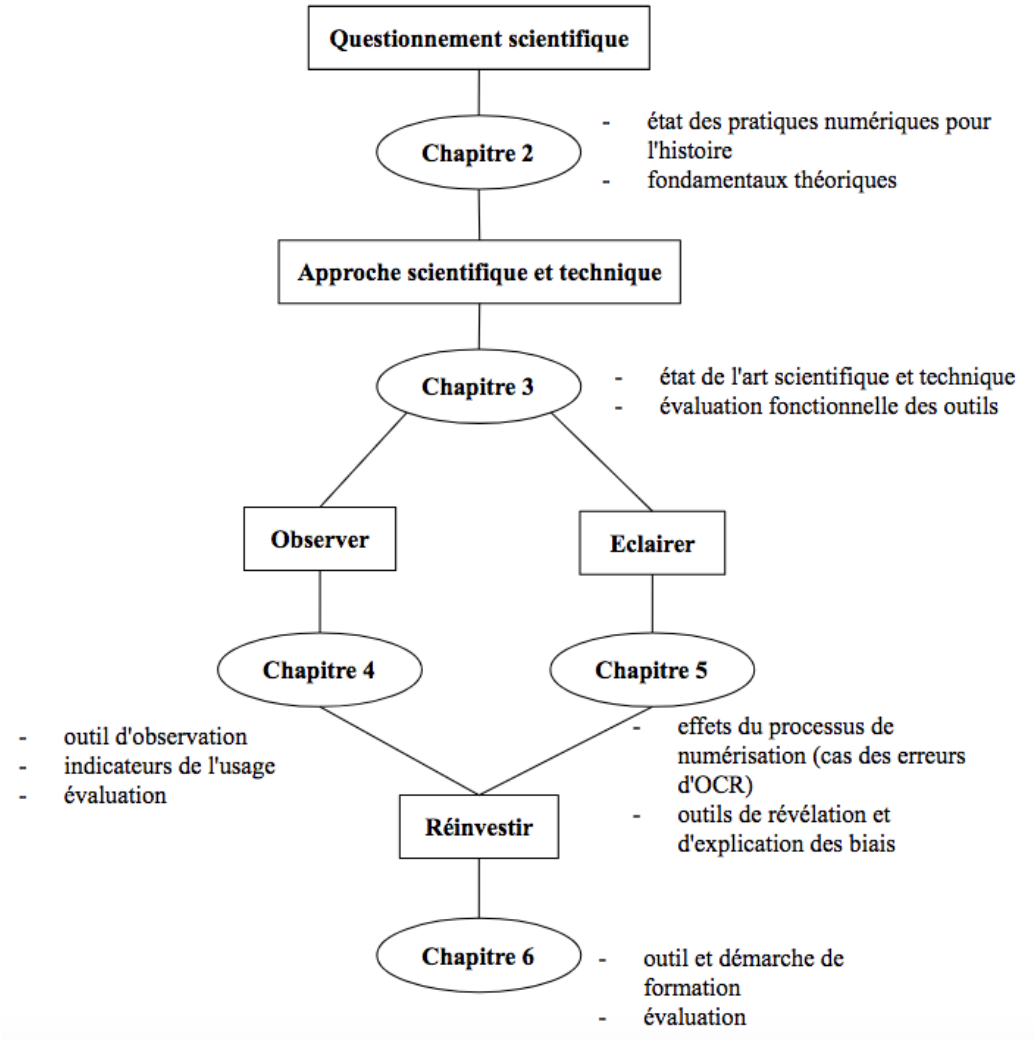
La problématique que nous étudions appelle un positionnement scientifique complexe, au carrefour de l'histoire et de l'informatique. Ces travaux se sont ainsi déroulés dans un cadre académique assez peu commun, associant autour de ce projet de recherche doctorale deux laboratoires de ces deux disciplines relativement éloignées. Il se justifie bien entendu pleinement au regard de notre questionnement scientifique et notre problématique. Cependant, il trouve d'abord sa source dans notre parcours universitaire combinant formation en histoire et en informatique, qui nous a permis de développer ses travaux dans un esprit interdisciplinaire qui dépasse, espérons le, la simple coloration d'une discipline par une autre. Il a enfin été rendu possible par une démarche de collaboration entre informatique et sciences humaines et sociales établie de longue date au laboratoire L3i et dans l'intérêt que portent les membres du CRHIA pour les objets d'études associant SHS et numérique.

Ces travaux de thèse se sont donc déroulés au sein du Laboratoire Informatique, Image

et Interactions (L3i) [6], en collaboration avec le Centre de Recherches en Histoire Internationale et Atlantique (CRHIA) [7], laboratoires de l'Université de La Rochelle [8] sous la direction de Pascal Estraillier (L3i) et Charles Illouz (CRHIA). Ils se sont principalement insérés dans les travaux menés par l'équipe Dynamique des systèmes et Adaptativité (e-Adapt) du L3i en interaction forte, toutefois, avec des recherches de l'équipe Images et Contenus (IC) du L3i.

Le manuscrit de cette thèse est organisé en cinq principaux chapitres, en dehors de la présente introduction, de la conclusion et des annexes. Leur organisation est présentée dans la figure 1.5. Le chapitre 2 s'attache à la construction d'un cadre scientifique et technique et examine les fondamentaux théorique et pratique de notre réflexion. Il contribue à positionner le numérique comme un acteur épistémologique et méthodologique en dressant un état de l'art et un état de la pratique du numérique pour l'histoire. Le chapitre 3 établit quant à lui l'état de l'art, scientifique et techniques des bibliothèques numériques que nous avons choisies d'étudier.

FIGURE 1.5 – Présentation schématique de l'organisation du manuscrit



6. EA 2118, Université de La Rochelle, http://l3i.univ-larochelle.fr
7. EA 1163, Universités de Nantes et de La Rochelle, https://www.crhia.fr
8. http://www.univ-larochelle.fr

Le chapitre 4 présente notre étude du contexte d'usage et sa démarche de validation expérimentale. Il contribue à la problématique par l'élaboration et la validation d'indicateurs pertinents de la pratique et la mise en évidence de comportements récurrents. Les résultats de ce chapitre peuvent être mis en perspective avec nos développements autour du contexte de production et d'exécution. Le chapitre 5 porte sur cet aspect de la problématique. Il s'intéresse plus précisément aux effets du processus de numérisation des contenus sur la recherche d'information, avec pour cas d'étude le problème des erreurs de reconnaissance optique de caractères.

La confrontation des résultats de ces deux chapitres contribue à la mise en évidence des biais que peuvent créer les algorithmes impliqués dans les processus de mise en ligne massive de documents historiques sur la pratique de l'historien. Le chapitre 6, s'emploie ainsi à les réinvestir et les évaluer au travers de notre démarche expérimentale de formation. Il propose un outil pédagogique visant à enseigner aux étudiants historiens les fondamentaux d'une approche critique de la recherche d'information en ligne. Pour se faire, il réemploie les indicateurs d'usage et les moyens d'identification des biais du contexte de production développés dans les chapitres précédents. Cette partie de notre travail contribue ainsi à l'effort de développement des programmes de recherche et de formation en humanités numériques par un dispositif et un outil pédagogique fondé sur l'expérimentation et l'observation directe des données et des mécanismes informatiques par les étudiants. Enfin, le chapitre 7 conclue ce manuscrit. Il en rappelle les principales contributions et présente les perspectives potentielles de développement.

1.5 Références

Afanador-Llach, M.-J., A. Rojas Castro, A. Crymble, V. Gayol, F. Gibbs, C. McDaniel, I. Milligan, A. Visconti et J. Wieringa, éd.. 2017, *The Programming Historian*. URL http://programminghistorian.org/. 9

Andersen, D. L. 2002, «Defining digital history», *Journal of the Association for History and Computing*, vol. 5, n° 1. 8

Audenaert, N. et R. Furuta. 2010, «What Humanists Want : How Scholars Use Source Materials», dans *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, ACM, New York, NY, USA, p. 283–292, doi : 10.1145/1816123.1816166. 6

Bloch, M. 1949, *Apologie pour l'histoire ou Métier d'historien*, A. Colin, Paris. 4

Bourdeloie, H. 2014, «Ce que le numérique fait aux sciences humaines et sociales», *tic&société*, vol. 7, n° N° 2. 5

Chiron, G., J.-P. Moreux, A. Doucet, M. Coustaty et M. Visani. 2017, «Erreurs OCR et biais d'indexation : impact sur les usages», dans *17ème conférence Extraction et Gestion des Connaissances, Atelier Journalisme Computationnel*, p. 69–73. 12

Cohen, D. 2017, «Roy's World», URL https://dancohen.org/2017/10/11/roys-world/. 2

Farge, A. 1989, *Le goût de l'archive*, Seuil, Paris. 4

Gooding, P. M. 2017, «A trace of this journey : Citations of Digitised Newspapers in UK PhD Theses», Montréal, p. 455–456. 5

Grandi, E. et E. Ruiz. 2012, «Ce que le numérique fait à l'historien.ne. Entretien avec Claire Lemercier», *Diacronie*, vol. 2, n° 10, p. 16. 5

Holley, R. 2009, «How good can it get ? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs», *D-Lib Magazine*, vol. 15, n° 3/4. 8

Jackson, A., J. Lin, I. Milligan et N. Ruest. 2016, «Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities», dans *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, ACM, New York, NY, USA, p. 103–106. 9

Kemman, M., M. Kleppe et S. Scagliola. 2013, «Just Google It - Digital Research Practices of Humanities Scholars», *arXiv :1309.2434 [cs]*. ArXiv : 1309.2434. 6

Lamassé, S. et P. Rygiel. 2014, «Nouvelles frontières de l'historien», *Revue Sciences/Lettres*, , n° 2, doi :10.4000/rsl.411. 3, 6

Meho, L. I. et H. R. Tibbo. 2003, «Modeling the information-seeking behavior of social scientists : Ellis's study revisited», *Journal of the American Society for Information Science and Technology*, vol. 54, n° 6, doi :10.1002/asi.10244, p. 570–587. 11

Milligan, I. 2013, «Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997–2010», *The Canadian Historical Review*, vol. 94, n° 4, p. 540–569. 5, 7

Milligan, I. 2014, «Le potentiel des données ouvertes pour l'histoire politique», *Revue parlementaire canadienne*, vol. 37, n° 4. 3

Mounier, P. 2015, «Histoire et numérique : une relation singulière et multiple», URL http://www.homo-numericus.net/article314.html. 5

Picard, F. 2011, *Contextualisation & Capture de Gestuelles Utilisateur : Contributions à l'Adaptativité des Applications Interactives Scénarisées*, thèse de doctorat, Université de La Rochelle. 9

Putnam, L. 2016, «The Transnational and the Text-Searchable : Digitized Sources and the Shadows They Cast», *The American Historical Review*, vol. 121, n° 2, p. 377–402. 4

Rhee, H. L. 2012, «Modelling historians' information-seeking behaviour with an interdisciplinary and comparative approach.», *Information Research*, vol. 17, n° 4. 11

Robertson, S. 2016, «The differences between digital humanities and digital history», *Istoriya*, vol. 7, n° 7 (51). 8

Rosenzweig, R. 2003, «Scarcity or Abundance ? Preserving the Past in a Digital Era», *The American Historical Review*, vol. 108, n° 3, doi :10.1086/529596, p. 735–762. 4

Rygiel, P. 2012, «L'enquête historique à l'ere numérique», *Revue d'histoire moderne et contemporaine*, vol. 58-4, n° 5, p. 30–40. xi, 2, 3

Sawadogo, D. 2016, *Architectures logicielles et mécanismes pour la gestion adaptative et consolidée de ressources numériques dans une application interactive scénarisée*, thèse de doctorat, Université de La Rochelle. 9

Tanner, S., T. Muñoz et P. H. Ros. 2009, «Measuring mass text digitization quality and usefulness», *D-lib Magazine*, vol. 15, n° 7/8, p. 1082–9873. 8

Traub, M. C., J. van Ossenbruggen et L. Hardman. 2015, «Impact Analysis of OCR Quality on Research Tasks in Digital Archives», dans *Research and Advanced Technology for Digital Libraries*, Springer, p. 252–263. 7, 12

Veyne, P. 1971, *Comment on écrit l'histoire*, seuil éd., Paris. 4

Wieviorka, M. 2013, *L'impératif numérique ou La nouvelle Ere des sciences humaines et sociales ?*, CNRS, Paris. 5

Xie, H. 2006, «Evaluation of digital libraries : Criteria and problems from users' perspectives», *Library and Information Science Research*, vol. 28, n° 3, p. 433–452. 8

Xie, H. I. 2008, «Users' evaluation of digital libraries (DLs) : Their uses, their criteria, and their assessment», *Information processing & management*, vol. 44, n° 3, p. 1346–1373. 8

Chapitre 2

Ressources numériques et histoire : état de l'art, état de la pratique

Résumé

Ce chapitre présente un état de l'art des démarches de recherche et principaux outils numériques utiles à la recherche en sciences humaines au sens large et en histoire en particulier qui impliquent, de près ou de loin, des processus informatiques de recherche et de gestion de l'information. Il se fonde sur une analyse des enquêtes d'usage déjà publiées et sur l'analyse de la lecture scientifique pour montrer les différentes formes que prennent aujourd'hui les relations entre l'informatique et l'histoire. Pour tenir compte de cette grande diversité, nous avons choisi d'adopter le cadre conceptuel de la ressource numérique, dont ce chapitre propose également les éléments de définition et d'état de l'art.

Sommaire

2.1 Introduction

Comme nous l'avons expliqué dans notre introduction (chapitre 1), nous avons choisi de nous intéresser en priorité aux moyens numériques d'accès aux sources de l'historien. Ce choix résulte de l'importance de ces outils numériques dans les usages mais également de la nature des informations manipulées. Rechercher et accéder à des sources en ligne et les manipuler sous une forme numérique a des conséquences sur toute l'activité de recherche, parce que l'on touche à la relation profonde qui unit l'historien et le matériau de base de son travail.

Le travail de l'historien sur la source est traditionnellement présenté, depuis Seignobos et Langlois tenants de l'histoire positiviste, comme un travail de critique [Seignobos et Langlois, 1899]. On distingue généralement critiques externe et interne du document. La critique externe est celle de la forme d'abord, le papier, l'écriture, *etc.* ainsi que celle du contexte, visant à établir l'authenticité du document et le replacer dans son contexte historique pour être à même de le comprendre. La critique interne est quant à elle celle du contenu et des événements qu'il relate. Cette présentation est bien entendu très réductrice de la relation entre l'historien et le document.

De nombreux auteurs ont écrit sur cette relation et ont permis de préciser la démarche scientifique de l'historien, parmi lesquels Marc Bloch, fondateur avec Lucien Febvre de l'École des Annales, Fernand Braudel, Paul Veyne, ou Arlette Farge [Bloch, 1949; Braudel, 1969; Veyne, 1971]. À la toute fin des années 1960, Michel Foucault [Foucault, 1969], expliquait que le document est à la base de la discipline, comme source première de l'histoire. Il en est la substance à partir de laquelle s'élaborent et se transforment les méthodes de l'historien.

> *Pas de malentendu : il est bien évident que depuis qu'une discipline comme l'histoire existe, on s'est servi de documents, on les a interrogés, on s'est interrogé sur eux ; on leur a demandé non seulement ce qu'ils voulaient dire, mais s'ils disaient bien la vérité, et à quel titre ils pouvaient le prétendre, s'ils étaient sincères ou falsificateurs, bien informés ou ignorants, authentiques ou altérés. [...] Le document était toujours traité comme le langage d'une voix maintenant réduite au silence, sa trace fragile, mais par chance déchiffrable. Or, par une mutation qui ne date pas d'aujourd'hui, mais qui n'est sans doute pas encore achevée, l'histoire a changé sa position à l'égard du document : elle se donne pour tâche première, non point de l'interpréter, non point de déterminer s'il dit vrai et quelle est sa valeur expressive, mais de le travailler de l'intérieur et de l'élaborer : elle l'organise, le découpe, le distribue, l'ordonne, le répartit en niveaux, établit des séries, distingue ce qui est pertinent de ce qui ne l'est pas, repère des éléments, définit des unités, décrit des relations*[1].

Michel Foucault explique ici avec force que le travail de l'historien dépasse de loin la seule lecture des sources et la seule extraction des informations qu'elles contiennent ou ne contiennent pas. L'historien travaille le document, lui fait subir de nombreuses critiques, le lie à d'autres sources. Ce travail d'élaboration prend corps dans les nombreux cadres théoriques de l'histoire, depuis l'histoire globale, jusqu'à la micro-histoire

1. Foucault 1969, p. 13-14

ou l'histoire marxiste. Ces positionnements théoriques conduisent à des questions de recherche et à l'invention, par l'historien, des sources qui permettront d'y répondre.

Cette relation multiple entre le chercheur et les sources de son travail se traduit, dans le cadre de notre problématique, par une chaîne de démarches de recherche et d'outils numériques à disposition des historiens, dont les bibliothèques numériques et moteurs de recherche ne sont qu'un maillon. Pour positionner notre réflexion sur ces outils au sein de leur écosystème, il nous est apparu important de faire, dans notre section 2.2, un état des lieux des démarches et techniques de recherche pour lesquelles l'informatique qui les supporte, depuis la recherche d'information jusqu'au web sémantique, joue un rôle majeur. En guise de bilan de cet état de la pratique, nous avons souhaité, dans notre section 2.3, montrer la centralité des outils sur lesquels nous nous focalisons pour la recherche en histoire et préciser les points fondamentaux de notre cadre conceptuel.

2.2 Quel numérique pour quelle histoire ? État de la pratique

Les relations entre histoire et numérique ne sont pas aussi récentes qu'il y parait. Elles sont le fruit d'une, voire de plusieurs traditions historiographiques remontant au milieu du XX^{e} siècle. La vigueur de ces relations a fait émerger l'histoire numérique, dont nous avons déjà parlé, et plus largement les humanités numériques, un champ de recherche au carrefour des sciences et sociales, des sciences de l'information et de la communication ou de l'informatique [2]. L'émergence de ces champs de recherche actent la démocratisation de l'usage de l'informatique et des réseaux par les chercheurs en SHS au sein de ces nouvelles communautés de pratiques. Pour autant, le seul usage de l'ordinateur en histoire et en SHS ne peut pas pour autant résumer à lui seul l'état de la pratique. Les humanités et l'histoire numérique sont de véritables activités de recherche pour lesquelles le numérique s'est, sinon imposé, du moins largement développé et pour lesquels il existe des outils et des projets scientifiques majeurs qui répondent à des problématiques particulières.

Cet état de la pratique ne saurait pour autant prétendre à une parfaite exhaustivité, des enquêtes ont été menées par le passé pour mesurer l'usage de tel ou tel dispositif numérique par les historiens. Notre objectif n'est pas de faire un nouvel état chiffré mais d'établir, par cet exercice de cartographie des outils, un espace de discussion commun et concret dans lequel peuvent s'insérer nos travaux et s'inscrire une réflexion conjointe entre l'histoire et les disciplines du numérique.

2.2.1 Enquêtes et taxonomies

Pour mener à bien l'état de la pratique dont nous souhaitions disposer, nous avons décidé, plutôt que de mener notre propre sondage, avec toutes les incertitudes et difficultés

2. Parmi les ouvrages généraux et guides utiles à la découverte du champ de recherche des humanités numériques, on peut citer, parmi d'autres, l'ouvrage collectif dirigé par Schreibman, Ray et Unsworth, 2016, l'ouvrage de Warwick, Terras et Nyhan, 2012 et le travail, en français, de Le Deuff, 2014. On peut également consulter le récent article de Clavert et collab., 2017, qui montre la diversité des profils de chercheurs impliqués dans les humanités numériques.

d'une telle entreprise, de passer en revue les sources disponibles. Au delà des études déjà parues sur cette thématique, nous avons ainsi pu identifier deux sources principales permettant une étude des pratiques actuelles du numérique chez les historiens et plus largement les chercheurs en SHS. Dans un premier temps, nous avons étudié les taxonomies qui ont pu être élaborées et publiées jusqu'à présent. Dans un second temps, nous avons souhaité croiser leurs résultats avec des données issues du terrain. Nous nous sommes donc appuyés sur les enquêtes et sondages, éventuellement complétées d'entretiens, qui ont déjà été menées auprès de différentes populations de chercheurs. Ces sources, prises indépendamment, sont parcellaires et n'apportent qu'une vue partielle. Leur confrontation, en revanche, peut nous aider à obtenir une vue plus globale de l'écosystème dans lequel prennent place les ressources numériques que nous étudions, bien que naturellement, elle ne remplace par une étude de grande ampleur.

Catégorisation des pratiques de recherches numériques : la taxonomie TaDiRAH

Le travail de catégorisation des pratiques numériques de recherche en Sciences Humaines et Sociales le plus abouti ayant été mené jusqu'à présent se présente sous la forme d'une taxonomie. Appelée TaDiRAH (*Taxonomy of Digital Research Activities in the Humanities*), elle a été développée par un groupe de travail de DARIAH (*Digital Research Infrastructure for the Arts and Humanities*) [Romary, 2014] [3]. Comme l'ont rappelé Borek et collab. [2016], cette initiative est née d'un besoin de classification des différentes activités de recherche des humanités numériques, dans le but de promouvoir et faciliter le partage des méthodes comme des objets d'étude. TaDiRAH est ainsi issue de plusieurs initiatives différentes.

La première d'entre elles a été la constitution d'une bibliographie partagée [4] "*Doing Digital Humanities*" créée à l'initiative de la branche allemande de DARIAH, DARIAH-DE. Ce travail pionnier a permis d'initier une réflexion sur les mots-clés et les catégories les plus à même de décrire, comme l'indique le nom de cette bibliographie, la « pratique » des humanistes numériques. En parallèle, le site DiRT [5] établissait, par le biais d'un wiki, une liste, elle aussi classée, d'outils utiles aux humanités numériques. Ce site a par la suite était repris et amélioré par un projet d'infrastructure dédiée aux humanités numériques, le projet Bamboo [Dombrowski, 2014] [6]. La jonction de ces deux initiatives a abouti à la création de la première version de la taxonomie TaDiRAH [Borek et collab., 2014]. Depuis, d'autres approches, par exemple fondées sur l'analyse des syllabus de formation en humanités numériques ont permis de la critiquer et de la compléter [Sula et collab., 2017]. D'autre part, plusieurs traductions de la taxonomie originale en anglais ont été proposées, en allemand [Baillot, 2016] ou en français à l'initiative de l'Association Francophone des humanités Numériques, Humanistica [7].

3. La taxonomie TaDiRAH est accessible à l'adresse : http://tadirah.dariah.eu/vocab/index.php dans sa version web et http://tadirah.dariah.eu/vocab/sparql.php dans sa version SparQL. Le site de DARIAH est quant à lui disponible à l'adresse : https://www.dariah.eu/

4. Cette bibliographie, riche aujourd'hui de plus de 1100 notices en différentes langues est accessible à cette adresse : https://www.zotero.org/groups/113737/

5. Toujours accessible à cette adresse : https://digitalresearchtools.pbworks.com/w/page/17801672/FrontPage

6. Dont la version actuelle est disponible à l'adresse : http://dirtdirectory.org/

7. La traduction française est disponible via Github à cette adresse : https://github.com/

Tableau 2.1 – Récapitulatif de la catégorie « activités de recherche » de la taxonomie TaDiRAH.

Activités de recherche	
Acquisition	conversion, reconnaissance de données, découverte , collecte, imagerie, enregistrement, transcription
Création	design, programmation, traduction, développement web, écriture
Enrichissement	annotation, nettoyage, édition
Analyse	analyse de contenu, analyse de réseaux, analyse relationnelle, analyse spatiale, analyse structurelle, analyse stylistique, visualisation
Interprétation	contextualisation, modélisation, théorisation
Stockage	archivage, identification, organisation, conservation
Dissémination	collaboration, commentaire, communication, crowdsourcing, publication, partage
Méta-activités	évaluation, construction de communauté, donner un aperçu, gestion de projet, enseignement et apprentissage

La taxinomie TaDiRAH est constituée de trois catégories principales :

— **Activités de recherche :** Cette catégorie, dont le détail complet est présenté dans le tableau 2.1 regroupe des actions de recherche qui peuvent s'appliquer à différents objets de recherche. Ainsi, comme le mentionne la description de la catégorie dans la taxonomie, un éditeur de texte porterait les étiquettes « Écriture », l'activité de recherche pour laquelle il est utile, mais également « Code » et « Texte », les objets d'étude pour lesquels il a un intérêt.

— **Objets de recherche :** Cette catégorie recense environ 35 objets de recherche des humanités numériques. Ces objets sont divers, depuis les médias jusqu'aux sources primaires, en passant par les infrastructures, les logiciels ou les métadonnées. Ces objets représentent la diversité du champ d'étude, qui interroge aussi bien les objets classiques des SHS que leurs outils et pratiques du numérique ou leurs cadres institutionnels.

— **Techniques de recherche :** Cette dernière catégorie établit quant à elle une liste des méthodes employées par les chercheurs en humanités numériques. Elle met en avant des processus algorithmiques d'analyse de données, des méthodes de représentation graphique, de recherche d'information ou de développement informatique. Cette liste met également en évidence la place importante prise par la problématique de la préservation et la conservation des objets et des résultats de la recherche dans cette taxonomie.

Enquêtes et sondages

Plusieurs enquêtes sur les pratiques numériques des historiens, voire des chercheurs en SHS ont donc été menées par le passé. Parmi elles, nous avons relevé l'enquête réalisée

`Humanistica/TaDiRAH`, le site de l'Association Francophone des Humanités Numériques, Humanistica est quant à lui disponible à l'adresse : http://www.humanisti.ca/

par Emilien Ruiz en 2015, qui s'adressait spécifiquement aux historiens [Ruiz, 2015]. Cette enquête se fonde sur une extension d'une enquête précédente à l'attention des étudiants uniquement. Malgré la modestie de cette entreprise, revendiquée par l'auteur, cette enquête a obtenu un nombre significatif de réponses qui s'élève à 353, en un mois. Le questionnaire, soumis aux répondants par internet, via des listes de diffusion, des blogs et des réseaux sociaux comportait un peu plus d'une dizaine de questions, relatives aux outils utilisés, essentiellement en matière de gestion documentaires et d'écritures. Ces thématiques, importantes dans notre contexte, ont ainsi retenu notre attention.

Nous avons par ailleurs étudié les résultats d'un second sondage, connu sous le nom *Practices4Humanities*, dont une partie des résultats a été rendue publique en 2016 [Baillot, 2016]. Ce sondage a quant à lui était mené auprès de la communauté des humanités numériques. Les répondants sont des chercheurs, des archivistes ou des bibliothécaires, de tous les niveaux, dans une aire géographique cette fois restreinte à la région berlinoise. L'objectif de l'enquête était de qualifier la relation qui peut exister entre un usage de l'ordinateur et le fait qu'un chercheur se présente ou non comme relevant des humanités numériques. Le nombre de réponses au formulaire, là encore diffusé en ligne, a été une nouvelle fois assez élevé puisqu'il atteint 123 réponses, ce qui permet de commencer à envisager ce sondage comme une étude qui, bien que non généralisable, dépasse tout de même la simple étude de cas.

Enfin, nous avons pris en compte les résultats issus d'enquêtes plus qualitatives, menées sous forme d'entretiens auprès de différents chercheurs. Bien entendu, ces enquêtes sont focalisées sur des points particuliers de la pratique et souvent intéressées à l'étude des sciences humaines et sociales au sens large. Si nous ne prétendons pas avoir identifié toutes ces enquêtes de manière exhaustive, certaines ont particulièrement retenu notre attention parce qu'elles se focalisaient sur les points les plus centraux de notre problématique et étaient elles aussi positionnées dans un écosystème complexe. Ainsi, certaines des catégories que nous évoquons dans cet état de la pratique pourront faire appel aux résultats de ces enquêtes, en particulier aux travaux de Audenaert et Furuta, 2010, de [Bulger et collab., 2011] et de Kemman et collab., 2013.

2.2.2 Constitution et gestion des corpus

Dans une recherche historique, la première étape pour laquelle les outils du numérique peuvent avoir un rôle important est la constitution du corpus documentaire. De plus en plus de sources primaires sont disponibles en ligne. Pour les sources qui ne sont pas numérisées ou exposées sur le web, il existe au moins, dans l'immense majorité des cas, un catalogue disponible sur internet qui permet à distance d'obtenir la liste et la description des documents intéressants. Il en va de même pour la documentation secondaire, ouvrage et articles scientifiques qui sont désormais, au moins pour les articles, systématiquement disponibles dans les bibliothèques numériques auxquelles la majorité des institutions de recherche donnent accès. Nous dressons donc ici le portrait de ces outils numériques utiles, voire indispensables, à la constitution des corpus ainsi que de ceux, très important également, dédiés à leur gestion, les gestionnaires de bibliographies ou de bases de données.

Outils de recherche documentaire et d'accès aux sources

Malgré le développement exponentiel des archives numérisées, les déplacements en archives restent comme nous l'avons vu importants pour les historiens. Le travail en ces lieux n'est pas, pour autant, resté hermétique aux outils numériques. En effet, parmi ces outils généralement considérés comme faisant partie de la palette des historiens, sur lesquels Emilien Ruiz a enquêté en 2015, trois seulement passaient la barre d'un usage régulier par au moins 50% des répondants. De ces trois outils, un seul semblait faire véritablement l'unanimité et faisait assurément partie des outils de base des historiens interrogés. Il s'agissait de l'appareil photo numérique, dont près de 90% des répondants ont déclaré se servir [Ruiz, 2015]. Tous les historiens, ou presque, lorsqu'ils se déplacent en archives, le font désormais équipés d'un dispositif de capture photo, véritable appareil photo numérique ou tout simplement téléphone portable. Les archivistes eux-mêmes confirment cette pratique, à en croire [Potin, 2011] qui écrivait en 2011 :

> *« une autre numérisation a déjà commencé dans l'ambiance feutrée des salles de lecture des dépôts d'archives. Celle-là est réalisée au quotidien par des centaines d'appareils photographiques personnels, dont le cliquetis automatique constitue désormais la symphonie discrète d'un « dépouillement » qui ne berce plus le silence des archives de la même manière. Tel un chasseur-cueilleur numérique, l'historien engrange, non plus des fiches, mais l'image des documents eux-mêmes, associant ou reconstituant des fragments de fonds entiers. Les récoltes personnelles ainsi menées constituent autant de « corpus » d'archives dispersés, dont la mise en œuvre, du nommage des fichiers à leur articulation en « dossiers » en partie factices, est établie par chacun dans l'intimité fragile de son disque dur. »*

L'usage de l'appareil photo numérique dans ce chapitre dévolu aux outils du numérique peut paraître curieux. Cet objet est de nos jours d'une grande banalité et n'est pas nécessairement celui qui symbolise le mieux ce que l'on peut généralement entendre par le mot « numérique ». Pourtant, il est fondamental de s'y intéresser dans un contexte comme le nôtre. Le simple usage de l'appareil photo numérique par les historiens lors de leurs séjours en archives fait en effet émerger beaucoup des problématiques liées à l'évolution des pratiques avec le développement du numérique. Derrière l'usage apparemment banal de ces dispositifs de capture, ce sont d'abord des documents issus de la numérisation qui sont produits. En terme numérique, il devient nécessaire de disposer d'un panel d'outils associés, pour la capture, le transfert, le stockage voire la retouche, l'extraction ou l'indexation d'information.

L'introduction de l'appareil photo numérique dans les centres d'archives vient également bouleverser la relation que les historiens ont longtemps entretenu avec ces lieux qu'Arlette Farge a si bien décrits. Les salles de lecture sont moins fréquentées et surtout moins longtemps par les historiens. Par le passé, bien au delà d'être des lieux de consultation, elles étaient également des lieux de travail, des lieux où l'analyse des sources commençait dès la transcription des documents et la prise de notes. L'appareil photo numérique renouvelle cette vision classique du travail en archive et réduit son coût. Le temps sur place n'est plus dédié au travail sur les documents, il est « optimisé » par la capture à la chaîne de sources historiques. La temps de l'analyse ne débute plus

aux archives, il est renvoyé à plus tard. Comme le remarquait récemment Sean Takats, ce temps de travail est devenu fragmenté et itératif. La relation entre l'historien et l'archive physique est plus éphémère qu'auparavant [8].

En plus de cette « autre numérisation » permise par l'appareil photo numérique, l'historien peut avoir accès depuis son poste de travail à une grande quantité de sources primaires et secondaires numérisées par diverses institutions. Les bases de données bibliographiques de littérature secondaire sont ainsi largement consultées par les historiens. 60% des répondants de l'enquête d'Émilien Ruiz affirmaient y accéder au moins une fois par semaine, 30% de plus au moins une fois par moi. Ces outils permettent d'accéder aux références voire au texte complet de la littérature secondaire, livres ou articles scientifiques. La plupart des bibliothèques universitaires, outre leur catalogue, fournissent un accès aux bases de données des éditeurs, parmi lesquels JSTOR, Cairn.info, *etc.* L'usage de ce type de catalogue facilite un accès à la documentation secondaire qu'il aurait auparavant été très coûteux de rassembler.

Les bases de données en ligne et bibliothèques numériques de sources primaires semblaient, dans la même enquête, être légèrement moins utilisées que pour la documentation secondaire. 50% des répondants déclaraient en faire usage au moins une fois par semaine et 30% au moins une fois par mois. Ces chiffres légèrement inférieurs s'expliquent par le fait que toutes les spécialités historiques ne disposent pas de sources largement numérisées. La numérisation ne concerne en effet pas tous les types de sources et privilégie largement l'écrit, manuscrit ou imprimé au détriment d'autres sources, plus complexes à mettre en ligne, objets archéologiques, témoignages, *etc.* Sans doute ne faut-il pas exagérer l'importance de ces moyens numériques d'accès aux sources pour tous les historiens, mais en fonction des sources manipulées et peut-être de la « technophilie » ou de la génération de chercheurs ou futurs chercheurs dont on parle, ils sont très importants. Sébastien Poublanc rapporte ainsi ce qu'il ressort d'un exercice qu'il mène avec des étudiants de Master recherche, dont l'objectif est de les amener à raconter leur expérience de recherche au moyen de billets de blog et écrit [Poublanc, 2018] :

> *Pour ces historiens d'un genre nouveau, se rendre dans un dépôt d'archive se révèle impossible, voire totalement inutile. Toutes leurs sources sont numérisées : Gallica, Europresse ou la Library of Congress fournissent des banques de données bibliographiques, des sources primaires numérisées, des sources imprimées sous formes de livres, des métadonnées à interroger sous la forme de catalogue d'articles etc. Les étudiants travaillent en bibliothèque pour bénéficier des bases de données payantes – encore que les Espaces Numériques de Travail (ENT) leur permettent à présent de s'en affranchir* [9].

8. Nous empruntons ici les mots de Sean Takats prononcés lors d'une conférence, « La crise épistémologique des humanités numériques » à Toulouse, 6 février 2017. Le résumé de cette intervention est disponible à l'adresse http://framespa.univ-tlse2.fr/actualites/la-recherche/axes-thematiques/conference-de-sean-takats-la-crise-epistemologique-des-humanites-numeriques-483836.kjsp

9. Ce texte est extrait d'un livre collaboratif en cours d'écriture, il est donc sujet à changement au gré des commentaires que tout un chacun peut faire à l'auteur. Saisissons l'occasion de souligner le problème que pose la présente note de bas de page en matière de citation de ce type de travaux collaboratif et mouvant, permis par le numérique. Si l'on se pose la question de la citation des sources numérisées depuis longtemps (voir [Graham, 2001]), sans avoir véritablement tranché le débat, les documents collaborativement élaborés ne manquent pas de renouveler la problématique.

L'activité de recherche d'information est particulièrement cruciale pour la pratique de l'historien. Face à l'explosion du volume des sources numérisées, il est fondamental de permettre une recherche pertinente dans les contenus numériques. Cela ne se limite pas simplement à extraire de ces documents numériques des descripteurs pertinents, il faut également faire en sorte de pouvoir les retrouver de manière efficace. De nombreux travaux se focalisent ainsi sur les techniques de classement automatique des documents numériques ou sur l'étude de l'expression des besoins d'information des utilisateurs. Il s'agit par exemple de développer des procédures et des algorithmes capables de comprendre au mieux une requête exprimée par l'utilisateur, que ce soit sous forme de mots-clés, d'expression en langage naturel ou sous la forme d'un autre document [10].

L'enquête menée par [Kemman et collab., 2013] confirme que l'usage des bases de données est largement répandu parmi les chercheurs et qu'elles sont majoritairement utilisées pour la documentation secondaire, les sources textuelles et les sources iconographiques. Cette enquête met également en avant que les points d'entrée principaux vers ces bases de données sont les moteurs de recherche généraux, tel que Google, et leurs variantes spécialisées dans certains type de document (Google Image, ou Google Scholar). Les plateformes spécialisées viennent en seconde position, que ce soient les plateformes de sources secondaires tel que JSTOR ou de sources primaires telles que Gallica ou *KB Digital Resources* [11].

En revanche, cette étude montre que les agrégateurs de sources, primaire comme secondaire, sont significativement moins utilisés que les moteurs de recherche généraux pour accéder aux contenus. Ces agrégateurs, pourtant jugés cruciaux pour la recherche en SHS [Pinol, 2011] ont pour objectif de proposer un accès unifié aux références et aux documents primaires ou secondaires dispersés sur des dizaines, voire des centaines de plateformes différentes.

Ces systèmes s'appuient sur des protocoles d'échange d'information, reposant sur des métadonnées, pour agréger, indexer et classer des ressources [12]. Très régulièrement les systèmes moissonnent les sources externes et enregistrent leur référence au sein d'un catalogue unique. Ces moissonneurs sont dotés de fonctionnalités puissantes permettant d'éviter l'enregistrement de doublons de références et de désambiguïser les références pointant vers les mêmes objets, mais décrites différemment. Ils donnent en quelques clics, et sur la base de mots-clés, accès aux ressources d'un nombre très important de catalogues tiers. Europeana donne par exemple accès, en 2017, à plus de 54 millions de ressources collectées dans toute l'Europe.

Ces corpus numériques sont décrits par des métadonnées et environnés d'outils destinés à en faciliter l'usage, à l'image de ce qui est pratiqué dans le monde des archives

10. Sur ce point, il est possible de comparer un document, une image fournie par un utilisateur par exemple avec le contenu d'une base de connaissance. La requête est alors construite en fonction d'une extraction de l'information de ce document et est comparée avec les informations extraites préalablement des documents contenus dans la base. Parmi ce type de méthode, on peut citer le *Word spotting* sur laquelle travaillent des chercheurs en vision par ordinateur (*computer vision*) [Almazán et collab., 2014]

11. KB est l'acronyme de *Koninklijke Bibliotheek*, la bibliothèque nationale des Pays-Bas, qui fournit une vaste collection de documents numérisés https://www.kb.nl/en/digital-resources

12. Parmi ces protocoles figurent le Z39.50 développé pour l'échange de notices entre les bibliothèques, progressivement abandonné au profit de protocoles plus performants et plus génériques, le protocole OAI-PMH par exemple.

physiques. Ceci dit, de nombreux auteurs ont souligné que ces corpus numériques ne reposent pas sur des normes stables, aussi bien du point de vue du format de ces métadonnées que des vocabulaires utilisés pour les décrire. Si l'on se limite au format RDF, il existe un nombre très important d'ontologies permettant de décrire spécifiquement les archives historiques [13] qui sont utilisées en parallèle de vocabulaires plus généraux, tel que *Dublin Core* (DC) par exemple.

Ces vocabulaires existent en grand nombre et avec des niveaux de spécialisations très variable. Les outils sont donc bien présents, mais les contextes de production des archives numériques sont d'une grande complexité. Il est nécessaire de respecter les principes de l'archivistique, de renseigner autant que possible le contexte historique du contenu numérisé, tout en ayant à faire face, très souvent à une grande hétérogénéité des documents. Il en résulte que le travail nécessaire pour établir un modèle adapté peut s'avérer très important si l'on souhaite éviter un grand empilement de vocabulaires divers, rendant le modèle difficile à maintenir.

L'un des problèmes majeurs de la mise en ligne des archives et l'effet de décontextualisation qu'elles subissent. L'absence de métadonnées à même de décrire le contexte de conservation physique du document est un des problèmes auxquels peuvent avoir à faire face les historiens lors de l'utilisation de moyens d'accès numérique au source. Des développements récents en matière de description des contenus visent ainsi à mieux tenir compte du contexte du document décrit. Un groupe de travail de l'*International Council on Archive* (ICA) a ainsi conçu le modèle *Record in Context*, dont font partie le modèle RIC-CM et l'ontologie RIC-O [14]. Ces initiatives visent d'abord à clarifier les formats et outils de mise à disposition et de visualisation des archives numériques et à amoindrir les effets de décontextualisation subies par les sources. D'autre part, elles permettent de proposer de nouveaux outils de visualisation des données archivistiques centrés sur l'utilisateur [Clavaud et Château-Dutier, 2017].

Moyens d'accès à de nouvelles sources

L'introduction du numérique dans la pratique de l'historien a donc favorisé l'accès aux documents en offrant de nouveaux moyens de collecte de la documentation conservée dans les archives ou produite par la recherche. Il lui permet également d'être un acteur de la collecte de nouvelles archives. Ces nouvelles pratiques sont généralement regroupées sous le terme de *crowdsourcing* dont la traduction française est délicate. Dans son acception large, le terme est traduit par l'expression « production participative » par la commission générale de terminologie et de néologie [15] qui en donne la définition suivante : « Mode de réalisation d'un projet ou d'un produit en faisant appel aux contributions d'un grand nombre de personnes, généralement des internautes ». Cette définition s'applique par exemple à des projets de développement *open source* ou à l'élaboration de connaissances partagées dont l'exemple le plus connu est évidemment l'encyclopédie en ligne Wikipedia. Dans le contexte qui nous occupe, elles désignent des projets de collecte ou d'enrichissement de données historiques, de manière collaborative

13. Les normes CIDOC-CRM ou FRBR par exemple se sont vues doter de modèles conceptuels et d'ontologies

14. Les résultats de ce groupe de travail et les modèles et vocabulaires produits sont disponibles sur le site de l'ICA à l'adresse : https://www.ica.org/fr/egad-ric

15. JORF n°0179 du 5 août 2014, pp. 12995, texte n°91.

et par l'entremise de plateformes numériques. Le *crowdsourcing* n'a pas été traité dans les enquêtes que nous avons citées, des études ont toutefois été menées pour en mesurer l'intérêt et explorer leurs conséquences [Warwick et collab., 2012].

Les opérations de collecte visent principalement à recueillir des archives conservées en dehors des centres d'archives institutionnels. Parmi les projets les plus connus, nous pouvons citer *The September 11 Digital Archive*, dont l'objectif est de collecter les témoignages ou les documents des témoins des attentats du 11 septembre [16]. La plateforme enregistre ces informations pour que les historiens puissent à terme avoir accès à cette source en provenance directe des témoins de l'événement.

Outre la collecte de données, les démarches de *crowdsourcing* sont également utilisées pour enrichir de l'information déjà archivée. Certaines bibliothèques permettent aux utilisateurs d'ajouter des tags sur certaines données. La *New York Public Library*, par exemple, propose à ses utilisateurs, à travers son projet *What's on the menu*, de transcrire l'information contenu dans des cartes de restaurants du XIX^e^ siècle et du début du XX^e^ siècle. Plus d'1.3 millions de plats, sur quelques 17000 menus ont été transcrits en 2017 depuis le début du projet. Cette opération de transcription massive, qui n'aurait pas pu être menée par un processus de reconnaissance de caractères, compte tenu de la complexité des documents, est d'un intérêt historique majeur pour l'étude des habitudes alimentaires durant cette période. Ce type d'initiative rencontre un certain succès bien qu'il pose des questions spécifiques [Brennan et Mills, 2011] et qu'il reste pour l'instant relativement confidentiel. Les démarches de *crowdsourcing* montrent toutefois comment le numérique participe à l'extension du territoire de l'historien, lui permettant d'inventer et d'exploiter de nouvelles sources.

Parmi ces nouvelles sources, il est important de mentionner que le développement des réseaux sociaux et le flux de données qu'ils génèrent peuvent constituer une importante source pour l'histoire. Des travaux récents ont montré des résultats intéressants, sur des problématiques strictement historiques [Clavert, 2017; Papastamkou, 2017] ou sur des problématiques de structuration de la recherche ou de mise en scène du discours historique [Grandjean, 2016, 2017]. En parallèle du développement des analyses fondées sur les données des réseaux sociaux, d'autres travaux, dont ceux de Ian Milligan [Milligan, 2014] ont montré le potentiel des données ouvertes pour l'analyse historique.

Comme les nouvelles ressources dont nous venons de parler, les archives du Web semblent être un des terrains les plus prometteurs de l'histoire contemporaine [Schafer et Thierry, 2015]. Par « archives du web » on entend généralement l'entreprise qui vise à enregistrer régulièrement l'état des sites web et à en garder une trace qui devra permettre à tout un chacun de revenir à une ancienne version d'une page web, qu'elle soit ou non toujours accessible. Bien entendu, cet idéal est loin d'être atteint. Le volume de données que constitue aujourd'hui le web est tel qu'il est en pratique impossible d'en enregistrer un instantané. Que dire alors d'un enregistrement régulier des changements de toutes les pages web ? Le projet Internet Archive, ainsi que diverses institutions s'emploie tout de même à archiver le web [17]. Pour accomplir cette

16. Cette plateforme est accessible à cette adresse : http://911digitalarchive.org

17. Le projet le plus connu en la matière est Internet Archive et sa "*Wayback Machine*" (https://archive.org/web/) qui permet relativement aisément de visualiser un état ancien d'un site web. En France, c'est à l'INA et à la BNF que sont dévolues les missions de collecte du « dépôt légal du

mission, une sélection des sites web à conserver est opérée et des sauvegardes automatiques sont effectuées régulièrement, le délai entre deux sauvegardes est variable et est à la discrétion des organismes effectuant ces sauvegardes. Bien entendu, la fréquence de sauvegarde est un problème important, puisque rien ne garantit que du contenu ne sera pas ajouté, puis supprimé dans l'intervalle entre deux sauvegardes et donc ne sera pas perdu. Par ailleurs, de nombreux sites ne sont pas du tout sauvegardés [Milligan, 2016].

Ce dernier cas est probablement le plus révélateur des changements qui touchent la position et le rôle de l'historien dans un contexte numérique. Au delà de pouvoir consommer différemment les sources, dans leur version numérique plutôt que physique, il peut être un acteur de la collecte de nouvelle source, comme le montrent les démarches de *crowdsourcing*. Plus encore, il peut tenir un rôle majeur dans la préservation des sources en devenir, celles du web. Les politiques de conservation des données du web ou des données issues du mouvement de l'*open data* sont encore en gestation et l'historien doit y tenir un rôle de premier plan si nous souhaitons garantir la conservation et l'exploitabilité de ces nouvelles sources.

Les nouveaux moyens d'accès aux sources ou les nouvelles données exploitables se traduisent par des activités de recherches identifiées. Dans la taxonomie TaDiRAH, ces activités sont clairement repérables et prennent place dans plusieurs catégories (voir tableau 2.1). Ces activités ne sont pas nécessairement nouvelles au sens strict, en revanche, elles peuvent faire appel à des techniques particulières qui n'ont pas de sens en dehors d'un contexte numérique. La définition de l'activité de collecte mentionne par exemple aussi bien le recensement bibliographique académique, que l'extraction de sites internet (en anglais *scrapping*). De la même manière, l'activité d'annotation est très liée aux objets numériques puisqu'il s'agit de les décrire pour en faciliter l'accès par les chercheurs et par les machines. Pour l'histoire, l'un des enjeux majeur est aujourd'hui d'être en capacité de gérer une masse d'information, sources primaires, secondaires ou données de recherche de plus en plus vaste.

Gestion des données

Parmi les outils largement adoptés par les historiens figurent fort heureusement des outils qui satisfont cet usage, en particulier les gestionnaires de bibliographie [18]. De nombreuses formations à leur usage sont régulièrement organisées par diverses institutions, témoignant de la vigueur de leur usage [19]. Ces logiciels servent à collecter, classer et enrichir un catalogue de références bibliographiques. Ils autorisent une visualisation très rapide dans une seule interface de l'intégralité de la documentation collectée lors de la recherche. Loin de se limiter à la gestion des sources secondaires, ils permettent d'indexer dans sa base de connaissance un grand nombre de types de documents différents, du livre au manuscrit en passant par les archives ou les articles scientifiques. Ils sont par ailleurs dotés de fonctionnalités puissantes d'enrichissement des métadonnées

web » (http://www.institut-national-audiovisuel.fr/collecte-depot-legal-web.html).

18. En témoigne par exemple l'enquête d'Émilien Ruiz, dans laquelle 50% des répondants ont indiqué utiliser ce type de logiciel [Ruiz, 2015] ou la série d'entretiens menée dans [Watkins, 2013].

19. Parmi ces nombreuses formations, citons celles régulièrement organisées par le blog Zotero francophone (http://zotero.hypotheses.org) ou l'association HackYourPhd (http://hackyourphd.org/).

des références bibliographiques, de prise de notes et d'ajout de fichiers. Les dernières versions de ces outils proposent même une synchronisation du catalogue entre plusieurs machines et un partage en ligne des bibliographies [20].

Leur principal intérêt réside dans leur capacité à générer automatiquement des bibliographies dans le style choisi par le chercheur qui sont compatibles avec les documents rédigés dans des traitements de texte ou écrits en LaTeX. Le gestionnaire de bibliographie le plus célèbre est probablement Zotero, développé depuis 2006 par le Roy Rosenzweig Center for History and New Media de l'Université Georges Mason. D'abord disponible sous forme d'extension du navigateur web Firefox, il est aujourd'hui disponible sous forme d'un logiciel indépendant compatible avec la plupart des systèmes d'exploitation [21]. Il est entièrement libre et open source. Pour toutes ces raisons, il est d'une grande utilité pour l'historien, son usage est par ailleurs très répandu dans les activités nécessitant de maintenir des bibliographies importantes, que ce soit en recherche ou dans d'autres secteurs d'activité publics comme privés. D'autres logiciels, à l'instar de Mendeley, propriété du groupe Elsevier rendent le même type de service.

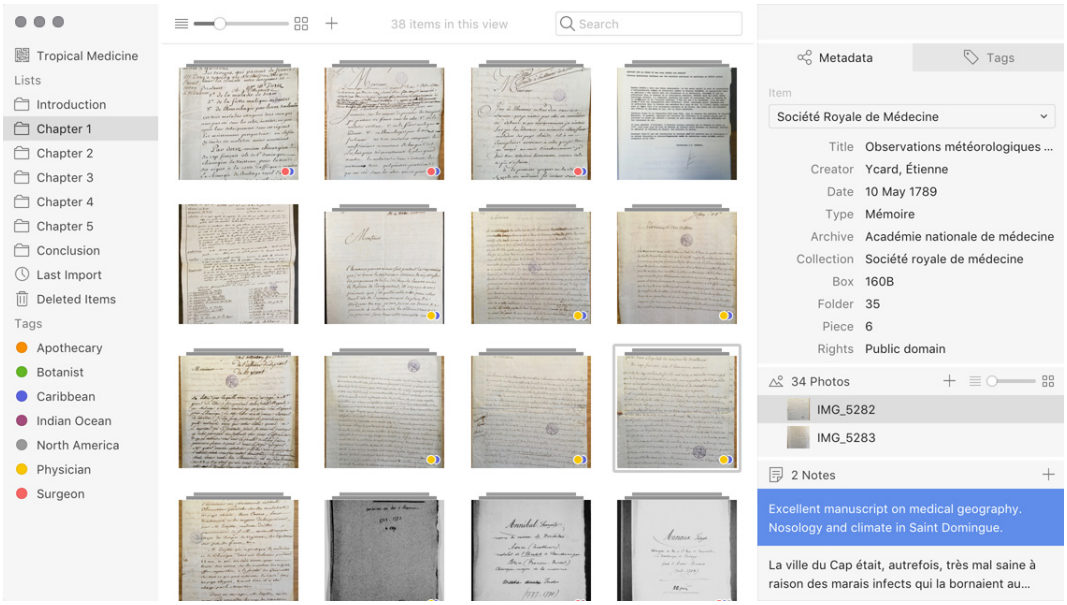
Des outils plus spécifiques sont également développés pour des usages plus précis. C'est par exemple le cas pour la problématique des photographies de sources prises en archives, dont la gestion peut s'avérer complexe. L'exploitation de ces données mobilise en effet une grande variété de compétences et d'outils pour étiqueter, annoter ou transcrire ces images numériques. Le rapport de Rutner et Schonfled explique par exemple que faire le lien entre ces images et les notes de recherche constitue désormais un challenge important [Rutner et Schonfeld, 2012, p. 9]. Des outils commencent néanmoins à prendre en compte ces problématiques. Le RRCHNM développe par exemple l'outil Tropy (voir figure 2.1), dont l'objectif est de répondre à ces problématiques et faciliter la recherche, l'exploitation et le partage de ces photographies [22].

20. Zotero propose cette fonctionnalité via sa plateforme https://www.zotero.org/.

21. Zotero est compatible avec Windows, MacOs et Linux, le code source, maintenu par le RRCHNM est disponible sur Github à l'adresse : https://github.com/zotero/zotero.

22. L'outil Tropy est en cours de finalisation par le RRCHNM et est disponible à l'adresse https://tropy.org/. Cet outil est, comme les autres outils développés par le RRCHNM, libre et « open source », son code source est accessible à l'adresse : https://github.com/tropy/tropy.

FIGURE 2.1 – Vue de l'interface de Tropy (RRCHNM)



Tropy permet, entre autres, de classer les photographies par catégorie, les annoter, les étiqueter et propose un moteur de recherche.

2.2.3 Analyse des sources et production du discours historique

Comme nous venons de le voir, l'usage du numérique est répandu pour la construction et la gestion de corpus. Il est également présent lors des phases d'analyse, de production et de diffusion du discours historique. La taxonomie TaDiRAH (voir tableau 2.1) mentionne en effet diverses approches depuis l'analyse de contenu, jusqu'à la visualisation. Ces pratiques sont plus exigeantes sur le plan technique, nécessitent souvent l'usage de l'ordinateur via son terminal, en ligne de commande donc. Ces usages sont ainsi moins répandus. Il est difficile de mesurer l'usage de ces technologies, l'enquête d'Emilien Ruiz montre un usage très restreint, par quelques répondants, d'outil tels que « l'analyse textuelle, lexicométrie, textométrie », les « analyses statistiques » ou les « langages informatiques ». Ces catégories très larges ne permettent pas de décrire plus finement l'usage.

Ceci étant, des initiatives témoignent non seulement de l'intérêt de ces techniques sur le plan historique, mais également de la maîtrise technique dont font preuve certains usagers, chercheurs ou équipes de recherche. Parmi ces initiatives, il nous faut citer le site *"The Programming Historian"* [AFANADOR-LLACH et collab., 2017] qui offre de nombreux tutoriels techniques sur des sujets variés essentiellement portés sur l'analyse de données. Le nombre de ressources présentes sur cette plateforme augmente rapidement, témoignant de la bonne marche du projet et de son impact [23]. Nous présenterons ici quelques exemples montrant l'état de la pratique sur cet aspect du numérique pour la recherche en histoire.

Lorsque l'on parle de méthodes informatiques pour analyser des sources, on entend

23. Le site (https://programminghistorian.org/) possède désormais l'ISSN 2397-2068, faisant, s'il était nécessaire, la preuve de son intérêt majeur.

souvent méthodes quantitatives, du moins en France, en référence à la tradition historiographique de l'histoire sérielle. Les grandes thèses d'histoire sérielle en France ont en effet marqué l'historiographie par le volume des données recueillies et formalisées dans ces travaux, qui n'ont rien à envier à ce qu'il est possible de faire avec les outils d'aujourd'hui. Ces grandes thèses ont permis, comme l'a montré Jean-Philippe Genet dans son texte *Histoire, Informatique, Mesure* dès 1986 [Genet, 1986], de faire émerger la notion de mesure dans des disciplines pour lesquelles elle était inhabituelle.

Pour prendre un exemple, on peut penser à l'immense travail de Pierre et Huguette Chaunu, dont l'accumulation de données a représenté près de 3900 pages en « petits caractères très denses » et dont le manuscrit de thèse, comptant lui même 3400 pages n'était, pour reprendre les mots de Pierre Chaunu, que « le commentaire un peu long » [Chaunu, 1987, pp. 81-82]. L'histoire sérielle, ou histoire quantitative, avait pour objectif de produire de l'histoire depuis des « séries » de données. Des données de prix du blé ou de mouvement de navires dans tel ou tel port ont par exemple été extraites des sources. Ces indices microscopiques, mis en fiches, en séries et en tableaux, devenaient progressivement porteurs de sens [24].

Cette histoire, qui a connue son apogée dans les années 1970 en France, trouvait naturellement dans les fiches et les tableaux son outil de travail privilégié. Elle pose la question de l'adéquation entre question de recherche et outil et format de données. En ce sens, le format tabulaire n'usurpe pas le nom de base de données que l'on peut lui donner. La structuration interne est en effet dirigée par un important effort de modélisation. Chaque ligne de tableau représente souvent une entité dont les attributs sont décrits dans des colonnes. Ce type de format se prête particulièrement bien à certaines activités intellectuelles. Il peut s'agir de lister des sources ou des entités trouvées dans les sources afin de synthétiser l'information. Les tableaux ont été et sont toujours régulièrement employés par les historiens.

Si l'outil tableau s'est imposé à Pierre Chaunu et à bien d'autres historiens, c'est avant tout parce que l'approche retenue, essentiellement statistique, imposait un format qui pouvait faciliter le traitement des données en séries. Comme le remarquait plus récemment Philippe Rygiel [Rygiel, 2012], les histoires produites par les historiens ne diffèrent pas seulement par leur contenu ou les thèses qui y sont mises en avant mais également par la manière dont elles sont produites. De la reproduction presque maniaque de la source, à la manière d'Arlette Farge, à l'extraction d'information opérée par les tenants de l'histoire quantitative, on observe une diversité d'approche qui trouve aujourd'hui son prolongement dans une grande diversité d'outils fournis par l'informatique, qui dépasse de loin les seules problématiques quantitatives.

Les méthodes numériques d'analyse de textes font partie de cette diversité d'approche. Bien que cela soit à nuancer en fonction des périodes étudiées et des méthodes de recherche adoptées, une grande partie des sources accessibles à l'historien sont des sources textuelles. Tout naturellement, des approches numériques visant à les analyser se sont développées, depuis les premiers travaux utilisant l'ordinateur pour l'analyse de texte du père Roberto Busa. Cette recherche visant à indexer l'intégralité de

24. Sans faire ici une histoire exhaustive de ce courant historique, on peut citer l'important ouvrage de Pierre Chaunu [Chaunu, 1978]. L'ouvrage de synthèse sur l'historiographie française dirigé par Delacroix, Dosse et Garcia [Delacroix et collab., 2007] dresse un portrait précis de l'ouvrage quantitative et de son rôle important pour le développement de l'historiographie française du XX^{e} siècle.

l'œuvre de Thomas D'Aquin a été menée à partir de 1949 en collaboration avec IBM [25]. Elle est régulièrement citée comme la première expérience d'une approche d'humanités numériques, bien que cette filiation quasi mythique soit contestée par de nombreux chercheurs [Scheinfeldt, 2014].

De nos jours, il existe de nombreuses approches d'analyse de texte fondées sur différentes techniques, textométrie, lexicométrie ou encore apprentissage automatique [26]. Elles visent à étudier le texte comme un ensemble structuré en utilisant des méthodes statistiques quantitatives ou qualitatives. Ces méthodes ont été popularisées par les travaux de Franco Moretti rassemblés dans son ouvrage *Distant Reading*. Ce chercheur a en effet mené beaucoup d'études impliquant des méthodes dites de lecture distantes pour éclairer, avec un certain succès, l'histoire du roman. L'appellation lecture distante s'oppose, sur la forme, à la lecture proche (en anglais *close reading*), la lecture attentive des documents par le chercheur lui même. Cette opposition n'est qu'une opposition de pure forme, comme l'a expliqué Frédéric Clavert, pour qui ces deux formes de lectures n'ont d'intérêt, en tout cas pour l'historien, que si on les rapproche. Ces deux lectures n'ont pas le même usage et ne répondent pas aux mêmes objectifs. L'application de la lecture distante à un corpus peut être un bon moyen de sélectionner des sources auxquelles appliquer une lecture proche [Clavert, 2014].

Il existe de nombreux outils qui permettent d'appliquer à des corpus de sources ce type de méthodes d'analyse de texte. Le plus célèbre et utilisé d'entre eux est *Voyant Tools*, développé par Stéfan Sinclair et Geoffrey Rockwell [27]. Cet outil, très facile d'accès puisque disponible sous la forme d'une simple application web, permet de produire rapidement des analyses quantitatives sur un corpus de texte. Il est capable de produire, comme nous pouvons le voir à la figure 2.2 de nombreuses représentations graphiques, depuis les nuages de mots jusqu'à des courbes de fréquence de termes plus élaborées. Dans le même ordre d'idée, sans prétendre toutefois à l'exhaustivité, nous aurions également pu citer le logiciel IRaMuTeQ, reposant sur le langage de programmation R, très utile pour les traitements statistiques, développé par le Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales (LERASS) [28]. Il permet de produire des représentations similaires à celles de *Voyant Tools* mais également des calculs de similarité entre segments de texte.

D'autres approches sont fondées sur des outils d'apprentissage automatique et des approches probabilistes. Elles sont plus difficiles d'accès que les méthodes statistiques évoquées plus haut mais sont utiles pour certaines tâches. Elles peuvent par exemple être utilisées pour trouver des sujets récurrents dans des grands corpus de texte. Ces sujets se présentent sous la forme d'une liste de mots-clés qui donnent une indication du contenu des documents. À titre d'exemple, cette méthode a été utilisée par [David J et Sharon, 2006] pour extraire une représentation des sujets traités par un journal du XVIII^{e} siècle et étudier leur évolution.

Les méthodes d'analyse de texte sont donc employées dans des tâches de recherche très diverses qui dépassent de loin le simple décompte de tel ou tel terme dans un corpus.

25. Le résultat de ce travail de 30 ans est aujourd'hui accessible en ligne à l'adresse http://www.corpusthomisticum.org/.

26. Ces méthodes ne sont pas seulement utiles à la recherche, on trouvera par exemple un état de l'art des approches pour le data-journalisme dans [Flaounas et collab., 2013].

27. Accessible publiquement à l'adresse : https://voyant-tools.org/.

28. IRaMuTeQ est libre et *open source*, il est téléchargeable à l'adresse http://iramuteq.org/.

FIGURE 2.2 – Vue de l'interface de Voyant Tools

Elles permettent de tenir compte de plus grands volumes de texte, de chercher des liens sémantiques difficiles à trouver manuellement ou encore de trouver par lecture distante des sources intéressantes, ensuite soumises à une lecture critique. La valeur de ces outils d'analyse se révèle souvent au travers des représentations graphiques qu'ils permettent de produire. Ces représentations sont elles aussi nombreuses en fonction des méthodes utilisées, depuis de simples courbes jusqu'à des nuages de mots, mais elles sont surtout les véritables outils d'interprétation du chercheur.

Encore une fois, nous ne pouvons présenter toutes ces représentations tant celles-ci sont nombreuses, comme l'a montré Johanna Drucker dans son ouvrage intitulé *Graphesis : visual forms of knowledge production* [DRUCKER, 2014]. L'utilisation de certaines d'entre elles, dont fait mention la taxonomie TaDiRAH, témoigne cependant d'activités de recherche répandues dans les humanités numériques. L'analyse de réseaux désigne les approches ainsi que les outils numériques qui permettent de représenter, sous formes graphiques la plupart du temps, les relations des entités les unes envers les autres. Elles peuvent par exemple donner les moyens de repérer l'importance d'une personne dans un réseau social ou de représenter des communautés identifiées dans un ensemble d'entités. Parmi les outils utiles à ces constructions visuelles, le logiciel *Gephi* est probablement le plus important. Ce programme, libre et *open source* permet de manipuler aisément des graphes et d'en produire des représentations graphiques élégantes. L'un des projets les plus ambitieux intégrant l'analyse de réseaux est le projet *Mapping the Republic of Letters*[29]. Ce projet, qui vise à étudier les correspondances d'intellectuels a déjà donné lieu à des publications faisant un usage pointu de l'analyse de réseaux [WINTERER, 2012].

Comme l'a montré Claire Lemercier, les outils d'analyse de réseaux permettent de produire des représentations des réseaux fondées sur des données plus que sur des intuitions, mais c'est l'usage de concepts comme la centralité ou la cohésion « qui

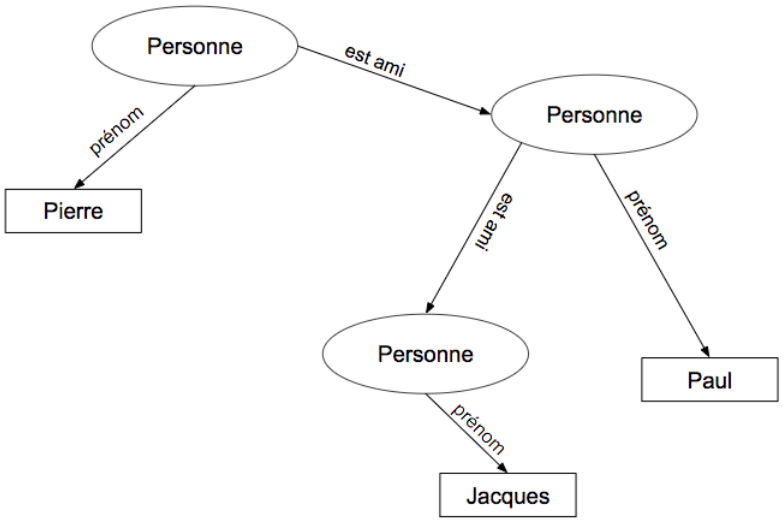
29. Ce projet est développé à l'Université de Stanford, avec de nombreux partenariats internationaux. Les détails et réalisation du projet sont visible sur le site http://republicofletters.stanford.edu.

fait entrer dans l'analyse de réseaux, plus que celui des logiciels, possibles ou non selon les sources » [Lemercier, 2005, p. 94]. C'est donc bien l'ambition historique, la problématisation ou les sources qui dirigent vers ce type d'outils numériques plutôt que d'autres. Pour autant, des projets tel que *Mapping the Republic of Letters* montrent bien l'influence des outils et des compétences du numérique pour la formulation des questions de recherche aussi bien que pour leur réalisation.

Pour l'informatique, ces méthodes se traduisent le plus communément par le traitement de graphes[30]. Ces objets, composés de nœuds (ou sommets) et de liens (ou arrêtes) sont utiles à un grand nombre de tâches informatiques et peuvent représenter beaucoup de type de données. On peut imaginer représenter, tout simplement, un réseau de personnes, qui deviennent les nœuds du graphe, reliés par des caractéristiques. De cette manière si deux personnes, Pierre et Paul sont amis, et que le lien représente l'amitié, le graphe sera composée de deux nœuds et d'un lien entre ces deux personnes. Si, une troisième personne, Jacques, est ami avec Paul, alors le graphe sera composé de trois nœuds pour seulement deux liens. Cette représentation permet par exemple de visualiser rapidement que Paul se trouve au centre du réseau ou encore qu'il est possible que Pierre et Jacques se soient un jour ou l'autre rencontrés, par l'intermédiaire de Paul. La figure 2.3 représente ces relations sous forme de graphe. On peut les enregistrer au format *RDF*, le format du web sémantique. Ce format traite les données sous forme de triplets : le sujet (la ressource à décrire), le prédicat (l'information décrite) et l'objet (la valeur de l'information). Dans le cas décrit ici :

— La classe « Personne » représente la ressource à décrire.
— « Prénom » est le prédicat, il a pour objet personne et pour valeur le nom de la personne.
— La valeur du prénom, Pierre par exemple, est l'objet.

Figure 2.3 – Exemple de réseau sous forme de graphe



Bien entendu, cet exemple est le plus simple que nous pourrions trouver. Cependant, à l'heure des réseaux sociaux et des millions de connexions, directes ou indirectes entre

30. Ces objets ont été théorisés. La « théorie des graphes » et leur manipulation est une des composantes majeure de la formation des informaticiens.

des personnes, des contenus, des services, etc. La recherche sur les graphes est très active. Elle cherche par exemple des moyens de traiter des grands graphes dans des temps de calcul raisonnables [Gonzalez et collab., 2012]. Il peut également s'agir de trouver des algorithmes permettant de détecter dans ce type de réseau des communautés par des moyens algorithmiques [Blondel et collab., 2008]. Comme nous le verrons dans un de nos prochains chapitres [31], ils peuvent également servir à représenter des données « brutes », comme du texte, afin d'opérer certains calculs ou certaines transformations. La modélisation des réseaux et l'analyse de graphe sont des terrains sur lesquels s'expriment déjà la collaboration entre chercheurs en SHS et informaticiens. À titre d'exemple, le programme transversal interdisciplinaire « Modélisations des Réseaux en Histoire (MOREHIST) » organise des journées d'étude rassemblant diverses spécialités de recherche intéressées à la modélisation et à l'analyse de réseaux [32].

Depuis quelques années, sous l'impulsion de plusieurs grandes institutions telles que la Bibliothèque Nationale de France (BNF) ou Europeana, se développent des versions sémantiques des moteurs de recherche que nous avons déjà évoqués. La BNF met ainsi à disposition le site Data BNF, en parallèle de Gallica et de son catalogue général. Europeana dispose aussi d'un point d'accès sémantique [33]. Parmi les outils spécifiquement dédiés aux SHS, le moteur de recherche Isidore permet un accès à ses données liées, Open Edition a également récemment ouvert la version sémantique de sa plateforme d'accès aux revues en ligne Persée [34]. Dans le contexte de l'usage pour des problématiques historiques, ces bases de données liées autorisent la conception de requête qu'il n'est pas possible d'envisager sans les technologies sémantiques. Ces requêtes sont exprimées dans des langages informatiques spécialement dédiés à cet usage, tel que SparQL [35]. Sans détailler exhaustivement le fonctionnement de ces technologies, sur lesquelles nous reviendrons, elles permettent d'extraire des informations qui peuvent s'avérer pertinentes pour l'historien.

2.2.4 Collaboration de recherche entre histoire et informatique

Les outils, approches et méthodes dont nous venons de parler montrent l'étendue de la relation entre la recherche en histoire et le numérique. Bien que, comme nous l'avons vu,

31. Voir le chapitre 5.

32. Le site internet du réseau MOREHIST se trouve à l'adresse : http://sms.univ-tlse2.fr/accueil-sms/le-labex-sms/seminaires/modelisation-des-reseaux-en-histoire-morehist-/.

33. Les collections de la BNF sous forme de données liées peuvent être interrogées à l'adresse http://data.bnf.fr/. Europeana est accessible à l'adresse : http://www.europeana.eu. Son accès sémantique de type SparQL Endpoint se trouve à cette adresse : http://sparql.europeana.eu/.

34. Isidore est accessible à l'adresse suivante : https://www.rechercheisidore.fr/ dans sa version moteur de recherche et à cette adresse dans sa version sémantique (SparQL Endpoint) : https://www.rechercheisidore.fr/sqe.. Persée, le portail de revue en ligne édité par Open Edition est accessible à cette adresse en version http://www.persee.fr/ et à cette adresse en version sémantique : http://data.persee.fr/.

35. SPARQL est un langage et un protocole d'échange de données conçu pour extraire, modifier ou créer de l'information dans des bases de données sémantiques, développé et maintenu par le W3C. Il opère sur des données formalisées en *Resource Description Format (RDF)*, son acronyme signifie **S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage. La norme de description de SPARQL est accessible à l'adresse https://www.w3.org/TR/sparql11-query.

les niveaux d'utilisation des différents outils soient très variables, du plus systématique au quasi confidentiel, il existe des approches numériques pour toutes les activités de l'historien. Cette relation se matérialise parfois tout simplement par l'usage d'un objet générique, comme l'appareil photo, qui répond à un besoin matériel. Pour d'autre outils, la relation est plus complexe lorsque la recherche est utilisatrice autant que conceptrice de son propre outil, comme c'est le cas avec les logiciels développés par le RRCHNM, qui s'est fait une spécialité d'outiller l'historien et les sciences humaines au sens large.

Dans d'autres cas, les outils sont issus de collaboration plus profondes et résultent de démarches de recherche en informatique pour lesquelles l'historien tient un rôle prépondérant. En recherche d'information par exemple, bien que certaines approches sont purement informatique, algorithmique ou statistique, l'historien peut avoir un rôle majeur dans le développement et l'évaluation des travaux. En effet, ces derniers nécessitent souvent une base de connaissance « de l'expert », qui permet de construire des jeux de données indispensables à l'entraînement des algorithmes pour une tâche particulière de description des contenus ou de recherche d'information. La collaboration entre informaticiens et historiens aboutit à la construction de modèles de description stables et réemployables pour des jeux de données similaires. Une telle collaboration a par exemple été au cœur du projet NAVIDOMASS[36] dont un des objectifs était de permettre la recherche dans une vaste collection de lettrines datant des débuts de l'imprimerie. Les historiens et historiens de l'art du Centre d'Études Supérieures de la Renaissance (CESR) ont fourni la connaissance et le vocabulaire de description et de classement qui ont permis de concevoir des techniques d'extraction et d'indexation des images de lettrines [Coustaty et Ogier, 2015].

Des collaborations de cette nature ont également été menées à grande échelle pour la numérisation des sources et l'extraction d'information. Les projets de numérisation et reconnaissance de la presse ancienne ont impliqué et impliquent toujours des collaborations fortes entre informaticiens, historiens et archivistes et donnent des résultats qui sont intégrés dans les outils de recherche de sources. Le projet *Europeana Newspaper*, financé par l'Union Européenne entre 2012 et 2015 a par exemple favorisé l'accès en texte intégral à près de 12 millions de pages de journaux numérisés, devenant le premier contributeur à la bibliothèque numérique européenne Europeana.[37]. A la suite de ce travail, le projet *newsEye*, également financé par l'Union Européenne dans le cadre des financements Horizon 2020 adoptera une démarche de collaboration forte entre histoire et informatique. Elle aura cette fois pour objectif de favoriser l'exploitation et la visualisation des données extraites de la presse ancienne sur des cas historiques précis de l'histoire des migrations jusqu'à celle des conflits armés de l'Europe contemporaine.

Le rôle que tient ici l'historien dans la collaboration interdisciplinaire n'a pas vocation a être joué uniquement a posteriori, lorsqu'il s'agit d'évaluer la qualité des traitements mis en place par l'informatique ou la praticité d'usage d'un outil. La réflexion de l'historien, sa connaissance et ses usages des sources doivent être intégrés au plus tôt dans la recherche comme l'expertise qui guide le travail de l'informaticien. Sans faire une liste exhaustive des nombreuses problématiques qui imposent une telle collaboration, le cas des archives du web est particulièrement signifiant.

36. Projet financé par l'Agence Nationale de la Recherche (ANR), entre 2006 et 2009.

37. Les détails des différentes réalisations du projet sont accessibles à l'adresse :http://www.europeana-newspapers.eu/. Le rapport final du projet est quant à lui disponible à l'adresse http://europeananewspapers.github.io/.

En effet, bien que plusieurs institutions « archivent » déjà le web, c'est à dire enregistrent l'état de certains sites web à intervalle régulier, de nombreux verrous scientifiques restent à lever pour garantir l'exploitation future de ces données. Le volume extraordinaire de ces enregistrements contraint à une sélection des pages archivées. L'indexation de ces contenus ne peut donc être que partielle. L'archivage seul ne suffit pas, il faut fournir les outils d'interrogations de ces archives, qui, pour l'heure, sont très limités. L'historien a un rôle a joué aussi bien dans les choix qui sont opérés en terme d'archivage que pour la conception d'outils d'interrogation pertinents. Des travaux ont ainsi révélé les failles des choix qui ont été faits jusqu'à présent en matières d'archivage [Milligan, 2016]. Des équipes pluridisciplinaires ont été constituées regroupant historiens, informaticiens et archivistes pour répondre à ces problématiques. Elles œuvrent à fournir des moyens de recherche d'information qui pourront permettre un accès à ce qui sera, selon toutes les vraisemblances, une des sources majeures des futures générations d'historien [Jackson et collab., 2016].

L'exploitation de ces données, comme beaucoup d'autres, passent par des mécanismes de recherche et d'accès à l'information. Les méthodes numériques de traitements exigent des corpus et des données adaptés [Baillot, 2015]. Il est inutile d'archiver une telle quantité de pages web s'il n'existe aucun moyen permettant de les exploiter rigoureusement. De la même manière, les outils d'analyses de texte sont puissants et autorisent la création de visualisations sopistiquées, mais ils n'ont que peu d'intérêt si les documents soumis sont lacunaires ou dégradés. Pour dépasser ces problématiques et assurer l'exploitation et la conservation future des sources numériques comme des résultats de la recherche, les collaborations entre les disciplines des SHS et celles du numériques sont naturellement fondamentales [Idmhand et Walter, 2014]. Elles requièrent cependant un positionnement conceptuel global, au carrefour des documents, des usages et des traitements numériques.

2.3 Positionnement et cadre théorique de notre étude

Ces relations entre l'histoire et le numérique, qu'elles soient ou non issues de problématiques scientifiques, ont des conséquences sur les activités de recherche. La liste même parcellaire d'outils, d'approches ou de collaborations que nous avons dressée témoigne néanmoins des impacts du numérique sur la recherche dont toute réflexion sur les usages devrait tenir compte. À la lumière de l'état de la pratique, il apparaît ainsi que la relation entre l'historien et ses sources est profondément impactée par le développement du numérique. Elle est bouleversée dans sa temporalité et dans ses lieux. Les outils numériques favorisent une recherche plus éloignée des centres de conservation des archives, des allers et retours plus fréquents avec les sources et une recherche plus itérative qu'auparavant.

2.3.1 L'accès aux sources et aux données, une problématique centrale

Parmi tous les outils et les approches numériques pour l'histoire que nous avons évoqués, nous avons choisi de nous intéresser, comme nous l'avons écrit dans notre chapitre

1, aux bibliothèques numériques et aux moteurs de recherche de sources. Bien entendu leur importance n'est pas la même pour tous les chercheurs et il ne faut sans doute pas exagérer leur importance. Mais selon le type de sources manipulées et les habitudes de travail propres à chacun, ils sont désormais un des moyens d'accès privilégié aux sources de l'histoire. À ce titre, la qualité de leur fonctionnement est primordiale. Ce n'est cependant pas la seule raison pour laquelle nous avons souhaité nous intéresser à eux.

Dans l'état de la pratique décrit plus haut, il nous apparaît deux catégories d'outils. D'abord il y a ceux pour lesquels, quel que soit le service offert, le chercheur a la maîtrise de leur fonctionnement, de leur paramétrage et des traitements qu'ils font subir aux données. C'est le cas d'une large part des outils que nous avons évoqué, depuis les gestionnaires de bibliographies jusqu'aux outils de modélisation de réseaux, dont le fonctionnement est connu et configurable. Ensuite, il y ceux dont le fonctionnement est masqué et dont le chercheur n'a pas la maîtrise parce qu'ils reposent sur des choix conçus par d'autres. C'est le cas des bibliothèques numériques, moteurs de recherche et bases de données de sources en ligne.

Bien entendu, leur fonctionnement repose sur des bonnes pratiques et des efforts importants de normalisation pour la description des contenus auxquels ils donnent accès. Mais il règne toujours, dans ce domaine, une grande confusion. Un travail important reste à mener pour proposer dans un environnement numérique des schémas de description adaptés à la complexité des archives [Baillot, 2016; Pitti et collab., 2018]. Par ailleurs, certains des traitements nécessaires au bon fonctionnement des bibliothèques numériques, que ce soient les algorithmes de recherche d'information, de classement des résultats ou d'extraction d'information, ne relèvent ni du savoir de l'archiviste ni de la connaissance de l'historien mais sont au contraire de la compétence de l'informaticien.

Malgré cette limite, ces bonnes pratiques existent et garantissent la rigueur et la qualité de l'information présente dans les bibliothèques numériques des institutions patrimoniales. En revanche, lorsque les systèmes utilisés pour accéder aux sources sont des systèmes commerciaux, comme Google, la problématique est bien différente. Ce type de système, même dans leur version dédiée aux chercheurs[38], sont des boites noires. Pourtant, si l'on s'en tient à l'enquête de [Kemman et collab., 2013] ou aux travaux de [Milligan, 2013], ils sont très utilisés pour accéder aux sources. Les conséquences de cet usage, que nous avons pour l'essentiel déjà évoquées, et qui ont été très bien résumées par [Clavert, 2012, 2013] sont nombreux et prêchent pour un effort d'étude et d'explication des biais de la recherche d'information pour l'historien.

Par ailleurs, il est intéressant de remarquer que les traitements informatiques qui opèrent au sein des systèmes de recherche d'information sont essentiellement des processus d'analyse de document. Ils ont à ce titre un plus large potentiel. Ils sont aussi employés par les approches d'analyse de texte, pour la lecture distante ou pour des calculs lexicométriques par exemple. Ils peuvent l'être également pour les analyses de réseaux dans la mesure ou les outils de recherche d'information font aujourd'hui une large place aux données sémantiques et à la mise en relation de documents ou d'objets extraits de ces documents [Pouyllau, 2012]. S'intéresser à leurs effets dans le contexte des moteurs de recherche de sources, c'est aussi s'intéresser à certains outils d'analyse employés par l'histoire et les humanités numériques et donc observer leur

38. Google Scholar par exemple, https://scholar.google.fr.

biais potentiellement aussi de ce point de vue.

Enfin, s'il fallait une raison supplémentaire de s'intéresser à ces objets numériques, nous avons déjà dit dans notre introduction (voir chapitre 1) que les problématiques de formation et d'enseignement étaient au cœur de l'histoire numérique. Or, les outils sur lesquels nous avons choisi de nous focaliser comptent parmi ceux pour lesquels les besoins de formation sont importants et qui font partie du socle commun de formation aux outils numériques qu'appelait de ses vœux Émilien Ruiz dès 2011 [Ruiz, 2011].

Les bibliothèques numériques et les moteurs de recherche sont donc, du point de vue de la pratique de recherche, des objets centraux en tant que point de jonction entre le chercheur et le document. Du point de vue technique, ils le sont également. Leur objectif général est de traduire l'expression d'un besoin d'information en une liste de résultats la plus pertinente possible. Ils jouent en ce sens un rôle d'intermédiaire qui entretient une relation aussi bien avec les documents qu'avec les utilisateurs. Les deux aspects sont de notre point de vue aussi important l'un que l'autre, les effets de bord que produisent les moteurs de recherche ne peuvent s'expliquer par la seule boite noire algorithmique qui parcourt, traite et ordonne les documents. La relation entre l'utilisateur et le système est également fondamentale.

Cet aspect du problème nous a été justement rappelé par Carlo Ginzburg, dans un très court texte intitulé *Conversation avec Orion*. Il y raconte son expérience avec le catalogue de la bibliothèque de l'Université de Californie (UCLA) où il enseigne et écrit :

> *« Orion (prononcer à l'anglaise Oraion) est le nom du logiciel qu'utilise le catalogue en ligne de la Research Library de l'Université de Californie à Los Angeles (Ucla). Par extension, Orion — aujourd'hui remplacé par une version qui prétend être plus avancée, Orion 2 — a fini par désigner le catalogue lui-même. En ce qui concerne l'informatique, je suis malheureusement un analphabète. Ma pratique d'Orion repose sur quelques commandes de base peut-être mal effectuées. Je dis « peut-être » car j'ai l'impression que les catalogues de bibliothèque (les catalogues électroniques ne dérogent pas à la règle) ont été conçus de tout temps pour permettre à ceux qui les utilisent de ne trouver que ce qu'ils cherchent. Moi-même, je les utilise ainsi. Mais je les utilise également très souvent dans un objectif autre, sinon opposé : celui de trouver ce que je ne cherche pas du tout, voire même ce dont je ne soupçonne pas l'existence. Il s'agit de quelque chose d'assez évident. Si le hasard, comme nous le rappelle avec autorité Dionisotti, est la règle qui préside à la recherche de l'inconnu, il semble évident que le chercheur doive s'efforcer de multiplier les possibilités en procédant à tâtons. [Ginzbourg, 2006, p. 129] »*

Le récit de Carlo Ginzburg est loin d'être anecdotique, c'est en partie à la suite de ces recherches dans *Orion* qu'il écrira, dix ans plus tard, son ouvrage sur les *Benandanti* [Ginzbourg, 1980]. Ce que Ginzburg explique relève d'un usage « détourné[39] », un cas où l'intelligence stratégique de l'utilisateur dépasse les intentions du système [Merzeau, 2013]. C'est la relation entre le chercheur et l'objet technique qui se manifeste

39. On trouvera un autre exemple significatif d'un tel détournement dans [Bénel, 2014].

ici et qui montre la complexité de l'étude des impacts des objets techniques, dont font partie les outils numériques, pour la pratique de la recherche.

Cette complexité du positionnement des bibliothèques numériques, à la fois objets centraux pour la recherche et corps intermédiaires sur le plan technologique, impose d'établir un cadre conceptuel qui puisse permettre de tenir compte de tous les aspects de la problématique. Ce cadre conceptuel, construit autour de la notion de ressource numérique, a été expliqué dans notre introduction (voir chapitre 1). À ce stade, il est tout de même nécessaire de revenir sur les principaux éléments de sa définition et de son rôle pour les développements qui vont suivre.

2.3.2 Données, documents, ressources numériques : état de l'art et définitions

Pour produire leurs travaux de recherche, les chercheurs sont donc amenés à utiliser, voire à concevoir, des ressources numériques. L'acception de base du mot « ressource » désigne, selon l'académie française « ce qui peut fournir ce dont on a besoin » [40], le dictionnaire Oxford donne une définition très proche pour le mot de langue anglaise "*resource*" [41]. L'ajout des mots « numérique » ou "*digital*" précise le contexte d'application, mais ne change pas le sens initial du terme « ressource ».

Cependant, cette expression se doit tout de même d'être définie formellement. Ces définitions très générales masquent en effet les logiques de création et de diffusion de ces objets complexes situés dans des environnements sociaux et technologiques bien identifiés. Les définitions générales que nous avons citées soulèvent ainsi les notions du besoin et de l'utilité, qui ne trouvent pas de traductions évidentes dans un contexte numérique et qui ont suscité de nombreux travaux scientifiques.

L'informatique et les sciences de l'information ont ainsi produit une réflexion sur le sens du terme « ressource numérique » visant essentiellement à établir des distinctions claires sur la granularité des objets qui sont manipulés dans le monde numérique. En effet, si nous n'apportons pas plus de précisions aux définitions que nous venons d'évoquer, une ressource peut désigner tout à la fois une donnée, un document, un fichier. Or, le besoin de clarification de chacun de ces termes s'est progressivement imposé au cours des années 1990 avec le développement du « World Wibe Web ». Bien que certains des concepts que nous évoquons s'ancrent dans une perspective historique plus vaste, nous limiterons volontairement notre analyse de l'état de l'art, sur le plan chronologique, de l'émergence du Web à aujourd'hui.

Le terme « ressource » n'apparaît pas dans les premiers documents de spécification du Web. Son premier usage date de Juin 1994 lorsque Tim Berners-Lee, père du Web, rédige un mémo donnant vie à la syntaxe *Universal Resource Identifier (URI)* [Berners-Lee, 1994] visant à donner à la communauté des premiers développeurs du Web un moyen de nommer et d'adresser ce que l'on appelle alors indistinctement les « objets »

40. Définition de la version 9 du dictionnaire de l'académie française, en ligne : http://www.academie-francaise.fr/le-dictionnaire/la-9e-edition.

41. Définition du dictionnaire Oxford en ligne accessible à l'adresse https://en.oxforddictionaries.com/definition/resource.

du Web[42]. L'URI, dont la signification sera par la suite modifiée dans le RFC 1738 [Berners-Lee et collab., 1994] pour signifier *Uniform Resource Identifier* et ses dérivés *Unifom Resource Locator (URL)* et *Uniform Resource Name (URN)* jetteront les bases de l'usage systématique du terme « ressources » pour tous les objets du Web, accessibles par le réseau, sans pour autant définir précisément ce que recouvre le terme.

Si le Web a donc été rapidement lié au concept de ressource, la définition en est toutefois restée très vague. Elle ne s'est précisée que quelques années plus tard, en 1998. Une ressource est alors définie de la manière suivante [Berners-Lee et collab., 1998] :

> *A resource can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), and a collection of other resources. [...] The resource is the conceptual mapping to an entity or set of entities, not necessarily the entity which corresponds to that mapping at any particular instance in time. Thus, a resource can remain constant even when its content—the entities to which it currently corresponds—changes over time, provided that the conceptual mapping is not changed in the process.*

Les ressources ont alors pris un sens plus précis, désignant ce qui est identifiable, mais se limitant à faire le lien vers l'entité cible. Cette entité est définie comme un fichier, un ensemble de fichiers ou un service. Elle est évolutive à travers le temps, à l'image du service de météo du jour, cité en exemple par Tim Berners-Lee. Dans ce cas, c'est le service qui constitue la ressource, pas les données météorologiques.

Cette définition, témoin des problématiques des créateurs du Web, est toujours d'actualité si l'on songe au développement du web sémantique. Elle précise des notions importantes, en particulier le caractère nécessairement identifiable d'une ressource, mais elle est loin d'être complètement satisfaisante. Le sens donné au terme « ressource » reste avant tout technique. Cette définition originelle a donc continué à évoluer dans le courant des années 2000. Sa signification s'est rapproché du mot « ressource » tel que les dictionnaires précédemment cités le définissent.

La définition du terme ressource numérique a donc progressivement inclus les notions de besoin et d'utilité, pour rendre un service positionné socialement et technologiquement dans un environnement : le contexte. Afin de prendre en compte ces notions et s'extraire d'une définition trop strictement technique, des travaux se sont orientés sur les contextes de production et d'usage des ressources numériques. Une distinction majeure a ainsi été établie en tenant compte, non plus seulement des aspects technologiques d'identité et de retrouvabilité sur le réseau, mais également des logiques de conception et d'utilisation prévue des ressources numériques. Ces réflexions sur les processus éditoriaux à l'œuvre sur le Web ont ainsi permis de distinguer les notions de **données**, de **documents** et de **ressources** numériques.

Données numériques : Lorsque des objets numériques ne sont soumis à aucun contexte éditorial, que ce soit en terme de production ou d'interprétation, on parle de **données numériques** [Lainé-Cruzel, 2004].

42. Les traces des premiers échanges des fondateurs du web sont disponibles dans des *Request For Comments (RFC)*. Ces mémos, généralement assez courts, sont les principaux témoins des échanges entre les différents acteurs des débuts du Web.

Documents numériques : Les documents numériques sont, quant à eux, organisés autour d'une structure stable garantissant une interopérabilité entre producteur et consommateur [Bachimont, 2004; Davenport et Prusak, 1997; Pédauque, 2006]. Outre cette nécessaire structuration, les documents numériques sont issus de logiques de production et de médiation également stables. C'est cette logique de conception tournée vers la production qui préside à la création de documents numériques, ils sont authentifiables et ont valeur de preuve [Lainé-Cruzel, 2004]. On observe ici aisément le glissement intellectuel qui conduit à privilégier des logiques éditoriales, voire juridiques, à des questionnements technologiques. Ce glissement de sens ne se limite pas à un processus intellectuel de définition. Il impacte également l'informaticien spécialiste de la gestion électronique de documents, dont l'attention doit désormais se porter sur l'authentification, la stabilité et plus généralement sur tous les aspects liés à la sécurisation des documents.

Ressources numériques : Dans ce contexte, les ressources numériques se différencient des documents par leur logique de conception, entièrement tournée vers l'usage cette fois. Les ressources rendent des services, elles renseignent mais ne peuvent pas avoir valeur de preuve. Elles peuvent donc être évolutives [Lainé-Cruzel, 2004].

La distinction des termes « données », « documents » et « ressources » en fonction de choix éditoriaux et de contexte de production et d'usage n'est pas la seule approche permettant de donner une substance à ces concepts dans un environnement numérique. Jean-Michel Salaün, par exemple, proposait dans un ouvrage de 2012 [Salaün, 2012] une réflexion fondée sur une analyse historique et bibliothéconomique du concept de document qu'il décrit dans trois dimensions, forme, contenu et fonction de transmission.

La forme définit le document comme un objet fini, portable, transférable, agissant comme une promesse. La seule forme peut nous permettre de déterminer l'utilité éventuelle du document.

Le contenu témoigne de « la variabilité de la valeur d'un document, qui dépend de son interprétation, par définition changeante selon les contextes » [43].

Le medium , dernière dimension, exprime les relations qu'un document a avec les autres et montre sa portabilité dans le temps aussi bien que dans l'espace.

L'intérêt majeur de cet ouvrage est de montrer que le développement du Web s'inscrit dans une tradition documentaire plus vaste imposant l'idée que le Web est un média « comme un autre » qui, après d'autres, vient modifier les conditions de production et de consommation des documents, allant jusqu'à créer un phénomène de « redocumentarisation » [Salaün, 2007]. Il n'est pas de notre ressort de développer plus avant ces questions qui dépassent de loin notre travail. Cependant, sans retrouver les termes « données » ou « ressources », mais en pensant le document numérique dans toutes ses dimensions, on retrouve dans cette réflexion ce qui semble faire l'essence des ressources numériques. Ce sont des objets complexes, mouvants, porteurs d'intention éditoriale par le producteur et d'intention d'exploitation par l'utilisateur, dont la valeur dépend fortement de leur contexte d'usage.

43. Salaün 2012, chap. 4.

Cette idée n'est pas sans rappeler l'approche instrumentale de Pierre Rabardel, qui proposait en 1995 [Rabardel, 1995] une distinction entre un objet matériel et un objet matériel qui s'inscrit dans le cadre d'un usage. Cet objet, un outil par exemple, devient un instrument dès lors qu'il est utilisé et donc qu'il a un usage bien déterminé. Il existe donc dans cette logique une différence majeure entre l'objet à l'issue de sa conception et l'objet à l'issue de son utilisation. Ce bagage conceptuel peut être utile pour investiguer le passage entre les logiques de production et d'usage d'une ressource numérique. La communauté de recherche qui travaille sur les ressources numériques à vocation pédagogique a par exemple pu montrer l'écart entre la conception et l'usage de ces ressources. Cet écart s'explique en grande partie par les difficultés d'application des technologies et normes de description des ressources pédagogiques sur le terrain [Contamines et collab., 2003].

Sur la base des réflexions sur la nature et les différents contextes influant les ressources, la communauté e-éducation a été amenée à formuler des recommandations de modifications des normes et des schémas conceptuels de description des ressources [44]. Des extensions ont par ailleurs pu être proposées pour mieux prendre en compte les contextes de production et d'usage des ressources afin d'en faciliter l'accès pour les utilisateurs finaux [Sawadogo et collab., 2014]. Ces travaux ont par ailleurs permis d'aboutir à des définitions du terme « ressources numériques » qui tiennent mieux compte des aspects contextuels que nous évoquons, sans pour autant négliger les problématiques d'implémentation technologiques. Le ressource numérique, composant structuré des points de vue intellectuel et technologique, s'est ainsi vu augmentée des notions d'intention d'usage et d'opérateurs associés, sans lesquels il semblait effectivement difficile, de permettre un usage effectif des ressources numériques [Sawadogo, 2016].

2.4 Conclusion

L'état de la pratique que nous avons mené dans ce chapitre, bien que sans doute encore très parcellaire puisque focalisé sur les outils et démarches impliquant de près ou de loin l'informatique de la recherche et de la gestion de l'information, nous a tout de même permis de mesurer la vigueur des relations entre les sciences humaines et sociales et le numérique. Ces relations se matérialisent par d'importants projets de recherche menés conjointement ou par la mise à disposition d'outils aboutis et performants. Les enquêtes publiées jusqu'à maintenant montrent toutes que, pour les thématiques qui nous concernent, le numérique s'est imposé. Il est fondamental pour la collecte de l'information aussi bien pour les sources primaires que secondaires et se développe progressivement pour l'analyse et pour l'écriture.

Tous les outils que nous avons évoqués supposent une adaptation, celle du chercheur quoi doit parfois développer de nouvelles compétences, celles des sources dans d'autres cas, qui doivent être transformées pour se plier aux exigences des instruments numériques. Le passage des sources comme des pratiques de recherche du monde physique vers le numérique n'est en tout cas pas neutre. Les adaptations nécessaires sont semble-t-il de plusieurs ordres, en fonction de l'instrumentation numérique impliquée. Certains

44. Dans ce cadre, il est essentiellement question de la norme de description des ressources pédagogiques LOM, voir Contamines et collab. 2003, p. 16-17.

des outils numériques que nous avons décrits sont développés directement pour la recherche et sont pensés en ce sens. Les travaux et enquêtes publiés dans la littérature scientifiques pointent le besoin de formation qu'ils suscitent pour une grande partie de la communauté de recherche, mais ne relèvent pas de risques importants à les utiliser. En revanche, d'autres outils sont à la fois plus génériques et plus problématiques.

C'est en particulier le cas des moyens numériques d'accès et de collecte des sources. À haut niveau, certains d'entre eux, comme l'appareil photo ou les bibliothèques numériques, contribuent par exemple à favoriser une surcharge informationnelle [45] problématique pour le chercheur [Blair, 2003]. D'autres, à l'image des moteurs de recherche généraux ou spécialisés posent problème parce que leur fonctionnement technique est inconnu ou parce qu'ils se fondent sur des données parcellaires ou dégradées.

Ces biais potentiels pour la recherche sont quoi qu'il en soit très dépendants du contexte dans lequel les outils qui les génèrent sont utilisés. La diversité des usages et des profils de chercheurs [Clavert et collab., 2017], comme des documents étudiés, rend illusoire l'objectif de trouver une réponse technique générique à tous les cas possibles. La grande variété des outils utilisés montre bien la multiplicité des approches, par l'historien, de la gestion des informations. Il ne s'agit pas de vouloir uniformiser les pratiques, ce qui serait réducteur, mais bien de tenter d'en maitriser les effets de bords ou, tout du moins de permettre d'en comprendre les biais. C'est ici que prend tout son sens le cadre conceptuel dont nous avons fait l'état de l'art. Les définitions des concepts de données, de documents ou de ressources sont intéressantes parce qu'elles matérialisent l'écart qui peut exister entre deux visions de la ressource numérique, le point de vue utilisateur et le point de vue du système.

Dans le cadre de nos travaux, les bibliothèques numériques et moteurs de recherche sur lesquels nous nous focalisons sont des ressources et peuvent être analysés à ces différents niveaux de lecture. D'un côté, du point de vue de leur conception, elles ont pour objectif de rendre disponible en ligne un nombre grandissant de documents et faciliter leur accès. D'un autre côté, celui de l'usage, elles sont des instruments de recherche essentiels qui permettent l'accès au matériau de base de l'historien. C'est de l'écart entre des contextes d'usage que notre état de la pratique montrent dans leur diversité et des contextes de production difficiles d'accès que naissent les biais que nous souhaitons révéler.

2.5 Références

Afanador-Llach, M.-J., A. Rojas Castro, A. Crymble, V. Gayol, F. Gibbs, C. McDaniel, I. Milligan, A. Visconti et J. Wieringa, éd.. 2017, *The Programming Historian*. URL http://programminghistorian.org/. 33

Almazán, J., A. Gordo, A. Fornés et E. Valveny. 2014, «Word spotting and recognition with embedded attributes», *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, n° 12, p. 2552–2566. 28

45. Ce concept, traduit de l'anglais *information overload*, est né dans les années 1960 et désigne la surabondance de l'information que l'on constate de nos jours dans de nombreuses activités [Eppler et Mengis, 2004].

Audenaert, N. et R. Furuta. 2010, «What Humanists Want : How Scholars Use Source Materials», dans *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, ACM, New York, NY, USA, ISBN 978-1-4503-0085-8, p. 283–292, doi :10.1145/1816123.1816166. 25

Bachimont, B. 2004, «Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle», *Mémoire de HDR*. 45

Baillot, A. 2015, «Reconstruire ce qui manque – ou le déconstruire ? Approches numériques des sources historiques», dans *Entre sources, données et réseaux (programme interdisciplinaire MOREHIST)*, Toulouse, France. 40

Baillot, A. 2016, «Qui sont les humanistes numériques ? Intervention lors de la table ronde de #ephn2016», . 23, 25, 41

Bénel, A. 2014, «Quelle interdisciplinarité pour les « humanités numériques » ?», *Les Cahiers du numérique*, vol. 10, n° 4, doi :10.3166/LCN.10.4.103-132, p. 103–132. 42

Berners-Lee. 1994, «Universal Resource Identifiers in WWW, RFC 1630», . 43

Berners-Lee, T., R. Fielding, U. Irvine et L. Masinter. 1998, «Uniform Resource Identifiers (URI) : Generic Syntax, RFC 2396», . 44

Berners-Lee, T., L. Masinter et M. McCahill. 1994, «Uniform Resource Locators (URL), RFC 1738», . 44

Blair, A. 2003, «Reading strategies for coping with information overload ca. 1550-1700», *Journal of the History of Ideas*, vol. 64, n° 1, p. 11–28. 47

Bloch, M. 1949, *Apologie pour l'histoire ou Métier d'historien*, A. Colin, Paris. 21

Blondel, V. D., J.-L. Guillaume, R. Lambiotte et E. Lefebvre. 2008, «Fast unfolding of communities in large networks», *Journal of statistical mechanics : theory and experiment*, vol. 2008, n° 10, p. P10 008. 38

Borek, L., J. Perkins, C. Schöch et Q. Dombrowski. 2014, «Building bridges to the future of a distributed network : From DiRT categories to TaDiRAH, a methods taxonomy for digital humanities», dans *Proceedings of the International Conference on Dublin Core and Metadata Applicatons*. 23

Borek, L., J. Perkins, C. Schöch et Q. Dombrowski. 2016, «TaDiRAH : a Case Study in Pragmatic Classification», *Digital Humanities Quarterly*, vol. 10, n° 1. 23

Braudel, F. 1969, *Écrits sur l'histoire*, Flammarion, Paris. 21

Brennan, S. A. et K. Mills. 2011, «Why Collecting History Online is Web 1.5», URL http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=47. 30

Bulger, M. E., E. T. Meyer, G. De la Flor, M. Terras, S. Wyatt, M. Jirotka, K. Eccles et C. M. Madsen. 2011, «Reinventing Research ? Information Practices in the Humanities», SSRN Scholarly Paper ID 1859267, Social Science Research Network, Rochester, NY. 25

Chaunu, P. 1978, *Histoire quantitative, histoire sérielle*, A. Colin. 34

Chaunu, P. 1987, «Le fils de la morte», dans *Essais d'ego-histoire*, gallimard éd., Maurice Agulhon et Pierre Nora, Paris, p. 61–107. 34

Clavaud, F. et E. Château-Dutier. 2017, «Une Preuve de concept pour la sémantisation et la visualisation orientée utilisateur de données archivistiques», dans *Digital Humanities 2017 Conference Abstracts*, Montréal, p. 195–197. 29

Clavert, F. 2012, «Le miroir de nos faiblesses ?», URL https://histnum.hypotheses.org/1103. 41

Clavert, F. 2013, «Camera obscurans», URL http://histnum.hypotheses.org/1853. 41

Clavert, F. 2014, «Lecture des sources historiennes à l'ère numérique», URL http://histnum.hypotheses.org/1061. 35

Clavert, F. 2017, «Présentation : Les réseaux sociaux numériques comme sources primaires de l'historien.ne», URL https://histnum.hypotheses.org/2671. 30

Clavert, F., J. Daniel, H. Fleckinger, M. Grandjean et F. Idmhand. 2017, «Histoire et humanités numériques : nouveaux terrains de dialogue entre les archives et la recherche», *La Gazette des Archives* , vol. 245, n° 1, p. 121–134. 22, 47

Contamines, J., S. George et R. Hotte. 2003, «Approche instrumentale des banques de ressources éducatives», *Sciences et Techniques Educatives*, vol. 10, hors s, p. p. 157–178. 46

Coustaty, M. et J. Ogier. 2015, «Graph matching versus bag of graph : a comparative study for lettrines recognition», dans *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, p. 356–360, doi :10.1109/ICDAR.2015.7333783. 39

Davenport, T. H. et L. Prusak. 1997, *Information ecology : Mastering the information and knowledge environment*, Oxford University Press on Demand. 45

David J, N. et B. Sharon. 2006, «Probabilistic topic decomposition of an eighteenth-century American newspaper», *Journal of the American Society for Information Science and Technology*, vol. 57, n° 6, p. 753–767. 35

Delacroix, C., F. Dosse et P. Garcia. 2007, *Les courants historiques en France : XIXe-XXe siècle*, edition revue et augmentée éd., Folio, Paris. 34

Dombrowski, Q. 2014, «What ever happened to Project Bamboo ?», *Literary and Linguistic Computing*, vol. 29, n° 3, p. 326–339. 23

Drucker, J. 2014, *Graphesis : visual forms of knowledge production*, Harvard University Press. 36

Eppler, M. J. et J. Mengis. 2004, «The concept of information overload : A review of literature from organization science, accounting, marketing, MIS, and related disciplines», *The information society*, vol. 20, n° 5, p. 325–344. 47

Flaounas, I., O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis et N. Cristianini. 2013, «Research methods in the age of digital journalism : Massive-scale automated analysis of news-content—topics, style and gender», *Digital Journalism*, vol. 1, n° 1, p. 102–116. 35

Foucault, M. 1969, *L'archéologie du savoir*, Paris, Gallimard. 21

Genet, J.-P. 1986, «Histoire, Informatique, Mesure», *Histoire & Mesure*, vol. 1, n° 1, doi :10.3406/hism.1986.904, p. 7–18. 34

Ginzbourg, C. 1980, *Les Batailles nocturnes sorcellerie et rituels agraires en Frioul XVI-XVIII° siècle*, Editions Verdier, Lagrasse, France, 180 p.. 42

Ginzbourg, C. 2006, «Conversation avec Orion», *Matériaux pour l'histoire de notre temps*, vol. 82, n° 2, p. 129–132, ISSN 0769-3206. 42

Gonzalez, J. E., Y. Low, H. Gu, D. Bickson et C. Guestrin. 2012, «Powergraph : distributed graph-parallel computation on natural graphs.», dans *OSDI*, vol. 12, p. 2. 38

Graham, S. 2001, «Historians and Electronic Resources : A Second Citation Analysis», *Journal of the Association for History and Computing*. URL http://hdl.handle.net/2027/spo.3310410.0004.203. 27

Grandjean, M. 2016, «A social network analysis of Twitter : Mapping the digital humanities community», *Cogent Arts & Humanities*, vol. 3, n° 1, p. 1–14. 30

Grandjean, M. 2017, «Médias sociaux et mise en scène de l'histoire», dans *Digital Humanities 2017 Conference Abstracts*, Montréal. 30

Idmhand, F. et R. Walter. 2014, «Les sources numériques de la recherche dans les sciences humaines et sociales», dans *Le numérique bouleverse la donne : vers de nouvelles démarches scientifiques ?, Colloque annuel de l'Académie Sciences Lettres Toulouse*, Académie Sciences Lettres Toulouse, Toulouse, France. 40

Jackson, A., J. Lin, I. Milligan et N. Ruest. 2016, «Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities», dans *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, ACM, New York, NY, USA, p. 103–106, doi :10.1145/2910896.2910912. 40

Kemman, M., M. Kleppe et S. Scagliola. 2013, «Just Google It - Digital Research Practices of Humanities Scholars», URL http://arxiv.org/abs/1309.2434. 25, 28, 41

Lainé-Cruzel, S. 2004, «Documents, ressources, données : les avatars de l'information numérique», *Revue I3-Information Interaction Intelligence*, vol. 4, n° 1. 44, 45

Le Deuff, O. 2014, *Le temps des humanités digitales : la mutation des sciences humaines et sociales*, Fyp Editions, Limoges, ISBN 978-2-36405-122-5 2-36405-122-3. 22

Lemercier, C. 2005, «Analyse de réseaux et histoire», *Revue d'histoire moderne et contemporaine*, vol. 52, n° 2, p. 88–112. 37

Merzeau, L. 2013, «L'intelligence des traces», *Intellectica*, , n° 59, p. 115–135. URL http://f.hypotheses.org/wp-content/blogs.dir/888/files/2013/02/Merzeau_intelligence-des-tracesBAT.pdf. 42

Milligan, I. 2013, «Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997–2010», *The Canadian Historical Review*, vol. 94, n° 4, p. 540–569. Volume 94, Number 4, December 2013. 41

Milligan, I. 2014, «Le potentiel des données ouvertes pour l'histoire politique», *Revue parlementaire canadienne*, vol. 37, n° 4. 30

Milligan, I. 2016, «Lost in the Infinite Archive : The Promise and Pitfalls of Web Archives», *International Journal of Humanities and Arts Computing*, vol. 10, n° 1, doi :10.3366/ijhac.2016.0161, p. 78–94. 31, 40

Papastamkou, S. 2017, «Twitter as a source for the history of the present : the 2015 Greek referendum as a case study», dans *Digital Humanities 2017 Conference Abstracts*, Montréal. 30

Pinol, J.-L. 2011, «Une infrastructure pour les SHS : le TGE Adonis», *Revue d'histoire moderne et contemporaine*, , n° 58-4bis, doi :10.3917/rhmc.585.0090, p. 90–100. 28

Pitti, D., B. Stockting et F. Clavaud. 2018, «An introduction to "Records in Contexts" : an archival description draft standard», *Comma*, vol. 2016, n° 1-2, p. 173–188. 41

Potin, Y. 2011, «Institutions et pratiques d'archives face à la « numérisation ». Expériences et malentendus», *Revue d'histoire moderne et contemporaine*, , n° 58-4bis, doi :10.3917/rhmc.585.0057, p. 57–69. 26

Poublanc, S. 2018, «Les historiens numériques rêvent-ils d'archives électroniques ? | Le goût de l'archive à l'ère numérique», URL http://www.gout-numerique.net/table-of-contents/title-page/les-historiens-numeriques-revent-ils-darchives-electroniques/. 27

Pouyllau, S. 2012, «Les moteurs de recherche profitent aussi de la sémantique», *Documentaliste - Sciences de l'Information*, vol. 48, n° 4, doi :10.3917/docsi.484.0022, p. 36–37. 41

Pédauque, R. T. 2006, *Le Document à la lumière du numérique : forme, texte, médium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité*, C & F Editions. 45

Rabardel, P. 1995, *Les hommes et les technologies, approche cognitive des instruments contemporains*, Armand Colin, Paris. 46

Romary, L. 2014, «Sustainable data for sustainable infrastructures», dans *Facing the Future : European Research Infrastructures for the Humanities and Social Sciences*, SCIVERO Verlag. 23

Ruiz, É. 2011, «Vers un socle commun de formation aux outils numériques ?», URL http://www.boiteaoutils.info/2011/08/vers-un-socle-commun-de-formation-aux/. 42

Ruiz, E. 2015, «Les historien-nes et le numérique : usages et besoins de formation», URL http://www.boiteaoutils.info/2015/03/historiens-numerique/. 25, 26, 31

Rutner, J. et R. Schonfeld. 2012, «Supporting the Changing Research Practices of Historians», . 32

Rygiel, P. 2012, «L'enquête historique à l'ère numérique», *Revue d'histoire moderne et contemporaine*, vol. 58-4, n° 5, p. 30–40. 34

Salaün, J.-M. 2007, «Éclairages sur la redocumentarisation», URL http://blogues.ebsi.umontreal.ca/jms/index.php/post/2007/05/05/252-eclairages-sur-la-redocumentarisation. 45

Salaün, J.-M. 2012, *Vu, lu, su : les architectes de l'information face à l'oligopole du Web*, La Découverte, Paris. 45

Sawadogo, D. 2016, *Architectures logicielles et mécanismes pour la gestion adaptative et consolidée de ressources numériques dans une application interactive scénarisée*, thèse de doctorat, Université de La Rochelle. 46

Sawadogo, D., R. Champagnat et P. Estraillier. 2014, «User Profile Modelling for Digital Resource Management Systems», dans *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014.*, p. 7 – 11. 46

Schafer, V. et B. Thierry. 2015, «L'ogre et la toile. Le rendez-vous de l'histoire et des archives du web», *Socio. La nouvelle revue des sciences sociales*, , n° 4, doi : 10.4000/socio.1337, p. 75–95. 30

Scheinfeldt, T. 2014, «The Dividends of Difference : Recognizing Digital Humanities' Diverse Family Tree/s», URL https://foundhistory.org/2014/04/the-dividends-of-difference-recognizing-digital-humanities-diverse-family-trees/. 35

Schreibman, S., S. Ray et J. Unsworth, éd.. 2016, *A New Companion to Digital Humanities, 2nd Edition - Susan Schreibman, Ray Siemens, John Unsworth*, wiley éd., ISBN 978-1-118-68059-9. 22

Seignobos, C. et C.-V. Langlois. 1899, *Introduction aux études historiques*, Hachette et Cie, Paris. 21

Sula, C. A., S. E. Hackney et P. Cunningham. 2017, «A Survey of Digital Humanities Programs», dans *Digital Humanities 2017*, Montréal. 23

Veyne, P. 1971, *Comment on écrit l'histoire*, seuil éd., Paris. 21

Warwick, C., M. M. Terras et J. Nyhan. 2012, *Digital Humanities in Practice*, Facet Publishing in association with UCL Centre for Digital Humanities, Londres, ISBN 978-1-85604-766-1. 22, 30

Watkins, A. 2013, «Zotero for Personal Image Management», *Art Documentation : Journal of the Art Libraries Society of North America*, vol. 32, n° 2, p. 301–313. 31

Winterer, C. 2012, «Where is America in the Republic of Letters ?», *Modern Intellectual History*, vol. 9, n° 3, doi :10.1017/S1479244312000212, p. 597–623. 36

Chapitre 3

Bibliothèque numérique, fondements théoriques et contexte technologique

Résumé

Ce chapitre présente l'état de l'art scientifique et technique des objets centraux de notre travail, les bibliothèques numériques et les moteurs de recherche de sources. Dans un premier temps, il examine les fondamentaux théoriques et le fonctionnement général de ces systèmes. Dans un second temps, il présente les contraintes de mise en œuvre d'un tel outil dans ces différents contextes de production, d'exécution et d'usage. Face à ces contraintes, les choix de conception que nous avons opérés pour mettre en place la bibliothèque numérique indispensable à nos travaux sont décrits et expliqués. Enfin, ce chapitre propose une évaluation fonctionnelle de cet outil de recherche, en comparant ses fonctionnalités avec celles d'un échantillon de bibliothèques numériques actuellement en production, afin d'en assurer la représentativité.

Sommaire

3.1 Introduction

Parmi les ressources numériques régulièrement utilisées pour la production de connaissances historiques que nous avons recensées dans notre précédent chapitre, les bibliothèques numériques et les bases de données de sources primaires ou secondaires sont fondamentales et constituent l'un des moyens d'accès privilégié à l'information. Ce type de systèmes a fait l'objet de travaux visant à définir leurs contours théoriques et leurs objectifs technologiques dont il convient ici de faire un rapide état.

Le terme « bibliothèques numériques » (*Digital Libraries*) a émergé au début des années 1990 pour remplacer et fédérer des travaux qui parlaient alors de « bibliothèques électroniques » ou de « bibliothèques virtuelles ». Très schématiquement, il désigne tout système mettant en relation des utilisateurs, des contenus et des technologies d'accès à ces contenus [Larsen et Wactlar, 2004]. Une définition plus précise peut être trouvée dans les travaux de Fox et collab., 2012, p. 6-7. Ces études ont abouti à une définition autour du concept dit des « 5S » qui exprime les principes fondamentaux d'un système de bibliothèque numérique de la manière suivante :

Societies : Les bibliothèques numériques satisfont les besoins d'information des utilisateurs.

Scenarios : Elles fournissent un service d'information.

Structures : Elles organisent l'information pour la rendre accessible.

Spaces : Elles présentent l'information pour la rendre utilisable.

Streams : Elles communiquent l'information aux utilisateurs.

Une bibliothèque numérique est ainsi un système de médiation, elle est tourné vers l'usage, présente et communique une information organisée pour résoudre des besoins d'information. Elle a ainsi toutes les caractéristiques d'une ressource numérique au sens des définitions que nous avons données dans notre section 2.3 et est clairement positionnée entre les usages attendus des utilisateurs et les intentions éditoriales des concepteurs.

Les auteurs que nous venons de citer utilisent à dessein le terme « organisé » plutôt que celui de « structuré » pour différencier un système de bibliothèque numérique d'une base de données. Ces systèmes se placent ainsi entre le web, qui peut lui aussi répondre à un besoin d'information mais qui n'est pas organisé et donc chaotique, et la base de données, très structurée et plus fortement dépendante du contexte de production initiale. D'autres aspects technologiques entrent dans la définition. Une bibliothèque numérique doit être accessible en ligne, elle est donc naturellement dotée d'un accès web et d'une interface dédiée, ce qui est rarement le cas d'une base de données.

Ces éléments de définition encadrent bien le contexte d'usage, en terme de besoin et de technologie. D'autres définitions, dont Waters, 1998 insistent sur le rôle des bibliothèques numériques en terme de préservation des contenus. Il explique par ailleurs que ces systèmes ne se contentent pas de fournir un simple accès aux contenus mais qu'ils doivent fournir tout ce qui peut permettre à l'utilisateur de sélectionner, structurer, interpréter les contenus. Ces usages attendus, et avancés, d'une bibliothèque numérique reposent sur des contenus et des outils produits à destination d'une ou plusieurs communauté d'utilisateurs clairement identifiées.

De ce fait le contexte d'usage est très variable d'une bibliothèque numérique à une autre, de l'éducation au sens large, à la recherche, à l'information scientifique et technique, aux arts, *etc.* Les bibliothèques numériques sont donc bien souvent spécifiquement développées à destination d'une communauté et doivent être construites en fonction de besoins spécifiques, qu'il est indispensable de connaître et de décrire [1].

Pour répondre à tous ces besoins, les bibliothèques numériques sont supportées par des processus informatiques divers, en fonction de chacun des contextes qui les environne. La numérisation et l'extraction d'information tiennent par exemple un rôle central dans le contexte de production. Dans l'optique d'identifier et contribuer à réduire les biais qu'ils peuvent créer, nous devons regarder de plus près les différents processus à l'œuvre et en faire un état de l'art. Ce dernier est d'autant plus important que notre approche est largement expérimentale. Il est en effet un maillon essentiel de l'expression des besoins et du cahier des charges du prototype expérimental sur lequel est fondé l'essentiel de nos travaux. Ce chapitre présente donc ces deux volets de notre travail, état de l'art des bibliothèques numériques et expression des besoins fonctionnels de notre prototype expérimental. Pour chacun d'eux, nous examinerons les processus à l'œuvre dans les différents contextes environnants la ressource.

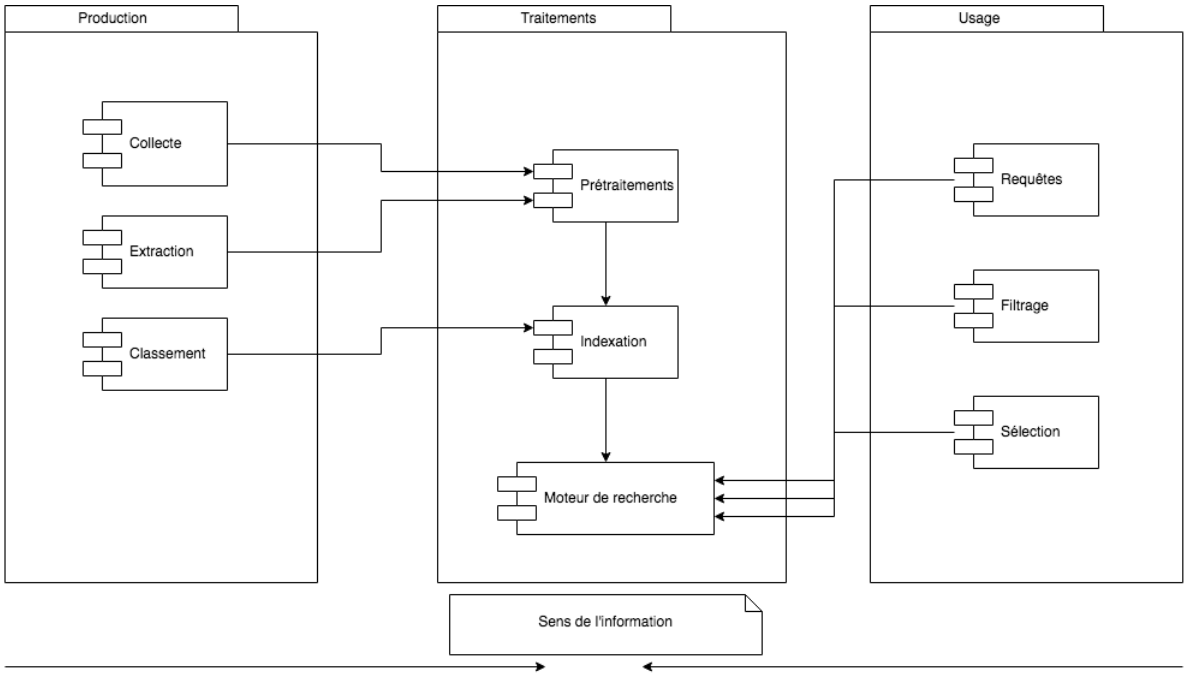
3.2 Bibliothèques numériques : état de l'art scientifique et technique

Une bibliothèque numérique, par sa fonction de médiation, donne accès à l'information, de la trouver et de l'exploiter, par un moyen ou un autre. Ces moyens reposent sur des traitements informatiques qui interagissent avec les données d'une part et les utilisateurs d'autre part. Les fonctionnalités de recherche d'information, matérialisées par le moteur de recherche, constituent l'un de ces traitements. Celui-ci, présent sur la grande majorité des plateformes permet de rechercher dans les contenus, données et métadonnées par l'intermédiaire de requêtes pouvant prendre des formes diverses, de la simple saisie d'un mot clé jusqu'à des opérations de recherches booléennes plus complexes. Il fait, comme le montre la figure 3.1, le lien entre des contenus et des demandes provenant de l'usage.

Pour faire ce travail, le moteur de recherche n'opère pas directement sur les contenus, mais sur des représentations informatiques intermédiaires, des index ou des bases de données par exemple. Ces représentations sont construites par des processus qui ne sont pas neutres et qui ont un impact sur le fonctionnement du système du point de vue de l'utilisateur. Il convient donc d'en dresser un tableau aussi complet que possible pour être en mesure de produire un prototype qui soit représentatif des bibliothèques numériques et des moteurs de recherche couramment utilisés par les historiens.

1. C'est par exemple ce qui a été fait pour les bibliothèques numériques dédiées à la musique par Cunningham et collab., 2003.

FIGURE 3.1 – Représentation schématique des flux de données utiles au fonctionnement d'un moteur de recherche



À gauche de ce diagramme sont représentées les opérations positionnées du côté de la production des contenus, à droite celles du côté du contexte d'usage. On observe le rôle de médiateur des traitements informatiques présentés dans cet état de l'art.

3.2.1 Contexte de production : capture, documentation et classement de l'information

Aussi complet qu'il puisse être, un système de bibliothèque numérique tel que nous venons de le décrire n'est qu'une ossature qui n'a pour seul rôle que d'accueillir des contenus et les rendre accessibles. Dans notre contexte, ce n'est pas l'intérêt historique à proprement parler qui nous importe, mais le processus qui permet de mettre en ligne des contenus qui, pour l'essentiel, ne sont pas nativement numériques [2].

À chacune des étapes de ce processus, des choix sont opérés. Ces choix ne sont pas toujours sans conséquence pour l'exploitation de ces contenus numériques en devenir. En nous fondant sur l'analyse de la littérature, nous mettrons en avant, dans cette section, les principales étapes de la dématérialisation des contenus qui impactent la construction de notre prototype. Nous mettrons particulièrement l'accent sur celles qui produisent des effets importants sur la pratique des utilisateurs et que nous étudierons plus en détail dans les chapitres 4 et 5.

2. C'est particulièrement vrai pour les documents utiles à l'étude de l'histoire, surtout pour les sources primaires. Notons qu'il existe toutefois des exceptions, il existe par exemple des équipes de recherche en histoire qui travaillent désormais avec des sources nativement numériques, en particulier les archives du web ou des courriels. Ces archives posent par ailleurs des problèmes d'accès spécifiques que nous ne traiterons pas ici, voir à ce sujet [JACKSON et collab., 2016].

Capture de l'information : numérisation et ocr

Dans un processus de mise en ligne de contenu numérique, l'étape de capture de l'information est bien entendu primordiale. Elle consiste en la numérisation elle même qui permet de passer du contenu physique ou analogique au contenu numérique. Cependant, pour que les contenus soient véritablement utiles, elle est, lorsque cela est possible, suivie d'étapes d'extraction d'informations. Ces étapes d'extraction peuvent être très diverses en fonction du média original. Depuis la reconnaissance optique de caractères pour le texte, jusqu'à la reconnaissance et l'extraction des paroles d'extrait sonore ou vidéo (*speech to text*), les méthodes et les technologies sont variables, tout comme la qualité des résultats produits. Dans cette diversité, nous avons choisi de nous intéresser au texte, qui est le média le plus répandu, sous sa forme imprimée ou manuscrite, dans les bibliothèques numériques les plus utilisées pour le travail sur les sources historiques.

Les projets de numérisation de documents anciens font appel à du matériel et des techniques de numérisation maîtrisées par des acteurs aussi bien publics que privés. Selon la nature des documents, les captures sont effectuées par l'intermédiaire de scanner de microfilms, d'appareil photographique numérique ou de scanners à plat, de taille et de capacité diverses. La qualité de la numérisation dépend évidemment de la résolution adoptée, du format d'enregistrement des images, des niveaux de couleurs capturés, noir et blanc, niveaux de gris ou couleurs. Les images peuvent par ailleurs être améliorée par des processus de réduction de bruit, de redressement, etc [Klijn, 2008].

De la qualité de la numérisation dépend déjà ce qu'il sera possible d'exploiter des documents. Naturellement, des processus tels que la reconnaissance optique de caractères fonctionnent bien mieux avec des images nettes, capturées en haute résolution et correctement redressées. Ces processus de numérisation sont de mieux en mieux maîtrisés, le coût modique du stockage informatique dont on bénéficie aujourd'hui permet par ailleurs de conserver les images dans des formats sans pertes d'information et avec un haut niveau de détail, du point de vue de la résolution et des informations de couleur. Il est important de se rappeler que cela n'a pas toujours été le cas. Certaines entreprises de numérisation peuvent dater du début des années 2000 et être loin des standards de qualité actuels. À titre d'exemple, le projet RESGATE[3], projet de numérisation de manuscrit de grande ampleur mené par la Bibliothèque Nationale du Brésil, que nous avons eu l'occasion d'analyser en détail, date de l'année 2001. À cette époque, les images ont été scannées en noir et blanc, dans une résolution relativement basse en comparaison des standards actuels (voir figure 3.2) et avec un niveau de bruit important qui les rendent extrêmement difficiles à exploiter informatiquement.

Ainsi, en matière de numérisation, les standards, les méthodes mais également les opérateurs ont beaucoup changé au fil du temps. Ces variations diverses aboutissent aujourd'hui dans les grandes bibliothèques numériques à des documents de qualité variable. Pour autant, ces grandes bibliothèques numériques ne fournissent que relativement peu d'informations sur la manière dont ont été numérisés les documents. L'absence de ces informations, comme le remarquait très justement Traub et collab. [2015], s'avère problématique pour l'utilisateur qui ne peut pas évaluer la qualité de la

3. Les informations détaillées de ce projet sont accessibles à l'adresse : https://bndigital.bn.gov.br/dossies/projeto-resgate-barao-do-rio-branco/.

FIGURE 3.2 – Exemple d'image du fonds Resgate

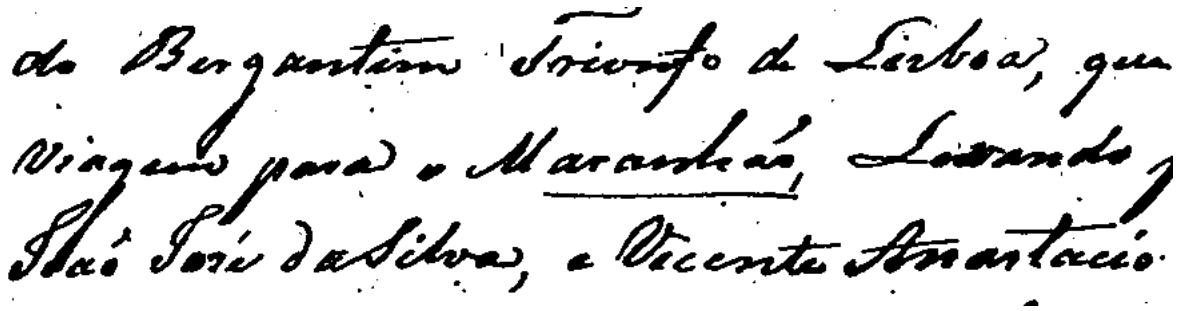


Image extraite d'un document de 1819, capturée dans sa taille réelle (100%).

reproduction numérique par rapport au document original. Par ailleurs, la qualité des processus d'extraction de contenus dépendant fortement de la qualité originale de la numérisation, il est également difficile d'évaluer la qualité de ces processus.

Lorsqu'un texte imprimé ou manuscrit est numérisé, il peut être soumis à un processus de reconnaissance optique de caractères (OCR) [4]. Ce processus algorithmique vise, comme son nom l'indique à reconnaître automatiquement les lettres, puis les mots qui composent un document afin d'en construire une version informatique manipulable. Sans faire ici un état de l'art complet sur l'OCR, rappelons tout de même brièvement qu'il s'agit d'un thème de recherche majeur en informatique et plus précisément en analyse de documents.

Les logiciels d'OCR, commerciaux ou *open source* atteignent aujourd'hui des résultats très satisfaisants, de l'ordre de plus de 99% de taux de réussite [ESKENAZI et collab., 2017] lorsque les documents qui leur sont soumis sont, selon la formule consacrée, relativement « propres ». Cette formule, typique du jargon informatique, signifie tout à la fois que le document a été imprimé avec une police de caractères connue, que l'impression et la numérisation ont été effectuées avec un niveau de qualité satisfaisant, en particulier sur le plan de la résolution. Évidemment, tous les documents sont loin d'atteindre ce niveau de qualité qui permet une très bonne reconnaissance automatique. Les taux de réussite sont alors très variables en fonction de la qualité et du type du document [5]. Les journaux anciens font partie des exemples les plus étudiés, parce qu'au centre de la politique de numérisation des grandes institutions nationales. Bien qu'étant des documents moins complexes à reconnaître que les documents manuscrits, leur état de conservation, leur mise en page et la qualité de leur impression rendent difficile la tâche de reconnaissance.

La recherche en informatique a donc pour objectif d'améliorer le taux de reconnaissance correcte des caractères de ce type de source. Différentes approches sont développées qui dépassent de loin la seule problématique de l'amélioration de la reconnaissance des caractères. La recherche travaille par exemples sur des approches permettant de corriger

4. L'acronyme OCR vient de l'anglais *Optical Character Recognition*, il est utilisé aussi bien en français qu'en anglais.

5. On trouvera une analyse des facteurs influençant la qualité des processus d'OCR depuis la phase de binarisation jusqu'à la reconnaissance dans les travaux de TANNER [2004] et de GUPTA et collab. [2007].

les erreurs générées par le processus d'OCR. Différentes techniques sont élaborées, depuis la correction par utilisation de modèle d'erreurs [Lee et Smith, 2012], de données de suggestion [Bassil et Alwani, 2012], jusqu'à la prise en compte du contexte environnant chaque mot pour pouvoir déterminer si celui-ci a été, ou non, correctement reconnu. Ces approches, sur lesquelles nous reviendrons, présentent l'avantage de corriger les erreurs mais également de les détecter et donc d'en faire état et d'en limiter l'impact.

Les erreurs dont nous parlons se traduisent dans la pratique de différentes manières. À grande échelle d'abord, elles sont sources de problèmes pour une analyse en grand volume de texte. Comme le remarquent des travaux récents, les approches *« big data »* ont le soucis de la quantité des données qu'elles manipulent mais ne tiennent que trop peu compte de la qualité [Chiron et collab., 2017]. Certaines approches statistico-lexicales ou d'apprentissage automatique, telle que LDA [6], servant à repérer des récurrences thématiques dans des corpus manipulent des volumes très importants de documents. Or, même dans les cas où les taux de reconnaissance de caractères sont bons, sur des milliers, voire des millions de pages, le nombre de mots mal reconnus peut s'avérer très important. Ici, c'est donc le volume de documents manipulés qui rend problématique les erreurs de reconnaissance de caractère.

Ces erreurs se manifestent également à des échelles bien plus réduites, dans la pratique des utilisateurs. Intuitivement, tout un chacun peut comprendre qu'un mot mal reconnu ne peut naturellement pas correspondre avec une requête, bien formulée, elle, par un utilisateur. Dans le pire des cas, des documents peuvent ainsi devenir inaccessibles. Chiron et collab. [2017] rapportent un cas différent de mots « pris pour un autre ». Le lexicographe Alain Rey explique par exemple qu'une recherche dans le corpus de presse du XIX^{e} siècle de Gallica sur le terme « gadget » renvoie à des occurrences du terme « budget » mal reconnu par l'OCR [7]. La confusion entraînée est ici manifeste.

Les moyens à disposition des utilisateurs qui pourraient leur permettre de mesurer le niveau de qualité de la reconnaissance optique de caractères auquel a été soumis un document sont limités. Certaines bibliothèques numériques fournissent une valeur numérique représentant le taux de reconnaissance par mot, pour certains des documents qu'ils conservent. Cette information est généralement disponible à la consultation de chacun des documents en « mode texte », comme dans le lecteur de Gallica (voir figure 3.4). Ces indicateurs sont calculés sur la base des informations fournies par l'algorithme d'OCR lui même qui indique, pour chaque caractère et pour chaque mot, un indice de confiance. Le logiciel d'OCR fournit les résultats dans un document XML de sortie, suivant généralement la norme ALTO [8] (voir figure 3.3). Ce fichier permet de garder une trace de la structure du document initial et de tous les résultats produits par l'OCR lors de l'analyse.

6. LDA, pour *Latent Dirichlet Allocation* est un modèle statistique décrit par [Blei et collab., 2003] très utilisé pour les tâches de *topic modelling*. On trouvera une explication et des cas d'application historique de ces technologies dans [Graham et collab., 2012] et [Yang et collab., 2011].

7. L'étymologie du mot « gadget » est par ailleurs discuté dans le contexte du XIXe siècle. On trouvera tous les détails dans « Alain Rey et Gallica : une grande histoire de mots » disponible sur le blog de Gallica à l'adresse : http://gallica.bnf.fr/blog/20102016/alain-rey-et-gallica-une-grande-histoire-de-mots.

8. La documentation du format ALTO est accessible à l'adresse : https://www.loc.gov/standards/alto/.

FIGURE 3.3 – Exemple d'une ligne de texte au format ALTO

```
<TextLine HEIGHT="25" WIDTH="322" VPOS="149" HPOS="189">
        <String WC="0.3463636339" CONTENT="Angleterre," HEIGHT="25" WIDTH="164" VPOS="149" HPOS="189"/>
        <SP WIDTH="14" VPOS="150" HPOS="354"/>
        <String WC="0.323333323" CONTENT="Italie" HEIGHT="20" WIDTH="73" VPOS="150" HPOS="369"/>
        <SP WIDTH="15" VPOS="155" HPOS="443"/>
        <String WC="0.2899999917" CONTENT="." HEIGHT="6" WIDTH="6" VPOS="165" HPOS="459"/>
        <SP WIDTH="15" VPOS="165" HPOS="466"/>
        <String WC="0.5899999738" CONTENT="-" HEIGHT="3" WIDTH="5" VPOS="165" HPOS="482"/>
        <SP WIDTH="16" VPOS="164" HPOS="488"/>
        <String WC="0.5500000119" CONTENT="." HEIGHT="5" WIDTH="6" VPOS="164" HPOS="505"/>
</TextLine>
```

La ligne notée ici en XML contient la chaîne de caractères « Angleterre, Italie . - . ». Les scores calculés par le logiciel d'OCR sont indiqués par l'attribut WC, les autres attributs indiquent le contenu et la position du mot.

Les indices de confiance de chaque mots, *Word confidence* (WC) sont additionnés pour obtenir un indicateur par page. Celui-ci, divisé par le nombre de mots de la page donne le taux moyen [9]. Des opérations de vérification de ce taux peuvent ensuite être effectuées [10].

La pertinence de ces indicateurs est toutefois très relative. Ils ne permettent par exemple pas de détecter quelles ont été les erreurs, ou tout du moins les erreurs fréquemment commises par l'OCR. Par ailleurs, ces indicateurs ne sont accessibles que pour un document à la fois, lors de la consultation. Il n'existe pas de tels indicateurs permettant à l'utilisateur de mesurer la qualité de reconnaissance globale des textes ayant servi à calculer les résultats de sa recherche d'information. En effet, TRAUB et collab. [2015] expliquaient que les indicateurs permettant de calculer les scores de confiance ont évolué depuis le début des grandes campagnes de numérisation. À titre d'exemple, la Bibliothèque Nationale des Pays-Bas a utilisé plusieurs dictionnaires pour valider le calcul de WC. Ces dictionnaires ont évolué, pour prendre en compte des formes anciennes de mots et affiner le calcul des scores, mais il n'est pas possible de savoir quel dictionnaire a été utilisé pour chaque document.

FIGURE 3.4 – Informations affichées par le lecteur de Gallica (BNF) en mode texte

À gauche le lecteur de Gallica fournit un indicateur de la qualité de la reconnaissance de caractère. Le lecteur de persee (OpenEdition), à droite, ne fournit pas d'indicateurs mais indique que le document a fait l'objet d'un processus d'OCR.

L'influence des erreurs d'OCR sur la pratique des utilisateurs est accentuée par le type

9. Ces indicateurs peuvent être diversement exprimés selon les plateformes et les logiciels utilisés pour effectuer le processus d'OCR, le calcul brièvement décrit ici est celui utilisé par la BNF, dont les détails sont accessibles à l'adresse http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.num_conversion_mode_texte.html.

10. La BNF exige de ses prestataires qu'ils vérifient le taux de qualité de l'OCR calculé par une vérification sur un échantillon de mot, selon les recommandations de la norme ISO 2859-1 (https://www.iso.org/fr/standard/1141.html).

de mots qui sont le plus généralement recherchés dans les bibliothèques numériques. Il est possible de montrer sans difficultés qu'une majorité des mots-clés saisis par les utilisateurs lors de leur recherche d'information sont, dans le jargon informatique, des « entités nommées ». Ces entités sont une catégorie de mots désignant des personnes, des lieux, des institutions, etc. Leur représentation importante dans les requêtes des utilisateurs s'explique par le caractère discriminant de cette catégorie de mots, qui sert de point de départ à bon nombre de nos recherches [Gooding, 2014]. Or, ces entités nommés sont plus sujettes aux erreurs d'OCR que les autres mots d'une langue déterminée.

Cet état de fait s'explique par les diverses techniques que les algorithmes de reconnaissance de caractères utilisent pour arriver à leurs fins. Les entités nommées ne faisant souvent pas partie des dictionnaires utilisés par l'OCR pour la correction, les processus de reconnaissance sont moins performants. Une étude récente a par ailleurs montré la différence importante qui existe entre l'usage de requêtes simulées et requêtes réelles sur les la capacité des moteurs de recherche à retrouver des documents dont le contenu a été soumis à un processus d'OCR [Traub et collab., 2016]. Les requêtes réelles contiennent en effet plus souvent des entités nommées que les requêtes simulées et souffrent plus des erreurs d'OCR. Ces raisons de performance et d'évaluation des OCR sur la reconnaissance des entités nommées expliquent pourquoi il est fondamental de s'attacher à l'étude de cette catégorie de mot, que nous traiterons en détail dans notre chapitre 5.

Les erreurs d'OCR sont régulièrement citées comme l'un des biais les plus importants des méthodes d'analyse impliquant le numérique (*distant reading*) en SHS [Gefen, 2015], il n'existe toutefois que peu d'études témoignant de leur impact. L'une d'elles, très étayée par un vaste jeu de donnée a été publiée en 2013 par l'historien canadien Ian Millian [Milligan, 2013]. Cette étude se focalise sur le rôle de la mise en ligne des bases de données de journaux numérisés sur les derniers développements de l'historiographie canadienne. Elle consacre une part importante de son développement à la problématique des erreurs de reconnaissance optique de caractères et en donne des exemples précis. L'auteur aboutit à la conclusion que l'historien se doit de développer une compréhension avancée de la manière dont fonctionne les moteurs de reconnaissance de caractère. Il ne doit pas la développer pour être en mesure d'éliminer ces erreurs mais pour en limiter les biais sur sa pratique.

Une bonne connaissance du fonctionnement de l'OCR et plus largement des bases de données et des moteurs de recherche permet d'intégrer à la méthodologie de l'historien les effets de bords que l'utilisation de ce type d'outil introduisent et de développer les « bonnes pratiques » d'usage. Parmi ces bonnes pratiques dont Ian Milligan se fait l'écho et qui sont directement liées à la problématique des erreurs d'OCR, il y a la nécessité pour l'historien de mener une réflexion sur son besoin d'information. De cette réflexion il pourra tirer une plus grande variété de mots-clés, fondés sur des synonymes ou des orthographes alternatives qui lui permettront de minimiser les biais d'OCR.

Documentation de l'information

Les matériaux bruts numérisés, à eux seuls, ne sont pas d'une grande utilité. Dès lors que leur volume devient relativement important [11], il est nécessaire de leur adjoindre des moyens de les exploiter et de retrouver leur contenu. Ces informations qui décrivent les données principales forment les métadonnées d'un contenu. Le terme métadonnées, en anglais *metadata*, désigne des données « sur des données », des données qui en décrivent d'autres [12]. Les métadonnées sont aujourd'hui omniprésentes, mais ont toujours été utiles au bibliothécaire et à l'archiviste qui ont toujours fait usage de moyens de décrire les contenus dont ils avaient la charge. Ainsi une fiche cartonnée rassemblant les informations sur un document particulier est un ensemble de métadonnée au même titre que le sont les informations automatiquement générées lorsque nous prenons une photo avec un smartphone ou que nous écrivons un courrier électronique.

En matière de moteur de recherche et de bibliothèques numériques, les métadonnées ont un rôle majeur, à tel point qu'elles peuvent devenir à elles-seules une source d'information. Le plus brillant exemple à ce sujet est probablement Dbpedia, la base de données qui recueille l'intégralité des métadonnées de l'encyclopédie en ligne Wikipédia et qui est une des plus grandes, sinon la plus grande collection d'information, sous forme de métadonnées, du web. Une grande majorité des requêtes sont ainsi faites dans des champs de métadonnées lorsque l'on utilise une bibliothèque numérique. C'est même souvent la seule source d'information lorsque le contenu lui même n'est pas directement recherchable, s'il s'agit d'image ou de son ou si les documents sont trop anciens pour être reconnus de manière automatique par un processus de reconnaissance optique de caractères. On distingue classiquement trois niveaux de métadonnées :

— Les métadonnées techniques indiquent quels sont les caractéristiques physiques ou numériques d'une ressource, sa taille, son poids informatique, etc. Elles servent également à déterminer la manière dont doit être décodée ou lue une ressource par un programme, sur la base de l'encodage informatique de ses caractères par exemple ou l'ordre de sa pagination, etc.

— Les métadonnées descriptives servent quant à elles à décrire le contenu du document. Elles indiquent si une ressource est un texte, quel est son auteur, son titre, son genre littéraire ou encore peuvent donner une indication de son contenu sous forme de mots-clés. Lorsqu'il s'agit d'images, elle peuvent brièvement décrire la scène, etc. Ces métadonnées sont pour partie subjective et, à ce titre, posent des problèmes spécifiques.

— Les métadonnées administratives enfin, servent à classer les ressources, peuvent indiquer une côte ou une localisation par exemple. Elles ont par ailleurs souvent un rôle juridique indiquant les conditions d'utilisation des ressources qu'elles décrivent, leur licence éventuelle, etc.

11. À cet égard, une simple collecte, à l'aide d'un appareil photo numérique, par une seule personne peut générer plusieurs centaines d'images. Cette masse de documents numériques devient rapidement inexploitable sans, a minima, un système de classement pertinent et des outils spécifiquement dédiés [Mullen, 2016], tel que le logiciel Tropy, actuellement développé par le RRCHNM et disponible à l'adresse : https://tropy.org.

12. Le terme a été utilisé pour la première fois par [Bagley, 1968] dans un article sur le thème de la programmation informatique.

Les métadonnées et les données, lorsqu'elles ont pu être extraites, forment donc l'intégralité de ce qui peut être recherché par un moteur de recherche. Ces informations ne sont pas, bien entendu, d'une fiabilité parfaite. Dans un premier temps, il est utile de rappeler que les métadonnées, à l'échelle d'un corpus, peuvent être lacunaires. Il n'est pas rare que certains champs soit présents pour certaines ressources et absents pour d'autres. Naturellement, cet aspect parcellaire des métadonnées peut avoir un impact non négligeable sur l'exploitation des ressources.

Nous avons déjà dit, par ailleurs, que l'origine de ces métadonnées pouvait être une description manuelle d'une ressource. Il s'agit, dans le cas de sources historiques, de descriptions faites par des archivistes qui sont, par nature, subjectives. Il est probable que la description d'un document ne recouvre pas l'intégralité de son potentiel et que la seule interrogation de ces descriptions lors d'une recherche conduise à dissimuler des ressources pertinentes pour la démarche de recherche d'un utilisateur. Ce problème trouve un écho particulier dans un champ très utilisé de métadonnées, le champ sujet. Décrire un contenu par une série de mots-clés censés représenter son sujet est un problème bien connu des bibliothécaires. Si le problème se pose dans les rayons des bibliothèques, il se pose également dans les métadonnées d'une bibliothèque numérique.

L'indexation par sujet requiert l'usage d'un référentiel commun entre les diverses institutions qui cataloguent et indexent des ressources. La construction de ces référentiels est un travail titanesque et de longue haleine, mais il est indispensable. Il s'agit ici de s'assurer qu'un mot choisi comme sujet d'une ressource recouvre bien la même réalité pour tous les acteurs susceptibles de l'utiliser. En France, le référentiel qui fait autorité est celui développé et maintenu par la BNF, le référentiel « Répertoire d'autorité-matière encyclopédique et alphabétique unifié » (RAMEAU) [13]. Ce modèle d'indexation est largement utilisé en France par la majorité des grandes bibliothèques, bibliothèques nationales et bibliothèques universitaires. Cependant, cet effort de normalisation se heurte à des problématiques de conversion entre les mots-sujets utilisés dans les anciens catalogues et le référentiel actuel. L'indexation par sujets n'est ainsi pas toujours disponible pour les ressources les plus anciennes [14]. Enfin, bien que le développement de ces référentiels soit effectué de manière concertée sur le plan international, de nombreuses différences subsistent entre les pratiques des différents référentiels couramment utilisés [15].

Classement de l'information

Les données et les métadonnées dont nous venons de parler n'ont qu'une existence numérique. Pour autant, une bibliothèque numérique ne peut se dispenser de tous liens avec l'origine physique des documents. C'est particulièrement le cas pour le classement physique des documents dont il convient de garder une trace lors du passage au nu-

13. Les détails du référentiel sont disponibles à l'adresse http://rameau.bnf.fr Ce référentiel a fait l'objet de travaux universitaires évaluant sa pertinence et ses effets sur la recherche de l'information. Voir en particulier [Tomic, 2006] et [Josué Seckel, 2006].

14. En l'occurence, à la Bibliothèque Nationale de France, seuls les ouvrages postérieurs à 1980 ont bénéficié systématiquement d'une indexation RAMEAU. Tous les ouvrages plus anciens n'ont pas encore pu bénéficier d'une rétro-conversion.

15. À titre d'exemple, le référentiel américain est le *Library of Congress subject headings* (LCSH) accessible à l'adresse https://www.loc.gov/aba/cataloging/subject/.

mérique. L'image de « boîtes » de documents est ici intéressante. Tous les historiens peuvent raconter des anecdotes sur la manière dont ils ont trouvé par hasard une « pépite » dans une boite d'archive qu'ils avaient demandée. L'un des principes fondateurs de l'archivistique, le respect des fonds, érige en règle de modifier le moins possible le classement dans lequel les documents ont été versés aux archives.

Ce classement est bien souvent porteur de sens. Dans le cas où les archives ont été numérisées et suivant le moyen d'accès au document, il est parfois difficile, voire impossible, de consulter les documents de la même boîte. Lorsque l'on accède à un document par le moyen d'une recherche textuelle ou d'une recherche dans des métadonnées, l'information du classement originel du document n'est pas toujours présente et les moyens pour rebondir dans ce classement ne sont pas non plus toujours mis à la disposition du chercheur. Bien entendu, les centres d'archives nationaux et locaux, par exemples les archives départementales françaises suivent les principes fondamentaux de l'archivistique et permettent généralement de parcourir les ressources numérisées par le moyen de différents cadres de classement tout en offrant, pour certains fonds disposant de métadonnées, des fonctions de recherche d'information (voir figure 3.5) [16]. Ces problématiques de classement sont donc fondamentales pour les bibliothèques numériques et donc pour la construction de notre prototype.

FIGURE 3.5 – Exemple d'interface permettant une navigation par cadre de classement et moteur de recherche d'information.

L'interface présentée ici est issue du site de consultation des fonds notariés des Archives Départementales de Charente Maritime. On peut observer que l'interface permet d'obtenir les résultats en fonction de la pertinence ou du cadre de classement.

3.2.2 Traitements

Les opérations qui ont lieu lors de la phase de production des contenus, en vue de leur passage vers la mise en ligne, que ce soit la numérisation ou l'extraction d'informations manipulables (OCR) ne représentent qu'une partie des processus informatiques que nous devons considérer. Pour présenter les contenus, la bibliothèque numérique fait également appel à des traitements qui s'opèrent à la fois sur les données issues des contenus et sur celles provenant des utilisateurs. Leur configuration résulte de choix de conception qu'il nous faut pouvoir reproduire dans l'optique d'évaluer leurs conséquences pour la pratique de l'historien.

16. L'interface présentée est accessible à l'adresse : http://www.archinoe.net/console/plugin/ad17/html/recherche/inventaires.html.

Prétraitement des données

Le moteur de recherche d'une bibliothèque numérique compare, comme nous l'avons déjà dit, des requêtes constituées de plusieurs critères, avec des contenus. Derrière cette opération, qui semble simple au premier abord, se cachent pourtant des opérations complexes. Les textes, tout comme les requêtes des utilisateurs sont en effet des objets de natures très diverses et une simple comparaison consistant à chercher une ou plusieurs occurrences de la requête dans le texte ne suffit pas. Imaginons une recherche ou un utilisateur fournit la requête constituée des mots « chants de noël ». Si le moteur de recherche se contente d'une simple comparaison, la version au singulier du mot « chants », est par exemple ignorée. L'expression « chants des fêtes de noël », qui pourrait elle aussi convenir à la recherche de l'utilisateur subit le même sort. Nous pourrions enfin imaginer qu'inclure à la recherche les termes, « chanter » ou « chanteur.euse » pourrait également s'avérer intéressant, dans l'optique de trouver des résultats.

On peut dès lors imaginer que tenir compte de toutes les variations possibles d'une requête est infiniment plus complexe qu'une simple recherche d'occurrences. D'abord, il faut tenir compte des règles grammaticales des langues. Un mot peut, sans que cela ne doive avoir un impact sur la recherche d'information être singulier ou pluriel, féminin ou masculin. Il existe ainsi une étape de prétraitement, sur laquelle nous reviendrons plus en détail, qui consiste à réduire un mot à sa forme de « racine ». Après un tel processus de racinisation [17], les mots « grand, grands, grande, grandes » par exemple sont tous réduits à la forme « grand ». Le moteur de recherche gagne ainsi en exhaustivité, du point de vue du nombre de résultats.

Par ailleurs, dans la requête que nous donnions précédemment l'expression « chants des fêtes de noël » contient 5 mots rendant la comparaison avec l'expression « chants de noël » inutilement complexe. Les mots « de, des » n'apportent pas d'information directement utile à la recherche. Une étape de préparation des données est donc souvent dédiée à leur suppression permettant des comparaisons plus simples entre requêtes et texte indexé [18].

Il existe ainsi de nombreuses étapes de prétraitement, plus ou moins complexes, qui transforment les données dans le but d'optimiser la recherche d'information, que nous ne citerons pas toutes ici mais sur lesquelles nous reviendrons en détail dans nos chapitres 4 et 6. Les bibliothèques numériques font naturellement usage de ces méthodes qui, si elles ont un intérêt pour le moteur de recherche peuvent aussi avoir, nous le verrons, d'importants effets de bords pour certaines pratiques de recherche d'information.

Indexation et traitement des requêtes

Que ce soit pour une recherche dans des métadonnées ou en plein texte, il est inconcevable pour un moteur de recherche de traiter systématiquement, et à chaque requête, tout un ensemble de contenus pour d'évidentes raisons de coût computationnel. L'opération d'indexation vise, comme son nom l'indique à construire un index des termes constituants les métadonnées et les données d'une ressource. À l'image de l'index classique d'un ouvrage, celui-ci référence les mots associés à des ressources. Il est ainsi

17. En anglais *stemming*, voir section 6.3.1.
18. Il s'agit de l'étape de suppression des mots-vides ou *stop words* en anglais.

nettement plus simple de les parcourir ensuite et de trouver rapidement quels documents sont associés à tel ou tel mot.

Cependant, le moteur de recherche, pour classer les documents, ne se contente pas, bien souvent, de compter le nombre de fois ou un mots est associé à un document pour déterminer sa « pertinence ». Il faut en effet tenir compte, entre autres, de la longueur du contenu, du nombre total de documents, etc. Des algorithmes spécifiques que nous utiliserons, sont dédiés à cette tâche. L'algorithme le plus connu est TF-IDF ($TF \times IDF$), décrit par [Ramos et others, 2003; Salton et Buckley, 1988], dont l'objectif est de déterminer l'importance d'un mot. Celle-ci est estimée en fonction des caractéristique suivantes :

— *TF (Term Frequency)* : la fréquence d'un mot t dans un document donné d. Plus t est présent dans d plus il décrit son contenu et peut donc être considéré comme important [19] ;
— *IDF (Inverse Document Frequency)* : l'inverse du nombre de documents qui contiennent, au moins une fois, le mot t. Plus le mot t est rare dans le corpus, plus sa valeur descriptive est considérée comme importante et donc plus IDF à une valeur élevée.

L'algorithme TF-IDF existe dans beaucoup de variations applicables à différents contextes et différents types de ressources indexables. Ainsi d'une fonction TF brute telle que $TF = f_{t,d}$ certaines variations proposent une fonction $TF = 1 + \log(f_{t,d})$ qui vise à amoindrir les écarts des scores obtenus entre documents. L'objectif est ici de diminuer l'impact de la fréquence d'apparition d'un mot sur la pertinence, car elle n'est pas systématiquement un bon indicateur, un document contenant 10 fois plus un mot t sera dans ce cas plus important, sans l'être 10 fois plus.

Dans le même ordre d'idée, d'autres algorithmes, fondés sur des modèles probabilistes existent visant à diminuer ce défaut de TF-IDF, c'est le cas, en particulier de de BM25 [Mitev et collab., 1985; Pérez-Iglesias et collab., 2009; Robertson et collab., 1995] qui tend à devenir l'algorithme le plus souvent utilisé dans les moteurs de recherche des bibliothèques numériques. À la différence de TF-IDF, BM25 limite TF à une valeur maximale qui ne pourra être dépassée même en cas d'une grande fréquence d'apparition d'un mot. La taille des documents est également prise en compte de manière différente [20]. Comme nous le verrons, les algorithmes de pertinence jouent un grand rôle dans la recherche d'information dans certains contextes. Il est indispensable de pouvoir les maîtriser si l'on souhaite étudier, de près ou de loin, le fonctionnement d'une bibliothèque numérique.

3.2.3 Contexte d'usage : évaluation fonctionnelle

Si nos objectifs de recherche nous conduisent à mettre en place une bibliothèque numérique, celle-ci ne peut pas se contenter de permettre de manipuler des algorithmes d'indexation, des données et des métadonnées. Comme nous l'avons constaté dans la section

19. La taille des documents entre également en ligne de compte, ce n'est pas la même chose de considérer un mot présent 10 fois dans un livre, que dans un article de blog.
20. Dans un document long, la fréquence TF d'un mot augmentera moins rapidement que dans un document plus court.

3.1, une bibliothèque numérique, c'est aussi un ensemble de fonctionnalités orientées vers l'utilisateur. Afin de nous assurer que notre plateforme est représentative des bibliothèques numériques fréquentées par les historiens, il est apparu nécessaire de mener une évaluation fonctionnelle. Cette démarche d'évaluation, très simple, a pour objectif de confronter les fonctionnalités disponibles dans différents systèmes, afin d'identifier les fonctionnalités essentielles d'un système par rapport à un autre et ses éventuels manquements. Cette évaluation est construite sur la base d'une liste de fonctionnalités reconnues comme utiles aux chercheurs et étudiants en sciences humaines et sociales [Audenaert et Furuta, 2010; Buchanan et collab., 2005].

Fonctionnalités évaluées

Afin de mener cette évaluation fonctionnelle et déterminer les fonctionnalités clés que doit fournir une plateforme expérimentale cohérente, nous avons analysé quatre bibliothèques numériques de nature différente mais toutes utiles à des recherches d'information pour l'étude de l'histoire. Ces plateformes sont Gallica, Europeana, Isidore et Jstor. Pour chacune d'entre elles, nous avons observé la présence ou l'absence des fonctionnalités de recherche d'information et de présentation des contenus suivantes :

— **Recherche textuelle simple** : Cette fonctionnalité, apparaissant comme une fonctionnalité de base, se traduit par la présence dans l'interface d'une « barre de recherche » qui permet la saisie de requêtes textuelle.

— **Recherche textuelle avancée** : En parallèle de la fonctionnalité précédente, certaines plateformes peuvent fournir des formulaires de recherche avancée qui permettent de cumuler plusieurs champs de recherche et d'utiliser des opérateurs de recherche complexes en une seule requête.

— **Filtrage par catégorie (facettes)** : Le filtrage par facette est également une fonctionnalité importante d'une bibliothèque numérique aboutie. Elle permet lors d'une requête ou depuis une page de résultats de filtrer ceux-ci en fonction de divers champs, souvent des catégories, par exemple géographiques, de langues, etc. ou par date. Ces fonctionnalités sont également très présentes sur les sites de e-commerce et sont par ce biais connus de la plupart des utilisateurs.

— **Recherche par thématique (index)** : Certaines plateformes fournissent, à la place, ou en parallèle des fonctionnalités de recherche par mots-clés, un index ou des catégories de recherche. Cela peut prendre la forme d'un index alphabétique, chronologique, thématique, ou d'expositions en ligne.

— **Recherche en plein texte** : Il s'agit d'évaluer la possibilité dans une plateforme de recherche en plein texte dans les contenus, en plus de la possibilité de chercher dans les métadonnées.

— **Recherche par entités nommées** : Cette fonction permet de recherche spécifiquement des contenus par le biais d'entités particulières, dites entités nommées. Ce sont des mots dont on peut déterminer la valeur sémantique, personne, lieu ou date par exemple, ce type d'interrogation du système peut se traduire par des questions simples du type *qui, quoi, où ?*

— **Présentation et normalisation des métadonnées** : Comme nous l'avons déjà expliqué, les métadonnées sont fondamentales pour le fonctionnement d'une

bibliothèque numérique. Elles constituent un point d'accès aux données, parfois le seul, et doivent ainsi être présentées par les plateformes et normalisées [21].

— **Recommandation** : Les fonctionnalités de recommandation sont un moyen intéressant d'accéder à l'information. Comme nous le verrons, ils fonctionnent souvent sur des fondamentaux obscurs, difficiles à comprendre pour l'utilisateur. Du reste, elles font partie des fonctionnalités désormais classiques d'une bibliothèque numérique aboutie.

— **Interrogation programmatique du catalogue (API ou Web sémantique)** : Le nombre souvent très important de contenus disponibles dans les bibliothèques numériques en font des ressources parfois difficiles à traiter manuellement. Il devient alors intéressant de pouvoir disposer d'un accès programmatique, par API ou interrogation sémantique [22], pour accéder aux données et les traiter par des méthodes de *distant reading*.

— **Affichage du contenu en plein texte** : L'affichage du contenu en plein texte, quand elle est possible, est une fonctionnalité intéressante pour faciliter le traitement des documents.

Plateformes évaluées

Nous avons évalué la présence ou l'absence des fonctionnalités que nous venons de décrire dans les grandes bibliothèques numériques suivantes :

Gallica est la bibliothèque numériques de la Bibliothèque Nationale de France (BNF), mise en ligne en 1997. À l'origine Gallica était une interface destinée à être consultée sur les terminaux de la bibliothèque nationale, en appoint aux collections classiques. Avec le développement du Web, Gallica est devenue une très importante plateforme présentant gratuitement des millions de documents [23]. Cette interface, visible à la figure 3.6, donne essentiellement accès aux collections de la BNF sans toutefois s'y limiter [24]. Gallica est une interface très aboutie, fournissant de très nombreuses fonctionnalités, font la recherche et l'affichage du plein texte, lorsque les collections s'y prêtent.

Europeana est une bibliothèque numériques lancée par la Commission Européenne en 2008 [25]. Elle a pour objectif de valoriser le patrimoine culturel européen en ligne à destination du grand public, de la recherche ou de l'éducation. Europeana est un méta-moteur. Il agrège, moissonne et centralise au niveau européen des contenus disponibles dans les plateformes dispersées dans les différents États membres au travers de nombreux partenariats. Europeana est par ailleurs un acteur majeur de la recherche sur les thématiques liées aux bibliothèques numériques et à

21. Il faut signaler toutefois que certains algorithmes sont désormais capables d'extraire de l'information depuis des contenus non structurés et être utile à la la recherche et au classement de l'information [Hinze et collab., 2015; Ojokoh et collab., 2015]. Malgré tout, la qualité des métadonnées reste pour l'heure un enjeu central de la qualité d'une bibliothèque numérique.

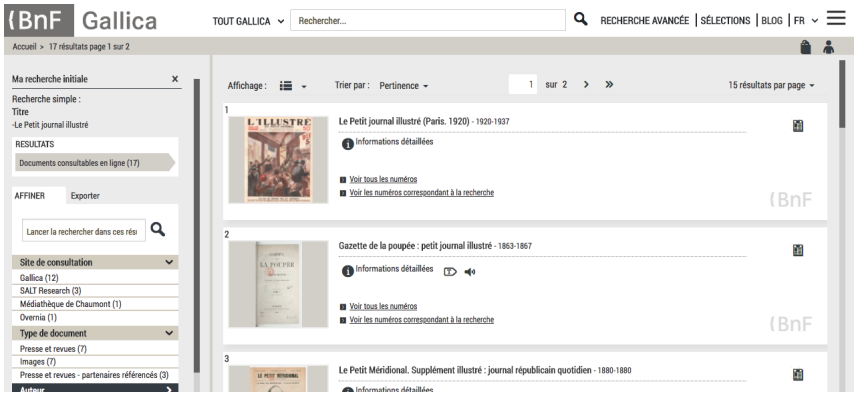
22. Il peut s'agir par exemple d'export RDF/XML ou JSON-LD des données de la bibliothèque numérique, ou de *SPARQL Endpoint*.

23. À l'heure actuelle plus de 4 millions de documents, accessibles via http://gallica.bnf.fr

24. La BNF passe en effet des partenariats permettant à d'autres institutions, bibliothèques ou centre d'archives de diffuser leurs collections via Gallica.

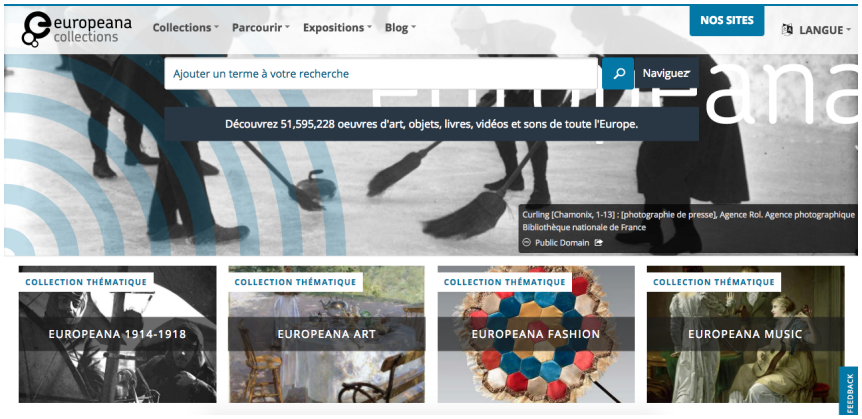
25. http://www.europeana.eu.

FIGURE 3.6 – Interface de recherche de Gallica



la diffusion numérique du patrimoine. Le site indexe un très grand volume de documents [26], ce qui se traduit souvent par des résultats de recherches pléthoriques. L'approche adoptée par Europeana, bien qu'elle ait été critiquée [ERWAY, 2009], consiste à filtrer progressivement les nombreux résultats issus de la requête initiale par de nombreux mécanismes, dans une approche *top-down*. L'une des particularités du méta-moteur européen est de ne pas fournir aux utilisateurs de formulaire d'interrogation avancée. Cependant, des requêtes complexes sont possibles par le biais de l'utilisation des opérateurs booléens (*AND, OR, etc.*) ou des opérateurs d'entités nommées (*what, who, etc.*). Europeana possède par ailleurs d'intéressantes fonctionnalités de navigation par catégories et organise régulièrement des expositions thématiques, telle que l'on peut le constater sur l'interface de la figure 3.7.

FIGURE 3.7 – Interface de recherche et d'exposition de Europeana



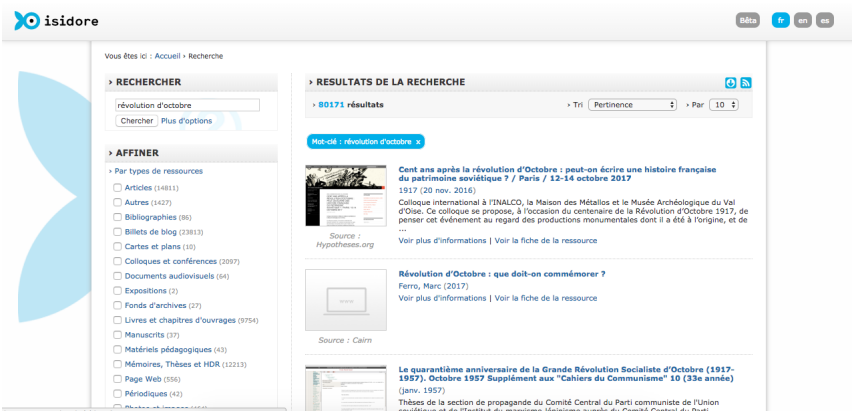
Isidore est à l'instar d'Europeana un méta-moteur qui indexe un grand nombre de ressources à destination de la recherche en sciences humaines et sociales. Il est développé depuis 2009 par la Très Grande Infrastructure de Recherche (TGIR) Huma-Num [27]. Isidore est une plateforme qui dispose d'une armement algorithmique important. Les métadonnées ou le texte intégral des contenus indexés sont analysés pour en extraire des concepts. Ces derniers sont mis en correspondance

26. À ce jour, près de 50 millions.

27. Isidore est accessible à l'adresse https://www.rechercheisidore.fr/ et indexe actuellement environ 5 millions de contenus issus de plus de 6000 sources différentes (voir figure 3.8).

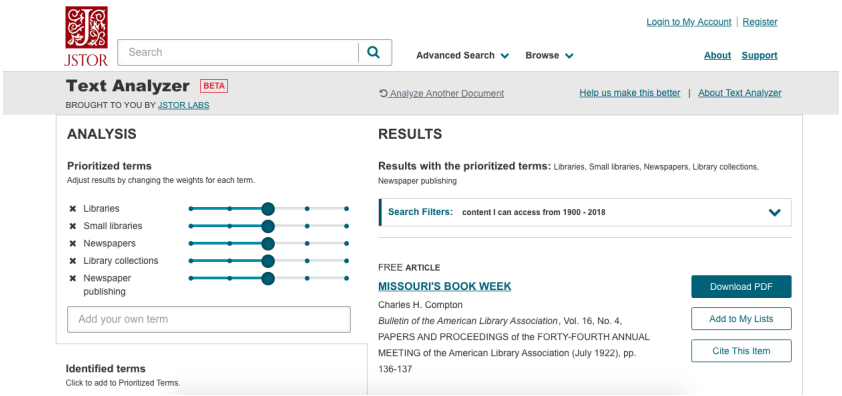
avec des thésaurus et servent à indexer les documents en fonction de ces concepts. Cet effort d'indexation permet de fournir des fonctionnalités d'enrichissement et de recommandation que n'offrent pas d'autres plateformes. La recherche dans le plein texte est également possible bien que relativement complexe du fait de l'absence de fonctions de *highlighting* mettant en lumière les extraits correspondant à la requête saisie par l'utilisateur. La recherche par facette est également disponible et autorise un filtrage *a posteriori* des résultats de requête.

FIGURE 3.8 – Interface de consultation des résultats de recherche d'Isidore



Jstor est une bibliothèque numérique de publications scientifiques lancée en 1995 [28] qui fournit un accès à plus de 10 million d'articles scientifiques et quelques sources primaires. Elle est accessible par abonnement, généralement souscrits par les bibliothèques universitaires des établissements de recherche et d'enseignement supérieur. JSTOR fournit la plupart des fonctionnalités essentielles de recherche d'information. Il est possible de chercher en plein texte autant que dans les métadonnées. Il n'est en revanche pas possible d'utiliser des opérateurs de recherche par entités. Cependant, JSTOR fournit une fonctionnalité qui permet d'utiliser un texte source comme critère de recherche. Par comparaison entre ce texte et l'index, la bibliothèque numérique est capable de trouver des résultats qu'il est ensuite possible de filtrer. Cette interface est visible à la figure 3.9.

FIGURE 3.9 – Interface *text analyzer* de JSTOR



28. JSTOR est accessible à l'adresse https://www.jstor.org.

Tableau 3.1 – Résumé des fonctionnalités offertes par les plateformes évaluées

	Gallica	Europeana	Isidore	Jstor
Recherche textuelle simple	✓	✓	✓	✓
Recherche textuelle avancée	✓	×	×	✓
Filtrage par catégorie (facettes)	✓	✓	✓	✓
Recherche par thématique (index)	✓	✓	✓	✓
Recherche en plein texte	✓	×	✓	✓
Recherche par entités nommées	✓	✓	✓	×
Présentation des métadonnées	✓	✓	✓	✓
Recommandation	×	✓	✓	×
Interrogation programmatique	✓	✓	✓	✓
Affichage du contenu en plein texte	✓	×	✓	✓

Les quatres bibliothèques numériques évaluées sont différentes et ne présentent pas toujours les mêmes fonctionnalités, bien qu'elles possèdent toutes l'essentiel des fonctionnalités attendues. Le présent tableau résume l'état des fonctionnalités, présentes (✓) ou absentes (×) de chaque plateforme.

Les résultats de cette évaluation fonctionnelle, résumés dans le tableau 3.1 montrent que toutes les fonctionnalités évaluées sont présentes dans au moins deux plateformes. Aucune d'entre elles toutefois n'en met à disposition l'intégralité. Afin d'être représentative des capacités réelles des plateformes modernes de recherche d'information, notre plateforme doit donc disposer de l'essentiel de ces fonctionnalités. Certaines d'entre elles sont toutefois dépendantes des contenus proposés. Il est bien entendu plus simple de proposer certaines de ces fonctions lorsque les contenus sont modernes, articles scientifiques ou autre, que lorsqu'il s'agit de contenus anciens dont l'extraction peut être difficile. La conception et le travail autour des contenus sont en effet tout aussi importants que celui sur la bibliothèque numérique.

3.2.4 Bilan de l'état de l'art : expression des besoins fonctionnels

L'état de l'art que nous venons de présenter nous permet d'exprimer les besoins fonctionnels du prototype expérimental indispensable à notre démarche de recherche. Dans l'optique d'être en capacité d'observer et mesurer les effets des différents choix de production des contenus et de conception du système sur la pratique, il nous faut pouvoir reproduire ces conditions et les faire varier au grès de nos besoins expérimentaux. Par ailleurs, notre démarche requiert de pouvoir gérer et observer tous les aspects du système depuis la production, jusqu'aux usages. Notre prototype a donc été conçu pour respecter les points suivants :

— **Stockage des contenus et de leurs représentations :** afin de pouvoir faire varier les conditions de production des contenus, de mener par exemple des recherches dans des contenus originaux avec et sans processus OCR, le prototype doit avoir accès indistinctement aux contenus. Il doit disposer d'un composant permettant de les sauvegarder et d'y accéder.

— **Stockage des métadonnées :** de la même manière le système doit avoir la capacité de sauvegarder et d'organiser les métadonnées de chaque contenu. Il doit pouvoir gérer les éventuels lacunes et permettre d'inclure ou d'exclure rapidement certains type de métadonnées pour la recherche d'information. Ces métadonnées doivent être décrites sous une forme normalisée, correspondant aux standards en la matière.

— **Établir et respecter des cadres de classement :** le prototype n'est pas un simple moteur de recherche, qui interroge ou classe indistinctement des contenus. Il doit disposer d'un modèle de données qui permette d'organiser les contenus sous la forme de collections respectant des cadres de classement divers, qu'ils soient thématiques, chronologiques, *etc.*

— **Simuler différents traitements et pré-traitements :** le système doit permettre de faire varier l'algorithmique de traitement des informations, aussi bien du côté des contenus que du côté des usages. Il doit par exemple permettre de paramétrer et faire varier aisément les règles de prétraitement des contenus et des requêtes ainsi que les paramètres des algorithmes de classement des résultats.

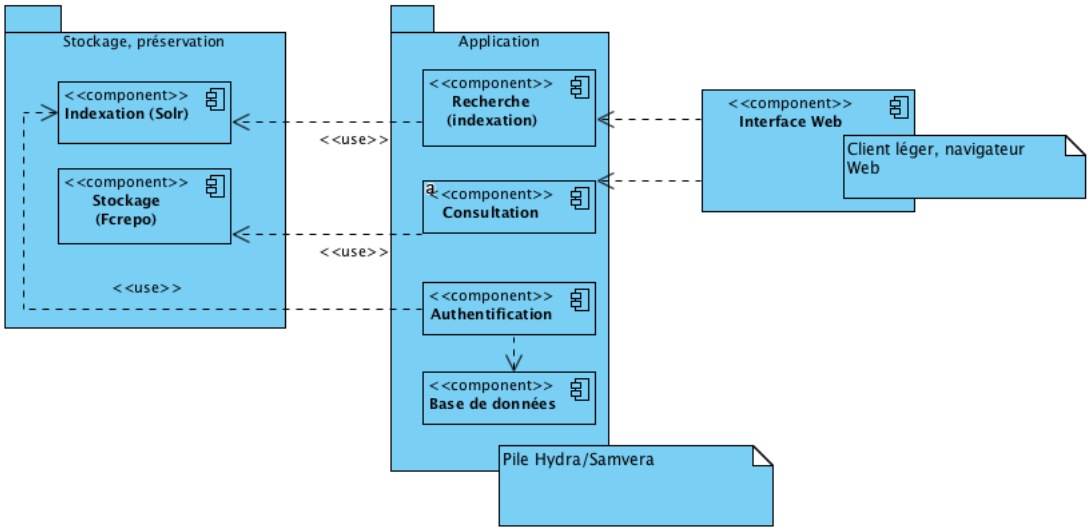
— **Offrir des moyens d'interactions représentatifs :** en terme d'interface et de fonctionnalités, le système doit être capable d'offrir la majorité des outils de recherche d'information disponibles dans les bibliothèques numériques couramment utilisés par notre public.

— **Stabilité et performance :** pour être représentatif des systèmes actuellement en production et ne pas influencer les résultats expérimentaux, notre prototype doit être stable et suffisamment rapide, même s'il intègre des fonctions expérimentales d'observation.

3.3 Plateforme expérimentale pour observer, éclairer et réinvestir : solutions retenues et implémentation

Pour les besoins expérimentaux de notre étude, compte tenu des contraintes fonctionnelles que nous venons de présenter, notre choix s'est porté sur la pile technologique Hydra [29] qui permet de mettre efficacement en place une bibliothèque numérique depuis le stockage et la préservation des contenus jusqu'à l'interface de consultation. Cette solution relativement souple laisse libre choix au développeur du modèle de données et des vocabulaires qu'il souhaite mettre en place et autorise donc une adaptation rapide du contexte de production aux usages à analyser. Par ailleurs, cette pile technologique facilite le développement et le paramétrage de nombreuses fonctions de recherche d'information, depuis la recherche textuelle jusqu'à la recherche par facette.

FIGURE 3.10 – Architecture globale de notre plateforme expérimentale



3.3.1 Architecture générale

Comme nous pouvons le constater sur la figure 3.10, notre plateforme expérimentale repose sur plusieurs composants, responsables des fonctionnalités principales d'une bibliothèque numérique :

Stockage et pérennisation Les données de notre plateforme sont enregistrées à l'aide du logiciel *open source* Fedora Commons (Fcrepo) [30]. Il s'agit d'un système de dépôt de fichiers numériques (*file repository*) qui a la responsabilité d'enregistrer les contenus. Il a également un intérêt pour l'accès à l'information [ZHANG et collab., 2012] et fait également l'objet d'étude de performance dans différents contextes de déploiement [XIE et collab., 2016]. Fcrepo est un outil s'exécutant sur un serveur, accessible par une API Restful. Il a pour responsabilité le stockage des fichiers principaux et de l'ensemble des fichiers dérivés, telles que les miniatures, ou fichiers convertis dans d'autres formats pour des raisons de compatibilité, etc. En parallèle des fichiers, nous enregistrons également dans Fcrepo les métadonnées complètes des contenus indexés par notre plateforme. Fcrepo autorise l'enregistrement dans de nombreux formats de métadonnées, XML ou RDF par exemple.

Indexation et recherche d'information L'indexation des contenus de notre plateforme repose sur le logiciel Solr, lui aussi *open source* et maintenu par la Fondation Apache. Cet outil repose sur un autre projet de la Fondation Apache, Lucene, une bibliothèque logicielle implémentant un grand nombre d'algorithmes utiles à la recherche d'information textuelle [31]. Solr implémente par dessus cette bibliothèque un serveur et des fonctions de haut niveau facilitant l'usage de ces fonctionnalités

29. La pile technologique Hydra a récemment changé de nom et est désormais appelée Samvera. Toutefois, elle reste pour l'heure connue de la communauté sous son nom précédent, par commodité nous nous référerons à cette pile sous son ancien nom : Hydra.
30. Fedora Commons est disponible à l'adresse : http://fedorarepository.org.
31. Les détails de Lucene et Solr sont accessibles à l'adresse : http://lucene.apache.org/solr.

de recherche. Solr est donc à l'image de Fcrepo un outil exécuté par un serveur et accessible par API. Solr n'a pas vocation à enregistrer l'intégralité des données ni même des métadonnées. Nous l'utilisons en lui fournissant uniquement les métadonnées utiles à la recherche d'information, titre, auteur, etc. Solr manipule donc moins de données et est optimisé pour les tâches de recherche, il est de ce fait extrêmement performant.

Interface de consultation L'interface de consultation que nous avons développée repose principalement sur la bibliothèque logicielle Blacklight [32] développée et maintenue par un consortium international regroupant en particulier de nombreuses univeristés. Ce logiciel est très utilisé et est à la base de bibliothèque numérique renommées, parmi lesquelles celle de l'Université américaine de Stanford qui met à disposition de vastes corpus de sources primaires [33]. Blacklight fournit une interface minimale avec le système d'indexation Solr et est aisément modifiable pour faciliter l'usage de Fcrepo. Il met par ailleurs à disposition un ensemble de fonctions logicielles permettant de développer rapidement une interface de recherche incluant recherche textuelle, navigation par catégorie, affichage et consultation des résultats d'une recherche. Blacklight étant principalement développé dans le langage Ruby et le *framework* RubyOnRails [34], nos développements sur les modèles d'interfaçage entre Blacklight, Solr et Fcrepo ainsi que sur l'interface utilisateur sont également écrits dans ce langage, appuyé naturellement des langages du web, HTML, CSS et Javascript.

3.3.2 Gestion du contexte de production et des traitements

Modèle et structure de données et de métadonnées

Comme l'établit la définition des bibliothèques numériques que nous avons exposée, ces systèmes sont organisés. Les contenus sont généralement traités comme des collections que le concepteur peut créer, enrichir, modifier. Les raisons qui président à la construction d'une telle collection sont à la discrétion du concepteur de la ressource, elles peuvent être thématiques, linguistiques, chronologiques, etc. Naturellement, les contenus peuvent appartenir à plusieurs collections et peuvent eux mêmes être composites, voire hétérogènes. Autrement dit, il est fréquent qu'un contenu soit composé de plusieurs fichiers, du même type ou de plusieurs types différents. À titre d'exemple on peut imaginer un manuscrit, sous forme de fichier image d'une part et sous forme de fichier texte, issu d'une transcription d'autre part. Ce manuscrit peut très bien appartenir à une collection thématique, une problématique historique par exemple, ou une collection géographique ou encore chronologique. Ces considérations liées au caractère organisé d'une bibliothèque numérique imposent de réfléchir à un modèle de données adapté qui réponde à ces exigences.

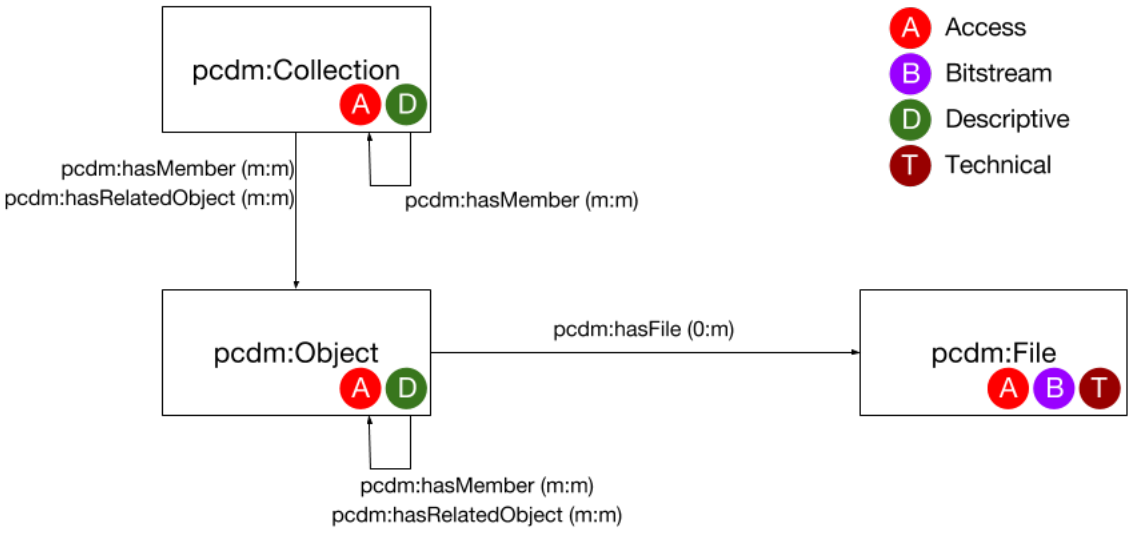
Par ailleurs, ces éléments de cahier des charges fonctionnel se complètent d'impératifs

32. Le site du projet Blacklight donne tous les détails à l'adresse : http://projectblacklight.org/.
33. À titre d'exemple et parmi les nombreuses ressources numériques mises à disposition par l'université Stanford, la collection numérique d'archive de la révolution française est accessible à cette adresse : https://frda.stanford.edu/.
34. La documentation de Ruby est disponible à l'adresse : https://www.ruby-lang.org/, celle de RubyOnRails à l'adresse : http://rubyonrails.org/.

techniques. En effet, notre objectif est de produire une bibliothèque numérique capable non seulement de respecter les bonnes pratiques d'indexation de contenus patrimoniaux, mais également de les organiser, les afficher les distribuer. Il est donc apparu nécessaire de se pencher sur la structuration du modèle de données le plus adapté à la pile technologique retenue. La principale contrainte était d'articuler avec un haut niveau de fiabilité les systèmes sur lesquels repose la pile, Fcrepo et Solr. Le système doit être en capacité d'adresser à l'un ou l'autre de ces outils logiciels les requêtes nécessaires aux fonctionnalités fournies aux utilisateurs. Pour tout ce qui concerne la recherche d'information et le traitement des résultats, le système doit s'adresser à Solr, pour obtenir le détail des métadonnées et accéder aux contenus, il doit alors contacter Fcrepo. Ces problématiques sont valables aussi bien pour la consultation des données que pour leur intégration dans le système.

En termes d'implémentation, l'idéal était donc de mettre en place un modèle unique disposant, dès lors qu'il est chargé des accès aux deux systèmes sous-jacents qui puissent automatiquement contacter le bon système en fonction du type d'information qu'il est nécessaire de récupérer. Par ailleurs ce modèle devait être en capacité de permettre une organisation cohérente des contenus, pour refléter aux mieux la définition des bibliothèques numériques. Nos recherches nous ont conduits vers des études et développements effectués en ce sens par un groupe du consortium développant la pile logicielle Hydra. Ce groupe a élaboré un modèle appelé *Portland Common Data Model (PCDM)*[35] utilisé avec succès dans différents contextes [Cowan et collab., 2015; Wilcox, 2016] dont les principes généraux sont exposés à la figure 3.11.

Figure 3.11 – Principes généraux du modèle PCDM.



Dans ce modèle, dont les détails sont accessibles à l'adresse https://github.com/duraspace/pcdm/wiki, les éléments notés A (droits d'accès), B (fichiers), D (métadonnées descriptives) et T (métadonnées techniques) désignent le type de métadonnées associé à chaque composant du modèle.

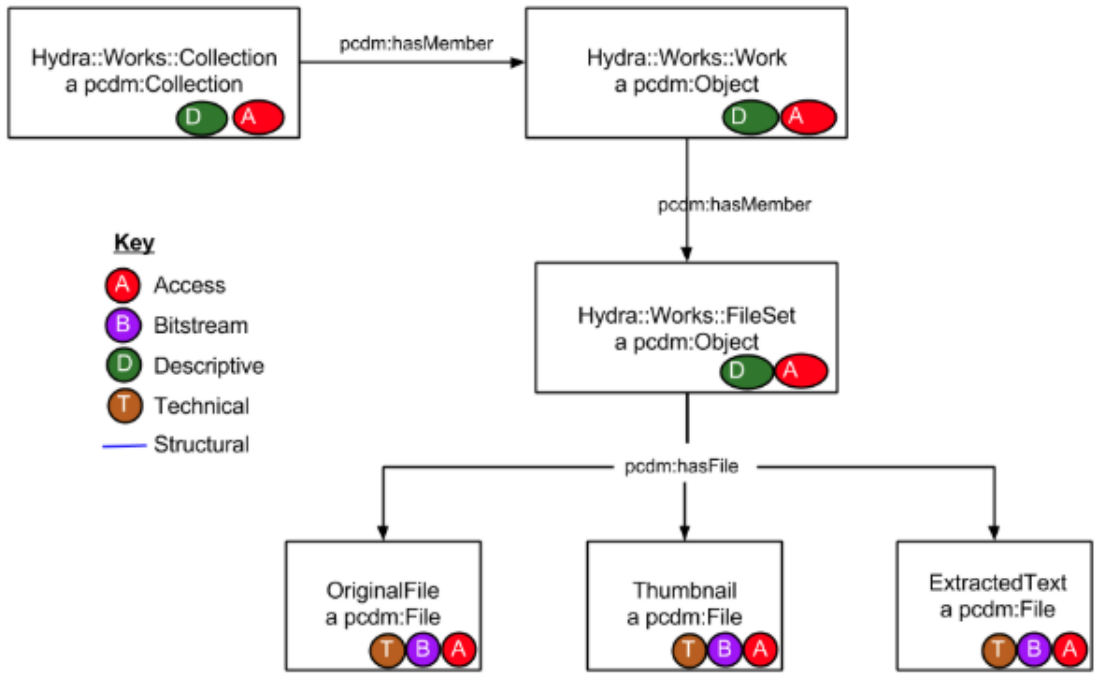
Ce modèle présente les relations entre les différents composants qui organisent les don-

35. L'historique du développement et les détails de PCDM, dont est extraite la figure 3.11 sont accessibles à l'adresse : https://github.com/duraspace/pcdm/wiki.

nées de la bibliothèque numérique. Les trois classes, *Collection*, *Object* et *File* possèdent des caractéristiques différentes. Les deux premières ne peuvent contenir que des métadonnées descriptives et possèdent des droits d'accès. La dernière possède également des droits d'accès mais se limite à des métadonnées techniques et aux fichiers. Le principal intérêt de ce modèle se trouve naturellement dans les relations entre ces trois classes. Une *Collection* contient un ou plusieurs *Object*, qui eux mêmes peuvent contenir des *File*. Il est important ici de remarquer que le modèle autorise qu'un *Object* en encapsule un autre et qu'une *Collection* puisse également contenir une autre collection. Enfin, rien n'empêche que ces relations soient multiples et qu'un *Object* puisse appartenir à plusieurs collections et ainsi de suite. Si l'on revient un instant aux exigences que nous avons décrites, ce modèle convient parfaitement. Pour l'exemple mentionné plus haut, notre manuscrit devient un *Object* et est ainsi décrit par des métadonnées descriptives. Il encapsule deux instances de *File*, l'une pour l'image et l'autre pour le texte transcrit. Il peut enfin appartenir indistinctement à des collections thématiques ou géographiques qui seront des instances de *Collection* et seront décrites indépendamment par leurs métadonnées descriptives.

Toutefois, dans le cas ou plusieurs fichiers composent un objet, il se peut que nous voulions pouvoir renseigner des métadonnées descriptives au niveau de ces fichiers. Dans notre exemple, l'auteur de la transcription par exemple. Ce problème est résolu par l'implémentation du modèle PCDM que nous avons choisi d'utiliser, dont les détails sont visibles à la figure 3.12. Cette implémentation du modèle, dénommée Hydra-Works est écrite en Ruby. Elle introduit les notions de *Work* et de *Fileset* qui étendent le concept d'*Object* de PCDM en ajoutant un niveau de souplesse au modèle, particulièrement pour les objets complexes composés de fichiers, ou jeux de fichiers multiples [36].

FIGURE 3.12 – Principes généraux de l'implémentation Hydra-works du modèle PCDM.



36. Le dépôt github de cette implémentation est accessible à l'adresse : https://github.com/samvera/hydra-works.

Cette implémentation et ce modèle nous permettent d'assurer l'organisation des données de notre plateforme expérimentale, nous nous référerons donc, dans les développements qui suivent, aux termes de ce modèle lorsque cela sera nécessaire. Ces instances possèdent des propriétés qui décrivent les métadonnées descriptives et techniques ainsi que les droits d'accès.

Nous avons choisi d'exprimer ces propriétés au format RDF, nativement supporté par Fcrepo. Ce format d'expression des métadonnées nous laisse libre de choisir les vocabulaires de description les plus adaptés aux métadonnées que l'on considère. Dans le cadre du développement de cette plateforme expérimentale, nous n'avions pas l'objectif de lier nos contenus à des plateformes externes. Toutefois, notre choix s'est tout de même portée vers le format RDF parce que celui-ci est aujourd'hui, avec le développement du web sémantique et des données liées, le format de description le plus utilisé, vers lequel sont en train de migrer de nombreuses institutions et acteurs des bibliothèques numériques [Clavaud et Château-Dutier, 2017].

En terme de vocabulaire, là encore, nous avons choisi d'utiliser autant que possible des ontologies reconnues. Selon les cas qui se sont présentés lors de nos expérimentations, et donc selon le type de contenu intégré à la plateforme ces vocabulaires ont été amenés à évoluer. Cependant, nous avons systématiquement limité le nombre d'ontologies utilisées et décrits nos métadonnées à l'aide de *DC*[37] pour les plus générales, *MODS*[38] pour les plus spécifiques et quelques autres ontologies telles que *FOAF*[39] pour décrire des personnes.

Règles d'indexation

L'indexation des documents repose dans notre plateforme expérimentale sur Apache Solr. Ce dernier, intégrant la bibliothèque logicielle Lucene, dispose de très nombreuses règles qui permettent de paramétrer la façon dont l'on souhaite indexer les données et donc la manière dont seront retrouvés les documents. Nous reviendrons en détail sur ces règles d'indexation, qui résultent pour l'essentiel de choix de conception des concepteurs de la ressource et qui se veulent parvenir à un compromis entre exhaustivité et pertinence des résultats de recherche.

Ces règles s'appliquent, par exemple, à l'algorithme de pertinence choisi et à ses paramètres. Pour rappel, ces algorithmes fournissent différents scores de pertinence qui peuvent être pondérés afin d'aboutir au score final, qui détermine la position d'un contenu parmi tous les contenus possibles. Une de ces pondérations est classiquement mise en place pour donner plus ou moins de poids au type de métadonnées dans lesquels un mot, ou une expression a été trouvé. Le concepteur peut ainsi considérer que si un terme de la recherche d'un utilisateur est présent dans le titre d'un contenu, alors ce contenu est plus pertinent que si le terme se retrouve dans le résumé ou dans le texte.

Notre plateforme rend possible ce type de paramétrage, comme nous le verrons en

37. DC (*Dublin Core*) est maintenu par la *Dublin Core Metadata Initiative* et spécifiée à l'adresse : http://dublincore.org/documents/2008/01/14/dc-rdf.

38. MODS (*Metadata Object Description Schema*) est maintenu par la *Library of Congress* et décrite à l'adresse : https://www.loc.gov/standards/mods/modsrdf.

39. FOAF (*Friend of A Friend*) est conçue par le W3C, les spécifications sont disponibles à l'adresse : http://xmlns.com/foaf/spec.

détail, mais la problématique principale ici est de définir un comportement « standard », qui reflète la réalité des systèmes actuellement mis en place et utilisés par les historiens et étudiants en histoire pour leurs recherches. Nous avons donc étudié ces paramètres, à la fois sur la base de travaux spécifiquement menés sur cette question et en étudiant les paramètres choisis par des bibliothèques numériques existantes, lorsque ces informations sont publiques, ce qui est loin d'être toujours le cas. Nous avons par ailleurs adapté les règles d'indexation aux langues des documents que nous utilisons pour nos expérimentations. Les paramètres que nous avons adoptés sont, sauf exception, ceux définis par Europeana, qui donne un accès publique aux règles d'indexation et de recherche de sa bibliothèque numérique [40].

3.3.3 Fonctionnalités et interface

L'interface de la plateforme présente les fonctionnalités essentielles d'une bibliothèque numérique. Dans un premier temps l'interface propose les fonctionnalités de recherche par l'intermédiaire d'une barre de recherche textuelle. Cette fonction minimale se complète d'une fonctionnalité de recherche par facette qui permet de chercher par catégorie (langue, type, chronologie, etc.). Ces deux fonctionnalités peuvent fonctionner de concert, dans ce cas, la recherche par facette fait office de fonction de filtrage des résultats issus de la recherche textuelle. Par ailleurs le système est capable d'indiquer l'impact de la sélection d'une catégorie sur le nombre de résultats de recherche. L'ensemble de ces fonctionnalités est visible sur la figure 3.13.

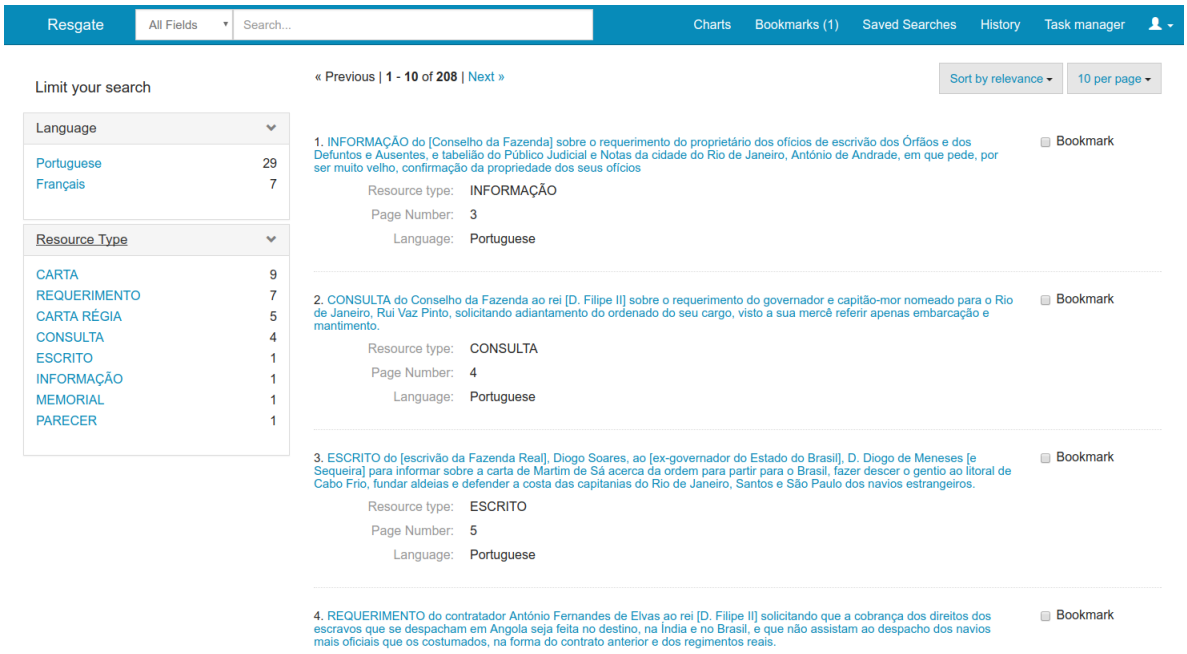
Notre plateforme expérimentale fournit naturellement une liste de résultats à laquelle nous nous référerons par la suite sous le terme *Search Engine Results page* (SERP). Celle-ci est paramétrable par le concepteur de la ressource et peut potentiellement afficher toutes les métadonnées indexées. La liste des champs de métadonnées affichées dans la SERP est toutefois généralement limitée aux informations permettant à l'utilisateur de choisir les contenus les plus utiles, titre, résumé, langue, etc. La SERP de notre plateforme peut afficher un nombre variable de résultats dont le classement peut également être modifié par l'utilisateur qui peut par exemple choisir de lister les résultats par ordre alphabétique plutôt que par pertinence (voir figure 3.13).

Lorsqu'un utilisateur sélectionne un contenu dans la SERP, la plateforme affiche une page affichant les détails de ce contenu. Cette page est, là encore, paramétrable et peut afficher si nécessaire l'intégralité des métadonnées dont le système dispose. Cette page permet par ailleurs de passer rapidement aux résultats précédents et suivants et donnent à voir le contenu en miniature. Cette page présente également un outil permettant de transférer par mail la référence du document et d'obtenir la référence complète du contenu à des fins de citation. C'est enfin depuis cette page que l'utilisateur peut accéder à la visualisation du contenu. L'ensemble de ces fonctionnalités sont visibles sur la figure 3.14.

Nous avons également doté notre plateforme de fonctionnalités permettant d'afficher et de télécharger les contenus. Si l'utilisateur souhaite consulter le contenu, la plateforme affiche le document (voir figure 3.15) avec différentes fonctions de visualisations,

40. Le dépôt Github suivant donne en effet accès aux nombreux fichiers de configurations utilisés par Europeana : https://github.com/europeana/search.

FIGURE 3.13 – Interface de recherche et de consultation des résultats.



Sur cette interface, la barre de recherche est située en haut de l'interface. La recherche par facette se trouve à gauche de la fenêtre. La zone principale affiche les résultats de la recherche d'information et les outils de classement utilisables par l'utilisateur.

pagination, zoom, etc. Par ailleurs, si le contenu s'y prête, cette fenêtre permet de basculer entre le contenu original et son extraction textuelle. Depuis cette extraction, si sa qualité et son volume sont compatibles, il est également possible d'opérer quelques analyses simples sur cette extraction, calcul de Ngram par exemple.

Enfin, la plateforme dispose de fonctions permettant à l'utilisateur d'enregistrer et de retrouver aisément les recherches qu'il a déjà menées. Les fonctions de mise en favoris sont disponibles sur les pages SERP et détail du résultat. Cette fonction ne concerne que les contenus individuels. Pour ce qui est des recherches, l'historique de recherche est sauvegardé automatiquement. Depuis cette historique l'utilisateur peut sélectionner les recherches qu'il souhaite sauvegarder de manière pérenne, ainsi qu'effacer l'intégralité des données (voir figure 3.16).

Ainsi, en plus de ces fonctionnalités de sauvegarde des recherches ou de lecture des contenus, notre plateforme fournit l'essentiel des fonctionnalités que nous avions identifiées dans notre section 3.2.3. Le tableau 3.2 présente un résumé de ces fonctionnalités.

FIGURE 3.14 – Interface de consultation des détails d'un contenu.

Sur cette interface, la zone principale affiche les métadonnées complètes. La miniature est affichée dans la zone de gauche, les outils d'export et le bouton vers la page d'affichage des contenus sont situés à droite de la fenêtre.

FIGURE 3.15 – Interface de consultation des contenus (contenu original).

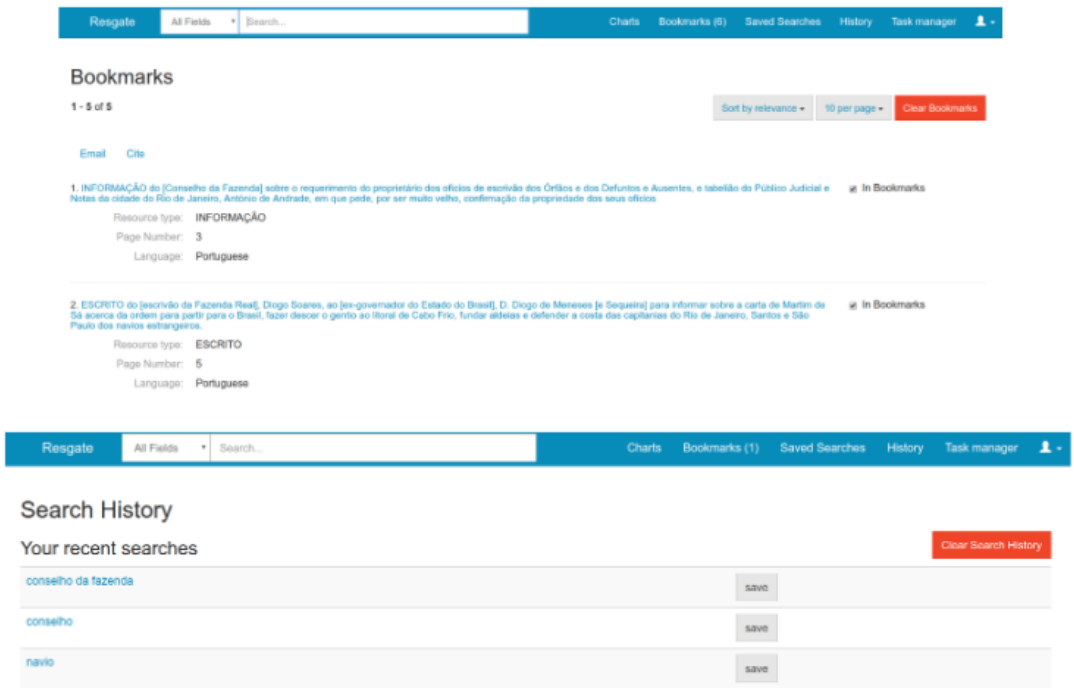
Ici, les pages du contenu sont visibles dans la zone principale de la fenêtre. Selon le type de fichier, différentes options (zoom, pagination, etc.) sont disponibles. Cette interface permet également de télécharger le contenu original ou d'afficher l'extraction textuelle du contenu lorsque le contenu a pu être extrait.

Tableau 3.2 – Résumé des fonctionnalités offertes par notre plateforme expérimentale

	Plateforme expérimentale
Recherche textuelle simple	✓
Recherche textuelle avancée	✓
Filtrage par catégorie (facettes)	✓
Recherche par thématique (index)	×
Recherche en plein texte	✓
Recherche par entités nommées	×
Présentation des métadonnées	✓
Recommandation	✓
Interrogation programmatique	✓
Affichage du contenu en plein texte	✓

Au regard des jeux de données et de l'objectif de nos expérimentations, il ne nous a pas été systématiquement possible de mettre en place les fonctionnalités marquées par un ×.

Figure 3.16 – Interface de consultation des favoris et de l'historique de recherche.



En haut, l'interface de consultation des favoris. Un clic sur l'un de ces favoris permet d'accéder directement au document. En bas, l'interface de consultation de l'historique de recherche. Les boutons « save » permettent de sélectionner les recherches à sauvegarder de manière pérenne.

3.4 Conclusion

Dans ce chapitre, nous avons pu voir que les bibliothèques numériques sont des objets logiciels et fonctionnels complexes. Leur définition théorique est très vaste, témoignant de leur large champ d'activité et leur importance primordiale en matière de recherche d'information et de documents primaires ou secondaires. Notre objectif étant d'observer et de réfléchir à leur impact sur la pratique quotidienne, il nous fallait disposer d'une plateforme expérimentale cohérente avec les standards des systèmes existants.

Pour ce faire, nous avons mené une évaluation fonctionnelle qui nous a permis de déterminer, dans un premier temps, une liste de fonctionnalités majeures utiles à l'activité de recherche d'information pour la pratique de l'histoire. Ces fonctionnalités constituent un large éventail d'outils d'accès et de filtrage de l'information, à l'heure où le nombre de documents disponibles sur de telles plateformes numériques explose. Dans un second temps, nous avons évalué la présence ou l'absence de ces fonctionnalités sur des bibliothèques numériques reconnues et utilisées par la communauté de recherche sur laquelle nous nous focalisons. Nous avons pu observer que ces plateformes fournissaient toutes la plus part des fonctionnalités attendues et qu'elles pouvaient ainsi satisfaire à des expressions complexes du besoin d'information.

Par ailleurs, nous avons également pu observer qu'un système de bibliothèque numérique ne se résume pas à un ensemble, plus ou moins important, de fonctionnalités techniques « visibles » de l'utilisateur, et donc évaluable par une simple évaluation fonctionnelle. Pour qu'un tel système puisse fonctionner de manière adaptée, un travail important doit être mené sur les données elle-mêmes. De la constitution des métadonnées à l'extraction d'information depuis les contenus, par exemple par des processus d'OCR, ces opérations ne sont pas neutres et ne peuvent fournir que des vues partielles de la réalité.

Ainsi, que ce soit au niveau du système et des traitements subis par les données, avant leur indexation comme pendant la recherche d'information ou du côté de l'utilisateur à travers les fonctionnalités des bibliothèques que nous avons évaluées, les biais potentiel qui peuvent nuire à la pratique de recherche d'information dans une discipline comme l'histoire sont nombreux. Afin de les observer et de les évaluer, nous avons mis en place une plateforme expérimentale qui respecte, aussi bien du point de vue du traitement des données et de l'indexation que de celui des fonctionnalités proposées les standards des grandes plateformes existantes. Cette plateforme peut ainsi être la base de nos travaux visant à montrer l'impact des biais, du côté des utilisateurs, comme de celui du système qu'induisent l'usage de tels outils de la boîte à outils des historiens et futurs historiens.

3.5 Références

Audenaert, N. et R. Furuta. 2010, «What Humanists Want : How Scholars Use Source Materials», dans *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, ACM, New York, NY, USA, p. 283–292, doi : 10.1145/1816123.1816166. 70

Bagley, P. R. 1968, «Extension of programming language concepts», cahier de recherche, University City Science Center Philadelphia. 65

Bassil, Y. et M. Alwani. 2012, «Ocr post-processing error correction algorithm using google online spelling suggestion», *arXiv preprint arXiv :1204.0191*. 62

Blei, D. M., A. Y. Ng et M. I. Jordan. 2003, «Latent dirichlet allocation», *Journal of machine Learning research*, vol. 3, n° Jan, p. 993–1022. 62

Buchanan, G., S. J. Cunningham, A. Blandford, J. Rimmer et C. Warwick. 2005, «Information seeking by humanities scholars», dans *International Conference on Theory and Practice of Digital Libraries*, Springer Berlin Heidelberg, p. 218–229. 70

Chiron, G., J.-P. Moreux, A. Doucet, M. Coustaty et M. Visani. 2017, «Erreurs OCR et biais d'indexation : impact sur les usages», dans *17ème conférence Extraction et Gestion des Connaissances, Atelier Journalisme Computationnel*, p. 69–73. 62

Clavaud, F. et E. Château-Dutier. 2017, «Une Preuve de concept pour la sémantisation et la visualisation orientée utilisateur de données archivistiques», dans *Digital Humanities (DH) 2017*, Montréal, p. 195–197. 80

Cowan, W., R. Floyd et D. Pierce. 2015, «"Contentless" Digital Collections», *Indiana University Digital Collections Services*. 78

Cunningham, S. J., N. Reeves et M. Britland. 2003, «An ethnographic study of music information seeking : implications for the design of a music digital library», dans *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, IEEE Computer Society, p. 5–16. 58

Erway, R. 2009, «A view on Europeana from the US perspective», *Liber Quarterly*, vol. 19, n° 2. 72

Eskenazi, S., P. Gomez-Kramer et J.-M. Ogier. 2017, «A comprehensive survey of mostly textual document segmentation algorithms since 2008», *Pattern Recognition*, vol. 64, p. 1–14. 61

Fox, E. A., M. A. Goncalves et R. Shen. 2012, «Theoretical foundations for digital libraries : The 5s (societies, scenarios, spaces, structures, streams) approach», *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 4, n° 2, doi :10.2200/S00434ED1V01Y201207ICR022, p. 1–180. 57

Gefen, A. 2015, «Les enjeux épistémologiques des humanités numériques», *Socio. La nouvelle revue des sciences sociales*, , n° 4, doi :10.4000/socio.1296, p. 61–74, ISSN 2266-3134. 64

Gooding, P. 2014, «Exploring Usage of Digital Newspaper Archives through Web Log Analysis : A Case Study of Welsh Newspapers Online», *Digital Humanities 2014*. 64

Graham, S., S. Weingart et I. Milligan. 2012, «Getting Started with Topic Modeling and MALLET», *Programming Historian*. URL https://programminghistorian.org/lessons/topic-modeling-and-mallet. 62

Gupta, M. R., N. P. Jacobson et E. K. Garcia. 2007, «OCR binarization and image pre-processing for searching historical documents», *Pattern Recognition*, vol. 40, nº 2, p. 389–397. 61

Hinze, A., C. Taube-Schock, D. Bainbridge, R. Matamua et J. S. Downie. 2015, «Improving Access to Large-scale Digital Libraries ThroughSemantic-enhanced Search and Disambiguation», dans *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, ACM, New York, NY, USA, p. 147–156, doi :10.1145/2756406.2756920. 71

Jackson, A., J. Lin, I. Milligan et N. Ruest. 2016, «Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities», dans *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, ACM, New York, NY, USA, p. 103–106. 59

Josué Seckel, R. 2006, «Sur quelques fausses perspectives de l'indexation par sujet (Rameau) et de l'Opac. L'exemple du catalogue BNOpale-plus (Bibliothèque nationale de France)», *Matériaux pour l'histoire de notre temps*, vol. 82, nº 2, p. 92–94. 66

Klijn, E. 2008, «The current state-of-art in newspaper digitization», *D-Lib Magazine*, vol. 14, nº 1/2, p. 1082–9873. 60

Larsen, R. L. et H. D. Wactlar. 2004, «Knowledge Lost in Information : Report of the NSF Workshop on Research Directions for Digital Libraries, June 15-17, 2003, Chatham, MA, National Science Foundation Award No. IIS-0331314», cahier de recherche, University of Pittsburgh, Pittsburgh. 57

Lee, D.-S. et R. Smith. 2012, «Improving book OCR by adaptive language and image models», dans *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, IEEE, p. 115–119. 62

Milligan, I. 2013, «Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997–2010», *The Canadian Historical Review*, vol. 94, nº 4, p. 540–569. Volume 94, Number 4, December 2013. 64

Mitev, N. N., G. M. Venner et S. Walker. 1985, *Designing an online public access catalogue : Okapi, a catalogue on a local area network*, The British Library. 69

Mullen, A. 2016, «Untangling the Mess : Researchers' Photo Practices», . 65

Ojokoh, B. A., O. M. Omisore et O. W. Samuel. 2015, «Automatic Classification of Research Documents Using Textual Entailment», dans *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, ACM, New York, NY, USA, p. 251–252, doi :10.1145/2756406.2756960. 71

Pérez-Iglesias, J., J. R. Pérez-Agüera, V. Fresno et Y. Z. Feinstein. 2009, «Integrating the probabilistic models BM25/BM25f into Lucene», *arXiv preprint arXiv :0911.5046*. 69

Ramos, J. et others. 2003, «Using tf-idf to determine word relevance in document queries», dans *Proceedings of the first instructional conference on machine learning*, vol. 242, p. 133–142. 69

Robertson, S. E., S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford et others. 1995, «Okapi at TREC-3», *Nist Special Publication Sp*, vol. 109, p. 109. 69

Salton, G. et C. Buckley. 1988, «Term-weighting approaches in automatic text retrieval», *Information processing & management*, vol. 24, n° 5, p. 513–523. 69

Tanner, S. 2004, «Deciding whether optical character recognition is feasible», *King's Digital Consultancy Services*. 61

Tomic, Y. 2006, «L' « a-chronologie » de Rameau», *Matériaux pour l'histoire de notre temps*, vol. 82, n° 2, p. 95–95, ISSN 0769-3206. 66

Traub, M. C., J. van Ossenbruggen et L. Hardman. 2015, «Impact Analysis of OCR Quality on Research Tasks in Digital Archives», dans *Research and Advanced Technology for Digital Libraries*, Springer, p. 252–263. 60, 63

Traub, M. C., T. Samar, J. van Ossenbruggen, J. He, A. de Vries et L. Hardman. 2016, «Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus», dans *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, ACM, New York, NY, USA, p. 7–16. 64

Waters, D. J. 1998, «Transforming Libraries Through Digital Preservation», *Collection Management*, vol. 22, n° 3-4, doi :10.1300/J105v22n03_09, p. 99–111. 57

Wilcox, D. 2016, «A Linked Data Approach to Digital Newspapers with Fedora and PCDM», *IFLA*, p. 6. 78

Xie, Z., Y. Chen, J. Speer et T. Walters. 2016, «Evaluating cost of cloud execution in a data repository», dans *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*, IEEE, p. 247–248. 76

Yang, T.-I., A. J. Torget et R. Mihalcea. 2011, «Topic modeling on historical newspapers», dans *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, p. 96–104. 62

Zhang, H., M. Durbin, J. Dunn, W. Cowan et B. Wheeler. 2012, «Faceted Search for Heterogeneous Digital Collections», dans *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, ACM, New York, NY, USA, p. 425–426, doi :10.1145/2232817.2232924. 76

Chapitre 4

Observer le contexte d'usage : outil d'observation et indicateurs du comportent de recherche d'information

Résumé

Dans l'objectif de confronter l'usage d'une bibliothèque numérique avec ses logiques de fonctionnement, il nous est indispensable de disposer de moyens permettant d'observer et de mesurer cet usage. Ce chapitre présente donc l'état de l'art des modèles conceptuels et des techniques d'observation des pratiques de recherche d'information. Devant les limites opérationnelles des modèles existants, ce chapitre propose un outil et une méthode d'observation de l'usage fondés sur des traces brutes et des indicateurs de plus haut niveau. Cette mécanique d'observation et les indicateurs de la pratique qu'elle produit sont évalués expérimentalement dans le contexte de tâches de recherche en histoire.

Sommaire

4.1 Introduction

Ce chapitre présente notre étude du contexte d'usage d'un système de recherche d'information pour l'historien. Ce contexte est le reflet de la pratique des utilisateurs dans le système. Il regroupe les activités d'expression du besoin d'information, sous forme de requêtes la plupart du temps, ainsi que les activités d'analyses des résultats du système et la consultation des contenus. Bien entendu, il est déterminant pour la qualité des résultats produits par le système. À ce titre, il est un maillon fondamental de notre schéma conceptuel, si l'on souhaite vérifier l'adéquation entre l'usage du système et les présupposés qui ont induit les choix de conception sur lesquels celui-ci repose.

L'objectif de ce chapitre n'est donc pas de déterminer si telle ou telle pratique peut, à elle seule, être source de biais méthodologiques pour l'historien. Il s'agit plutôt de se donner les moyens d'observer les pratiques dans le contexte global de la ressource et de les confronter avec les autres processus à l'œuvre dans la bibliothèque numérique. Il n'est ainsi pas question de recommander une stratégie de recherche dans l'absolu, mais de mettre en place une méthode qui, dans des contextes de production ou d'exécution donnés, puisse permettre de révéler les biais méthodologiques.

Cet objectif repose sur l'observation et la modélisation du comportement de recherche d'information des utilisateurs. L'état de l'art présenté dans ce chapitre fournit des modèles généraux ainsi que des modèles plus spécifiquement orientés vers les chercheurs en sciences humaines et particulièrement en histoire. Ce sont des modèles de recherche d'information globaux, qui tiennent compte de tous les aspects de la documentation du chercheur, depuis la bibliographie d'un ouvrage physique, jusqu'à la documentation numérique. Leur développement s'est fondé sur des données collectées via des enquêtes quantitatives et qualitatives. Pour autant, si ils sont d'un grand intérêt sur le plan conceptuel, il le sont peu sur le plan fonctionnel dans la mesure où ils ne fournissent pas d'indicateurs précis des différentes stratégies de recherche d'information des utilisateurs.

Pour proposer une étude du contexte d'usage qui puisse satisfaire nos objectifs de recherche, nous avons donc souhaité :

— mettre en place une mécanique d'observation de la pratique des utilisateurs dans la bibliothèque numérique ;

— définir une méthodologie de traitement des traces résultantes de cette observation ;

— vérifier la pertinence des indicateurs produits, autrement dit leur capacité à décrire l'activité de l'utilisateur ;

— confronter ces indicateurs avec les contextes de production et d'exécution et identifier des biais liés au comportement des utilisateurs.

Ce chapitre présente donc l'état de l'art, à travers différents modèles, de l'analyse du comportement de recherche d'information en section 4.2. Il se poursuit, en section 4.3, par la description et l'explication de la mécanique d'observation que nous avons mis en œuvre dans notre contexte de bibliothèque numérique. Cette stratégie s'appuyant sur des indicateurs issus d'observation de terrain, elle s'accompagne d'une démarche de

validation expérimentale et d'une présentation des résultats, respectivement en section 4.4 et 4.5.

4.2 Comportement de recherche d'information : état de l'art

L'observation et la modélisation des utilisateurs et de leurs activités dans un système est un des défis majeurs de l'informatique et une thématique de recherche sur le plan international rassemblant de nombreux chercheurs autour de conférences et de publications importantes [1]. Ce champs de recherche a donc un périmètre très vaste et impose des choix de lecture et de présentation de l'état de l'art.

Dans notre contexte, pour faire les choix les plus adaptés à nos objectifs, nous nous focaliserons en priorité sur les travaux qui ont été spécifiquement développés pour le domaine de la recherche d'information. Nous présenterons d'abord les grands modèles qui visent à conceptualiser et définir ce qu'est la recherche d'information. Puis nous affinerons la lecture de l'état de l'art en direction des activités de recherche d'information scientifique, particulièrement pour les sciences humaines et sociales. Enfin, nous préciserons ces modèles conceptuels par l'analyse de leur application aux systèmes numériques de recherche d'information.

4.2.1 Modèles conceptuels de recherche d'information

La recherche a produit de nombreux travaux en matière d'étude du comportement dans un contexte de recherche d'information. Ces études, débutées dans les années 1980 et 1990, ont abouti à la conception de nombreux modèles s'intéressant à des parts plus ou moins larges de la population et à des contextes de recherche d'information eux aussi très divers. Sur la base de ces modèles originaux se sont construits des modèles plus fins intégrants des travaux et des réflexions issus de disciplines extérieures aux sciences de l'information et à l'informatique. Ces deux disciplines sont en effet à l'origine des principaux fondements de l'état de l'art de la modélisation des processus de recherche d'information.

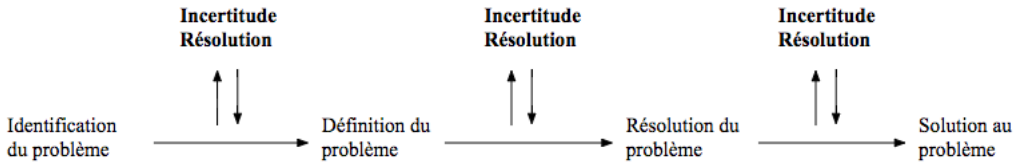
Dans le champs des sciences de l'information, les modèles souvent qualifiés de modèles d'*information behavior* (IB) se focalisent sur la manière dont nous cherchons et utilisons l'information au sens large. Ces modèles intègrent de nombreux facteurs humains mais également sociaux ou environnementaux. Ils ont en effet pour objectif de définir ce que sont des besoins ou des recherches d'information, dans toute leur complexité. Les premières briques de réflexion sur ces sujets ont ainsi été posées par Wilson [1981]. À la suite de ces travaux, de nombreuses approches ont été proposées visant à mieux comprendre comment qualifier le besoin d'information et fixer les cadres d'une analyse du comportement de recherche d'information. Ce dernier, en effet, ne peut être caractérisé comme l'écrivait Wilson en 1999, par la seule approche quantitative, parce que compter « le nombre d'entrées dans une bibliothèque, le nombre d'abonnements à un

1. À titre d'exemple, l'une des conférences majeures de ce domaine de recherche est la conférence internationale *User Modelling and Application* (UMAP) qui se tient tous les deux ans.

journal ou le nombre d'éléments cités dans un article n'apportent pas de réelle valeur pour le développement d'une théorie et la compréhension de la pratique [2] ».

Les modèles développés depuis les années 1980 mobilisent, comme nous l'avons déjà dit, une grande diversité d'approches. Certains sont très généraux et offrent une vue globale de la problématique. Wilson, que nous avons déjà cité, a ainsi proposé une vision fondée sur un modèle en forme de problème et résolution (voir figure 4.1). Le problème est défini dans ce cas comme une incertitude et la résolution comme la levée de cette incertitude, grâce à l'acquisition d'information. Ce modèle est itératif et montre qu'entre l'identification du problème et sa solution, l'acquisition d'information est régulière et intervient à chaque étape du processus cognitif.

FIGURE 4.1 – Modèle « problème-résolution » de Wilson [WILSON, 1999]



Le modèle de Wilson est un modèle générique de haut niveau, d'autres sont plus spécifiques. L'un des modèles les plus importants de la littérature, a été développé par David Ellis à la fin des années 1980 et au début des années 1990 [ELLIS, 1989, 1993; ELLIS et collab., 1993]. Ce modèle propose une vision du comportement de recherche d'information comme un ensemble de comportements liés et interagissants entre eux dans le but d'atteindre un objectif. Il est en ce sens beaucoup plus précis que le modèle précédent.

Ce modèle a été conçu avec l'idée de réduire le fossé entre l'utilisateur et le système, en opérant un déplacement de focale des travaux de recherche et des systèmes de recherche d'information vers les utilisateurs [3]. Avec une méthode empirique [4], Ellis a ainsi constitué un modèle d'activités de recherche d'information, tout particulièrement focalisé sur la pratique des chercheurs en sciences humaines et sociales. Son modèle se présente sous la forme de six étapes potentiellement reliées entre elles. Une étape peut inspirer la suivante sans que cela soit nécessairement systématique. Les étapes du modèles de Ellis sont ainsi définies :

— *Starting :* cette étape correspond dans le modèle à toute activité qui initie le processus de recherche d'information, qu'il s'agisse, par exemple, d'un projet de recherche, d'une discussion avec un collègue, une question de cours, etc ;

— *Chaining :* cette étape décrit les activités qui consistent à suivre, de référence en référence, une chaîne de documents. Il peut s'agir de suivre une chaîne de citation, un lien hypertexte, etc ;

2. [WILSON, 1999].

3. Cette idée, décrite par [NORMAN et DRAPER, 1986] est devenue classique. Elle exprime l'espace entre le système et l'utilisateur par la métaphore d'un fossé que l'on doit chercher à réduire. L'utilisateur devant mieux comprendre le système, et le système l'utilisateur, pour que chacun puisse s'adapter au mieux au fonctionnement de l'autre.

4. La construction du modèle résulte de l'observation empirique des stratégies employées par 60 membres de l'Université de Sheffield lors de leur recherche d'information scientifique [ELLIS, 1987].

— *Browsing :* il s'agit ici d'activités qui consistent à consulter les ressources dont on a déjà estimé le potentiel pour son besoin d'information, analyser une bibliographie thématique par exemple ;

— *Differentiating :* cette étape est dédiée à l'activité d'évaluation des différences entre les divers contenus pour en mesurer la pertinence ou la qualité ;

— *Monitoring :* à l'aide d'une base de connaissance existante, l'activité de *monitoring* vise à suivre les développements d'une thématique, d'un champ de recherche, etc ;

— *Extracting :* cette activité consiste à dépouiller des sources d'information de manière fine et systématique.

Si le modèle de Ellis est particulièrement important, c'est qu'il a donné lieu à de nombreuses études à sa suite, comme l'a très justement mis en lumière Savolainen [2016]. D'autres approches existent néanmoins. Le modèle développé par Kuhlthau [1988], explique ainsi les processus de recherche d'information comme des processus complexes, mêlant cognitif, physique et affectif et permettant de mettre en lumière, par exemple, le rôle de l'incertitude dans l'abandon de certaines recherches d'information.

Les travaux que nous avons brièvement évoqués ont ainsi permis, au début des années 2000 ; de définir plus précisément les différentes strates du processus de recherche d'information. Ils ont été définis et s'entendent, à plusieurs échelles, de la manière suivante [Wilson, 2000] :

Information Behavior : cette expression regroupe la totalité du comportement humain face à l'information, que la confrontation à l'information soit intentionnelle, comme c'est le cas en situation de recherche d'information, ou fortuite, comme c'est le cas dans une discussion du quotidien par exemple.

Information Seeking Behavior : l'expression *information seeking behavior* désigne au contraire un accès intentionnel à l'information, comme une conséquence d'un objectif à remplir, d'un besoin d'information à satisfaire. Il inclut l'accès à l'information physique, en bibliothèque par exemple, comme numérique, via des moteurs de recherche.

Information Searching Behavior : l'*information searching behavior* est plus précis que le concept précédent et opère à un niveau de granularité plus fin. Il désigne le fait de chercher l'information avec une stratégie, qui se traduit dans un environnement numérique par la saisie d'une requête complexe ou par un clic sur un résultat de recherche. Il se traduit également, sur le plan intellectuel, par la succession d'opérations mentales visant à distinguer et sélectionner l'information pertinente et utile de celle qui ne l'est pas.

Information Use Behavior : enfin, l'expression *information use behavior* désigne l'acte intellectuel ou physique d'intégrer l'information préalablement jugée pertinente dans son propre système d'information, comme peuvent l'être le fait de prendre une note sur un document particulier ou citer un document dans son travail.

Nous pouvons noter ici une difficulté de traduction de ces expressions en français, en particulier pour exploiter la différence ténue qui existe entre *information seeking* et *information searching*. Les deux expressions peuvent se traduire par « chercher » ou

« rechercher » de l'information. Nous adopterons ici le terme générique « recherche d'information », mais à des fins de clarté nous adoptons les expressions anglaises lorsqu'une distinction précise s'impose. À ce stade, l'état de l'art nous permet de situer nos objectifs de recherche dans les strates de l'*information seeking* et de l'*information searching* et nous donne une idée plus précise des activités qui constituent notre contexte d'usage. Cependant, les modèles dont nous parlons souffrent d'une limite majeure. Ils ont pour l'essentiel été créés, en particulier le modèle de Ellis, avant l'explosion du Web et n'ont donc pas intégré dans leur développement la dimension numérique.

Des travaux ont ainsi été menés dans les années 2000 et 2010 pour intégrer la dimension numérique au modèle de Ellis, tout en gardant une population étudiée similaire. Le travail de Meho et Tibbo [2003] a ainsi permis de confirmer, d'une part, la validité du modèle de Ellis dans un contexte intégrant plus largement le numérique et, d'autre part, de l'augmenter. L'étude a en effet permis de décrire de nouvelles caractéristiques de la recherche d'information qui sont pour l'essentiel liées à l'*Information Use Behavior*. Il s'agit des caractéristiques : *analyzing*, *synthesizing*, *writing*, *networking*, et *information managing*.

Le web a ainsi fait émerger des pratiques nouvelles, que Ellis n'avait pas pu mettre en évidence. La masse d'information désormais accessible par le numérique a par exemple fait émerger l'activité de gestion de l'information (*information managing*). La mise en évidence de l'activité de *networking*, qui consiste à maintenir une relation proche avec des personnes travaillant dans des domaines proches du sien a également était facilitée par le numérique.

Cependant, l'apport principal du modèle de Meho et Tibbo [2003] a été l'introduction d'un second niveau conceptuel au modèle de Ellis. Celui-ci était jusqu'alors limité à une liste de caractéristiques de la pratique de recherche d'information sans véritable relation d'ordre et de classement. Le modèle de Meho et Tibbo introduit un classement de ces différentes caractéristiques regroupées en quatre grandes étapes : *searching*, *accessing*, *processing* et *ending* :

— *searching* : correspond à la période de recherche d'information durant laquelle les ressources pertinentes sont identifiées par des moyens de recherche d'information traditionnels ou numériques ;

— *accessing* : cette étape est une étape intermédiaire entre l'activité précédente et l'étape de *processing* ;

— *processing* : cette étape est le moment de l'analyse et de la synthèse des information récoltées ;

— *ending* : cette étape marque la fin du processus.

Ce modèle n'est pas linéaire, des allers-retours sont possibles entre les différentes étapes. Par ailleurs, ces étapes peuvent contenir des caractéristiques communes. À titre d'exemple, les activités de *chaining* ou de *differentiating* du modèle de Ellis sont identifiées dans les deux premières étapes. À la suite de ces travaux, une étude similaire a été conduite en 2012 par Rhee [2012]. Cette étude, fondée là encore sur des enquêtes mais également sur une analyse très fine de texte disciplinaire et de syllabus de cours a porté uniquement sur une discipline, l'histoire [5].

5. Cette étude a également employé la méthode de l'open-coding, qui consiste à étiqueter des

Le modèle produit par Rhee [2012] introduit trois nouvelles caractéristiques spécifiques à la pratique de recherche d'information en histoire. Il s'agit de :

— *orienting* : cette caractéristique décrit une activité qui consiste à déterminer quels sont les dépôts de ressources potentiellement intéressants pour son objet d'étude, depuis les bibliothèques spécialisées jusqu'aux bases de données spécifiques ;

— *constructing contextual knowledge* : il s'agit ici de l'activité spécifique à la construction d'éléments de contexte, particulièrement importante en histoire, où le point de départ de la recherche peut être un évènement, un lieu, etc. et le champ de la recherche très large ;

— *assessing* : cette activité, elle aussi spécifique à l'histoire, concerne la vérification et la critique de l'information, qu'il s'agisse de vérifier un fait, d'authentifier une source, etc.

Dans ce modèle, la place des systèmes numériques de recherche d'information semble particulièrement importante, surtout pour les activités de construction d'éléments de contexte et de vérification de l'information. Le modèle de Rhee conserve par ailleurs les mêmes grandes étapes que le modèle précédent de Meho et Tibbo, comme on peut le voir à la figure 4.2. Enfin, il témoigne lui aussi des allers et retours possibles entre ces grandes étapes.

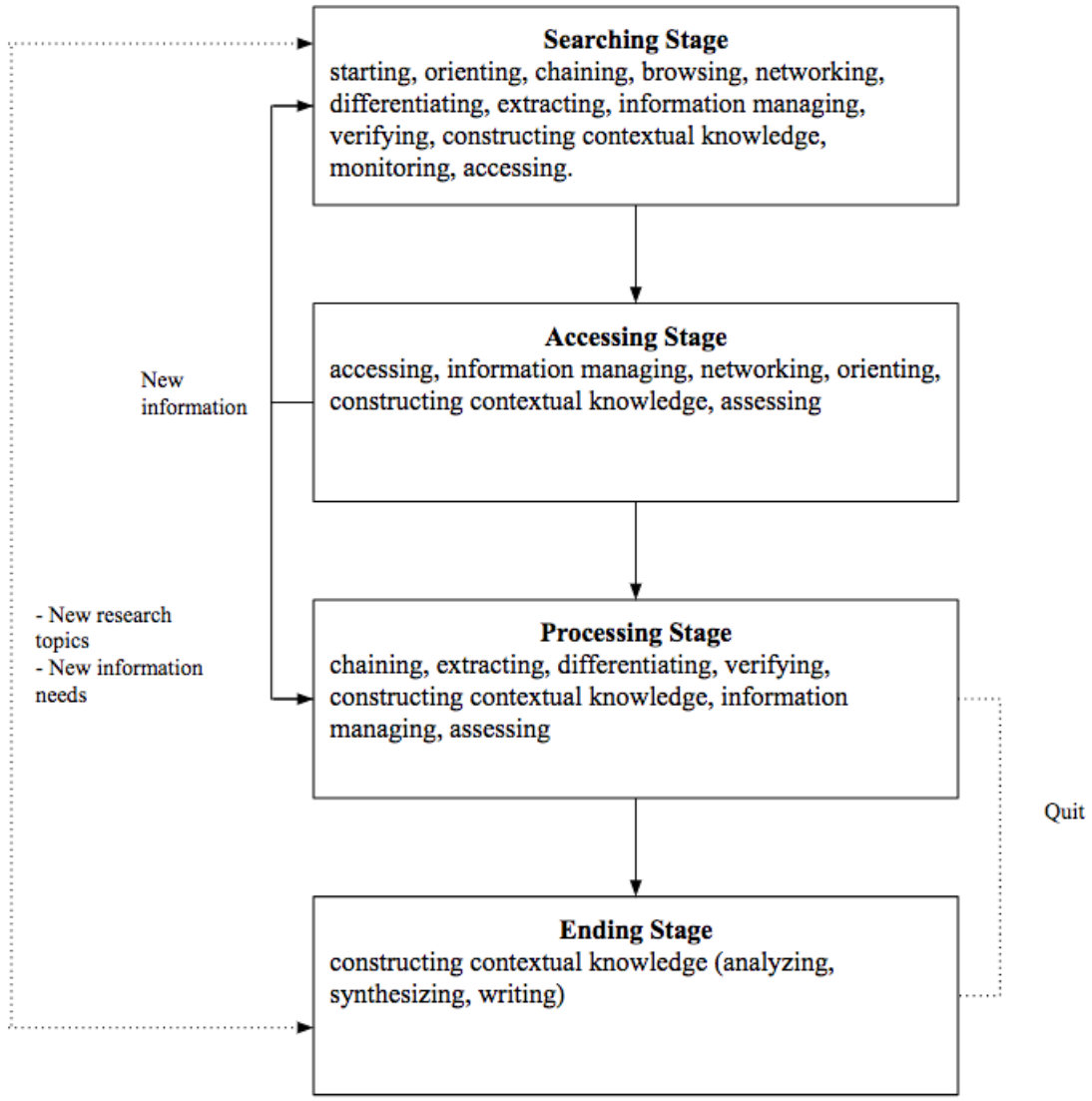
Ce modèle, qui intègre la dimension numérique et qui s'intéresse de près à l'histoire semble donc nous indiquer où se situent, dans le processus global de recherche d'information, les pratiques que nous souhaitons observer. Cependant, il ne nous permet pas d'identifier et de catégoriser plus finement les tâches qui peuvent être effectuées dans le système. Une caractéristique telle que *constructing contextual knowledge*, bien que clairement positionnée dans un processus plus vaste, reste imprécise. Certaines de ces caractéristiques peuvent se traduire dans le système par des recherches simples, s'il s'agit par exemple de vérifier un fait ou trouver une source. Elles peuvent également prendre la forme de séries de requêtes plus complexes, s'il s'agit cette fois d'acquérir des éléments de contexte.

Cette imprécision est symptomatique de la principale limite des modèles que nous venons d'évoquer. Ils sont difficiles à traduire par des indicateurs et des métriques issus de l'observation d'interactions entre utilisateurs et système numériques de recherche d'information. Cette difficulté explique probablement pourquoi les modèles conceptuels n'ont pas été immédiatement exploités pour aider à la conception et l'évaluation de systèmes de recherche d'information, bien qu'ils aient été à l'origine conçus pour cela [Colbert et collab., 1997].

C'est seulement à la fin des années 2000 que des travaux ont été menés en ce sens. Les modèles conceptuels ont ainsi été utilisés à des fins d'évaluation, voire de conception de système de recherche d'information. Makri et collab. [2008a] ont ainsi étudié de près le comportement de recherche d'information de juristes à l'aide du modèle conceptuel de Ellis. Ils ont ensuite appliqué leurs découvertes à l'évaluation de la qualité du système [Makri et collab., 2008b]. Ces tentatives d'application des modèles conceptuels ont montré clairement qu'ils ne se suffisent pas à eux mêmes pour évaluer la qualité d'un système de recherche d'information. Il est nécessaire de leur adjoindre d'autres modèles

passages de texte pertinents par catégorie, pour ensuite en établir des synthèses utiles à l'analyse.

FIGURE 4.2 – Modèle de Rhee de l'activité de RI en histoire [RHEE, 2012]



Le modèle reprend les étapes principales du modèle de [MEHO et TIBBO, 2003], nous pouvons noter les relations séquentielles et allers-retours possibles entre ces étapes.

et concepts, souvent spécifiques à une discipline, si l'on souhaite expliquer tout ou partie de la pratique des utilisateurs.

Dans notre contexte, nous éprouvons la même limite. Les modèles conceptuels, en particulier celui de [RHEE, 2012], nous enseignent que plusieurs étapes et plusieurs activités peuvent se traduire par des recherches d'information dans des bases de données ou des bibliothèques numériques. Pour autant, aucun des modèles que nous avons décrits ne permet de descendre à un niveau de détail plus fin. Aucun ne permet, par exemple, de discriminer ce qui est de l'ordre d'une recherche simple ou au contraire complexe. Rien ne nous permet, surtout, d'évaluer la pertinence de la stratégie employée par les utilisateurs pour résoudre un besoin d'information, au regard des autres contextes de la ressource.

Pour résoudre ce problème, les travaux de Marchionini nous ont semblé particulière-

Tableau 4.1 – Composants du modèle de tâche de Marchionini

Recherche simple (lookup)	Tâches exploratoires	
	Apprentissage	Investigation
- recherche de fait - recherche de ressource connue - navigation - vérification - réponse à une question	- acquisition de savoir - compréhension - comparaison - agrégation	- analyse - exclusion / négation - synthèse - évaluation - découverte - planification - transformation

ment pertinents. Ce modèle décrit dans Marchionini [1995] et Marchionini [2006] fournit une nomenclature des tâches de recherche d'information de bas niveau, qui s'appliquent aux étapes des modèles conceptuels que nous venons de voir. Les opérations intellectuelles qui aboutissent à interroger un système de recherche d'information (RI) sont classés en deux grandes catégories, présentées dans le tableau 4.1, les tâches dites *lookup* concernent des opérations de recherches simples, de la recherche d'un fait, jusqu'à la réponse à une question simple par exemple. La deuxième catégorie est dite « exploratoire » [White et collab., 2006]. Elle regroupe des tâches intellectuellement plus coûteuses, qui appellent des stratégies de recherche plus complexes. Cette catégorie est elle-même divisée en deux sous-catégories, d'égale importance. Ces tâches exploratoires sont par exemple constituées de recherche visant à acquérir un nouveau savoir, à synthétiser de l'information ou encore à découvrir un champ de recherche.

Le modèle de tâche de Marchionini est plus simple que les modèles précédents. Toutefois, de notre point de vue, il les complète plus qu'il ne les remplace, les tâches que ce modèle met en avant sont les composantes atomiques des différentes activités des modèles précédents. Les tâches « recherche de fait » ou « réponse à une question » sont par exemple l'expression, en termes opérationnels, d'activités de plus haut niveau comme la « construction du savoir contextuel » ou la « gestion d'information » du modèle de [Rhee, 2012]. Pour nous, qui souhaitons disposer d'indicateurs d'usage, ces tâches sont importantes parce qu'elles sont la formalisation conceptuelle qui est la plus proche de la réalité des besoins et de la pratique des utilisateurs dans le système.

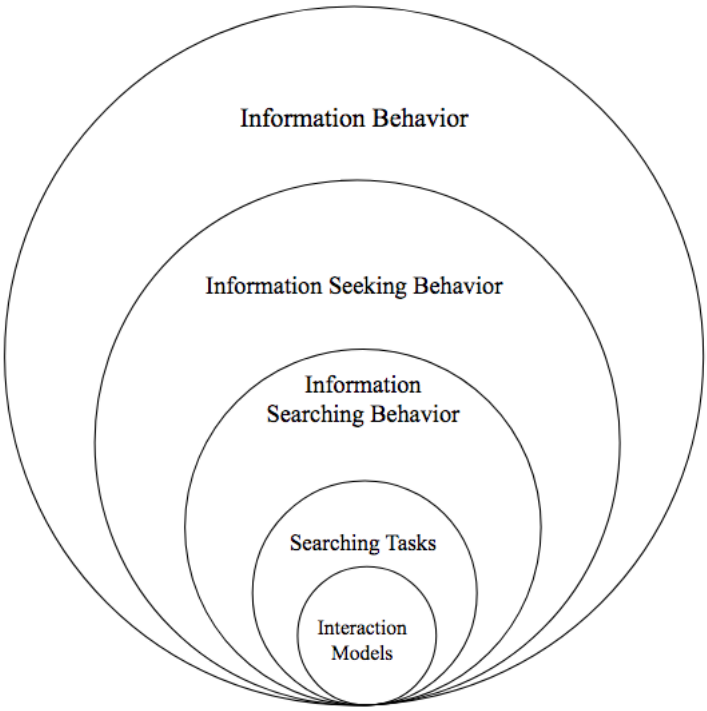
4.2.2 Modèles d'interactions

À tous les niveaux du processus de recherche d'information, de la recherche jusqu'au traitement de l'information, sont donc impliqués des tâches de recherche d'information simples, ou plus complexes. Nous faisons l'hypothèse que ces tâches sont caractérisables par des indicateurs issus de l'observation de la pratique des utilisateurs. Cette pratique ne peut s'observer que par l'intermédiaire des interactions qui ont lieu entre l'utilisateur et le système de recherche d'information.

Afin d'observer les comportements des utilisateurs d'un système de recherche d'information il faut aussi identifier les différents niveaux d'interactions possibles. Ces derniers constituent le niveau d'analyse le plus fin de notre modèle conceptuel (voir figure 4.3). La recherche scientifique, en particulier informatique, a par le passé produit des tra-

vaux ayant abouti à des modèles de ces niveaux d'interactions entre l'utilisateur et le système qu'il nous faut considérer.

FIGURE 4.3 – Niveaux du modèle conceptuel issu de l'état de l'art.



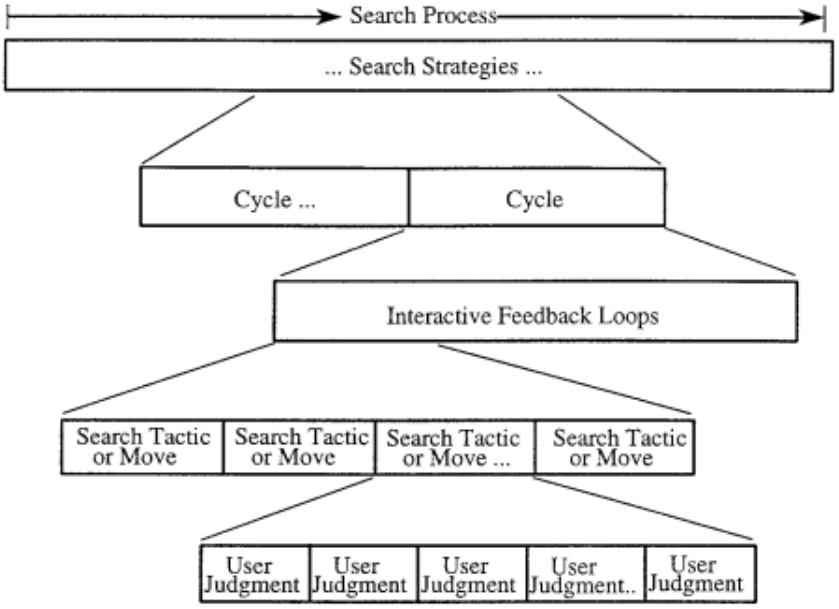
Aux niveaux conceptuels qui nous intéressent, il existe en effet plusieurs modèles d'interactions, ceux-ci, sans s'exclure mutuellement, sont complémentaires. Ainsi, le modèle défini par SPINK [1997] nous semble le plus générique. Ce modèle, décrit par la figure 4.4 présente un schéma général des interactions avec un système d'*information retrieval* (IR)[6]. Chaque recherche (*Search Strategy*) est constituée de plusieurs cycles, contenant ce que l'auteur nomme des *feedbacks loops* (boucles de retour d'information interactives), qui sont composées d'*inputs* utilisateurs et de réponse du système d'IR. Ces *feedbacks loops* sont donc issues d'interactions de l'utilisateur avec le système. Celles-ci varient et supposent des jugements sur la sortie fournie par le système amenant l'utilisateur à modifier éventuellement sa stratégie.

Un second modèle nous apparaît comme étant pertinent dans notre contexte, il s'agit du modèle en couches développé par SARACEVIC [1996], décrit à la figure 4.5. Ce modèle, qui semble se fonder en partie sur l'idée du fossé sémantique dont nous avons déjà parlé, est décrit en deux niveaux principaux, la partie utilisateur (cognitive), et les ressources informatiques. L'interface jouant ici un rôle de médiation entre les deux niveaux. Notre méthode d'observation doit ainsi permettre de collecter les traces de ces interactions à chacun des différents niveaux, de recueillir les *inputs* des utilisateurs d'une part et les retours du système d'IR d'autre part.

Pour observer le comportement des utilisateurs en situation de recherche d'information, plusieurs méthodes peuvent être mises en œuvre en fonction des objectifs de l'étude et

6. Traduction littérale de « Recherche d'Information (RI) » en français.

FIGURE 4.4 – Modèle de Spink des interactions en *information retrieval* [SPINK, 1997]

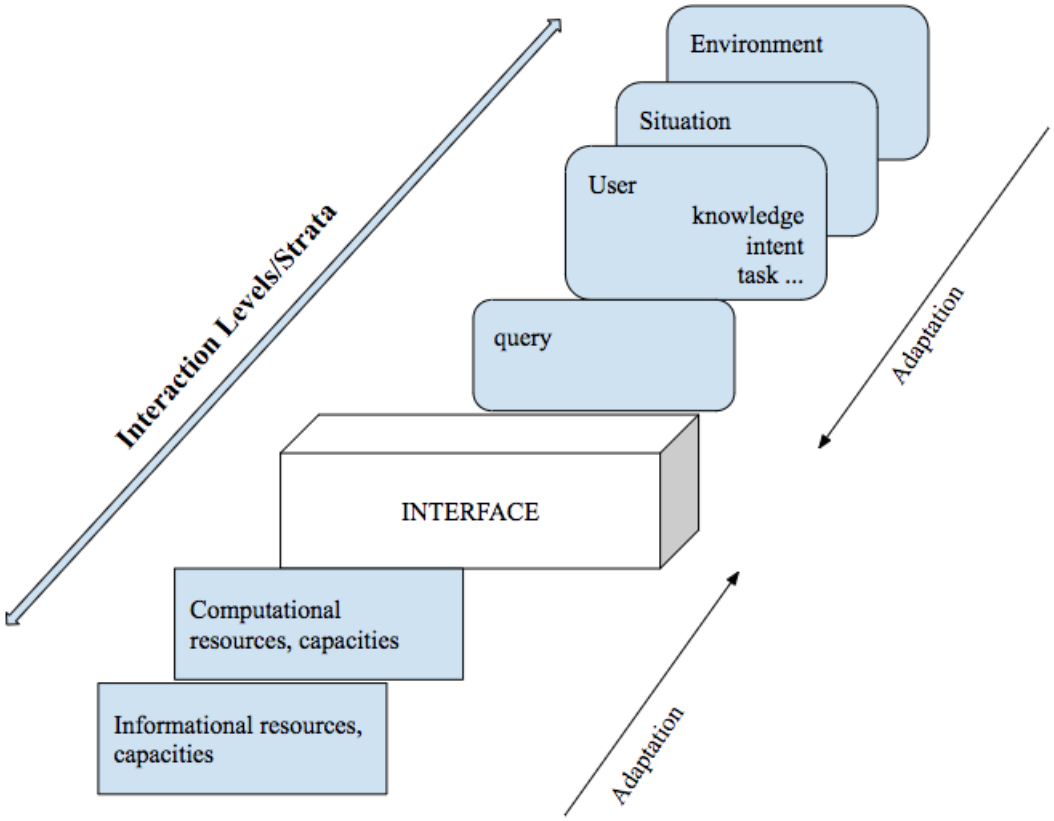


du niveau conceptuel que l'on souhaite adresser. La méthode la plus générale associe les techniques d'enquêtes quantitatives et qualitatives. Cette méthode, par exemple mise en œuvre par [BUCHANAN et collab., 2005] et [KEMMAN et collab., 2013] permet d'étudier les usages de haut niveau des bibliothèques numériques. Le volet quantitatif permet, entre autres, de mesurer les usages des différents outils disponibles, de comparer les moyens numériques par rapport aux méthodes traditionnelles de recherche d'information, en bibliothèques ou en archives. Le volet qualitatif a quant à lui vocation à permettre une étude visant la compréhension des outils dont l'usage est déclaré par les utilisateurs. La méthode ne permet toutefois pas de définir des métriques fines de l'usage de tel ou tel outil et reste cantonnée à des problématiques relativement larges.

Une autre méthode, qui autorise une analyse plus fine de l'usage d'outils de recherche particuliers est celle de l'observation directe et de la prise de notes, par exemple employée par MAKRI et collab. [2008a]. Cette méthodologie, cohérente avec le modèle de Spink, consiste à observer les actions d'un utilisateur utilisant un outil, en regardant tout simplement ce qu'il fait et en lui demandant éventuellement d'expliquer à haute voix ses actions. Cette méthode peut également être employée au moyen d'une analyse d'un enregistrement vidéo de l'écran de l'utilisateur observé. Il est nécessaire dans ce contexte d'établir une grille d'analyse précise et de consacrer beaucoup de temps à l'observation, au visionnage ou à la transcription. Tout comme la méthode précédente, elle est hors-ligne et ne peut pas s'appliquer à nos objectifs puisqu'elle ne permet pas d'observer automatiquement les interactions des deux niveaux du modèle de Saracevic.

Afin de pouvoir définir des métriques bas niveaux, une approche, en particulier mise en application par [ATHUKORALA et collab., 2015], [PARRA et collab., 2014] et [SAWADOGO et collab., 2015] consiste à concevoir une surcouche d'un moteur de recherche existant. Cette surcouche, constituée d'une interface graphique et d'un ensemble de fonctions d'observation du comportement des utilisateurs, se connecte à un moteur de

FIGURE 4.5 – Modèle d'interaction en couches de Saracevic [SARACEVIC, 1996]



recherche existant au moyen d'une API et transmet les requêtes et les réponses. Cette méthode permet d'utiliser des systèmes de recherche existants et connus des utilisateurs tout en autorisant une observation fine des interactions avec le système. Elle a par ailleurs l'avantage d'être relativement peu coûteuse en termes de développement informatique puisque le moteur de recherche lui même est déjà fonctionnel et qu'il ne reste qu'à développer l'interface d'observation. Cependant, cette méthode ne permet pas de maîtriser les paramètres du moteur de recherche et n'autorise pas d'avoir accès à l'observation du système d'IR du modèle de Saracevic.

4.2.3 Bilan de l'état de l'art

Cet état de l'art montre que les modèles d'interactions et les méthodes d'observation employés dans la littérature scientifique sont importantes dans notre contexte, parce qu'elles conditionnent pour beaucoup ce qu'il sera possible ou non d'observer. Les modèles conceptuels de recherche d'information, qu'ils se fondent sur une logique « problème-résolution » ou sur une logique de tâche sont adaptés pour décrire, avec plus ou moins de précision, les intentions des utilisateurs en situation de recherche d'information. Ils sont une source précieuse pour comprendre les enjeux à relativement haut niveau, à l'échelle d'une discipline par exemple. En revanche, ils n'offrent pas le niveau de détail suffisant pour évaluer l'impact qu'un outil de recherche, ou plus subtilement encore son paramétrage, peut avoir sur une activité de recherche d'information.

Les modèles d'interactions que nous venons de décrire sont beaucoup plus adaptés à cet objectif. Ils offrent des repères pertinents pour l'observation des interactions entre l'utilisateur et le système. Pris isolément, ils font toutefois abstraction de l'environnement et du contexte d'usage. Pour être utile à notre objectif et témoigner des variations des différents contextes de la ressource numérique, nous avons choisi de les associer avec les modèles conceptuels de plus haut niveau.

Cette nécessaire convergence conceptuel ne règle toutefois pas la question de la méthodologie la plus adaptée à l'observation des utilisateurs en situation de recherche d'information. Nous avons évoqué plusieurs stratégies, depuis les démarches d'enquête qualitative fondée sur des transcriptions des actions jusqu'à des systèmes plus automatisés utilisant une surcouche de moteur de recherche existant. À la lecture de l'état de l'art et au regard de nos objectifs, nous avons les contraintes suivantes :

— Observer le comportement d'utilisateurs exécutant des tâches de recherche d'information, positionnée conceptuellement, dans une bibliothèque numérique ;
— Disposer, dans le système, de documents cohérents avec ce positionnement conceptuel qui permettent l'accomplissement des tâches ;
— Produire des indicateurs du contexte d'usage pouvant être confrontés avec d'autres contextes du système (production ou exécution).

Si l'on tient compte de ces impératifs, la pleine maîtrise du système apparaît fondamentale. Nous avons donc choisi de construire un prototype à même de nous permettre de faire varier autant que nous le souhaitons les différents contextes. Ainsi, il est possible de choisir des corpus adaptés aux tâches évaluées sans dépendre de systèmes extérieurs. Ce prototype nous offre également la pleine maîtrise du contexte d'exécution, il nous est donc permis de faire varier tous les paramètres que nous avons à notre disposition et observer leur conséquences, ce qui n'aurait pas été possible si nous nous étions contenté d'utiliser une surcouche logicielle à une bibliothèque numérique existante.

4.3 Mécanique d'observation

Notre bibliothèque numérique a ainsi été enrichie de fonctionnalités d'observation prenant place du côté de l'utilisateur comme du côté du système. Nous avons choisi d'observer toutes les interactions entre l'utilisateur et le système. Ces observations brutes sont des traces de l'activité. Elles sont constituées des expressions du besoin d'information de l'utilisateur et de ces actions visant à traiter l'information d'une part et des retours et résultats produits par le système d'autre part. Sur la base de ces événements, nous avons calculé des indicateurs de plus haut niveau, dont les valeurs sont comparables d'une tâche à une autre.

4.3.1 Modèle d'observation et de collecte des traces

Les traces, résultantes de l'activité des utilisateurs dans la bibliothèque numérique, peuvent être collectées par différents moyens. Lors de nos premières implémentations et expériences de collecte de traces, nous avions recueilli les traces à travers les logs

générés par le serveur de l'application. Cette méthode est très simple en apparence, il s'agit de tirer parti des informations utilisées d'ordinaires pour analyser les performances du système informatique ou identifier les erreurs que rencontrent le programme. Dans ces données brutes, il est possible d'extraire de l'information de plus haut niveau. À titre d'exemple, lorsqu'un utilisateur clique sur un document de notre bibliothèque numérique, le serveur de l'application écrit une chaîne de caractères (voir tableau 4.2). En extrayant ces informations, nous pouvions recalculer certains indicateurs de la pratique des utilisateurs.

Tableau 4.2 – Exemple de logs générés par le serveur

Timestamp	**Sévérité**	**détails de l'évènement (Json)**
1452592827395	INFO	"type" :"resourceOcrView", "origin" :"http ://url-event-origin/resourceId

Néanmoins cette méthode d'extraction de modélisation des traces depuis les logs s'avère complexe dans certains cas. Les traces, collectées de cette manière, subissent un effet de décontextualisation qui rend difficile, voire impossible, le calcul de certains indicateurs. En pratique, il par exemple complexe de retrouver et d'associer toutes les traces relevant d'une seule et même recherche d'information, surtout quand le système est utilisé par de nombreux utilisateurs en parallèle. Cet effet de décontextualisation, associé au coût d'ingénierie nécessaire pour maintenir le programme de modélisation des traces, nous a poussé à adopter une stratégie de collecte différente.

Afin de maintenir le contexte des traces que nous collectons, nous avons choisi de les agréger par session utilisateur. Une session débute par la connexion de l'utilisateur à la bibliothèque numérique et se termine lorsqu'il quitte celle-ci (voir figure 4.6). Durant cette session l'utilisateur effectue un ensemble de recherches. Une recherche (voir figure 4.7) est une situation qui débute par l'expression d'un besoin d'information sous la forme d'une requête saisie par l'utilisateur dans la barre de recherche ou bien par la sélection d'une ou plusieurs facettes. L'expression du besoin peut naturellement être une combinaison de ces deux paramètres. La fin d'une situation de recherche est déterminée par le début d'une autre recherche ou la fin de la session. Pendant une recherche, les traces que nous collectons sont représentées par des événements. Ces événements sont issus des différentes interactions entre l'utilisateur et le système durant l'exploitation des résultats d'une recherche. Ce modèle permet aisément de produire des indicateurs reposant sur une session, donc un ensemble de recherche ou sur une recherche particulière. Notre modèle de session s'apparente ainsi aux cycles du modèle de Spink 4.4 et la recherche à ses boucles interactives (*feedbacks loops*). Ce modèle permet par ailleurs d'observer les interactions, par l'entremise des événements, à tous les niveaux du modèle de Saracevic 4.5.

FIGURE 4.6 – Schéma d'une session

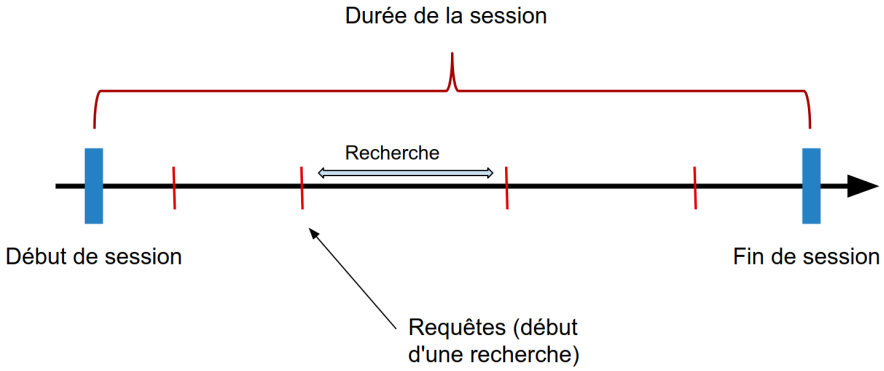
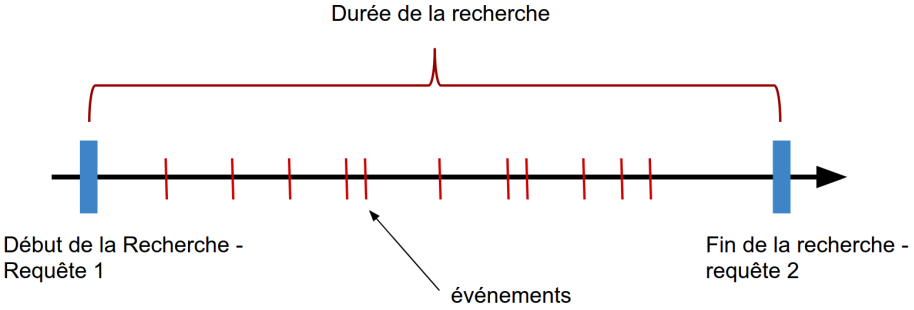


FIGURE 4.7 – Schéma d'une recherche



4.3.2 Événements tracés et collectés

Les événements sont les informations de plus bas niveau qui sont observées dans notre bibliothèque numérique. Comme nous l'avons expliqué, ils sont toujours contextualisés et prennent ainsi place dans une session et une situation identifiée. Ces événements sont recueillis au moyen d'observateurs logiciels placés dans le code de l'application qui ont la responsabilité d'enregistrer la trace d'un événement lorsque celui-ci se produit. La majeure partie de ces observateurs se trouvent du côté serveur de notre application.

En effet l'essentiel des interactions entre l'utilisateur et l'application produisent un appel au serveur du système qui est alors en capacité d'enregistrer la trace de cette interaction. Cependant, certains des événements que nous souhaitions observer n'occasionnent pas de tels appels et doivent donc être observés par le client de l'application, donc dans le navigateur web de l'utilisateur. Ce sont majoritairement des événements liés à l'interface homme machine. À titre d'exemple un *scroll*, c'est-à-dire le fait de descendre dans une page, n'est observable que du côté client de l'application. Pour ce type d'interactions, nous avons donc développé des observateurs logiciels côté client qui capturent les événements et les transmettent au serveur qui peut ainsi les enregistrer.

Ces événements variés peuvent être classés en trois groupes distincts desquels découlent des indicateurs spécifiques. Nous avons déterminé trois grandes catégories afin de faciliter l'analyse : l'expression du besoin d'information, l'exploitation des résultats de recherche et l'exploitation des ressources. La description complète de ces événements

et leur mécanique de collecte est présenté en annexe A.3.

Traces d'expression du besoin d'information

Cette catégorie regroupe les traces générées par l'utilisateur lorsqu'il exprime son besoin d'information dans notre application.

— Requête : expression du besoin de l'utilisateur, saisie sous forme de texte ;
— Filtrage : filtres sélectionnés par l'utilisateur (langue, type de contenu, *etc.*).

Traces d'exploitation des résultats de recherche

Dans cette catégorie sont présentés les événements liés à l'exploitation des résultats d'une recherche d'information.

— Résultats de recherche : liste des résultats produits par le moteur de recherche lors d'une expression du besoin d'information, requête ou filtrage ;
— Documents visibles : liste des résultats visibles à l'écran, c'est-à-dire qui ont été véritablement affichés et donc potentiellement vus par l'utilisateur ;
— Documents sélectionnés : document sur lequel l'utilisateur a cliqué et dont le détail des métadonnées a été vu ;
— Documents sélectionnés par recommandation : document dont le détail a été vu au moyen du mécanisme de recommandation ;
— Recherches sauvegardées : événement émis lorsque l'utilisateur choisi d'enregistrer les résultats d'une recherche.

Traces d'exploitation des documents

La dernière catégorie d'événements est celle qui est liée à l'exploitation des documents eux-mêmes, essentiellement à leur consultation.

— Documents mis en favoris : événement émis lorsqu'un document est mis en favoris depuis la liste des résultats de recherche ou depuis les pages de consultation de la description ou du contenu ;
— Documents consultés en version originale (*original*) : détail de la consultation d'un document dans sa version originale ;
— Documents consultés en plein texte : lorsque le document est consulté en plein texte, si cette option est disponible pour ce contenu, cet événement est émis par le système ;
— Documents téléchargés : événement qui enregistre le téléchargement d'un document par l'utilisateur, il est émis lorsque l'utilisateur clique sur le bouton de téléchargement.

4.3.3 Indicateurs calculés sur la base des événements

Les événements que nous venons de décrire n'ont à eux seuls que peu de potentiel. Ils sont les traces brutes de l'interaction entre l'utilisateur et le système. Sur cette base, nous avons choisi de définir et calculer des valeurs de plus haut niveau sémantique. Toujours dans l'hypothèse où ces indicateurs traduiraient une partie de l'usage du système, ils sont établis pour chaque session afin de les comparer entre elles.

Indicateurs de l'expression du besoin d'information

Les indicateurs de l'expression du besoin sont essentiellement fondés sur les événements « requête » et « filtre ». Nous faisons l'hypothèse que ces indicateurs et leurs métriques permettent de mesurer les différences entre les différentes expressions du besoin d'information et que ces différences expriment une évolution ou une rupture de l'activité de l'utilisateur dans l'application.

— Longueur des requêtes : L'indicateur « longueur des requêtes » est extrait des traces « requêtes » et « filtres ». Plus il est élevé, plus il indique une requête complexe, impliquant beaucoup de mots ou de filtres ;

— Durée de la session : Cet indicateur présente le temps passé par l'utilisateur dans le système durant une session (voir figure 4.6).

Indicateurs de l'exploitation des résultats de recherche

Les indicateurs d'exploitation des résultats de recherche ne concernent que les événements et les traces générés lorsque l'utilisateur consulte la liste des résultats de recherche ou les métadonnées détaillées d'une ressource. Notre hypothèse est qu'ils peuvent être intéressants pour mesurer l'intérêt suscité par une recherche d'information pour l'activité en cours de l'utilisateur et donc indiquer le type de cette activité.

— Nombre de documents visibles : le nombre de documents visibles est une somme calculée sur la base des traces « ressources visibles ». Cette somme peut être calculée pour une session comme pour une recherche d'informations. Elle donne une indication de l'intérêt d'un utilisateur pour les résultats de ses recherches. Une valeur importante peut indiquer que l'utilisateur a porté beaucoup d'intérêt pour une recherche spécifique. L'observation de l'évolution de cette valeur permet de mesurer l'intérêt d'une recherche particulière au regard de l'ensemble de la session ;

— Nombre de documents sélectionnés : cet indicateur est la somme des ressources sur lesquelles a cliqué l'utilisateur. Cet indicateur est généré sur la base des traces « ressources vues » et « ressources vues par recommandation ». Il donne également une indication de la pertinence d'une recherche d'information pour l'activité en cours de l'utilisateur ;

— Position des documents sélectionnés : sur la base des traces « documents sélectionnés », nous pouvons déterminer la position de la ressource dans la liste de résultats qui a conduit à sa consultation. Cet indicateur complète les valeurs des

ressources visibles et vues pour la mesure de l'intérêt d'une recherche au regard de l'activité de l'utilisateur.

Indicateurs de l'exploitation des ressources

La dernière catégorie que nous avons définie pour les indicateurs est naturellement celle qui concerne l'exploitation des ressources numériques. L'hypothèse est ici que ces métriques peuvent permettre de mesurer l'intérêt d'une ressource pour l'activité en cours d'un utilisateur.

— Nombre de documents consultés : parmi les indicateurs pertinents de l'activité d'exploitation des ressources, le nombre de ressources consultées est important. Il est calculé a partir des traces de consultation « ressources consultées en version originale » et « ressources consultées en plein texte ».
— Durée d'exploitation des résultats : Cette durée, calculée en seconde, est établie en cumulant le temps passé par l'utilisateur durant une situation d'exploitation des résultats de recherche.

4.4 Evaluation de la pertinence des indicateurs : démarche expérimentale

Les indicateurs que nous venons de décrire sont produits sur la base de traces d'usage. La seule manière de vérifier leur pertinence, c'est-à-dire leur capacité à décrire le contexte d'usage est de vérifier qu'ils sont modifiés significativement si ce contexte change, autrement dit si l'utilisateur change d'activité de recherche d'information. Ce type d'évaluation peut peut s'avérer être une tâche complexe. Si l'on sait par la littérature que certains indicateurs sont très révélateurs de comportements déjà bien décrits, l'évaluation de la pertinence de chacun des indicateurs une problématique qui ne peut être étudiée que par l'expérimentation [Athukorala et collab., 2015; Marchionini, 2006; Walkowski, 2017].

Le changement du contexte peut, si l'on suit l'état de l'art, être opéré en changeant de catégorie du modèle de tâche. L'objectif de la démarche empirique est alors de vérifier si les indicateurs de la pratique que nous avons décrits sont discriminants d'une catégorie de tâche à une autre, et dans quelle proportion.

4.4.1 Conditions expérimentales

Afin de mesurer la pertinence des indicateurs définis plus haut dans notre contexte, nous avons choisi de mener plusieurs expérimentations, avec des participants issus des formations en Sciences Humaines et plus précisément en histoire. Nous avons ainsi conduit des expérimentations avec un public de 36 étudiants de niveau Master sur une période de quelques semaines.

Nos expérimentations ont toutes été menées avec un protocole similaire. Les sessions d'expérimentation, d'une durée de 2h00, ont été organisées dans des salles équipées

en postes informatiques identiques, pourvus du même système d'exploitation et du même navigateur. Durant chaque session, les participants ont eu à effectuer différentes tâches de recherche d'information de catégories (*lookup* et *exploratory*) différentes. Leur comportement a été observé sur l'intégralité de ces tâches.

Les corpus expérimentaux ont été construits pour couvrir une diversité importante de type de documents allant du document iconographique aux articles de revue scientifique. La nature et la thématique générale de ces corpus a naturellement été adaptée au public que nous souhaitions observer. Les ressources constituant les corpus expérimentaux ont été récupérées depuis des bases de données en ligne libre de droits et disposaient de métadonnées complètes[7]. En fonction des divers contextes expérimentaux, nous avons indexé ces ressources dans notre plateforme en respectant des règles d'indexation adaptées à la nature et à la langue des documents.

4.4.2 Contraintes spécifiques

Comme toute démarche expérimentale impliquant des utilisateurs, de nombreux facteurs entrent en ligne de compte et impactent plus ou moins lourdement la pertinence de nos métriques. Le protocole expérimental doit ainsi tenir compte de tous ces facteurs si l'on souhaite valider la pertinence de notre approche. Nous avons ainsi tenté d'identifier les biais qui pouvaient s'avérer problématiques. La liste suivante décrit les facteurs que nous avons identifiés et dont les différentes expérimentations décrites tout au long de ce manuscrit ont dû tenir compte.

Facteurs environnementaux Parmi les facteurs qui ont une influence notable sur notre démarche expérimentale, le type d'environnement matériel et logiciel est important. Le type de machine utilisé pour accéder à l'application peut avoir un impact sur les interactions entre l'utilisateur et le système. Naturellement, nous pouvons immédiatement penser à la différence en matière d'interaction entre un ordinateur disposant d'un clavier et d'un dispositif de pointage et une tablette ou un smartphone tactile. Ce ne sont cependant pas les seules différences, la taille et la résolution de l'écran ont également un impact. Le type d'usage pour lesquels sont conçus les smartphones et tablettes, essentiellement utiles en situation de mobilité rendent plus complexes l'accès aux documents et leur pleine exploitation par les utilisateurs. Notre démarche visant à évaluer les comportements de recherche d'information dans le cadre d'une situation de travail, nous avons souhaité limiter l'expérience à l'usage d'un ordinateur. Nous avons, qui plus est, mené la majeure partie de nos expérimentations en environnement contrôlé dans lequel tous les utilisateurs disposaient du même matériel (dispositif de pointage et écran). Nous avons par ailleurs fait le choix d'encadrer les aspects logiciels, tous les utilisateurs disposant dans ce contexte du même système d'exploitation et du même navigateur web.

Maîtrise technique Outre les aspects matériels et logiciels, un des autres facteurs déterminants est le niveau de maîtrise technique des utilisateurs. Autrement dit, il

7. À titre d'exemple, les articles scientifique ainsi que des ressources graphiques ont été importés depuis la plateforme libre de droits HAL disponible à l'adresse https://hal.archives-ouvertes.fr/. D'autres ressources ont été récupérées depuis les plateformes Persée (http://www.persee.fr/) ou Europeana (http://www.europeana.eu).

s'agit du niveau de maîtrise de la pratique du web et de des outils informatiques dont disposent les utilisateurs. Ce niveau est variable et dépend de nombreux facteurs, mais a un impact non négligeable sur le comportement de recherche d'information [Saito et Miwa, 2001]. Il est bien entendu difficile de contrôler ce niveau de maîtrise, il est même souhaitable de pouvoir disposer en la matière d'une population d'utilisateurs relativement hétérogène. Cependant, un niveau de maîtrise minimal est nécessaire pour que les utilisateurs soient en capacité d'adapter leur comportement dans l'outil à la tâche qu'ils souhaitent accomplir. Lorsque cela s'est avéré possible, nous avons demandé aux utilisateurs utilisant nos outils de déclarer, au travers de quelques questions simples, leur niveau de maîtrise des outils informatiques. Néanmoins, pour certaines de nos expérimentations, avec des étudiants notamment, cela n'a pas toujours été possible. Pour autant, au regard des travaux de Saito et Miwa [2001], les utilisateurs sujets de nos expérimentations ont systématiquement pu être classés comme "*web experts*".

Connaissance du domaine Dans le même ordre d'idée que le point précédent, un des facteurs externes qui impacte notre démarche expérimentale réside dans le niveau de maîtrise du domaine ou du sujet des recherches effectuées. Naturellement, un utilisateur qui a une connaissance pointue d'une thématique de recherche aura une stratégie de recherche d'information différente, potentiellement plus efficace qu'un utilisateur qui connait moins, voire pas du tout le sujet. Ce niveau de connaissance du domaine influe sur de nombreux indicateurs numériques, en particulier les durées, mais également les observations de l'usage de certains outils, les fonctions de filtrage de requêtes et de résultats par exemple [Jenkins et collab., 2003; Liu et collab., 2016] [8]. Afin de contrôler ces biais expérimentaux, il est nécessaire de mener une réflexion sur les corpus de ressources numériques utilisés lors des phases d'expérimentations.

Facteurs linguistiques Notre application de bibliothèque numérique est capable d'indexer et d'afficher des ressources numériques dans de nombreuses langues. Les bibliothèques numériques les plus connues traitent d'une manière générale des ressources en plusieurs langues, bien qu'une langue domine souvent le corpus [9]. Cependant, il parait évident que l'usage de ces ressources dépend de la capacité des utilisateurs à les comprendre et les exploiter. Pour expérimenter les usages d'une bibliothèque numérique, il faut donc opérer un compromis entre diversité linguistique et accessibilité pour la population d'étude. Nous avons donc choisi de n'utiliser pour nos expérimentations que des ressources en langue française ou anglaise, langues largement maîtrisées parmi nos utilisateurs lors de nos expériences.

8. Par ailleurs, il est à souligner que la problématique de la connaissance du domaine n'est pas qu'un potentiel biais expérimental. Il peut-être une variable de l'utilisateur que l'on peut chercher à déterminer à des fins, par exemple, d'adaptation. Des modèles atteignant un niveau de réussite satisfaisant on ainsi été expérimentés [Liu et collab., 2016].

9. À l'exception d'Europeana, plateforme qui agrège des ressources à l'échelle européenne, qui dispose d'une large diversité linguistique, plus uniformément répartie que les plateformes nationales. Par ailleurs, il existe également des bibliothèques numériques, d'articles scientifiques en particulier, qui peuvent ne contenir que des documents en une seule langue, l'anglais le plus souvent. C'est par exemple le cas des plateformes IEEE Xplorer (https://ieeexplore.ieee.org/Xplore/home.jsp) ou ACM Digital Library (https://dl.acm.org/).

Tableau 4.3 – Tâches proposées pour l'expérimentation

Tâche	Catégorie	Caractéristiques impliquées	Détail
$T1$	lookup	recherche de fait	La tâche consiste à retrouver le détail d'un événement en ne disposant que d'une information parcellaire le décrivant.
$T2$	lookup	réponse à une question	La tâche consiste à répondre à une question simple en trouvant un des documents permettant d'y répondre
$T3$	lookup	recherche de ressource connue, vérification	La tâche consiste à retrouver un document précis en disposant de mots clés s'y référant
$T4$	exploratory	découverte, acquisition de savoirs, synthèse	La tâche consiste à élaborer une problématique de recherche cohérente avec les thématiques décelables dans la collection de ressources du corpus.

4.4.3 Tâches de recherche d'information

Afin d'expérimenter la pertinence des indicateurs qu'il est possible de calculer dans notre contexte expérimental et aux niveaux conceptuels qui nous intéressent, nous avons donné aux participants de l'expérience une série de tâche de recherche d'informations. Le détail de ces tâches est présenté dans le tableau 4.3. Nous avons choisi de donner des tâches de recherche simples (*lookup*) et une tâche de recherche exploratoire, effectué dans l'ordre de $T1$ à $T4$. Les tâches *lookup* étant simple et rapide à exécuter nous avons donner trois tâches de type légèrement différent au sens de Marchionini [Marchionini, 1995] mais d'égale difficulté, afin de pouvoir disposer de résultats homogènes et éviter des problèmes de démarrage à froid lors de l'expérimentation.

La liste des tâches est présentée dans une section dédiée de l'application, les participants disposaient pour chaque tâche d'un bouton permettant d'indiquer le début du processus de résolution de la tâche. Ils disposaient également d'un champ pour saisir les éléments de réponse à cette tâche et clore le processus. À l'issue de l'expérimentation, nous avons manuellement vérifié l'accomplissement de la tâche et la validité des réponses. Certaines réponses peu satisfaisantes ont été rejetées et les données des participants concernées exclues de l'évaluation des résultats.

4.5 Résultats et discussion

4.5.1 Méthode d'analyse des résultats

À l'issue des phases d'expérimentation, nous disposons d'une grande quantité de traces diverses. Chaque trace étant liée à un contexte, en l'occurrence une tâche, il est possible de comparer la valeur des indicateurs pour chacun d'entre eux. Autrement dit, il s'agit de déterminer si le type de tâche effectué sur la plateforme à un impact sur les indicateurs comportementaux que nous calculons. Pour ce faire, nous avons décidé

d'utiliser des tests statistiques dont l'objectif est précisément de mesurer si les différences entre deux jeux de données sont statistiquement significatives ou non. Ces tests, généralement développés pour le monde médical visent dans ce contexte à vérifier la probabilité que les différences entre deux séries d'analyses soient dues à des facteurs médicaux et non à une simple variation statistique.

L'objectif de ces tests est de vérifier une hypothèse, dite « hypothèse nulle » ou « hypothèse 0 », pour l'accepter ou la rejeter. Dans notre contexte, on peut par exemple tester l'hypothèse selon laquelle deux séries d'indicateurs de même nature, proviennent du même type d'activité. Le test permet de valider ou d'infirmer statistiquement cette hypothèse en produisant une probabilité de validité de l'hypothèse nulle. On considère alors qu'une probabilité inférieure à une certaine valeur, par exemple 0.05, dépendante du test utilisé, conduit à rejeter l'hypothèse nulle. Dans ce cas le rejet de l'hypothèse nulle nous amène à conclure que les séries d'indicateurs testés ne peuvent pas provenir du même type d'activité.

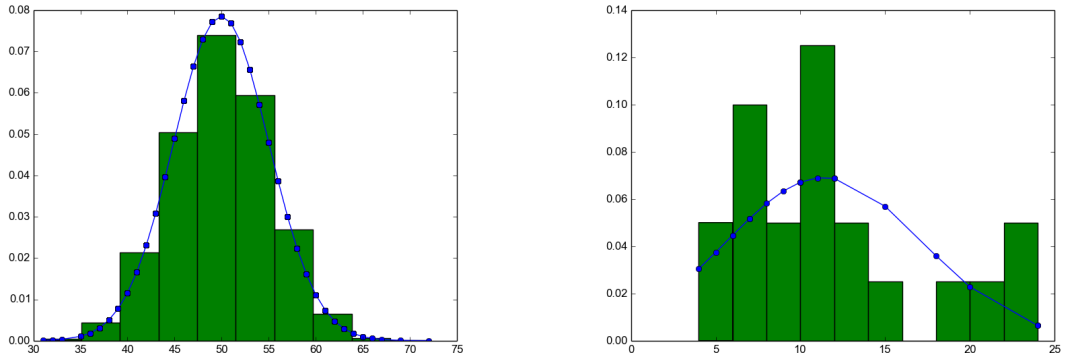
Sur cette base, si l'on compare deux séries d'indicateurs provenant de type d'activité que nous savons être différents et que l'hypothèse nulle « les deux séries de valeurs proviennent du même type d'activité » peut être rejetée, alors nous pouvons conclure que l'indicateur que nous évaluons est pertinent et qu'il permet de discriminer, du point de vue statistique, deux activités différentes. Si, au contraire, cette hypothèse nulle est validée, alors l'indicateur examiné n'est pas discriminant et ne permet pas de qualifier l'activité de l'utilisateur.

Les tests que nous avons choisis d'utiliser sont des tests majoritairement **non paramétriques**. Les tests non paramétriques sont à privilégier lorsque le modèle de données n'est pas soumis à de fortes contraintes, en particulier la normalité des distributions ou l'égalité des variances. Les données de nos indicateurs sont dans leur immense majorité dans ce cas de figure. Elles ne suivent pas une distribution normale et nous imposent des méthodes non paramétriques. Un exemple de distribution normale ainsi qu'un exemple de la distribution d'une série de l'indicateur *query length* sont visibles à la figure 4.8. Par ailleurs, du fait de la relative faiblesse de la taille de nos échantillons, il est difficile d'adopter une démarche paramétrique [10]. Nous ne postulons toutefois pas systématiquement que l'emploi d'un test non paramétrique s'impose pour l'intégralité des indicateurs que nous souhaitons évalué. Afin de vérifier la normalité de la distribution nous effectuons systématiquement un test dont l'objectif est précisément de vérifier la distribution des données. Les tests utilisés diffèrent selon le nombre de valeurs du jeu de données que l'on souhaite considérer. Pour les jeux de données contenant un grand nombre de valeurs (supérieures à 50), nous utilisons le test de normalité de D'Agostino et Pearson [D'Agostino, 1971; D'Agostino et Pearson, 1973]. Dans le cas où le jeu de données est moins important (inférieur à 50 valeurs) alors nous utilisons le test de Shapiro-Wilk [Shapiro et Wilk, 1965] dont des travaux récents ont montré l'efficacité pour des jeux de données de ce type [Razali et collab., 2011].

Ainsi, pour produire les résultats présentés dans les tableaux 4.6 et 4.7, nous avons utilisé le test des rangs signés de Wilcoxon (*Wilcoxon signed rank test*) décrit par [Wilcoxon, 1945] qui satisfait aux contraintes de nos jeux de données.

10. Il convient de préciser que l'emploi de tests non paramétriques peut constituer une limite dans la mesure ou les tests paramétriques sont généralement plus puissants que leur équivalent non paramétriques.

FIGURE 4.8 – Exemple de distribution normale et distribution d'un exemple de l'indicateur *Query length*



À gauche de la figure, un jeu de données suivant une distribution normale, à droite un jeu de données de l'indicateur *Query Length* ne suivant pas une telle distribution.

4.5.2 Évaluation de la pertinence des indicateurs

Les résultats expérimentaux moyens de l'ensemble des participants, pour chaque indicateur, calculés sur la base des traces décrites dans la section 4.3.2, sont présentés dans le tableau 4.4. Dans ce tableau, les résultats sont obtenus par l'observation de l'intégralité de la tâche, toutes les recherches faites par les utilisateurs du début de la tâche à sa résolution sont prises en compte.

Nous présentons par ailleurs, en parallèle des résultats sur l'ensemble de la tâche, les résultats sur la première recherche dans le tableau 4.5. Il est en effet établi par la littérature que la première recherche d'un utilisateur dans un moteur de recherche peut dans certain cas être très significative de l'activité qu'il est en train de mener [ATHUKORALA et collab., 2015] et donc être un critère de validation pertinent de la qualité d'un indicateur. Cette première recherche est par ailleurs intéressante à observer pour fournir des services spécifiques. Pour pouvoir, par exemple, adapter le système à l'activité de l'utilisateur, il faut en effet pouvoir disposer au plus tôt des variables permettant de discriminer l'activité et décider d'une logique d'adaptation.

Une simple lecture de ces tableaux permet déjà de mesurer les différences notables du comportement des utilisateurs que nous avons évalués entre les tâches $T1$, $T2$ et $T3$ de type *lookup* et la tâche $T4$ de type exploratoire (*exploratory*). Si l'on considère les résultats tenant compte de l'intégralité de la tâche, les valeurs pour $T4$ varient significativement des valeurs obtenues pour $T1$, $T2$ et $T3$. Pour le cas restreint de la première requête, là encore, nous pouvons observer des variations significatives sur l'essentiel des indicateurs. Nous pouvons toutefois noter que pour les indicateurs f'_4 et f'_6, il ne semble pas y avoir de variation significative.

Bien entendu, la seule lecture des valeurs moyennes brutes ne permet pas d'établir que les indicateurs comportementaux définis ici sont véritablement représentatifs de pratiques différentes, simulées dans notre cas par nos types de tâche. Nous avons donc soumis ces résultats brutes aux tests statistiques décrits dans la sous-section 4.5.1. Les tableaux 4.6 et 4.7 présentent les résultats de cette méthodologie pour la tâche

Tableau 4.4 – Résultats expérimentaux pour chaque indicateur, par tâche complète.

		simples				exploratoire
		$T1$	$T2$	$T3$	M	$T4$
longueur des requêtes	f_1	3.61	3.57	5.16	4.11	2.57
durée de la session	f_2	427.88	317.94	172.63	306.15	1464.27
docs. visibles	f_3	37.96	28.78	10.59	25.77	61.59
docs. sélectionnés	f_4	1.68	2.18	1.07	1.64	5.04
position des docs. sélectionnés	f_5	1.33	1.73	1.21	1.42	3.70
docs. consultés	f_6	1.21	2.07	1.07	1.45	5.20
durée d'exploitation	f_7	79.57	143.04	61.68	94.76	485.50

M est la moyenne des résultats des tâches simples (*lookup*).

Tableau 4.5 – Résultats expérimentaux pour chaque indicateur, pour la première recherche.

		simples				exploratoire
		$T1$	$T2$	$T3$	M	$T4$
longueur des requêtes	f'_1	4.72	3.72	5.32	4.59	2.88
durée de la session	f'_2	264.49	139.87	153.68	185.68	408.60
docs. visibles	f'_3	43.79	30.54	13.64	29.32	66.89
docs. sélectionnés	f'_4	0.89	1.04	0.93	0.95	1.14
position des docs. sélectionnés	f'_5	1.24	1.44	1.18	1.28	2.61
docs. consultés	f'_6	0.61	0.96	0.96	0.84	0.89
durée d'exploitation	f'_7	33.28	71.76	56.35	53.80	156.71

M est la moyenne des résultats des tâches simples (*lookup*).

complète et la première recherche.

Ces tableaux exposent donc une comparaison entre la tâche de type exploratoire ($T4$) et les tâches de recherches d'informations simples ($T1$,$T2$ et $T3$). Les résultats sont statistiquement significatifs pour des valeurs de p inférieures à 0,05 et sont dans ce cas matérialisées par des *. L'observation du tableau 4.6 concernant la tâche complète confirme nos impressions à la lecture des résultats brutes. Tous les indicateurs à l'exception de f_3 et f_7 pour $T2$ montrent de significatives différences dans les résultats. Sur la tâche complète, donc, nos indicateurs sont bien significatifs et varient entre différents types de pratiques de recherche d'information pour la population et le corpus considéré. Si l'on observe cette fois le tableau 4.7, nous pouvons constater que seuls les indicateurs f_1, f_2, f_3 et f_4 sont significatifs et peuvent être considérés comme des valeurs discriminantes d'une pratique de recherche. Les trois indicateurs restants ne sont, quant à eux, pas suffisamment représentatifs. Si notre objectif avait été de discriminer les types de tâche au plus tôt, nous n'aurions pas pû utiliser ces caractéristiques du comportement utilisateur. Individuellement, les résultats pour chacun des indicateurs peuvent ainsi s'interpréter de la manière suivante :

— ***Longueur des requêtes (f_1) :*** les résultats montrent que la longueur des requêtes est une caractéristique intéressante de la pratique des utilisateurs, aussi

TABLEAU 4.6 – Écart statistique entre une tâche exploratoire (T4) et trois tâches simples (*lookup*) (T1, T2 et T3) pour l'intégralité de la tâche.

	intégralité de la tâche (f)		
	T4		
	T1	T2	T3
longueur des requêtes (f_1)			
p	*	***	***
Z	-2.56	-4.05	-4.08
durée de la session (f_2)			
p	***	***	***
Z	-4.33	-4.55	-4.53
docs. visibles (f_3)			
p	**	0.18	*
Z	-2.64	-1.34	-2.09
docs. sélectionnés (f_4)			
p	**	***	***
Z	-3.21	-3.42	-4.2
position des docs. sélectionnés (f_5)			
p	***	***	***
Z	-3.39	-3.92	-3.51
docs. consultés(f_6)			
p	**	**	***
Z	-3.63	-3.21	-4.04
durée d'exploitation (f_7)			
p	***	0.28	***
Z	-3.39	-1.07	-4.6

Ce tableau présente les résultats comparant les deux types de tâches pour l'intégralité de la durée de la tâche. Le test utilisé est un test des rangs signés de Wilcoxon. Les valeurs p-values avec * sont statistiquement significatives avec * pour $p < 0.05$, ** pour $p < 0.01$ et *** pour $p < 0.001$.

bien sur la tâche complète que sur la première recherche. Les requêtes sont globalement plus longues et plus précises dans le cas des tâches de recherche d'information simple, où l'utilisateur à une idée claire de ce qu'il cherche.

— ***Durée de la session (f_2) :*** cet indicateur est significatif sur nos deux catégories de tâches. Les tâches dites exploratoires prennent nettement plus de temps que les tâches simples. Il faut noter que l'écart entre les deux types de tâches est également observable dès la première requête. Les utilisateurs prennent plus de temps à analyser les résultats et affiner leur recherche dans un contexte exploratoire, ces écarts sont visualisables sur la figure 4.9.

— ***Documents visibles (f_3) :*** cet indicateur est globalement significatif de la pratique, dans les deux catégories de tâches. Les utilisateurs consultent plus de résultats dans un contexte exploratoire, comme en témoigne également la figure 4.10

TABLEAU 4.7 – Écart statistique entre une tâche exploratoire (T4) et trois différentes tâches simples (*lookup*) (T1, T2 et T3) pour la première recherche.

	première recherche (f')		
	T4		
	T1	T2	T3
longueur des requêtes (f'_1)			
p	***	***	***
Z	-4.28	-3.33	-4.16
durée de la session (f'_2)			
p	0.1	***	***
Z	-1.61	-3.75	-3.78
docs. visibles (f'_3)			
p	0,06	**	**
Z	-1.87	-2.97	-3.74
docs. sélectionnés (f'_4)			
p	0.29	0.62	0.24
Z	-1.05	-0.49	-1.15
position des docs. sélectionnés (f'_5)			
p	***	***	0.05
Z	-2.25	-2.31	-1.96
docs. consultés (f'_6)			
p	0.09	0.65	0.62
Z	-1.66	-0.45	-0.49
durée d'exploitation (f'_7)			
p	*	0.13	0.16
Z	-2.46	-1.51	-1.4

Ce tableau présente les résultats comparant les deux types de tâches pour la première requête uniquement. Le test utilisé est un test des rangs signés de Wilcoxon. Les valeurs p-values avec * sont statistiquement significatives avec * pour $p < 0.05$, ** pour $p < 0.01$ et *** pour $p < 0.001$.

— ***Documents selectionnés*** (f_4) **:** comme nous pouvons intuitivement le percevoir, le nombre de ressources sélectionnées par l'utilisateur est un bon indicateur de la pratique si on l'observe suffisement longtemps. À l'aide de la seule première recherche, cet indicateur n'est pas statistiquement pertinent.

— ***Position des documents sélectionnés*** (f_5) **:** cet indicateur apparait comme étant très discriminant de la pratique, dans presque tous les cas, excepté la comparaison $T3/T4$ pour f'_5, où la valeur obtenue reste toutefois très proche du seuil de validation. Les utilisateurs sélectionnent presque toujours un des deux premiers résultats pour $T1$, $T2$ ou $T3$ alors qu'ils sélectionnent dans leur grande majorité des résultats entre les positions 1 à 5 pour $T4$, telle que le présente la figure 4.11.

— ***Documents consultés*** (f_6) **:** cet indicateur s'analyse de la même manière que f_4 et f'_4 et n'est significatif que si l'on considère une observation longue, de toute

la tâche.

— ***Durée d'exploitation (f_7) :*** cet indicateur, représentant le temps passé à consulter les documents, n'a par nature d'intérêt que pour toute la durée de la tâche. Il est, à l'exception de $T2/T4$ fortement significatif.

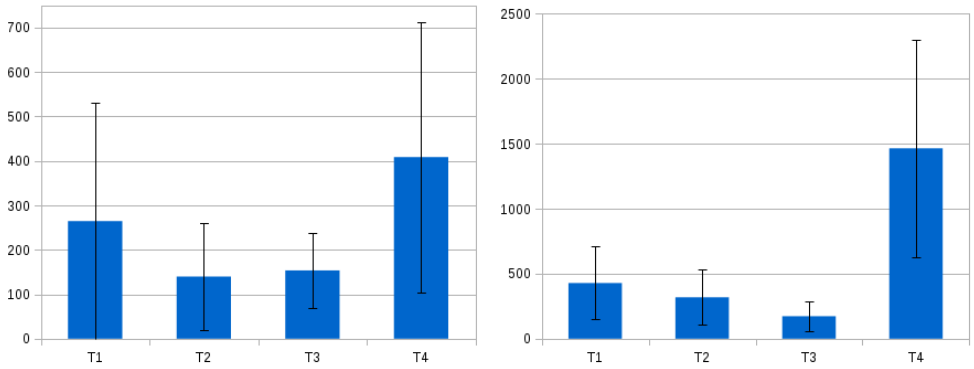
4.5.3 Principaux comportements mis en évidence

Si les indicateurs que nous venons de présenter sont bien significatifs du comportement de nos utilisateurs dans le cadre d'une recherche d'information, ils nous permettent également de tirer d'autres enseignements.

Phase d'apprentissage

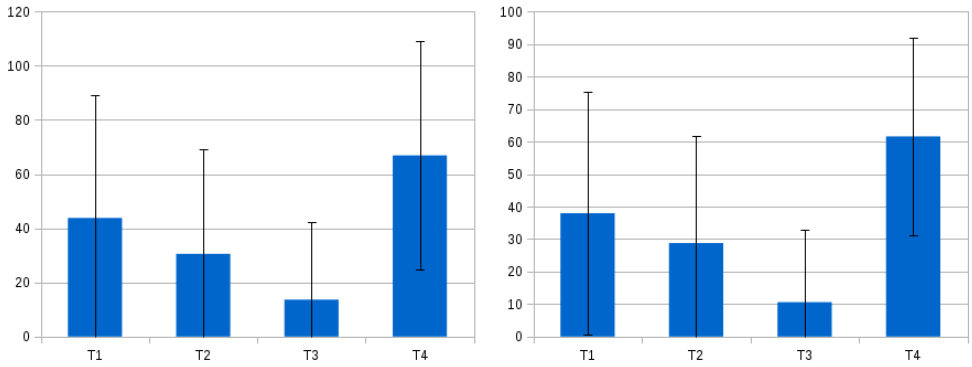
Sur la base de ces résultats, nous pouvons d'abord remarquer que pour trois tâches théoriquement très proches, les comportements des utilisateurs semblent varier si l'on s'en tient à nos indicateurs. Entre la tâche $T1$ et la tâche $T2$, pourtant toutes deux du type « recherche de faits », on observe quelques divergences. Il est assez difficile d'attribuer ces différences à un facteur de la tâche. Cependant, il est probable toutefois que certains de nos indicateurs soient soumis à des biais expérimentaux qu'il est important de noter. Ainsi, si l'on observe les résultats des indicateurs f_2 et f_3 sur nos trois tâches *lookup*, on constate que les résultats décroissent selon une pente relativement importante, visualisable par exemple sur les figures 4.9 et 4.10, où la première recherche est présentée à gauche et l'intégralité de la tâche à droite.

FIGURE 4.9 – Durée moyenne de la première recherche et de la tâche pour T_1, T_2, T_3 et T_4



Nous pensons que pour ces deux indicateurs l'explication de la variation significative de leur valeur est liée à une phase d'apprentissage. Les utilisateurs, bien qu'ayant bénéficié d'une courte introduction au fonctionnement de la plateforme ont pris du temps pour en découvrir le fonctionnement et les fonctionnalités. Le temps mis pour effectuer les tâches $T1$ et dans une moindre mesure $T2$ s'en ressent. Par ailleurs, la valeur moyenne de cet indicateur est trompeuse dans la mesure où l'écart type varie beaucoup, il est très large pour $T1$ et nettement plus réduit pour $T3$. Certains participants ont donc mis beaucoup de temps à effectuer la première tâche, probablement à cause d'un défaut de connaissance du système, alors que d'autres ont été plus rapides. En comparaison, la tâche $T3$ a été résolue nettement plus rapidement, par tous les utilisateurs. Il est

FIGURE 4.10 – Documents visibles pour T_1, T_2, T_3 et T_4 (résultats moyens).



probable que ceux qui ont pris le plus de temps pour découvrir la plateforme au début ont finalement rapidement trouvé leurs marques et ont été aussi efficaces que les utilisateurs plus aguerris.

Un second indicateur semble se comporter de la même manière, il s'agit du niveau de *scroll* (f_3) significativement plus important pour $T1$ et dans une moindre mesure pour $T2$ que pour $T3$. Là encore l'explication probable de la phase d'apprentissage peut s'appliquer. Les utilisateurs, découvrant la plateforme ont pu regarder plus de résultats dans l'optique de comprendre le fonctionnement de la plateforme. Si l'on compare les tâches $T1$ et $T2$ sur le plan statistique, les résultats présentés en annexe A.1 n'indiquent pas de différences majeures pour les indicateurs que nous venons d'évoquer. La distribution des indicateurs ne permet pas de discrimination statistique. Les utilisateurs moins expérimentés pour la tâche T1 le sont ainsi toujours pour la tâche T2, n'influençant pas significativement la distribution des résultats. Cependant, nous pouvons constater une différence dans le nombre de documents vus et la durée de consultation des documents, qui peuvent eux aussi être dûs, sans certitude toutefois, au niveau de maîtrise technique de la plateforme expérimentale. Ces effets de tâtonnement des utilisateurs ne sont pas à négliger dans un contexte expérimental.

Dans notre cas, bien que nous ayons identifié cette potentielle source de biais[11], il semble que la courte introduction au fonctionnement de la plateforme que nous avons mise en place avant nos expérimentations n'a pas suffit à la faire disparaître totalement. Ceci dit, il semble important de tenir compte de cet effet pour gérer les problèmes de démarrage à froid du système d'observation. Sur la base des indicateurs qui témoignent de cette phase de prise en main, il est sans doute possible d'exclure de l'observation des utilisateurs qui se trouveraient dans cette situation. Ces métriques pourraient par ailleurs être précieuses pour qui voudrait évaluer l'impact d'une démarche de formation sur une bibliothèque numérique ou détecter des utilisateurs peu aguerris, dans une logique d'adaptation.

Poids du classement des résultats par pertinence

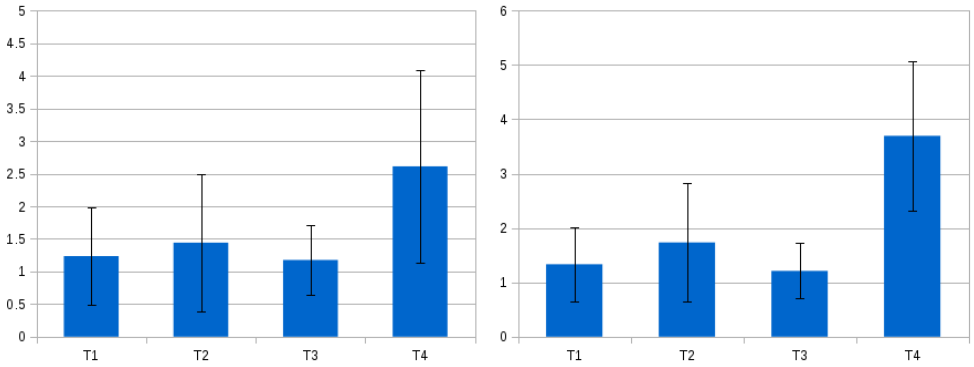
Malgré la phase de prise en main que nous venons de décrire, nous pouvons constater l'impact majeur de l'algorithme de classement par pertinence des résultats de recherche

11. Voir à ce sujet la section 4.4.2.

sur la pratique. Cet impact est visible à travers différents prisme. D'abord, le nombre de résultats consultés pour accomplir une tâche est faible. Si cela est plutôt logique dans une recherche d'information simple, dans un contexte exploratoire, cela devient plus problématique. Ces tâches ont en effet vocation à nécessiter une lecture transversale du corpus et une démarche de recherche relativement systématique, qui ne peuvent être menées à bien qu'en consultant un nombre important de ressources.

Pour notre tâche exploratoire $T4$, en effet, le nombre moyen de document consulté, c'est à dire ouvert en version originale ou texte intégral, est d'environ 5, ce qui est relativement faible au regard des consignes de la tâche et des documents que l'on peut raisonnablement considérer comme pertinent pour cette tâche. Pour autant, il faut noter que la tâche ne nécessitait pas nécessairement d'analyser finement un nombre important de documents, une lecture exhaustive des résultats de recherche pouvait être adaptée, au moins en début de tâche, pour affiner la problématique. Dans ce cas, si l'on ajoute aux documents véritablement ouvert, ceux donc les métadonnées complètes ont été consultés, alors les utilisateurs ont investigué 7,23 ressources en moyenne [12]. Néanmoins, on constate que l'indicateur de niveau de *scroll* est lui aussi relativement bas. Traduit en nombre de documents vus, cet indicateur nous permet de constater que les utilisateurs ont en effet affichés 7,2 résultats de la liste, en moyenne, par requête. Ce comportement qui consiste à limiter la consultation des résultats aux tous premiers résultats de la liste est visible aussi bien pour la tâche complète que pour la première requête. Ce comportement est par ailleurs partagé par une large majorité de la population étudiée, comme nous pouvons le constater par la faible amplitude de l'écart type à la figure 4.11.

FIGURE 4.11 – Positions moyennes des objets sélectionnés dans les résultats pour T_1, T_2, T_3 et T_4.



En sélectionnant peu de résultats par ailleurs souvent bien classés dans la liste, les utilisateurs observés ont accordé beaucoup de crédit à l'algorithme de pertinence dont on connaît depuis notre section 3.2.1 les potentiels biais. Notre plateforme expérimentale offre pourtant les moyens de réduire ces effets pour des tâches exploratoire. Les utilisateurs ont les moyens de réduire le nombre de résultats et préciser leur recherche, sans se limiter à ne considérer qu'un nombre faible de résultats dans la liste. Il est par exemple possible d'utiliser le filtrage par facette ou le classement par ordre alphabétique. En pratique, ces fonctions ont été très peu utilisées [13] augmentant de fait l'impact

12. Cette valeur est plus élevée que le nombre de documents cliqués dans la liste des résultats car elle inclut les documents consultés via les favoris ou les recherches sauvegardées.

13. Jamais pour le classement par ordre alphabétique et 33 fois pour le filtrage par facettes. À titre

de l'algorithme de pertinence sur la pratique de recherche d'information.

Une lecture fine des résultats nous a par ailleurs permis de voir émerger deux stratégies de recherche d'information pour l'accomplissement de la tâche $T4$. Ces stratégies ont pu être observées par les participants eux mêmes à l'issue de la séance sous la forme d'une interface matérialisant leurs différentes actions (voir figure A.3). La première stratégie employée a consisté à effectuer une requête, puis à consulter beaucoup de documents issus de la liste des résultats avant d'effectuer d'autres requêtes, en nombre limité. La seconde stratégie, quant à elle, a consisté en une requête initiale, suivi de la consultation de peu de documents, souvent un seul, puis de la saisie d'une nouvelle requête. Ces deux stratégies sont modélisées dans le schéma 4.12.

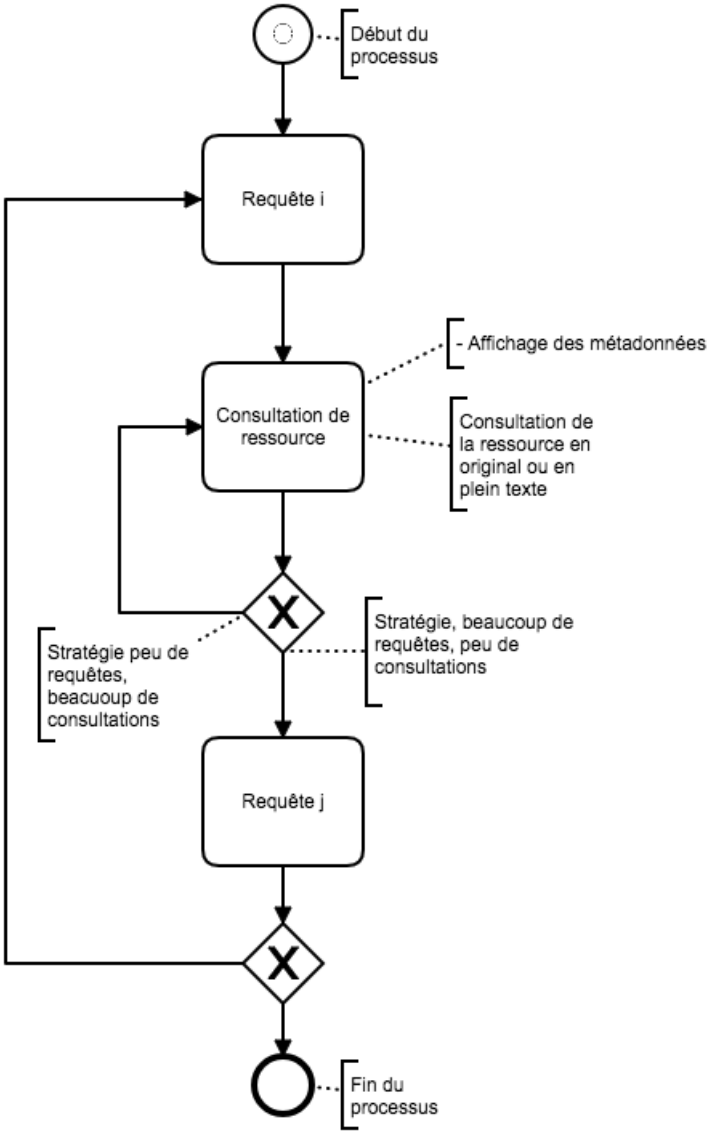
Bien que les participants semblent avoir employé deux stratégies relativement différentes, les résultats de la tâche sont apparus comme très homogène. Nous avons donc construit deux représentations graphiques afin de visualiser les documents utilisés par les participants pour résoudre la tâche[14]. La partie gauche de la figure 4.13 est un graphe représentant le réseau des documents utilisés par les participants pour résoudre la tâche. Les documents sont liés entre eux par un arc si au moins deux utilisateurs les ont utilisés, ils sont donc graphiquement plus proches si ils ont été utilisés par beaucoup d'utilisateurs. La partie droite représente le graphe des utilisateurs. Ceux-ci sont liés si ils ont consultés au moins un document en commun et sont donc graphiquement proches s'ils partagent beaucoup de ressources en commun.

Ces représentations montrent une pratique relativement homogène des participants. Ils ont été nombreux à avoir utilisé les mêmes ressources et donc à être fortement connectés dans ces graphes. Les utilisateurs à la périphérie du graphe sont ceux qui ont eu une pratique plus originale et une lecture plus transversale du corpus qui leur était proposé à l'image du participant mis en avant dans la figure 4.13 et les ressources qu'il a utilisées.

de comparaison 358 requêtes ont été effectuées dont 127 requêtes uniques dans le moteur de recherche. Les résultats complémentaires de cette tâche 4 sont accessibles dans le tableau annexe A.2.

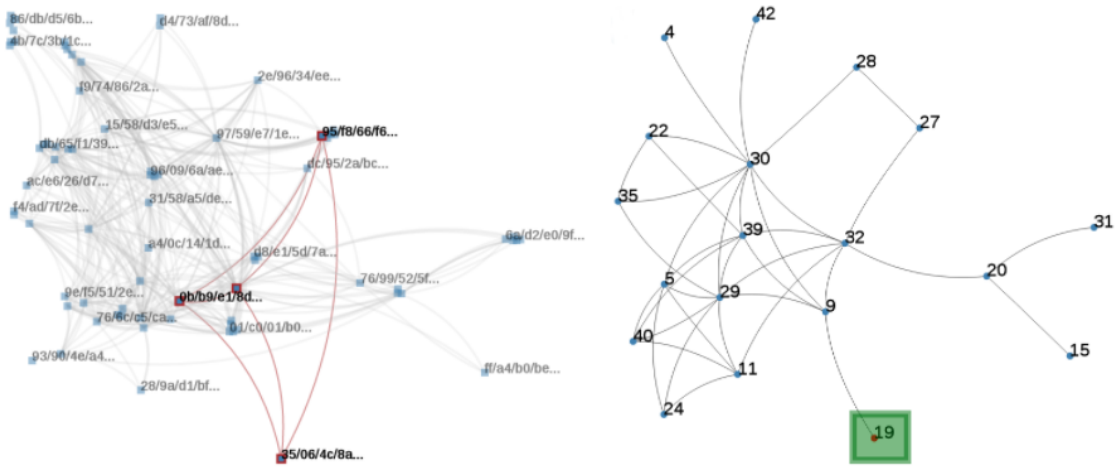
14. Ces représentations graphiques ont été construites à l'aide de l'outil *Detangler*, permettant de faire correspondre plusieurs représentations sous forme de graphes des mêmes données [Renoust et collab., 2015].

FIGURE 4.12 – Stratégies adoptées pour la tâche $T4$.



Modélisation BPMN des stratégies de recherche d'information pour la résolution de la tâche T4. À l'issue de la première requête et la première consultation de ressources, une partie des utilisateurs choisissent de saisir immédiatement une nouvelle requête (stratégie « beaucoup de requêtes, peu de consultations. »). À l'inverse, une autre partie choisit de de consulter d'autres ressources de la liste de résultats avant de saisir une nouvelle requête (stratégie « peu de requêtes, beaucoup de consultation »).

FIGURE 4.13 – Réseau des ressources utilisées par les participants pour la tâche $T4$.



À gauche de la figure le réseau des documents utilisés par les participants pour résoudre la tâche $T4$. À droite le réseau des utilisateurs. L'un d'entre eux est sélectionné pour mettre en lumière son utilisation des documents par rapport à la communauté des participants.

4.6 Conclusion

Dans ce chapitre, nous avions pour principal objectif de définir une méthode qui permette, depuis les interactions des utilisateurs dans la bibliothèque numérique, d'aboutir à des indicateurs de haut niveau à même de discriminer différentes pratiques de recherche d'information. Pour atteindre cet objectif, nous nous sommes fondés sur un état de l'art qui a montré la complexité, sur le plan conceptuel, des comportements de recherche d'information qui dépendent pour beaucoup du besoin d'information des utilisateurs et des objectifs qu'ils se fixent. Nous avons également eu l'occasion de constater les limites opérationnelles de ces modèles conceptuels qui restent difficiles à traduire sur le terrain par des valeurs observables. Les modèles d'interaction en recherche d'information que nous avons présentés nous ont permis de préciser cet aspect du problème en montrant la nécessité d'observer les comportements des utilisateurs en interaction forte avec le système utilisé. Nous avons donc pu conclure à la nécessité d'adopter une démarche expérimentale, fondée sur un prototype logiciel nous autorisant à recueillir des traces brutes de l'interaction du côté de l'utilisateur, comme du côté du système.

Par la suite, nous avons traduit ces traces brutes en indicateurs de haut niveau, comparables entre différentes tâches de recherche, suivant là le modèle conceptuel de [Marchionini, 2006], et différents utilisateurs. Pour vérifier l'hypothèse selon laquelle ces indicateurs traduisent différentes pratiques de recherche d'information, nous avons souhaité en évaluer la pertinence. Pour cela, nous avons mis en place une méthode d'évaluation statistique, permettant d'observer l'évolution de la valeur de ces indicateurs entre plusieurs types de tâches de recherche d'information. Cette évaluation est applicable en considérant l'intégralité de la tâche, comme une extraction de celle-ci, dans notre cas la première recherche.

Grâce à notre expérimentation impliquant plusieurs utilisateurs et plusieurs tâches, cette méthode d'évaluation a montré que les six indicateurs présentés sont statistiquement pertinents si l'on tient compte de l'intégralité de la tâche et que quatre d'entre eux le sont si l'on se limite à la première recherche ($f'_1, f'_2, f'_3 et f'_5$). Les variations de ces indicateurs témoignent certes qu'ils sont statistiquement significatifs pour discriminer plusieurs pratiques de recherche et différents besoins d'information, mais ils sont aussi utiles, indépendamment les uns des autres. Nous avons ainsi pu remarquer qu'ils pouvaient témoigner du niveau de connaissance technique de la plateforme, matérialisée par la courbe d'apprentissage mise en évidence.

Enfin, ces indicateurs sont aussi révélateurs d'une relative homogénéité des pratiques qui se traduit par une popularité de peu de documents, beaucoup consultés, au détriment des autres, très rarement utilisés. Ce constat s'explique par le fonctionnement de l'algorithme de pertinence, qui pour les requêtes générales effectuées par les participants en début de tâche, privilégiait une faible part du corpus disponible. Le poids de l'algorithme pris par le classement par pertinence pour une telle tâche exploratoire est problématique. Les participants n'ont que très peu cherché à s'extraire de ce mode de fonctionnement et à avoir une vision plus globale de la documentation disponible dans la plateforme.

Ici, ce sont les choix de conception qui s'expriment à travers le contexte d'exécution et une pratique inadaptée du côté de l'usage qui sont les principales sources de biais.

Il est donc nécessaire de réinvestir les indicateurs que nous avons produits pour permettre aux utilisateurs d'adapter leur pratique. Cependant, avant de traiter de cet aspect de notre problématique, objet de notre chapitre 6, nous prendrons d'abord en considération, dans notre chapitre suivant, le contexte de production. L'analyse de ce dernier est un préalable à la démarche de réinvestissement des résultats présentés ici que nous souhaitons être en capacité de mener. Elle requiert en effet de disposer de moyens permettant de révéler aux utilisateurs les conséquences sur leur pratique des biais induits par les processus transformant des documents physiques en leurs avatars numériques.

4.7 Références

Athukorala, K., D. Glowacka, G. Jacucci, A. Oulasvirta et J. Vreeken. 2015, «Is exploratory search different ? A comparison of information search behavior for exploratory and lookup tasks», *Journal of the Association for Information Science and Technology*. 100, 107, 112

Buchanan, G., S. J. Cunningham, A. Blandford, J. Rimmer et C. Warwick. 2005, «Information seeking by humanities scholars», dans *International Conference on Theory and Practice of Digital Libraries*, Springer Berlin Heidelberg, p. 218–229. 100

Colbert, M., C. Peltason, R. Fricke et M. Sanderson. 1997, «The application of process models of information seeking during conceptual design : the case of an intranet resource for the re-use of multimedia training material in the motor industry», dans *Proceedings of the 2nd conference on Designing interactive systems : processes, practices, methods, and techniques*, ACM, p. 73–81. 96

D'Agostino, R. B. 1971, «An omnibus test of normality for moderate and large size samples», *Biometrika*, p. 341–348. 111

D'Agostino, R. B. et E. S. Pearson. 1973, «Tests for departure from normality. Empirical results for the distributions of b2 and b1», *Biometrika*, vol. 60, n° 3, p. 613–622. 111

Ellis, D. 1987, *The derivation of a behavioural model for information retrieval system design.*, PhD Thesis, University of Sheffield. 93

Ellis, D. 1989, «A behavioural approach to information retrieval system design», *Journal of documentation*, vol. 45, n° 3, p. 171–212. 93

Ellis, D. 1993, «Modeling the information-seeking patterns of academic researchers : A grounded theory approach», *The Library Quarterly*, p. 469–486. 93

Ellis, D., D. Cox et K. Hall. 1993, «A comparison of the information seeking patterns of researchers in the physical and social sciences», *Journal of documentation*, vol. 49, n° 4, p. 356–369. 93

Jenkins, C., C. L. Corritore et S. Wiedenbeck. 2003, «Patterns of information seeking on the Web : A qualitative study of domain expertise and Web expertise», *IT & society*, vol. 1, n° 3, p. 64–89. 109

Kemman, M., M. Kleppe et S. Scagliola. 2013, «Just Google It - Digital Research Practices of Humanities Scholars», *arXiv :1309.2434 [cs]*. 100

Kuhlthau, C. C. 1988, «Developing a model of the library search process : Cognitive and affective aspects», *Rq*, p. 232–242. 94

Liu, J., C. Liu et N. J. Belkin. 2016, «Predicting information searchers' topic knowledge at different search stages», *J Assn Inf Sci Tec*, vol. 67, nº 11, doi :10.1002/asi.23606, p. 2652–2666, ISSN 2330-1643. 109

Makri, S., A. Blandford et A. L. Cox. 2008a, «Investigating the information-seeking behaviour of academic lawyers : From Ellis's model to design», *Information Processing & Management*, vol. 44, nº 2, p. 613–634. 96, 100

Makri, S., A. Blandford et A. L. Cox. 2008b, «Using information behaviors to evaluate the functionality and usability of electronic resources : From Ellis's model to evaluation», *Journal of the American Society for Information Science and Technology*, vol. 59, nº 14, doi :10.1002/asi.20927, p. 2244–2267. 96

Marchionini, G. 1995, *Information seeking in electronic environments*, Cambridge University Press. 98, 110

Marchionini, G. 2006, «Exploratory Search : From Finding to Understanding», *Commun. ACM*, vol. 49, nº 4, doi :10.1145/1121949.1121979, p. 41–46. 98, 107, 122

Meho, L. I. et H. R. Tibbo. 2003, «Modeling the information-seeking behavior of social scientists : Ellis's study revisited», *J. Am. Soc. Inf. Sci.*, vol. 54, nº 6, doi : 10.1002/asi.10244, p. 570–587. 95, 97

Norman, D. A. et S. W. Draper. 1986, *User centered system design ; new perspectives on human-computer interaction*, L. Erlbaum Associates Inc. 93

Parra, D., P. Brusilovsky et C. Trattner. 2014, «See what you want to see : visual user-driven approach for hybrid recommendation», dans *Proceedings of the 19th international conference on Intelligent User Interfaces*, ACM, p. 235–240. 100

Razali, N. M., Y. B. Wah et others. 2011, «Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests», *Journal of statistical modeling and analytics*, vol. 2, nº 1, p. 21–33. 111

Renoust, B., G. Melançon et T. Munzner. 2015, «Detangler : Visual analytics for multiplex networks», dans *Computer Graphics Forum*, vol. 34, Wiley Online Library, p. 321–330. 119

Rhee, H. L. 2012, «Modelling historians' information-seeking behaviour with an interdisciplinary and comparative approach.», *Information Research*, vol. 17, nº 4. xi, 95, 96, 97, 98

Saito, H. et K. Miwa. 2001, «A cognitive study of information seeking processes in the WWW : the effects of searcher's knowledge and experience», dans *Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on*, vol. 1, IEEE, p. 321–327. 109

Saracevic, T. 1996, «Relevance reconsidered», dans *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, p. 201–218. xi, 99, 101

Savolainen, R. 2016, «Contributions to conceptual growth : The elaboration of Ellis's model for information-seeking behavior», *Journal of the Association for Information Science and Technology*. 94

Sawadogo, D., C. Suire, R. Champagnat et P. Estraillier. 2015, «Adaptive Representation of Digital Resources Search Results in Personal Learning Environment», dans *Artificial Intelligence in Education - 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings*, *Lecture Notes in Computer Science*, vol. 9112, Springer, p. 562–565, doi :10.1007/978-3-319-19773-9_62. 100

Shapiro, S. S. et M. B. Wilk. 1965, «An analysis of variance test for normality (complete samples)», *Biometrika*, vol. 52, n° 3-4, p. 591–611. 111

Spink, A. 1997, «Study of interactive feedback during mediated information retrieval», *Journal of the american society for information science*, vol. 48, n° 5, p. 382–394. xi, 99, 100

Walkowski, N.-O. 2017, «Evaluating Research Practices in the Digital Humanities by Means of User Activity Analysis», dans *Digital Humanities (DH)*, Montréal. 107

White, R. W., B. Kules et S. M. Drucker. 2006, «Supporting exploratory search, introduction, special issue, communications of the ACM», *Communications of the ACM*, vol. 49, n° 4, p. 36–39. 98

Wilcoxon, F. 1945, «Individual Comparisons by Ranking Methods», *Biometrics Bulletin*, vol. 1, n° 6, doi :10.2307/3001968, p. 80–83. 111

Wilson, T. D. 1981, «On user studies and information needs», *Journal of documentation*, vol. 37, n° 1, p. 3–15. 92

Wilson, T. D. 1999, «Models in information behaviour research», *Journal of documentation*, vol. 55, n° 3, p. 249–270. xi, 93

Wilson, T. D. 2000, «Human information behavior», *Informing science*, vol. 3, n° 2, p. 49–56. 94

Chapitre 5

Éclairer le contexte de production : processus d'OCR et recherche d'information

Résumé

Après l'étude du contexte d'usage effectuée dans notre précédent chapitre, le positionnement conceptuel que nous avons adopté pour ce travail requiert l'analyse du contexte de production des bibliothèques numériques. Les processus qui conduisent un document physique vers sa version numérique sont néanmoins nombreux et complexes. Parmi ces processus, nous avons choisi d'étudier plus en détail la reconnaissance optique de caractères sur le cas particulier des entités nommées. Ces dernières sont en effet des plus utiles pour la recherche d'information et particulièrement soumises aux erreurs d'OCR. Plutôt que de chercher à corriger ces erreurs, nous proposons dans ce chapitre une méthode qui permet à l'historien de repérer les différentes formes orthographiques d'une même entité. Cette méthode est évaluée pour la pratique de la recherche d'information.

Sommaire

5.1 Introduction

L'état de la pratique que nous avons établi dans notre chapitre 2 témoigne du vaste champ d'intervention du numérique et de l'informatique dans les activités de recherche en histoire. Par ailleurs, notre chapitre 3 a montré que pour les bibliothèques numériques, cas particulier que nous avons choisi d'étudier, il existe de nombreuses sources de biais potentiels pour le chercheur. Ces ressources numériques essentielles de la production de l'histoire sont, comme nous l'avons déjà expliqué, environnées d'une instrumentation informatique qui opère dans les contextes de production et d'exécution et qui ne produit pas d'effets visibles. Du point de vue de la conception du système, ces traitements sont indispensables au bon fonctionnement de la ressource. En revanche, du point de vue de l'utilisateur, leurs effets restent silencieux et pour cette raison, ils sont problématiques.

Afin d'étudier globalement l'impact des bibliothèques numériques et des moteurs de recherche sur la pratique de recherche en histoire, il est donc indispensable de travailler cette problématique. De nombreux traitements peuvent potentiellement être générateurs de biais depuis le recueil et la gestion des métadonnées, jusqu'au classement de l'information. Parmi eux, nous avons choisi de prendre pour cas d'étude les processus de reconnaissance optique de caractères. Il s'agit en effet d'un traitement très régulièrement employé lors de la production des contenus ayant vocation à être mis en ligne dans les bibliothèques numériques. Il est particulièrement important parce que c'est lui qui rend possible la recherche de l'information au sein même du contenu d'un document. La qualité de ce processus est donc un enjeu majeur pour toutes les bibliothèques numériques qui offrent l'accès à des contenus textuels numérisés.

Aussi performants qu'ils puissent être, comme nous l'avons indiqué dans notre section 3.2.1, ces processus font des erreurs dont l'utilisateur n'a pas de moyen de prendre connaissance, en dehors d'indicateurs de qualité globaux dont la pertinence est toute relative. Le problème est d'autant plus important si l'on considère les entités nommées, ces mots qui désignent des personnes, des lieux ou des organisations. Cette catégorie de mots est en effet à la fois celle qui est la plus soumise aux erreurs de reconnaissance de caractères et celle qui est le point d'entrée principal dans un corpus [Gooding, 2014]. Nous avons donc pris le parti de nous focaliser tout particulièrement sur cette catégorie de mots.

En nous intéressant à ce problème, déjà bien maîtrisé dans notre contexte de recherche local, notre objectif n'était pas, comme c'est souvent le cas, de chercher à corriger les erreurs de reconnaissance optique de caractères. En cohérence avec notre objectif global, l'objet de ce chapitre est bien plus de mettre en lumière le problème et donner au chercheur des moyens de se l'approprier, plutôt que de chercher à corriger des mots et prendre le risque de créer une nouvelle boîte noire. Ce que l'on peut identifier comme des erreurs n'en sont parfois pas. Il se peut tout à fait, pour des documents historiques, qu'il s'agisse tout simplement d'une orthographe différente, une alternative valable et porteuse de sens d'un mot. Dans ce cas, une correction détruit une information potentiellement précieuse pour l'historien.

Notre parti pris a donc été différent, nous avons surtout voulu rendre transparente, ou tout du moins éclaircir, la boîte noire de la reconnaissance de caractères et par la même faciliter, voire sécuriser, du point de vue scientifique, l'usage de corpus issus de tels

procédés. Il se dessine pour l'informaticien les points fondamentaux d'une démarche de recherche dont la problématique peut être résumée ainsi : comment donner les moyens à l'historien de tenir compte de variations orthographiques de termes, sans connaissance a priori, et sans autre ressources que celles fournies par le contexte de production initial ?

Pour répondre à cette question, notre approche vise à identifier des mots ou de groupes de mots similaires. D'une manière schématique, elle a donc pour objectif de regrouper des entités qui, bien qu'elles ne soient pas écrites de la même manière, représentent la même chose. Ces groupes d'entités proches peuvent par la suite être présentées à l'utilisateur qui a la liberté de les inclure ou les exclure de son processus de recherche d'information. D'un point de vue méthodologique, dans notre contexte, la nature des documents manipulés impose l'usage de méthodes non supervisées pour atteindre cet objectif. Il n'est en effet pas possible de recourir à l'intervention manuelle, à des dictionnaires ou à d'autres ressources de ce type. En tenant compte de cette contrainte et sur la base de l'état de l'art, nous avons abordé la problématique en postulant que deux entités sont similaires lorsque :

— Les deux entités sont écrites, à quelques caractères près, de la même manière. Autrement dit leur distance d'édition est faible, comme ce peut-être le cas lors d'erreurs de reconnaissance de caractère ;
— Les deux entités partagent un contexte similaire, autrement dit les mots qui les entourent sont proches. Selon l'hypothèse distributionnelle [Firth, 1957], ces entités sont donc probablement proche du point de vue sémantique.

Le premier point de ce postulat ne pose pas véritablement de problèmes du point de vue informatique, puisqu'il existe de nombreuses techniques pour comparer des chaînes de caractères entre elles. En revanche, la seconde est plus problématique puisqu'il s'agit de trouver un moyen de comparer différents contextes entourant un mot. Pour atteindre cet objectif, la littérature scientifique nous offre plusieurs pistes répondant en partie à nos objectifs. Elles font pour une large part appel à des représentations informatiques, mathématiques et statistiques dont il convient, dans un premier temps, de faire l'état de l'art. Dans un second temps, nous expliquerons le fonctionnement de la méthode que nous avons développée et son évaluation. Enfin, nous présenterons les résultats et les moyens mis en place pour les visualiser et les présenter aux utilisateurs.

5.2 Cas d'étude : Rendre visible le bruit informationnel, le cas des erreurs d'OCR

5.2.1 Représentations informatiques du texte : état de l'art

La communauté de recherche en informatique qui s'intéresse au traitement automatique de la langue est généralement regroupée sous l'appellation *Natural Language Processing (NLP)* ou Traitement Automatique du Langage Naturel (TALN) en français. L'objectif général de recherche en NLP est de fournir des méthodes et outils permettant d'extraire de l'information, d'analyser voire de comprendre automatiquement les textes rédigés en langage naturel.

Le traitement informatique des langues naturelles existe depuis les années 1950. D'abord fondée sur les approches de grammaires formelles notamment développées par Chomsky 1956, dont les règles se sont avérées complexes et difficiles à manipuler, la recherche s'est tournée vers des modèles d'apprentissage automatique de la langue [Adamo, 1980]. Ce basculement a en particulier permis de générer automatiquement les règles de description des langages puis de glisser progressivement vers des modèles statistiques, plus souples et plus simples à manipuler.

Représentations vectorielles

Parmi ces représentations informatiques, la famille des représentations vectorielles est une des plus importantes et des plus utilisées. Ces techniques permettent de représenter tous les éléments d'un texte sous formes de vecteurs et de les manipuler sous cette forme. L'approche la plus basique en la matière est celle dite des « sacs de mots » dont une description détaillée figure en annexe A.4.1. Dans cette représentation, chaque mot est un élément du vecteur, où seule l'information de fréquence des mots est conservée. Autrement dit, un mot présent n fois dans le texte apparaîtra n fois dans le vecteur. Les informations relatives à l'ordre des mots ou la grammaire ne sont quant à elles pas conservées.

Ce type de représentation très simple est utilisée aujourd'hui pour certaines tâches d'indexation en recherche d'information, mais souffre d'inconvénients majeurs. Le principal défaut de cette représentation est son absence de codage sémantique. Alors même que deux mots peuvent être très proches du point de vue du sens, il arrive que les vecteurs qui les représentent soient quant à eux très différents. Par ailleurs, cette technique a pour effet de faire augmenter très rapidement la taille du vocabulaire causant des problèmes de calculs statistiques.

La décomposition atomique produite par la méthode des « sacs de mots » a par ailleurs le défaut de ne pas tenir compte du contexte environnant d'un mot, puisque l'information d'ordre n'est pas préservée. Or, si nous prenons un exemple simple, tel que les expressions « la maison blanche » et « la maison hantée », nous pouvons constater l'importance du contexte, qui dans ce cas permet immédiatement d'affirmer qu'il ne s'agit pas du même objet. L'utilisation des *n-grams* en lieu et place d'une décomposition atomique permet de pallier ce défaut. Le terme *n-grams* désigne une suite de n mots contigus. Décomposé en *bigrams (2-grams)*, notre exemple précédent donne les deux résultats « la maison » et « maison hantée ». Les modèles statistiques qui reposent sur les *n-grams* permettent ainsi d'établir la probabilité qu'un nombre n d'éléments se succèdent. Dans notre cas, que l'adjectif « hantée » succède au nom « maison » par exemple. L'usage des *n-grams* à l'échelle des mots permet donc de lever certaines ambiguïtés sémantiques. Cette méthode est par ailleurs célèbre par l'usage qu'en fait Google dans son outil N-gram Viewer [1].

La méthode des *n-grams* peut également être appliquée à l'échelle des caractères. Calculer la fréquence de certains *n-grams* rend par exemple possible de déterminer la langue d'un texte. À titre d'exemple, le *bigram th* est nettement plus présent en anglais que dans d'autres langues, alors qu'en français le *bigram* le plus présent est *de*, souvent présent comme préposition, mais également comme suite de lettre composant

1. Accessible à l'adresse : https://books.google.com/ngrams.

de nombreux mots comme *demande, diode, etc.* Les calculs de *bigram*, *trigram*, *etc* et l'analyse de leur fréquence peuvent permettre de mettre en lumière la loi de probabilité d'une langue ou d'un style particulier d'écriture [2].

Word embedding

Les modèles de type *Word embedding* [3], bien qu'utilisant également des représentations vectorielles, ont un fonctionnement différent. Cette famille de méthodes informatiques vise à apprendre de manière automatique des descripteurs caractéristiques de mots ou d'expressions du langage. Ces méthodes sont fondées sur l'idée que le sens d'un mot peut être déduit de son contexte. Autrement dit que deux mots différents partageant un contexte très proche auront très probablement un sens également proche [4]. Pour chaque mot du vocabulaire (défini de la même manière que pour les sacs de mots (voir A.4.1)), un vecteur représentatif de dimension relativement faible, de l'ordre de 100 à 1000, est établi. Lorsqu'une similarité calculée entre deux de ces vecteurs est forte, alors les deux mots sont sémantiquement proches. Ces méthodes ont de nombreuses application mais la première, très souvent donnée en exemple, est de répondre à des questions sémantiques du type :

— « homme » est au « roi » ce que « femme » est à ... « reine ».

Cela peut également être utilisé d'un point de vue plus syntaxique de la manière suivante :

— « voiture » est à « voitures » ce que « cheval » est à ... « chevaux ».

Les vecteurs qui permettent de répondre à ce type de problème peuvent être obtenus par le recours à deux méthodes différentes. La première, GloVe, décrite par Pennington et collab. 2014 consiste à rassembler dans une matrice les statistiques de co-occurrences entre termes et contextes constitués de plusieurs mots. Cette approche est de nature combinatoire, occasionnant la création d'un grand nombre de contextes qu'il est nécessaire de réduire en factorisant la matrice obtenue, pour extraire des vecteurs de plus faible dimension.

La seconde méthode, word2vec décrite par Mikolov et collab. 2013 permet d'obtenir des résultats similaires sur de nombreuses tâches de NLP, mais fonctionne de manière différente. Elle présente l'avantage d'être moins coûteuse en terme de calcul, la rendant plus pertinente dans notre contexte d'usage. Tout un chacun, doté d'un ordinateur personnel, peut en effet l'utiliser pour des corpus relativement vastes, avec des temps de calcul limités [5].

2. Ces modèles de probabilités sont en particulier déterminés par les modèles de Markov cachés (*Hidden Markov models*), dont les fondamentaux sont décrits dans [Baum et Petrie, 1966].

3. La traduction de cette famille de méthodes en français est « plongement de mots » ou « plongement lexical ». Elles sont toutefois très peu utilisées et pour des raisons de clarté nous utiliserons le terme anglais.

4. Cette idée a été popularisée par le linguiste britannique John Rupert Firth qui écrivait en 1957 : *You shall know a word by the company it keeps* [Firth, 1957].

5. Bien que la méthode GloVe soit dans les faits plus facile à paralléliser que cette méthode. Elle reste moins performante sur des ordinateurs disposant d'un nombre de processus parallèles limité, essentiellement du fait du coût d'apprentissage de chacun des vecteurs.

La méthode WORD2VEC se fonde sur un réseau de neurones[6]. Si l'usage général de ce type de réseau est souvent de faire des prédictions, en soumettant un cas nouveau à un réseau préalablement entraîné, WORD2VEC ne l'utilise pas de cette manière. La méthode utilise un réseau de neurones dans le seul objectif d'apprendre les poids composant le réseau et les vecteurs caractérisant chaque mot du corpus. Le réseau est donc entraîné à prédire des mots en fonction de leur contexte, mais cette tâche doit être vue comme un moyen plutôt qu'une fin. Cette phase d'apprentissage nécessite des exemples, qui sont obtenus dans le corpus par la construction d'une fenêtre glissante de taille *window_size*. Les mots contenus dans cette fenêtre forme la liste des mots contextes (*context*) et le mot principal (w), au centre (voir figure 5.1). Ce processus est répété à l'échelle choisie, pour chaque document, chaque paragraphe ou encore chaque phrase.

FIGURE 5.1 – Construction des exemples d'apprentissage

Ici le mot principal est encadré en bleu, les mots contextes en vert avec une taille *window_size* de valeur égale à 2.

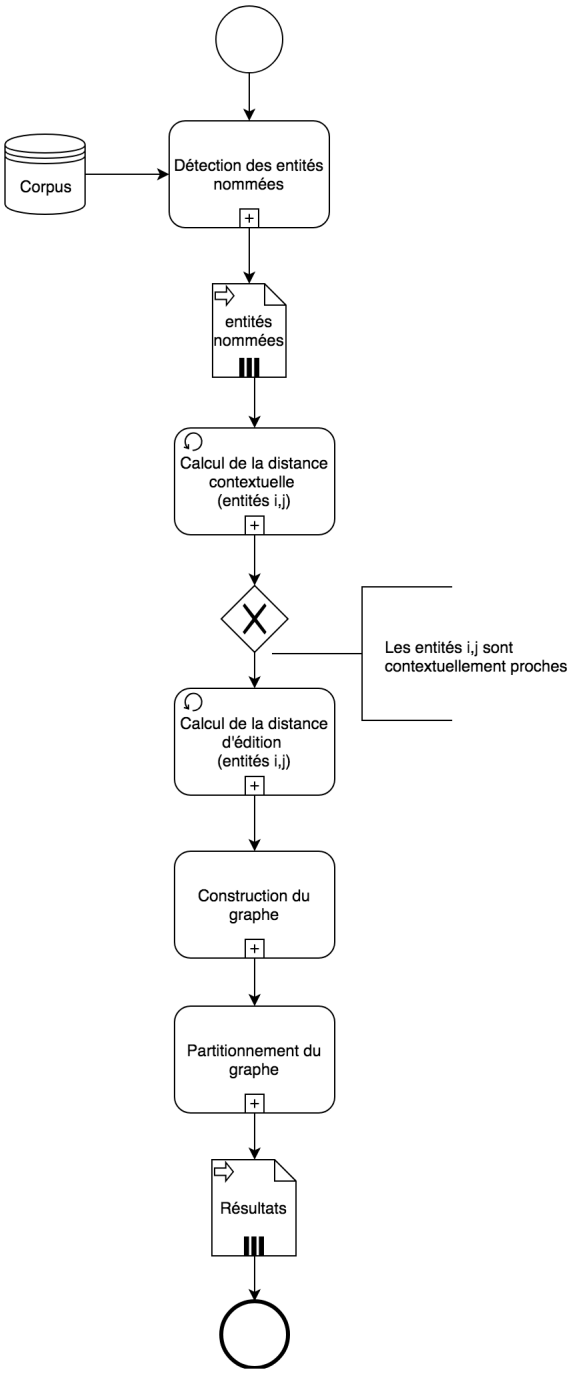
L'état de l'art fourni donc des moyens efficaces pour représenter informatiquement un mot dans son contexte. Ces représentations autorisent des comparaisons entre contextes et répondent donc, sur ce point, à nos besoins. Parmi les deux méthode de *word embedding* que nous avons brièvement décrites, nous avons choisi d'utiliser la méthode WORD2VEC pour son efficacité en terme de temps de calcul.

5.2.2 Description globale de la méthode développée

La méthode que nous avons développée repose sur plusieurs étapes résumées dans le schéma 5.2. La première consiste à repérer les entités nommées d'un corpus bruitées. Dans un second temps, la méthode génère des exemples d'apprentissage, puis entraîne le modèle de langage et détermine les vecteurs représentatifs pour chaque entité. Les étapes suivantes visent à comparer ces entités, en terme d'édition et de contexte, au moyen d'un graphe. Enfin, la méthode propose un partitionnement de ce graphe dans lequel chaque *cluster* regroupe les différentes formes d'une même entité.

6. Un réseau de neurones est une structure informatique, généralement présentée sous forme de couches, inspirée du fonctionnement des neurones biologiques. Dans cette structure, chaque couche est dotée d'une fonction particulière dont la sortie est ensuite transférée en entrée à la couche suivante.

FIGURE 5.2 – Diagramme BPMN des étapes successives de la méthode



Étape de détection d'entités nommées

La première étape de la méthode que nous avons mise en place est de détecter les entités nommées présentes dans un corpus. Cette étape de repérage peut naturellement être automatisée. Bien entendu, elle n'est pas parfaite, loin de là, surtout dans un contexte où les textes soumis au processus sont bruités. Néanmoins, il est toujours possible de repérer une large part de ces entités. Cette détection et cette classification est une des tâches classiques du traitement automatique des langues. De nombreux outils existent pour effectuer cette tâche. Ils peuvent être fondés sur un ensemble de règles d'analyse ou bien sur des mécanismes d'apprentissage automatique. Lorsque le texte est dégradé,

les méthodes qui utilisent les règles voient leurs résultats plus lourdement affectés que les méthodes fondées sur l'apprentissage automatique. En effet, aucun mécanisme ne permet dans leur cas de tenir compte du bruit dans le texte, celui-ci a donc un impact important. À l'inverse, les mécanismes et corpus d'apprentissage permettent, dans le cadre d'un apprentissage automatique, de limiter l'impact du bruit.

Fondés sur l'une ou l'autre de ces approches, les outils suivants sont reconnus par la littérature [Rodriquez et collab., 2012] :

— **BM Watson (Alchemy API)** qui permet, entre autres, d'annoter dans des textes des entités nommées. Cet outil a le défaut majeur d'être propriétaire et fermé. On ne connaît pas les détails de son fonctionnement, tout juste sait-on qu'il fait appel à des mécanismes d'apprentissage profond (*deep learning*) [7].

— **NeuroNer** utilise lui une approche fondée sur les réseaux de neurones pour détecter et classer les entités nommées. Le système peut être entraîné sur différents corpus [Dernoncourt et collab., 2017].

— **Stanford NER** qui fait partie de la suite Stanford NLP développée à l'Université Stanford, utilise quand à lui une approche statistique fondée sur les *Conditionnal Random Fields* (CRF) [8] [Finkel et collab., 2005].

Dans un contexte comme le nôtre, impliquant des textes dégradés, les outils **Stanford NER** sont globalement les plus performants [Rodriquez et collab., 2012]. Ceux-ci sont néanmoins très variables en fonction du corpus d'entraînement utilisé. L'outil Stanford NER a par ailleurs été utilisé lors du projet Europeana Newspaper et a montré des résultats intéressants. Le modèle d'entraînement utilisé lors de ce projet est par ailleurs publiquement accessible. Le système Watson (Alchemy API) étant fermé, on ne connaît pas les détails de son fonctionnement et ses résultats sont difficilement explicables.

Pour ces raisons de précision et de performance [9], nous avons utilisé les outils de Stanford NER, avec le modèle d'entraînement réalisé par le projet Europeana Newspaper [Neudecker, 2014] [10]. Néanmoins, rien n'empêche de choisir un autre système pour faire le travail de cette phase, si il donne de meilleur résultat pour un corpus particulier. À l'issue de cette détection, le système génère une liste d'entités contenant indistinctement des entités « propres » et des entités « bruitées ». À titre d'exemple, on peut trouver dans cette liste les termes *"lenrngrad", "leningrad", "lenlngrad"*. Dans ce cas, l'entité Leningrad a été trouvée dans sa version correcte et dans deux autres versions erronées.

7. La documentation des APIs de Watson est accessible à l'adresse : https://console.bluemix.net/developer/watson/documentation.

8. Les CRF sont une classe de modèle statistique, ils sont particulièrement adaptés au traitement de données sérielles, dont peut faire partie le langage naturel. Le fonctionnement des CRF a été décrit par Lafferty et collab. [2001].

9. La bibliothèque logicielle Stanford NER peut en effet sans problème être utilisée de manière parallèle pour accélérer le traitement de grands corpus. Il est *thread safe* depuis sa version 1.3.0 (voir https://stanfordnlp.github.io/CoreNLP/history.html).

10. Publiquement accessible à cette adresse : http://lab.kbresearch.nl/static/html/eunews.html.

Calcul de la distance contextuelle : création du modèle

L'objectif de la seconde étape du processus est d'établir et de représenter le contexte de chaque entité trouvée dans le corpus. Comme nous l'avons déjà dit, deux entités ayant le même sens ont toutes les chances d'être entourées d'un contexte lui aussi très proche. Si nous reprenons notre exemple précédent, le mot « leningrad », qu'il ait été ou non correctement reconnu par l'OCR, sera probablement suivi ou précédé par les mots « ville » ou « soviétique » par exemple. La seconde étape de la méthode calcule donc le contexte de chaque entité en cherchant les mots encadrant chaque entité.

Nous avons vu en section 5.2.1 que les représentations vectorielles se prêtent bien à représenter ce type de données. nous avons ainsi mis en application WORD2VEC qui permet au modèle d'apprendre les vecteurs représentatifs de chaque entité. Dans notre contexte, il est préférable de faire fonctionner WORD2VEC avec l'architecture *Skip-Grams*, plus à même de représenter correctement les contextes de mots rares, comme peuvent l'être les entités nommées [MIKOLOV et collab., 2013]. Par ailleurs, pour maximiser les chances de reconnaître un contexte cohérent, malgré le bruit, nous avons augmenté la taille de la fenêtre d'apprentissage [11]. Enfin, il convient de préciser que certaines entités sont composées de plusieurs mots, un prénom et un nom de personne par exemple. Dans ce cas, afin de faciliter l'établissement du contexte et la comparaison entre entités, les mots sont fusionnés en remplaçant le caractère d'espacement par un tiret-bas, de la manière suivante : *"prénom_nom"*. À l'issue de cette étape, nous disposons des entités nommées détectées dans le corpus, ainsi que de leur vecteur contextuel représentatif.

Définition de la similarité contextuelle et de la similarité d'édition

Sur la base des informations construites lors des étapes 1 et 2 de la méthode, il est possible de calculer deux similarités différentes entre les entités nommées. La première est une distante dite d'édition qui vise à mesurer la différence entre deux mots. Il est possible d'effectuer cette comparaison en calculant le nombre d'opérations nécessaires à transformer une entité a en une entité b. Le calcul de cette distance est expliqué en détail dans notre section 6.4.2, et particulièrement dans l'équation 6.1. Pour comparer les distances entre elles, il est toutefois nécessaire de normaliser le résultat. Ainsi, plus la similarité sim_{edit} est proche de 1, plus les termes sont proches.

Dans notre contexte toutefois, la distance de Levenshtein n'est pas nécessairement la plus adaptée du fait de la nature des entités nommées qui peuvent être composées de plusieurs mots. Une autre métrique, dite coefficient de Dice [DICE, 1945] est dans ce cas plus pertinente [12]. Le coefficient de Dice se fonde sur la comparaison de *bigram* entre deux expressions. À titre d'exemple le mot « information » sera décomposée en la suite de *bigram* suivante : « in, nf, fo, or, rm, ma, at, ti, io, on ». Le coefficient est défini comme le double du nombre de *bigram* partagés n_t entre deux expressions comparées i et j, rapporté à leur somme selon la formule 5.1.

11. La taille par défaut de cette fenêtre est 5, nous avons choisi de la porter à 15. La taille des vecteurs est par ailleurs fixée par défaut à 300 et le modèle est entraîné sur 30 itérations.
12. Ce coefficient est également connu sous le nom d'indice de Sorensen [SØRENSEN, 1948].

$$sim_{edit} = \frac{2n_t}{n_i + n_j} \tag{5.1}$$

Cette similarité se limite donc à tirer partie des seules entités nommées. Étant donné que nous disposons des vecteurs représentatifs du contexte de chacune d'elle, nous pouvons également calculer un indicateur de la similarité entre ces vecteurs, sim_{ctx}. La distance cosinus est la distance la plus adaptée à cette opération [Mikolov et collab., 2013] de comparaison entre vecteurs de ce type. En comparant les vecteurs représentatifs $\vec{e_i}$ et $\vec{e_j}$ des entités e_i et e_j d'un ensemble d'entités $\mathbb{E}$, on obtient un résultat compris entre -1, dans le cas où les vecteurs sont totalement différents et 1 si les vecteurs sont strictement égaux. Cette distance est formalisée par l'équation 5.2. À l'image de la distance d'édition et pour rendre possible la comparaison, il est nécessaire de formaliser cette distance pour que la valeur de sim_{ctx} soit contenu entre 0 et 1. Cette normalisation est opérée par l'intermédiaire de la formule 5.3.

$$\begin{aligned} cosinus : \mathbb{E} \times \mathbb{E} &\to [-1, 1] \\ (e_i, e_j) \mapsto cosinus(e_i, e_j) &= \frac{\vec{e_i} \cdot \vec{e_j}}{\|\vec{e_i}\| \times \|\vec{e_j}\|} \end{aligned} \tag{5.2}$$

$$\begin{aligned} sim_{ctx} : \mathbb{E} \times \mathbb{E} &\to [0, 1] \\ (e_i, e_j) \mapsto sim_{ctx}(e_i, e_j) &= \frac{\vec{e_i} \cdot \vec{e_j} + \|\vec{e_i}\| \times \|\vec{e_j}\|}{2 \times \|\vec{e_i}\| \times \|\vec{e_j}\|} \end{aligned} \tag{5.3}$$

Ainsi sont définies les similarités contextuelles et d'édition. La combinaison de ces deux similarités donne un couple de valeurs nommé $\mathbb{S}$ et défini dans la formule 5.4.

$$\begin{aligned} \mathbb{S} : \mathbb{E} \times \mathbb{E} &\to [0, 1] \times [0, 1] \\ (e_i, e_j) \mapsto \mathbb{S}(e_i, e_j) &= (sim_{ctx}(e_i, e_j), sim_{edit}(e_i, e_j)) \end{aligned} \tag{5.4}$$

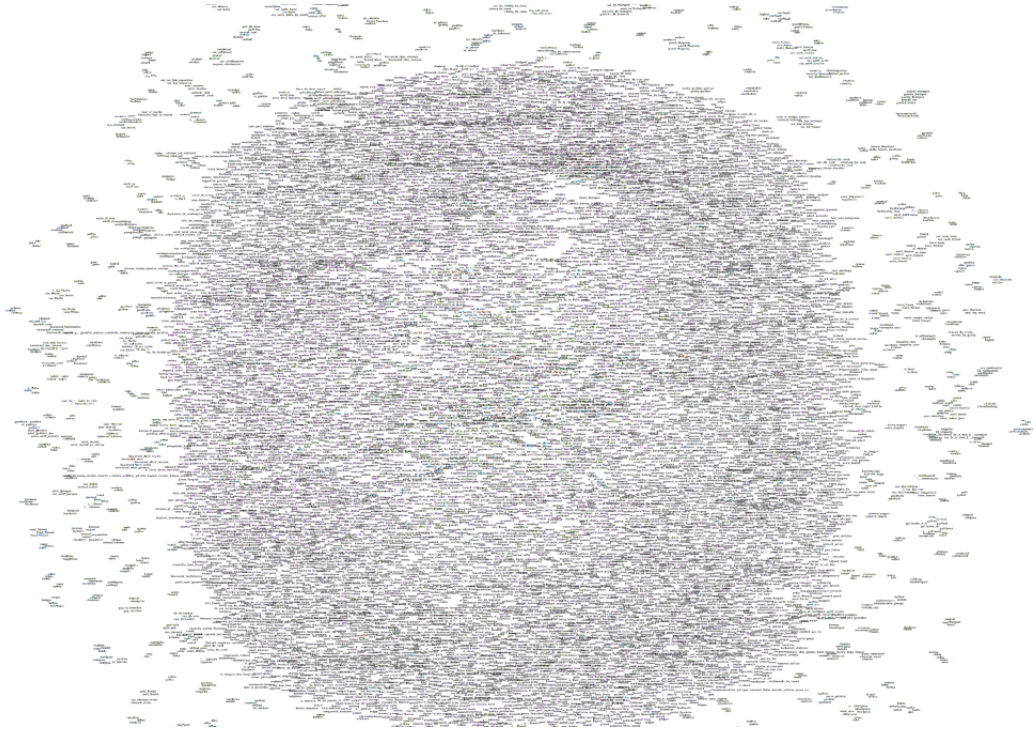
Construction du graphe

L'étape suivante consiste à construire un graphe qui connecte entre elles les entités nommées, sommet du graphe, en fonction de la valeur de bi-similarité $\mathbb{S}$. Le graphe qui en résulte ($\mathbf{G}_n^t$), dont le détail de la construction est présenté en annexe A.5.1 et un exemple à la figure 5.3, est un graphe pondéré non orienté. Les sommets de ce graphe sont connectés s'ils représentent des entités nommées similaires du point de vue contextuel (sim_{ctx}) et si leur similarité d'édition sim_{edit} dépasse un seuil t paramétrable. Les paramètres, n et t sont fixés expérimentalement et ont un rôle important pour limiter la taille du graphe généré lors de cette phase du processus.

Partitionnement

Le graphe construit à l'étape précédente de cette méthode doit ensuite être partitionné afin de déterminer des groupes d'entités proches, comme ceux visibles à la figure 5.4. Pour se faire, il est nécessaire d'appliquer sur le graphe un algorithme de *clustering*

FIGURE 5.3 – Exemple d'un graphe complet

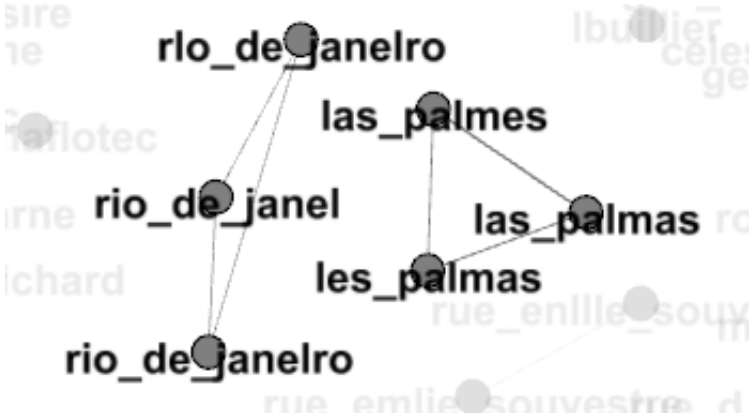


Cette figure représente un graphe complet des entités bruitées détectées.

qui a pour objectif de rassembler les sommets du graphe qui représentent un groupe correspondant à une entité. Il faut noter que nous ne pouvons pas connaître *a priori* le nombre de *cluster*. Par ailleurs, un même mot détecté et présent dans le graphe peut appartenir à plusieurs *cluster*. Un mot mal reconnu par l'OCR peut très bien être résolu en deux mots existants, de sens différent.

Ces deux contraintes nous ont conduit à sélectionner l'algorithme *weighted Clique Percolation Method* (CPMw) décrit par PALLA et collab. [2005]. Cet algorithme est capable de déterminer des communautés au sein du graphe par calcul des sous-graphes complets possédant k sommets, nommés k-cliques. Lorsque deux k-cliques sont adjacentes, si et seulement si elles partagent $k-1$ sommets, elles sont rassemblées pour former une communauté. La complexité du graphe qui résulte de l'étape précédente de la méthode, dans lequel certaines zones peuvent être fortement connectées, impose cependant de limiter le temps de calcul pour chaque sommet du graphe afin d'éviter des temps de calcul trop importants.

FIGURE 5.4 – Exemple de groupes de mots représentatifs d'entités



Ici, il est possible d'observer deux groupes distincts d'entités nommées correspondant à Rio de Janeiro et Las Palmas.

5.2.3 Résultats produits

À l'issue du traitement complet d'un corpus, la méthode produit des listes de mots. Chaque liste contient au minimum deux éléments et est considérée comme représentant la même entité. Cette sortie se présente sous la forme d'un fichier *Json*, dont un exemple est visible à la figure 5.5, facilement lisible et aisément réemployable dans une application.

FIGURE 5.5 – Exemple de résultats produits à l'issue de la méthode

```
1 [["rosa_luxemburg", "rosa_luxembourg", "rosa_luxembur", "rosa_luxernburg"],
2 ["ravensbruclr", "ravensbruck", "ravensbrucltt"],
3 ["sabne_et_loire", "saene_et_loire", "saone_et_loire"],
4 ["michel_alexandrovitch", "alexandrovitch", "mikhail_alexandrovitch"],
```

Les exemples de la figure précédente (5.5), ne posent, pour les lignes 1 à 3, pas de difficultés particulières, ce sont bien différentes versions d'une même entité qui ont été correctement regroupées. Le groupe visible à la ligne 4 pose cependant plus de problèmes. S'il est raisonnablement possible de penser que « michel » est bien la traduction française de « mikhail », rien ne garantie qu'il s'agisse bien du même personnage. Par ailleurs, l'entité présente sous la seule forme du nom « alexandrovitch » dans ce groupe ne garantie pas non plus que le personnage auquel elle renvoie soit le même que les deux autres versions. Si notre objectif avait été de corriger automatiquement ces entités, ce type de groupe de mots aurait été un véritable problème qui aurait nécessité une étape supplémentaire de désambiguïsation. Dans notre contexte, ce groupe n'est pas problématique dans la mesure où il signale la présence de différentes formes orthographiques d'un terme à l'utilisateur et lui laisse toute latitude d'en tenir compte ou de l'ignorer.

Toutefois, même dans ce contexte, la méthode peut dans certains cas générer beaucoup de bruit, si les paramètres n et t sont mal choisis. Il se peut par exemple, avec une valeur de t trop basse, autrement dit un seuil de similarité d'édition sim_{edit} insuffisant, que la méthode regroupe un grand nombre de personnage ayant le même prénom mais pas le même nom de famille. Le seul prénom en commun suffisant dans ce cas à calculer une valeur sim_{edit} dépassant le seuil t.

5.3 Impact de la méthode sur l'accessibilité des documents

Notre objectif de recherche nous impose de montrer l'impact du bruit généré par la reconnaissance optique de caractères sur la pratique. Il ne s'agit pas, comme c'est souvent le cas, d'évaluer la qualité de notre méthode par rapport à une autre, mais bien de montrer sa capacité à améliorer la qualité de la recherche d'information, du point de vue de l'utilisateur. Dans ce contexte, nous pouvons raisonner de la manière suivante. Dans le cas d'un corpus dit propre, ne comportant pas d'erreurs les documents sont tous accessibles de la même manière, leur présence ou leur absence ainsi que leur position dans les résultats ne dépendent que de la requête saisie et du paramétrage du moteur de recherche, en particulier des algorithmes de pertinence qu'il utilise. Dans un contexte bruité, en revanche, les erreurs que nous avons relevées ont un impact sur l'accessibilité des documents dans les résultats de recherche. Un document théoriquement accessible parce qu'il contient un ou des termes de la requête peut ne plus l'être, ou l'être plus difficilement, s'il contient des erreurs. Nous avons donc évalué l'impact de notre traitement sur cette accessibilité des documents les uns par rapport aux autres, qu'il nous faut dans un premier temps être capable de mesurer.

5.3.1 Méthode de mesure de l'impact sur l'accessibilité globale des documents

Une méthode de calcul de l'accessibilité des documents a été développée par Azzopardi et Vinay [2008]. Elle a été développée pour mesurer l'accessibilité des documents du point de vue de l'utilisateur et compléter les traditionnelles mesures de précision et de rappel de la recherche d'information. Elle produit un score d'accessibilité pour chaque document du système. Celui-ci doit reposer, à l'image de notre plateforme expérimentale, sur un corpus de documents indexés D que l'on peut interroger en exprimant une requête q. Cette interrogation produit une liste de résultats R_q classée en fonction de la pertinence selon q. L'ensemble des requêtes possibles est définie par Q et la popularité d'une requête est prise en compte en établissant un poids $\mathbf{O}_q$ pour chacune d'elle. La méthode de calcul du score de *Azzopardi et Vinay* est fondée sur l'idée qu'un document devrait être considéré comme très accessible et donc obtenir un score d'accessibilité élevé, si :

— Un grand nombre de requêtes q dans l'ensemble Q permettent d'accéder à ce document ou si les requêtes q, bien que peu nombreuses, sont populaires et possèdent un poids $\mathbf{O}_q$ élevé.

— Si le document est bien classé et obtient donc un rang faible, de valeur minimale 1 pour la première position, dans la liste R_q des résultats.

Ainsi, la fonction de calcul du score d'accessibilité est donnée dans l'équation 5.5.

$$r(d) = \sum_{q \epsilon Q} \mathbf{O}q \times f(k_{dq}, c) \tag{5.5}$$

Ici, la fonction $f(k_{dq}, c)$ est une fonction de coût dont le résultat est déterminé par k_{dq}, le rang du document d suite à une requête q et c le rang maximum que l'on souhaite considérer. Dans le travail de *Azzopardi et Vinay*, cette fonction renvoie la valeur 1 si k_{dq} est inférieur à c et 0 dans le cas inverse. Le paramètre c est ici fondamental dans la mesure où il est le critère discriminant qui nous permet de considérer ou non le document comme « accessible ». La méthode la plus évidente est de considérer c comme le nombre maximal de documents que les utilisateurs sont susceptibles de consulter lors de leur activité de recherche d'information. Dans ce cas, c est un entier arbitrairement fixé, par exemple, sur des bases expérimentales. Il est néanmoins possible de remplacer cette fonction et ce seuil par d'autres valeurs, selon ce que l'on souhaite mesurer.

Cette méthode permet donc de calculer un score d'accessibilité document par document. Cependant, afin de pouvoir évaluer l'accessibilité de chaque document au sein d'un corpus, il est nécessaire de représenter les mesures d'accessibilité plus globalement. Azzopardi et Vinay [2008] font dans cette optique usage de courbes dites de Lorenz, décrites dans [Lorenz, 1905]. Cette représentation graphique, issue des sciences économiques, permet d'associer les détenteurs d'une part x d'une grandeur quelconque, avec la part y de la grandeur détenue. Elle est en particulier utilisée pour mesurer les inégalités de partage des richesses dans une population. Dans une telle représentation, si 50% d'une population détient 50% de la richesse, alors la courbe de Lorenz passe par le centre de la figure représentant ainsi l'égalité du partage de la richesse. Cette représentation est aisément transposable à d'autres cas. Dans le nôtre, il s'agit de visualiser les écarts d'accessibilité des documents d'un corpus, un exemple de courbe de Lorenz dans ce contexte est donné à la figure 5.6.

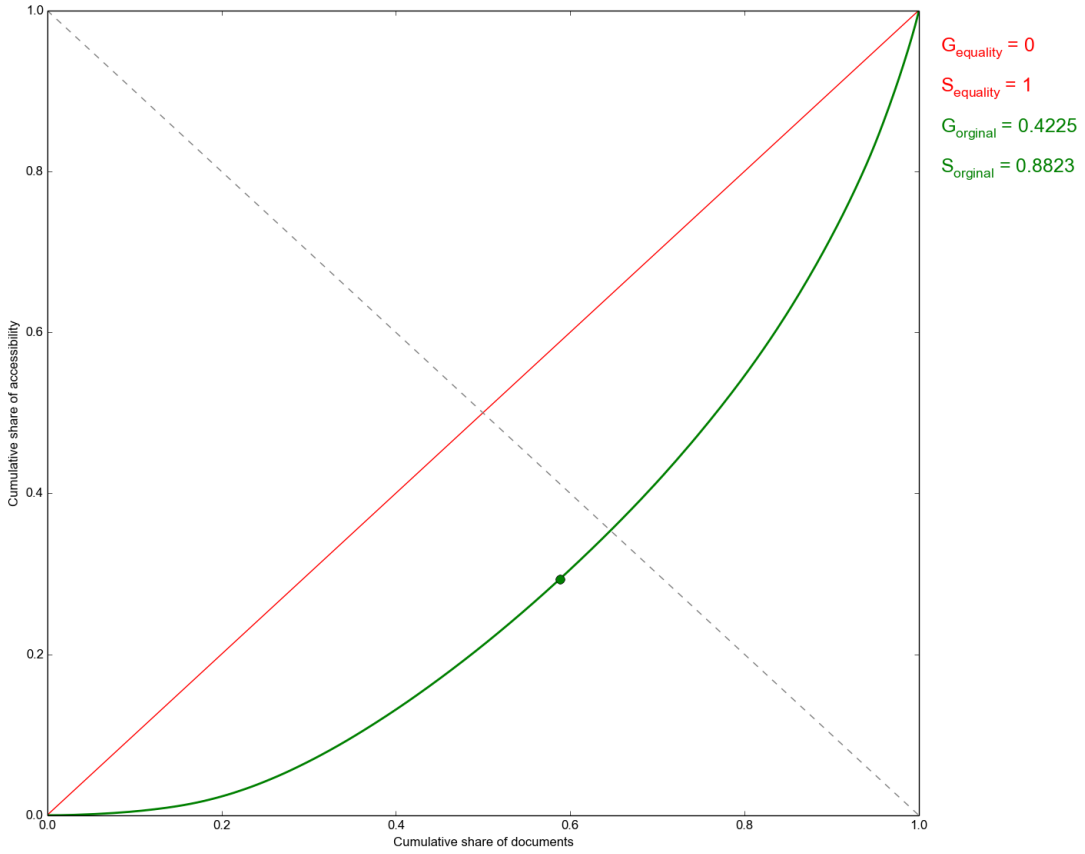
La construction de cette courbe est opérée en deux étapes. Dans un premier temps, les documents sont classés en fonction de leur score d'accessibilité obtenu par le calcul formalisé dans l'équation 5.5. La distribution cumulative des scores d'accessibilité peut ensuite être tracée. Ainsi, plus la distribution obtenue est proche de la distribution linéaire, plus les documents du corpus sont globalement et équitablement accessibles.

L'inégalité visualisable via l'aire entre les deux courbes de distribution peut être explicitée par le coefficient de Gini [Gastwirth, 1972]. Cette mesure, formalisée par l'équation 5.6 permet d'obtenir un coefficient G normalisé entre les valeurs 0 et 1. Dans notre contexte, plus le coefficient G s'approche de la valeur 0, plus les documents sont équitablement accessibles.

$$G = \frac{2\Sigma_{i=1}^{N} (2 \times i - N - 1) \times r(d_i)}{N\Sigma_{j=1}^{N} r(d_j)} \tag{5.6}$$

Cependant, l'indice de Gini ne permet pas à lui seul de tenir compte de la répartition des inégalités d'accessibilité des documents. À titre d'exemple, si la moitié d'un corpus est systématiquement accessible et l'autre moitié ne l'est jamais, alors $G = 0,5$. On obtiendra le même résultat si 25% seulement des documents partagent 75% de l'accessibilité et si les 75% de documents restants ne se partagent que les 25% d'accessibilité restante. La répartition des inégalités est dans ce cas bien différente, mais le coefficient G reste égal à 0,5. Pour mesurer la répartition des inégalités il est possible de calculer le coefficient d'asymétrie de Lorenz, noté S [Damgaard et Weiner, 2000] et formalisée par l'équation 5.7. Ce coefficient est défini par la somme du résultat de la fonction de

FIGURE 5.6 – Exemple d'une courbe de Lorenz



La courbe rouge représente le cas idéal, pour lequel tous les documents du corpus partagent chacun le même niveau d'accessibilité. La courbe verte représente le niveau d'accessibilité réel des documents du corpus. L'aire comprise entre les courbes rouge et verte représente visuellement le coefficient de Gini, noté G, mesurant l'ampleur des inégalités. Le point sur la courbe verte représente le coefficient d'asymétrie, où la tangente est parallèle à la courbe d'égalité (en rouge).

répartition des documents une fois ordonnés en fonction de leur accessibilité F et la fonction de répartition de cette accessibilité L pour μ l'accessibilité moyenne du corpus. Le détail des calculs de ces fonctions est présenté en annexe A.5.2.

$$S = F(\mu) + L(\mu) \tag{5.7}$$

Ainsi, le point présenté dans la figure 5.6 correspond à S et se situe au point où la tangente de la courbe de Lorenz est parallèle à la courbe représentant la distribution linéaire. Le point représenté par le coefficient S prend une valeur supérieure à 1 si il se situe au dessus de la ligne de symétrie, les documents les plus accessibles étant dans ce cas les plus impactés par les inégalités de répartition et inférieure à 1 s'il se situe en dessous, indiquant alors que les documents les moins accessibles sont aussi ceux qui souffrent le plus de l'inégalité de la répartition.

5.3.2 Construction des jeux de requêtes

Mesurer les valeurs qui permettent d'évaluer l'impact d'une méthode sur l'accessibilité des documents d'un corpus suppose au préalable de disposer d'un ensemble de requêtes, noté Q. Théoriquement, cet ensemble de requêtes est infini dans la mesure où il dépend des saisies fournies par les utilisateurs du moteur de recherche. Il est possible d'établir une liste de requête réelle effectuée par les utilisateurs, si ces données sont disponibles. Lorsqu'elles ne le sont pas, il est nécessaire de créer un jeu de requête simulées. Cette approche a par exemple été utilisée par Azzopardi et Vinay [2008]. Si l'on s'intéresse à tous les types de recherche d'information, et pas seulement aux entités nommées, le jeu de requête peut être construit par extraction de *bigrams*, *trigrams*, *etc.* Ces extraits sont ensuite classés par fréquence pour établir la popularité des requêtes.

Cette méthode induit un certain nombre de biais. D'abord, elle ne fait pas de différence dans le texte et calcule indistinctement des séries de mots. Cette méthode ne tient ainsi pas compte de la surreprésentation des entités nommées dans les requêtes réelles. Par ailleurs, la fréquence d'apparition de tel ou tel extrait du texte ne suppose pas un niveau de popularité identique dans des requêtes réelles, produites par des utilisateurs. Des travaux ont été menés sur cette question et ont montré des différences substantielles entre requêtes simulées et requêtes réelles pour des évaluations d'accessibilité [Traub et collab., 2016]. Cependant, procéder à des tests à l'aide de requêtes simulées reste un moyen efficace, même s'il n'est pas totalement représentatif de la réalité, pour mesurer les performances d'un système et mesurer les inégalités et les biais de recherche.

Dans le cas que nous traitons ici, le problème est toutefois plus simple. Étant donné que nous ne traitons ici que le cas des entités nommées, notre méthode n'a pas d'impact sur le reste du texte. Notre jeu de requête se fonde donc uniquement sur ces entités nommées. Les résultats n'ont donc d'intérêt que pour ce type de recherche d'information. Du reste, il est nécessaire de maintenir un jeu de requête réaliste. Pour ce faire, nous avons extrait la liste des entités reconnues par l'algorithme de détection d'entités nommées, que nous avons manuellement corrigée en nous référençant au texte initial si nécessaire.

À l'issue de cette correction, nous disposons d'une liste de requêtes, non bruitées, qui peuvent posséder ou non une ou plusieurs alternatives produites par notre méthode. Les mesures sont donc réalisées sur la base de cette liste. Les documents, une fois indexés dans la plateforme de recherche, sont interrogés pour chaque requête contenue dans la liste et produisent une première série de mesure d'accessibilité. Pour produire les mesures à comparer, nous vérifions pour chaque entité de la liste la présence d'alternatives potentielles. Lorsqu'elles existent, la requête envoyée au moteur de recherche inclue ces alternatives, modifiant ainsi les résultats produits et établissant de nouvelles mesures d'accessibilité.

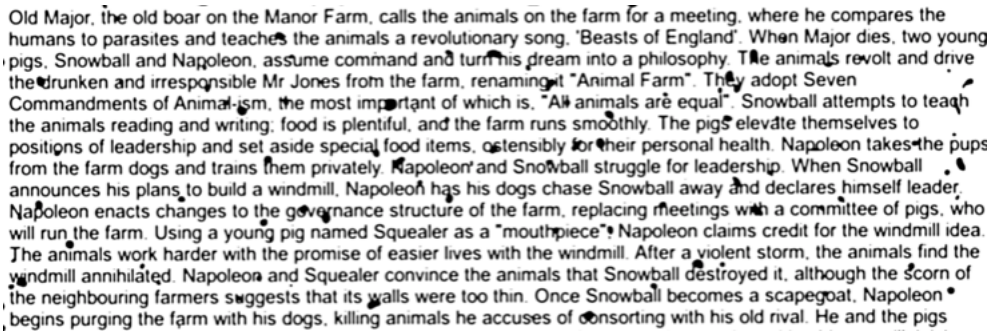
5.3.3 Corpus d'évaluation

Afin de mesurer l'impact de notre méthode et donc de erreurs d'OCR sur la pratique de recherche, il nous faut disposer d'un corpus de documents sur lesquels mener des expérimentations, établir des requêtes et produire les métriques dont nous venons de parler. Dans un premier temps, il s'agit de construire un corpus de documents sur

lequel il est pertinent de mener des recherches par entités nommées. Sélectionner de tels corpus est relativement simple, aussi bien dans les jeux de données classiques de l'état de l'art que sur les plateformes de documents historiques. Cependant, dans un cas idéal, ce corpus devrait être disponible en deux versions, corrigée et donc « propre » d'une part et brute et bruité d'autre part. Dans ces deux versions, ou du moins dans la version corrigée, la liste des entités nommées devrait être établie manuellement afin de disposer d'une liste de requête initiale.

Dans les faits, de tels corpus, réunissant toutes ces conditions, n'existent pas. Il se peut que certains n'existent qu'en version propre, car nativement numériques, d'autre encore existent bien dans les deux versions mais ne sont pas annotés ou encore ne remplissent pas les caractères représentatifs d'un document interrogeable dans une démarche historique. De ce fait, la première approche possible est de sélectionner un corpus propre, de le dégrader artificiellement, tel que l'exemple visible à la figure 5.7, et de le soumettre à un processus de reconnaissance de caractère. Cette approche permet de contrôler le niveau de bruit du texte traité et d'établir facilement une liste de requête initiale dans la mesure ou la version propre est disponible. Elle manque toutefois de réalisme, au regard du niveau de qualité atteint par les processus de reconnaissance et de correction appliqués par les grands acteurs du domaine.

FIGURE 5.7 – Exemple d'image artificiellement dégradée pour générer du bruit informationnel



Le bruit visible sur cette image est important, il est bien entendu possible de le réduire, de l'augmenter ou de faire varier le type de bruit utilisé.

La seconde approche, que nous avons retenue ici, consiste à travailler sur un corpus originellement dégradé. Ici, le bruit présent dans les textes est représentatif. Le principal problème est d'établir une liste de requêtes initiales comprenant un nombre suffisamment important d'entités nommées présentes dans le corpus. Ce processus nécessite une phase de vérification manuelle, fastidieuse mais inévitable. De tels documents peuvent être obtenus facilement. Les grandes bibliothèques, en particulier nationales, rendent disponibles de nombreux documents dans leur version en plein texte. C'est en particulier le cas de la BNF, via Gallica. Nous avons donc téléchargé depuis Gallica plusieurs corpus et utilisé la reconnaissance OCR de la BNF quand celle-ci était disponible (corpus A) ou soumis nous-même les documents à un tel processus quand elle ne l'était pas (corpus B). Un exemple de chacun de ces corpus est présenté en figure 5.8 [13].

13. Les exemples présentés sont accessibles aux adresses http://gallica.bnf.fr/ark:/12148/bpt6k76363707/f5 pour le corpus A et http://gallica.bnf.fr/ark:/12148/bpt6k717451x/f5 pour le corpus B.

— Corpus A : Presse française (Quotidien "Ce Soir", 1200 pages datées de l'année 1939.)
Cette publication est une publication quotidienne d'information générale, constituée de beaucoup de texte et relativement peu d'illustrations. La reconnaissance de caractères a été opérée par la BNF.

— Corpus B : Presse française (Hebdomadaire "Le Petit Journal Illustré", 11130 pages datées des années 1920 à 1937.)
Notre choix s'est arrêté sur « Le petit journal illustré ». Cet hebdomadaire créé en 1884 est d'abord paru comme supplément du dimanche du célèbre quotidien « Le petit journal » avant de changer de nom en 1920 pour devenir « Le petit journal illustré » jusqu'à la fin de sa parution en 1937. Ce titre est disponible via la Gallica, la bibliothèque numérique de la BNF [14]. Contrairement à la publication précédente, cette publication est beaucoup plus illustrée et propose moins de texte. Ces derniers étant par ailleurs très souvent narratifs et beaucoup moins liés à l'actualité, les entités nommées présentes sont d'une nature différente et en nombre plus limité.

Nous avons choisi de limiter notre corpus aux années durant lesquelles cette publication a porté le nom de « Le petit journal illustré » entre les années 1920 et 1937, ce qui représente environ 800 numéros pour un total de plus de 11000 pages. Ce titre n'a pas encore fait l'objet d'un processus d'OCR au sein de la BNF. Nous avons donc, via les APIs fournies par la BNF, télécharger l'intégralité de ces 800 numéros, page par page. Nous avons ensuite soumis l'intégralité de ces pages à un logiciel d'OCR [15] paramétré pour obtenir un résultat au format ALTO XML en conservant les indices de confiance de l'algorithme, duquel nous avons ensuite extrait le texte seul. Un extrait de cette publication et de la sortie texte de l'OCR sont visibles à la figure 5.9.

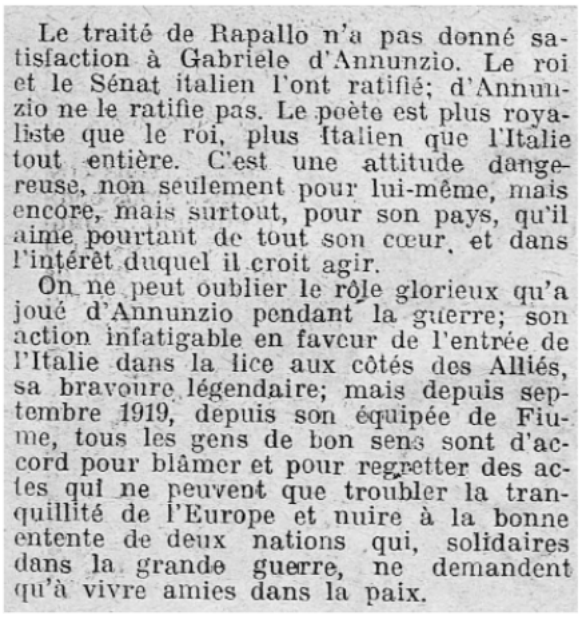
14. À l'adresse suivante : http://gallica.bnf.fr/ark:/12148/cb32836564q.
15. Nous avons utilisé le logiciel Abby Finereader dans sa version 11.

FIGURE 5.8 – Extrait des corpus A et B.

(a) Extrait du corpus A.

(b) Extrait du corpus B.

FIGURE 5.9 – Exemple de document du corpus et de la sortie du processus d'OCR



Le traité de Rapallo n'a pas donné satisfaction à Gabriele d'Annunzio. Le roi et le Sénat italien l'ont ratifié; d'Annunzio ne le ratifie pas. Le poète est plus royaliste que le roi, plus Italien que l'Italie tout entière. C'est une attitude dangereuse, non seulement pour lui-même, mais encore, mais surtout, pour son pays, qu'il aime pourtant de tout son cœur, et dans l'intérêt duquel il croit agir.

On ne peut oublier le rôle glorieux qu'a joué d'Annunzio pendant la guerre; son action infatigable en faveur de l'entrée de l'Italie dans la lice aux côtés des Alliés, sa bravoure légendaire; mais depuis septembre 1919, depuis son équipée de Fiume, tous les gens de bon sens sont d'accord pour blâmer et pour regretter des actes qui ne peuvent que troubler la tranquillité de l'Europe et nuire à la bonne entente de deux nations qui, solidaires dans la grande guerre, ne demandent qu'à vivre amies dans la paix.

```
Le traité de Rapallo n'a pas donné sa- .

tisfaction à Gabriele d'Annunzio. Le roi
et le Sénat italien l'ont ratifié; d'Anhun-
2zi0 ne le ratifie pas. Le poète est plus roya-
liste que le roi, plus Italien que l'Italie
tout.. entière. C'est une attitude dange-
reuse, non seulement pour lui-même, mais
encôre, mais surtout, pour son pays, qu'il
aime pourtant de tout son cœur, et dans
l'intérêt duquel il.croit agir.

On, né peut oublier le rôle glorieux qu'a
joué d'Annunzio pendant la guerre; son
action infatigable en faveur de l'entrée de
l'Italie dans la lice aux côtés des Alliés,
sa bravoure légendaire; mais depuis sep-
tembre 1919, depuis son équipée de Fiu-
me, tous les gens de bon sens sont d'ac-
cord pour blâmer et pour regretter des ac-
les qui ne peuvent que troubler la tran-
quillité de l'Europe et nuire à la bonne
entente de deux nations qui, solidaires
dans la. grande guerre, ne demandent
qu'à vivre amies dans la paix.
```

Extrait du numéro du 25 décembre 1920 de « Le petit journal illustré ».

5.4 Application au cas d'étude : impact sur le contexte d'usage

5.4.1 Résultats : impact sur l'accessibilité globale des documents

Les résultats présentés ici se fondent sur les mesures d'accessibilité que nous avons présentées. Pour chacun des corpus utilisé, une liste de requête a été déterminée. Elle permet de calculer le *Document Query Ratio (DQR)* qui est obtenu par le nombre de documents d'un corpus, divisé par le nombre de requêtes soumis à ce corpus [Bache et Azzopardi, 2010]. La valeur du *DQR* est importante. Pour une valeur de *DQR* faible, aux alentours de un, on peut s'attendre à ce que tous les documents du corpus aient de bonnes chances d'être présents un nombre de fois significatif dans les résultats de la recherche. À l'inverse, une valeur élevée, correspondant par exemple à beaucoup de documents pour peu de requêtes, réduit les chances de nombreux documents d'être accessibles dans les résultats de recherche.

À titre d'exemple, [Azzopardi et Vinay, 2008] utilise une valeur de *DQR* de 1. L'impact du nombre de requêtes, traduit par la valeur du *DQR* est bien entendu plus ou moins importante suivant la valeur du seuil c au delà duquel on considère un document comme accessible ou non. Une valeur très basse de c se traduira nécessairement par un nombre plus important de documents très peu accessibles. Le seuil c représentant le nombre de document qu'un utilisateur est susceptible de consulter, il peut être très variable en fonction des situations. Nous avons donc testé trois valeurs de c à 10, 20 et 50 documents consultés. Ces valeurs couvrent un large spectre de tâches de recherche d'information.

Par ailleurs, les travaux de Traub et collab. [2016] ont montré que l'algorithme de pertinence utilisé par le moteur de recherche dans un contexte de mesure d'accessibilité pouvait avoir une influence sur les résultats. Bien que, dans notre cas, nous n'évaluons que des requêtes fondées sur des entités nommées et pas des requêtes plus complexes, plus à même d'être impactées par un changement d'algorithme de pertinence, nous avons tout de même effectué nos expérimentations en variant ces algorithmes. Ainsi, les tableaux 5.2 et 5.1 présentent les valeurs mesurées de G et S, avec différentes valeurs du seuil c avec et sans les alternatives découvertes par notre méthode, pour les corpus A et B, avec *BM25* et *TF-IDF*.

Tableau 5.1 – Résultats à $c = 10$, $c = 20$ et $c = 50$ avec *BM25*

	A ($n = 1200$, $DQR = 1,16$)		**B** ($n = 11130$, $DQR = 15,7$)	
	alt	**no alt**	**alt**	**no alt**
c = 10	$G = 0,5550$ $S = 0,9162$	$G = 0,5587$ $S = 0,9231$	$G = 0,8257$ $S = 0,7282$	$G = 0,8586$ $S = 0,7937$
c = 20	$G = 0,5456$ $S = 0,9731$	$G = 0,5485$ $S = 0,9342$	$G = 0,7891$ $S = 0,6410$	$G = 0,8266$ $S = 0,7263$
c = 50	$G = 0,5397$ $S = 1,0177$	$G = 0,5427$ $S = 0,9980$	$G = 0,7502$ $S = 0,9216$	$G = 0,7879$ $S = 0,6251$

Tableau 5.2 – Résultats à $c = 10$, $c = 20$ et $c = 50$ avec *TF-IDF*

	A ($n = 1200$, $DQR = 1,16$)		**B** ($n = 11130$, $DQR = 15,7$)	
	alt	**no alt**	**alt**	**no alt**
c = 10	$G = 0,5506$ $S = 0,9151$	$G = 0,5497$ $S = 0,9379$	$G = 0,8225$ $S = 0,7250$	$G = 0,8587$ $S = 0,7935$
c = 20	$G = 0,54$ $S = 0,9880$	$G = 0,5390$ $S = 0,9426$	$G = 0,7874$ $S = 0,6388$	$G = 0,8267$ $S = 0,7263$
c = 50	$G = 0,5345$ $S = 1,0286$	$G = 0,5330$ $S = 1,0053$	$G = 0,7480$ $S = 0,9271$	$G = 0,7874$ $S = 0,6243$

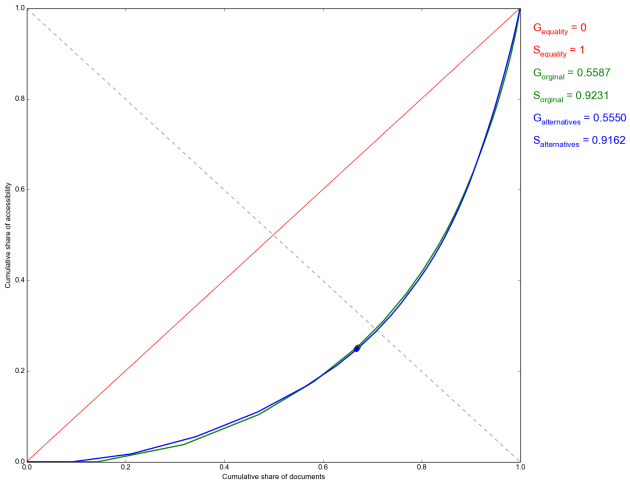
Les résultats sont également présentés sous la forme de courbe de Lorenz dont la construction a été décrite à la section 5.3. Pour le corpus A, la figure 5.10 présente les résultats avec et sans les alternatives pour les valeurs de $c = 10$, $c = 20$ et $c = 50$ avec l'algorithme de pertinence *bm25*. Les résultats avec l'algorithme *tf-idf* sont quant à eux présentés à la figure annexe A.6.

Pour ce corpus A, nous constatons que le coefficient G varie très peu entre les deux séries de résultat, avec et sans alternatives. Les inégalités touchent globalement plus les documents les moins accessibles (valeur de S inférieure à 1), mais là encore, très légèrement. Si l'on observe finement les courbes de Lorenz, nous pouvons toutefois constater que, quel que soit le seuil, l'accessibilité globale des documents les moins accessibles s'améliore légèrement. Quoi qu'il en soit, si l'on s'en tient à ces valeurs et à ces représentations, l'impact de la méthode est difficile à observer sur ce type de corpus. Cet état de fait s'explique par le DQR, très proche de 1. Dans ce cas, en effet, le nombre de requêtes s'approchant du nombre de documents, ces derniers ont beaucoup de chance d'être présents dans les résultats de recherche, qu'ils souffrent ou non du bruit lié au processus de numérisation.

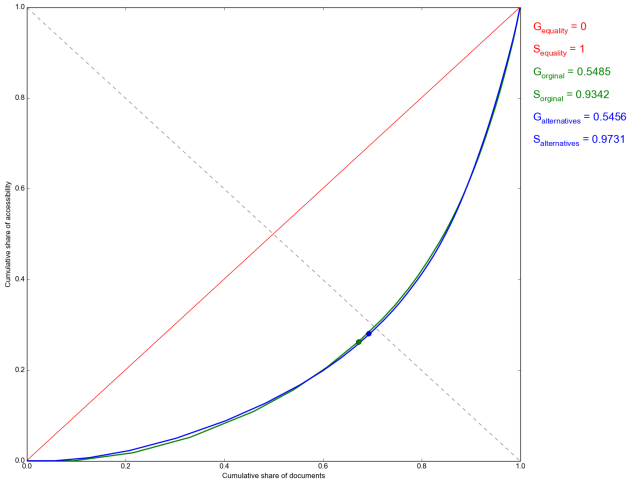
Cependant, comme nous l'avons déjà dit, à elles seules, les représentations en courbe de Lorenz et les valeurs du coefficient G ne suffisent pas à donner une vue complète de l'impact de la méthode. Ainsi, si l'on considère la distribution des documents en fonction du score d'accessibilité, représentée sur la figure 5.11, on peut constater des différences plus notables [16]. Sur ces représentations, le score, toujours un nombre entier, est en abscisse et le nombre de document ayant atteint ce score est en ordonnée. Ainsi, les valeurs représentées à la gauche de chaque graphique témoignent du nombre de documents qui sont très peu accessibles, voire totalement inaccessibles. Pour toutes les valeurs du seuil c, nous pouvons constater que le nombre de documents atteignant des scores élevés augmente. Par ailleurs, le nombre de documents très peu accessibles diminue sensiblement.

L'addition des alternatives dans les requêtes permet donc de diminuer le nombre de document qui ont très peu de chance d'apparaître dans les résultats de recherche. Cette diminution s'accompagnant également d'une augmentation du nombre de documents obtenant des scores élevés et étant donc fortement présents dans les résultats de recherche, nous pouvons comprendre pourquoi l'accessibilité globale, mesurée à l'échelle du corpus par le coefficient G, varie peu.

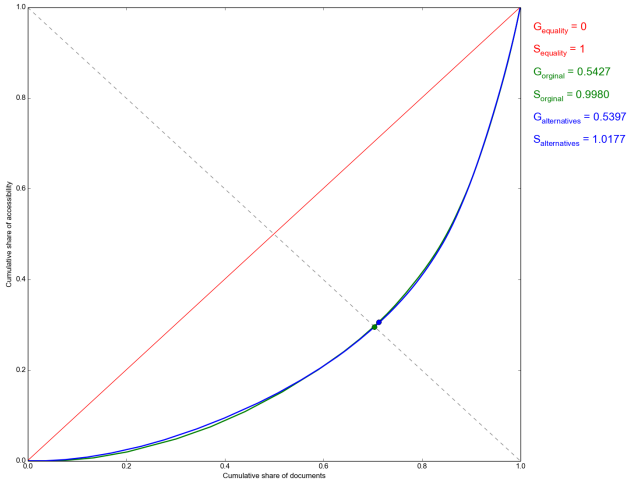
16. Nous représentons ici uniquement les résultats avec *bm25*, les résultats avec *tf-idf* étant très proches.



(a) c = 10

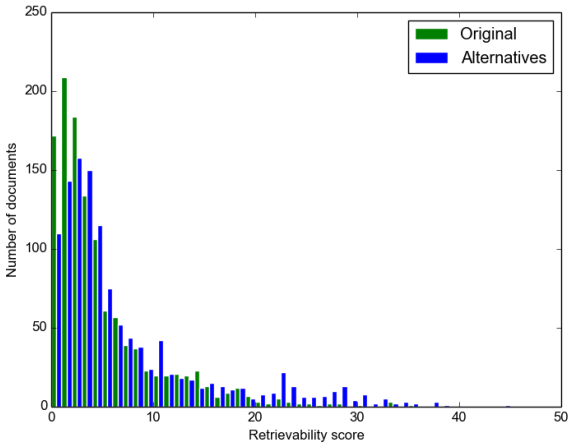


(b) c = 20

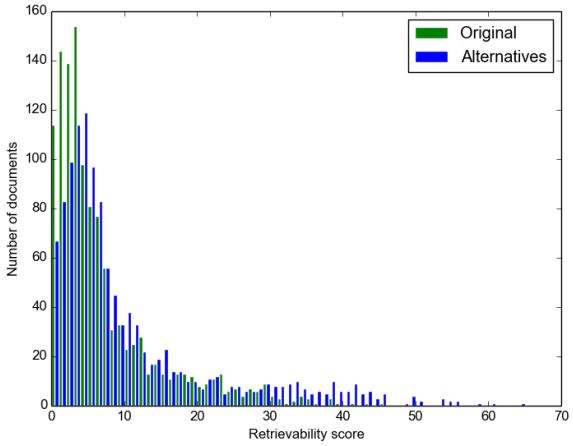


(c) c = 50

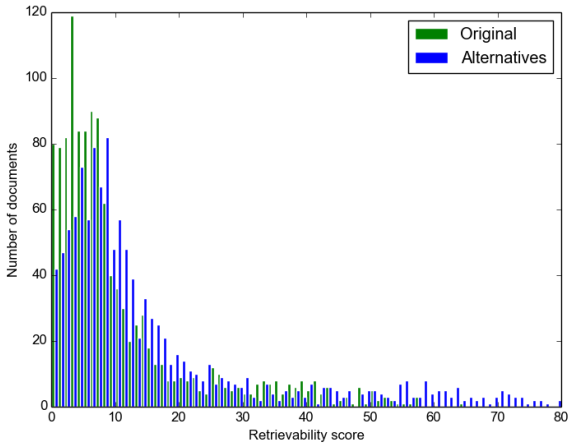
FIGURE 5.10 – Corpus A : Représentation de l'accessibilité pour c = 10, c = 20 et c = 50 avec *bm25*



(a) c = 10



(b) c = 20



(c) c = 50

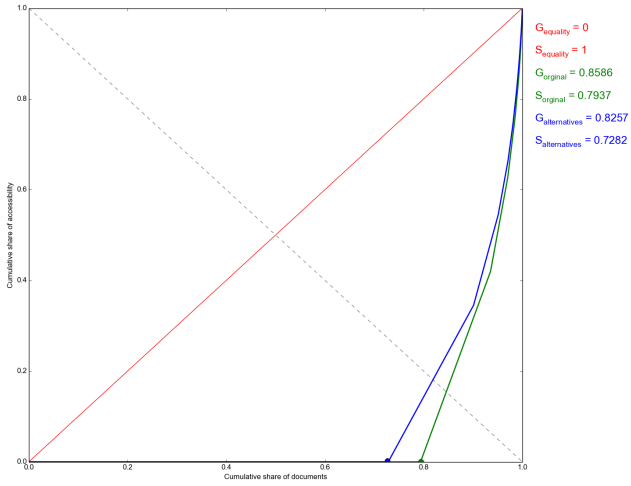
FIGURE 5.11 – Corpus A : Distribution des documents pour c = 10, c = 20 et c = 50 avec *bm25*

Comme il est possible de le visualiser dans les tableaux 5.1 et 5.2 ainsi que sur les figures 5.12 et 5.13, les résultats sur le corpus B ne se présentent pas de la même manière. Nous pouvons d'abord constater qu'avec ce corpus et notre jeu de requêtes, de très nombreux documents restent inaccessibles à ces seuils. Cet état de fait s'explique d'abord par la valeur élevée de DQR qui limite les chances de voir ressortir de nombreux documents dans les résultats de la recherche. Par ailleurs, les valeurs relativement faibles de c que nous considérons, pour un corpus de cette taille ($n = 11130$), produisent un effet de seuil important qui limite, de beaucoup, l'accessibilité de nombreux documents, n'obtenant que des scores très faibles pour notre jeu de requêtes. Nous pouvons constater cet effet de seuil si l'on compare les courbes présentées en figure 5.12 et la courbe présentée à la figure annexe A.8 pour laquelle nous avons supprimé le seuil c. Dans ce cas, le nombre de documents totalement inaccessibles diminue fortement.

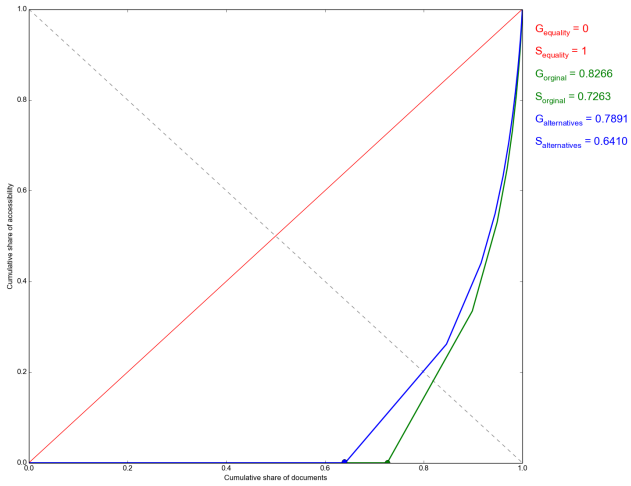
Pour la part de document qui obtiennent une valeur d'accessibilité, les valeurs du coefficient G sont cette fois significativement inférieures lorsque que l'on inclue les alternatives dans les requêtes soumises au moteur de recherche, aussi bien avec l'algorithme BM25, qu'avec l'algorithme TF-IDF [17]. Pour toutes les valeurs du seuil c les documents apparaissent globalement plus équitablement accessibles. Les documents obtenant les scores les plus faibles sont globalement plus présents dans les résultats de recherche qu'ils ne le sont sans alternatives. Nous pouvons confirmer ces résultats par les distributions du nombre de documents par score d'accessibilité présentées en figure 5.13 [18]. Sur ces représentations, le nombre de documents inaccessibles diminue alors que le nombre de documents obtenant des scores plus élevés augmente de manière relativement homogène.

Ainsi, que ce soit pour les corpus A et B, la méthode développée ici permet de mettre en lumière les biais produits par le processus de numérisation lors de recherche en plein texte impliquant des entités nommées. Les différences relevées sur le comportement du moteur si l'on inclue ou non les alternatives aux requêtes initiales témoignent de l'impact que peuvent avoir ces biais sur la pratique des utilisateurs. Si certains documents sont tout simplement rendus inaccessibles par la mauvaise qualité de l'OCR, celle-ci impacte également le calcul de la pertinence des documents. Les algorithmes utilisés pour calculer cette pertinence sont en effets fondés sur la fréquence d'apparition des termes de la requête dans le document.

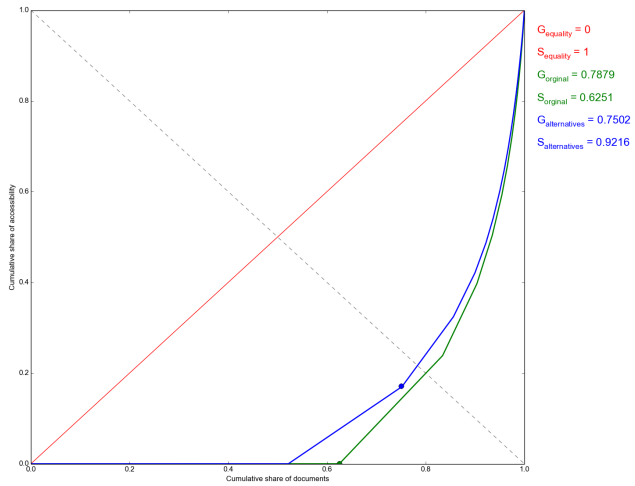
17. Dont les résultats sont présentés en annexe A.7.
18. Sur ces représentations, pour des raisons de clarté, l'échelle adoptée est logarithmique.



(a) c = 10

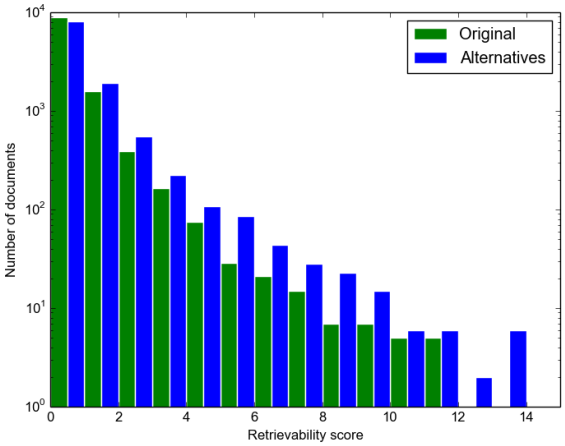


(b) c = 20

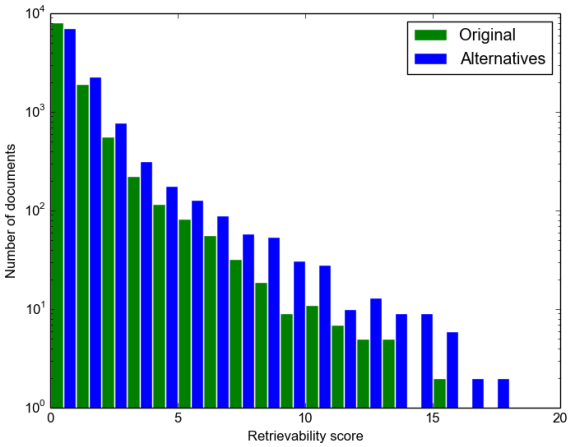


(c) c = 50

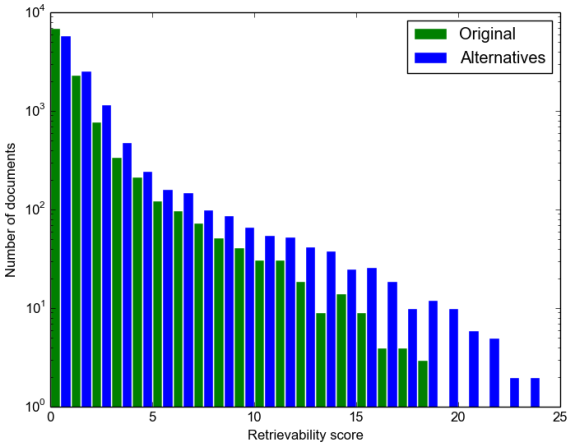
FIGURE 5.12 – Corpus B : Représentation de l'accessibilité pour c = 10, c = 20 et c = 50 avec *bm25*



(a) c = 10



(b) c = 20



(c) c = 50

FIGURE 5.13 – Corpus B : Distribution des documents pour c = 10, c = 20 et c = 50 avec *bm25*

5.4.2 Intégration aux ressources : implémentation en bibliothèque numérique

La méthode que nous avons présentée pourrait, comme nous l'avons déjà dit, être appliquée à la correction automatique d'erreurs d'OCR. Cet objectif, toutefois, supposerait des développements et des évaluations supplémentaires. En effet, à l'issue de notre méthode, il n'y a pas de solutions pour déterminer facilement quelle est, parmi toutes les versions d'une même entité, la version non bruitée, si elle existe. Certaines méthodes de désambiguïsation pourraient être appliquées pour atteindre cet objectif, comme les méthodes fondées sur des bases de connaissances extérieures [Bunescu et Paşca, 2006; Cucerzan, 2007]. Encore faudrait-il, dans ce cas, que les entités soient connues pour faire partie de ces bases de connaissance.

Dans notre contexte, la méthode a été avant tout mise en place pour mettre en lumière le problème des erreurs d'OCR et des variations orthographiques et pour donner des moyens à l'utilisateur de les visualiser et de les inclure dans sa stratégie de recherche d'information. Afin d'atteindre ces objectifs, nous avons produit deux implémentations différentes, une intégrée dans une bibliothèque numérique, l'autre autonome.

Le premier cas d'usage de cette méthode découle directement des travaux et des observations décrits plus haut, il s'agit d'un usage dans une bibliothèque numérique. Pour pouvoir fournir aux utilisateurs les moyens d'inclure les variations possibles d'une entité à sa stratégie de recherche, notre plateforme expérimentale fournit les fonctionnalités suivantes :

— proposer dès l'étape de saisie de la requête des variations possibles d'un terme ;
— donner la possibilité à l'utilisateur d'inclure tout ou partie de ces termes dans sa recherche ;
— signaler les entités nommées et leur différentes variations possibles lors de la consultation d'un document.

Il faut noter ici que dans le contexte d'une bibliothèque numérique, notre méthode est utilisée par le concepteur de la ressource qui a la responsabilité de paramétrer la méthode. Naturellement, les utilisateurs ne pourront pas modifier ces paramètres et donc agir sur les résultats de la méthode. Le choix des paramètres résultera donc d'un compromis des concepteurs de la ressource.

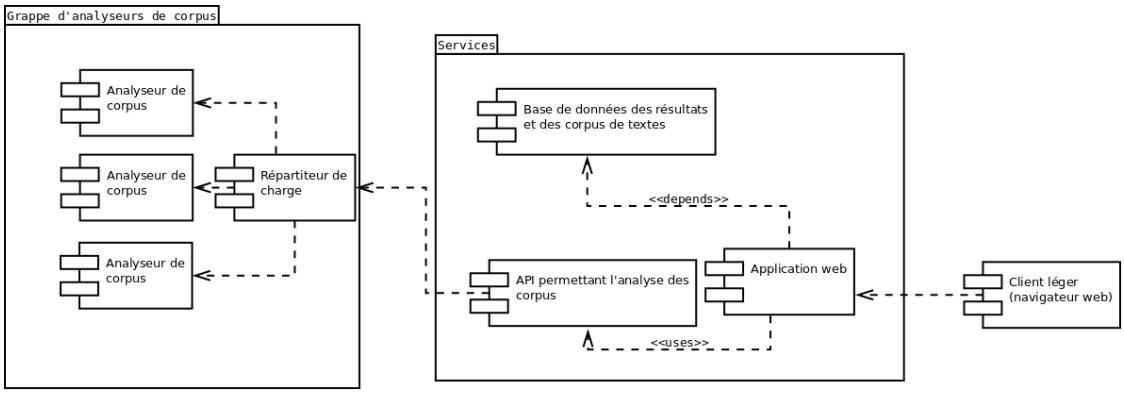
5.4.3 Impact sur la pratique : implémentation libre

Notre méthode, implémentée dans une bibliothèque numérique ne peut être utilisée que pour les corpus de texte indexés par cette bibliothèque numérique. Suite à des discussions avec nos collègues historiens, il est apparu qu'il pouvait être intéressant de l'ouvrir plus largement et de la rendre utilisable par tout un chacun. L'objectif, ici, est de permettre aux historiens de traiter facilement des corpus à leur disposition. Nous avons donc mis en place une application fondée sur notre méthode, capable de fonctionner seule, en dehors du contexte d'une bibliothèque numérique. Cette implémentation pose les contraintes spécifiques suivantes :

— facilité d'accès et d'installation ;
— import des corpus et gestion des corpus ;
— gestion du paramétrage ;
— gestion des temps de calcul ;
— lisibilité des résultats.

Afin de faciliter l'accès à notre méthode et simplifier au maximum son installation, nous avons choisi de mettre en place une application web. Ce choix présente l'avantage de ne pas imposer d'étapes d'installation et de réduire le risque d'erreurs durant le fonctionnement compte tenu du nombre relativement important de bibliothèques logicielles requises pour le bon fonctionnement de l'application. Il permet par ailleurs de faire évoluer et de garder à jour les outils beaucoup plus facilement. Par ailleurs, il autorise la possibilité d'effectuer les calculs coûteux sur des serveurs spécifiques. En contrepartie, il est nécessaire de maintenir des serveurs d'hébergement et de calcul et de gérer les processus concurrents si plusieurs utilisateurs lancent des calculs en parallèles. L'architecture de la solution est présentée à la figure 5.14.

FIGURE 5.14 – Architecture de la version autonome de l'outil.



Le choix d'une application web complexifie également la gestion des corpus, dans la mesure où ceux-ci doivent rester privé, l'application doit fonctionner avec un mécanisme d'authentification et d'espace personnel. L'import des corpus est également plus complexe puisqu'il faut passer par une étape de téléversement depuis le navigateur vers le serveur de l'application. Cette étape est visible à la figure 5.15.

Pour le reste des contraintes que nous avons évoquées plus haut, le choix d'une application web n'occasionne pas de difficultés particulières. Le paramétrage est effectué sur une page dédiée avant de lancer la procédure, dont la mise en forme est visible figure 5.16. Les paramètres laissés à la discrétion des utilisateurs sont les suivants :

— n : nombre de liens générés entre entités en fonction de sim_{ctx} ;
— t : seuil de validation de chaque entité en fonction de sim_{edit} ;
— calcul de sim_{edit} : méthode de calcul appliquée (distance de Levenshtein, coefficient de Dice) ;
— c : temps de calcul octroyé à l'algorithme de clustering par nœud (un temps plus long peut réduire le nombre de clusters approximatifs) ;

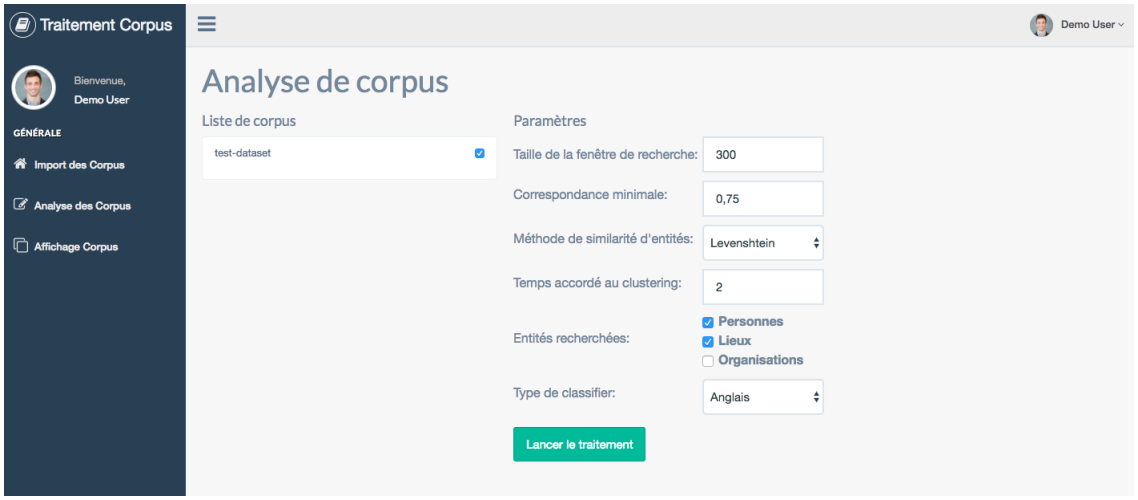
FIGURE 5.15 – Import des corpus dans la version autonome de l'application

— langue du corpus : langue du corpus source afin d'affiner la détection des entités nommées ;

— type de classifier : choix du classifier pour la détection des entités nommées ;

— entités calculées : type d'entités nommées pour lequel la méthode est appliquée (choix multiple entre personne, lieu et organisation).

FIGURE 5.16 – Paramétrage de l'analyse dans la version autonome de l'application



Une fois les traitements achevés, l'utilisateur peut visualiser les résultats. Il a en particulier l'opportunité de comparer plusieurs mots d'un même groupe dans leurs contextes respectifs, c'est à dire dans le texte original, pour être en mesure de vérifier qu'il s'agit bien du même terme, voire de comprendre l'origine d'une écriture différente, erreur d'OCR ou simple variation (voir figure 5.17). Il peut télécharger un fichier contenant les résultats pour les analyser dans un outil externe. Cette version autonome rend donc possible d'utiliser cette méthode à d'autres fins que la recherche d'information. Tous les cas de corpus comportant des entités nommées risquant d'être écrites de manière légèrement différentes peuvent être traités. Dans un contexte de transcription de source par exemple, particulièrement lorsqu'il y a de nombreux transcripteurs, cet outil peut être un moyen de trouver des noms interprétés de manière différente par les transcripteurs.

FIGURE 5.17 – Visualisation des variations détectées dans leur contexte

Cet écran permet de sélectionner deux entités et de comparer leurs différentes versions au sein de leur environnement, l'application précise le nom du fichier original et présente les extraits de texte dans lesquels ont été trouvées ces occurences.

5.5 Conclusion

En nous intéressant aux processus informatiques que nous situons dans le contexte de production à travers cette étude de cas, nous avions pour principal objectif de montrer que les traitements qui y sont opérés peuvent avoir des conséquences. Il s'agissait toutefois de dépasser ce seul constat général qui pour le cas de l'OCR peut être démontré par un simple alignement entre le résultat produit par le processus et une version manuellement établie du texte original, comme on peut le voir à la figure 5.18. Dans notre cas, il était plutôt question de montrer l'impact de ces traitements sur la pratique de la recherche d'information et fournir des outils qui permettent aux utilisateurs d'en prendre conscience et de les visualiser.

La méthode décrite dans ce chapitre répond à ces objectifs. Dans un premier temps, les visualisations et résultats chiffrés que nous avons produits sur les conséquences des erreurs d'OCR sur l'accessibilité sont un moyen de mesurer l'importance de ces biais du point de vue de l'utilisateur et pas seulement du point de vue du système. Dans un second temps, elle permet, sans a priori, de rechercher et regrouper les différentes versions d'une même entité et d'afficher à l'utilisateur, lorsqu'il saisit une recherche, des alternatives potentielles. Ces fonctionnalités rendent possible une recherche fondée sur des entités nommées avec une plus grande couverture de requêtes et minimise le risque de passer à côté de documents intéressants.

Cette méthode a toutefois ses limites. Elle est d'abord très dépendante des systèmes de détection d'entités nommées, qu'elle ne permet pas d'améliorer. Elle est par ailleurs coûteuse en temps de calcul pour les grands corpus. Enfin, elle peut, en cas de paramétrage inadapté, générer du bruit dans les résultats qui peut nuire à son efficacité

FIGURE 5.18 – Exemple d'un alignement OCR et vérité terrain

```
1 OCR:    De fait, l'@eutrier s'était cuit au jeu des affaires; sa maigre moustache,
2 GT :    De fait, @@Feutrier s'était cuit au jeu des affaires; sa maigre moustache,
```

Le processus d'alignement permet de comparer aisément un texte corrigé et le résultat produit par un traitement quelconque, les différences sont corrigés par l'introduction du caractère @.

globale. Sa principale limite est de ne fonctionner que sur des entités nommées. Cette limite dépend d'une contrainte que nous nous sommes nous-mêmes fixés au regard de notre contexte et de l'importance de cette catégorie de mot pour la recherche d'information. Cependant, rien n'empêche de l'appliquer à d'autres catégories de mots en remplaçant la détection d'entités nommées par un autre processus de détection. Il est par exemple possible de l'appliquer à une catégorie grammaticale, plutôt qu'aux entités nommées en utilisant un algorithme de détection adapté. Le volume de mots considéré doit cependant être limité si l'on veut éviter de coûteux temps de calcul.

Dans notre contexte, notre démarche a toutefois permis deux types d'implémentation, l'une intégrée à notre plateforme expérimentale de bibliothèque numérique, l'autre autonome. L'intérêt de la première est évidente au regard de nos objectifs de recherche, l'évaluation de son intérêt est présentée dans notre chapitre 6. La seconde est quant à elle utile pour d'autres usages que la recherche d'information. Nous avons déjà évoqué le cas de la transcription, mais elle peut être plus largement utile pour toute tâche impliquant une stratégie de lecture distante. Ces méthodes de recherche, que nous avons évoquées dans notre chapitre 2, nécessitent des corpus les moins bruités possibles. Qu'il s'agisse de trouver la première apparition d'un terme ou de produire une représentation de la fréquence de l'emploi de telle ou telle expression, des corpus propres sont indispensables pour limiter les biais de ces approches. Ainsi, pour des usages quotidiens de recherche d'information comme pour des approches plus techniques d'histoire numérique, la réflexion et l'outillage de la recherche en histoire face aux risques de biais technologiques doivent être poursuivis.

5.6 Références

ADAMO, J. M. 1980, «Fuzzy decision trees», *Fuzzy sets and systems*, vol. 4, n° 3, p. 207–219. 131

AZZOPARDI, L. et V. VINAY. 2008, «Retrievability : an evaluation measure for higher order information access tasks», doi :10.1145/1458082.1458157. 140, 141, 143, 147

BACHE, R. et L. AZZOPARDI. 2010, «Improving Access to Large Patent Corpora.», *Trans. Large-Scale Data-and Knowledge-Centered Systems*, vol. 2, p. 103–121. 147

BAUM, L. E. et T. PETRIE. 1966, «Statistical inference for probabilistic functions of finite state Markov chains», *The annals of mathematical statistics*, vol. 37, n° 6, p. 1554–1563. 132

BUNESCU, R. et M. PAŞCA. 2006, «Using encyclopedic knowledge for named entity disambiguation», dans *11th conference of the European Chapter of the Association for Computational Linguistics*. 154

Chomsky, N. 1956, «Three models for the description of language», *IRE Transactions on information theory*, vol. 2, n° 3, p. 113–124. 131

Cucerzan, S. 2007, «Large-scale named entity disambiguation based on Wikipedia data», dans *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 154

Damgaard, C. et J. Weiner. 2000, «Describing inequality in plant size or fecundity», *Ecology*, vol. 81, n° 4, p. 1139–1142. 141

Dernoncourt, F., J. Y. Lee et P. Szolovits. 2017, «NeuroNER : an easy-to-use program for named-entity recognition based on neural networks», *arXiv :1705.05487 [cs, stat]*. 135

Dice, L. R. 1945, «Measures of the amount of ecologic association between species», *Ecology*, vol. 26, n° 3, p. 297–302. 136

Finkel, J. R., T. Grenager et C. Manning. 2005, «Incorporating non-local information into information extraction systems by gibbs sampling», dans *Proceedings of the 43rd annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 363–370. 135

Firth, J. R. 1957, «Applications of general linguistics», *Transactions of the Philological Society*, vol. 56, n° 1, doi :10.1111/j.1467-968X.1957.tb00568.x, p. 1–14. 130, 132

Gastwirth, J. L. 1972, «The Estimation of the Lorenz Curve and Gini Index», *The Review of Economics and Statistics*, vol. 54, n° 3, p. 306–16. 141

Gooding, P. 2014, «Exploring Usage of Digital Newspaper Archives through Web Log Analysis : A Case Study of Welsh Newspapers Online», *Digital Humanities 2014*. 129

Lafferty, J., A. McCallum et F. Pereira. 2001, «Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data», *Departmental Papers (CIS)*. 135

Lorenz, M. O. 1905, «Methods of Measuring the Concentration of Wealth», *Publications of the American Statistical Association*, vol. 9, n° 70, doi :10.2307/2276207, p. 209–219. 141

Mikolov, T., K. Chen, G. Corrado et J. Dean. 2013, «Efficient estimation of word representations in vector space», *arXiv preprint arXiv :1301.3781*. 132, 136, 137

Neudecker, C. 2014, «Named Entity Recognition for digitised newspapers – Europeana Newspapers», URL http://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/. 135

Palla, G., I. Derényi, I. Farkas et T. Vicsek. 2005, «Uncovering the overlapping community structure of complex networks in nature and society», *Nature*, vol. 435, n° 7043, p. 814–818. 138

Pennington, J., R. Socher et C. Manning. 2014, «Glove : Global vectors for word representation», dans *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543. 132

Rodriquez, K. J., M. Bryant, T. Blanke et M. Luszczynska. 2012, «Comparison of named entity recognition tools for raw OCR text.», dans *The Conference on Natural Language Processing (KONVENS)*, p. 410–414. 135

Sørensen, T. 1948, «A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons», *Biol. Skr.*, vol. 5, p. 1–34. 136

Traub, M. C., T. Samar, J. van Ossenbruggen, J. He, A. de Vries et L. Hardman. 2016, «Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus», JCDL '16, ACM, New York, NY, USA, p. 7–16, doi :10.1145/2910896.2910907. 143, 147

Chapitre 6

Réinvestir : vers un outil dédié à l'enseignement de la recherche d'information pour l'historien

Résumé

Ce chapitre présente un outil et une démarche de formation à une approche critique de la recherche d'information pour l'historien. Le logiciel proposé, nommé Brightbox, réinvestit les résultats des chapitres précédents dans l'optique de démontrer aux étudiants les biais méthodologiques de la recherche d'information. Il est conçu pour être utilisé dans un cadre de formation présentielle et est largement fondé sur l'expérimentation des biais par les étudiants, apprentis historiens. Il permet dans un premier temps de montrer les conséquences de certains traitements des contextes de production et d'exécution. Dans un second temps, en réemployant les indicateurs d'usage que nous avons précédemment mis en place, il produit des représentations de l'usage qui servent de base à une réflexion sur les approches de recherche d'information les mieux adaptées aux différents besoins des utilisateurs.

Sommaire

6.1 Introduction

Durant les chapitres précédents, nous avons montré qu'en matière de recherche d'information en ligne, les biais méthodologiques auxquels doivent faire face les historiens ou les apprentis historiens peuvent être nombreux. Nous nous sommes intéressés à un cas particulier et mis en lumière le rôle majeur que doit tenir l'informaticien comme contributeur indispensable de la conception de ressources numériques pertinentes. Ceci étant, les problématiques que nous avons soulevées sont plus larges que le seul cas concret que nous venons de décrire, une réponse informatique ne peut à elle seule permettre de les résoudre. D'ailleurs, il semble important de rappeler qu'une réponse technique a toutes les chances de créer encore plus de difficultés si elle ne s'accompagne pas d'un volet heuristique, permettant aux utilisateurs de bien comprendre les effets du processus informatique mis en place.

Le dernier angle de réflexion que nos travaux ont adressé est donc, plus précisément, celui de la compétence et de la formation. À la lumière des premières expérimentations que nous avons menées et des solutions que nous avons mises en œuvre, il nous a semblé évident que le point de jonction entre les contributions de l'informatique et la pratique de la recherche historique devait être le terrain pédagogique. L'enseignement est en effet un des terrains sur lesquels se développent de nombreuses expérimentations et innovations portées par les humanités numériques [Bourgatte, 2017]. Ces formations sont généralement destinées aussi bien aux apprentis historiens, dans leur cursus universitaire, qu'aux chercheurs aguerris qui ressentent un besoin de formation au numérique. Nous avons adopté ce point de vue en gardant à l'esprit l'ambition de viser un public aussi large que possible, bien que nous ayons limité notre expérimentation à des étudiants en cours de formation pour des raisons pratiques.

Le savoir méthodologique de l'historien permet sans conteste de former les étudiants à la formulation de questions de recherche, à la détermination de problématiques et à la recherche et à la sélection des sources pertinentes pour leur étude. Néanmoins, il est clair désormais que ces activités passent de plus en plus par les moteurs de recherche et par l'accès numérique à l'information qu'elle soit primaire ou secondaire. Les phénomènes que nous avons tentés de montrer jusqu'à présent ont une prise réelle sur l'activité des étudiants et relèvent ainsi de l'*information literacy*, qui peut être traduit par littératie informationnelle en français. Cette expression désigne la capacité à comprendre et utiliser l'information aussi bien dans la vie courante que dans le cadre de travaux universitaires.

Bien entendu, les thématiques de l'accès à l'information dans un contexte numérique et ses conséquences peuvent être intégrées à la formation des historiens mais restent insuffisamment développées. Ce problème n'intéresse pas seulement l'enseignant en histoire, ou dans d'autres disciplines d'ailleurs, mais également le bibliothécaire, dont l'orientation des usagers dans la masse d'information à leur disposition fait partie des principales missions. Des études, menées dans le contexte des bibliothèques universitaires ont soulevé les difficultés des étudiants, au sens large, à comprendre les différents aspects d'un processus de recherche d'information. Dans un rapport édité par l'association américaine des bibliothèques, Duke et Asher, 2012 écrivait :

> *The majority of students – of all levels – exhibited significant difficulties that ranged across nearly every aspect of the search process. [...] Almost*

> *without exception, IWU students exhibited a lack of of understanding of search logic, how to build a search to narrow or expand results, how to use subject headings, and how various search engines (including Google) organize and display results. As one student mentioned while conducting a search of the library's online catalog, "Apparently you don't have much on rock and roll" not realizing that if she changed her search term (i.e. to rock music), she would have encontered many excellent sources for her assignment*[1].

Évidemment, c'est le défaut de formation qui est ici à interroger. Bibliothécaires autant qu'enseignants ont sans doute, comme le rappelait East, 2005; Kolowich, 2011, sur-estimé les compétences des étudiants en matière de recherche d'information nous conduisant une nouvelle fois à détruire le mythe des *digital natives*[2]. Si des outils numériques existent sous forme de cours en ligne ou de blogs [Williams, 2010], il reste de nombreux efforts à accomplir en matière de formation à la recherche d'information. Ces efforts dépassent nos objectifs, puisque c'est bien ici le processus global de recherche, aussi bien dans un environnement physique que numérique qui est en cause.

Dans notre contexte, sur le terrain pédagogique, notre contribution est naturellement plus limitée. À la suite de réflexions plus générales dont celle de [Cardon, 2013], nos chapitres précédents ont mis en lumière les problématiques techniques, méthodologiques, voire cognitives qui limitent notre compréhension de l'accès à l'information. La démarche et l'outil que nous proposons se situent au centre de la distinction entre *information seeking*, l'ensemble des processus et pratiques des utilisateurs pour répondre à un besoin d'information et *information retrieval*, les méthodes et techniques informatiques qui permettent au système de répondre à ces besoins, que nous avons déjà expliquées. C'est le fossé entre les besoins des utilisateurs et les technologies utilisées pour y répondre, matérialisé par les relations de causes à effets entre les critères de recherche (*inputs*) saisis par l'utilisateur, et les résultats (*outputs*) fournis par le système qui sont au cœur de la démarche que nous souhaitons présenter ici.

6.2 Objectifs de formation et fonctionnement général de Brightbox

6.2.1 Objectifs pédagogiques

L'outil que nous détaillons et la démarche de formation qui l'accompagne ont l'ambition de fournir les moyens de donner à voir et faire comprendre les traitements subis par les données, des *inputs* aux *outputs*, depuis la requête saisie ou le filtre sélectionné jusqu'à l'affichage et la consultation des résultats. Il a avant tout l'objectif de faire réfléchir les étudiants dans ce contexte de boite noire et de les amener à mener un exercice réflexif sur leurs pratiques de recherche. Plus précisément, nos objectifs pédagogiques peuvent être décomposés en quatre catégories, **transformation des données**, **effets**

1. Duke et Asher, 2012, p. 76.
2. Le constat est partagé par d'autres, par exemple par Moisson, 2011.

de boites noires algorithmiques, **comportement utilisateur** et enfin **indicateurs et visualisations**.

Transformations subies par les données

Les processus qui aboutissent à l'indexation de documents les rendant exploitables par un moteur de recherche se résument souvent à une série de transformations. Les données initiales, dans notre cas du texte, sont transformées par une série d'opérations algorithmiques qui produisent un texte, ou plutôt une suite de mots, représentative du document, facilement manipulable par les algorithmes de recherche d'information. Ces diverses opérations, que nous détaillerons plus précisément en section 6.3, ont un impact très important sur la recherche d'information. À titre d'exemple, la phrase « *Ces transformations impactent beaucoup les données et l'indexation* » peut devenir « *transform impact beaucoup don index* », une fois les mots trop communs retirés et les autres mots réduits à leur racine. Dans ce cas, on peut rapidement comprendre que le moteur de recherche ne pourra pas faire de différence entre « un index » et « l'indexation » par exemple. Il est ainsi important de donner à voir ces transformations aux étudiants car elles expliquent à elles seules une grande partie du fonctionnement du moteur de recherche.

Algorithmes de recherche d'information

Le deuxième aspect de la problématique qui mérite d'être présenté et expliqué aux étudiants est le mécanisme de recherche d'information et en particulier le classement des résultats de recherche d'information. En effet, les algorithmes de recherche d'information utilisés dans les moteurs de recherche textuels reposent sur des logiques relativement simple mais qui ont un impact considérable. Le tri par pertinence, souvent activé par défaut s'appuie sur ces logiques simples mais puissantes. Ici, il est important que les étudiants puissent visualiser les effets de certains des paramètres de ces algorithmes de classement de résultats. Il est nécessaire qu'ils comprennent que ces classements résultent de choix de conception qui leur échappent et dont les objectifs et le fonctionnement ne sont pour ainsi dire jamais précisés. Sur la base d'algorithmes simples (*TF-IDF*,*BM25*), l'objectif pédagogique se résume ainsi à montrer divers moyens de manipulation du classement des résultats au regard de divers objectifs. Il est donc question de montrer que certaines plateformes sont, du point de vue leur pertinence pour l'interprétation, particulièrement inadaptées à une pratique scientifique, qui demande clarté et reproductibilité de la recherche.

Comportement utilisateur en situation de recherche d'information

Notre troisième objectif pédagogique et de montrer aux participants qu'en plus des effets strictement dus au système informatique, la pratique de l'utilisateur à un impact sur la qualité d'une recherche d'information. La formulation des requêtes comme l'analyse des résultats et la sélection des documents pertinents sont fondamentaux. L'objectif est ici de révéler les pratiques et de les représenter statistiquement ou graphiquement. Sur la base de ces constats, il s'agit d'amener les étudiants à développer des stratégies

de recherche pertinentes en fonction d'un besoin d'information clairement défini. Cet objectif est fortement lié aux méthodes et aux exigences de la discipline d'origine des participants. Le discours pédagogique doit donc être adapté.

Indicateurs et visualisations

Enfin, notre démarche a pour objectif de sensibiliser les étudiants, particulièrement les étudiants en histoire, aux indicateurs et aux représentations visuelles. En effet, le développement de l'histoire numérique et plus largement des humanités numériques pousse l'historien à manipuler plus régulièrement qu'auparavant des représentations graphiques, statistiques ou cartographiques, comme nous avons pu le voir dans notre chapitre 2. Brightbox utilisant de telles représentations, par exemple des représentations de réseau, pour expliquer les problématiques auxquelles nous nous attaquons, il est pertinent d'expliquer comment sont construites ces représentations. Leur compréhension par les étudiants peut en être facilitée parce que les visualisations qui servent à l'explication sont issues d'une expérience tout juste terminée.

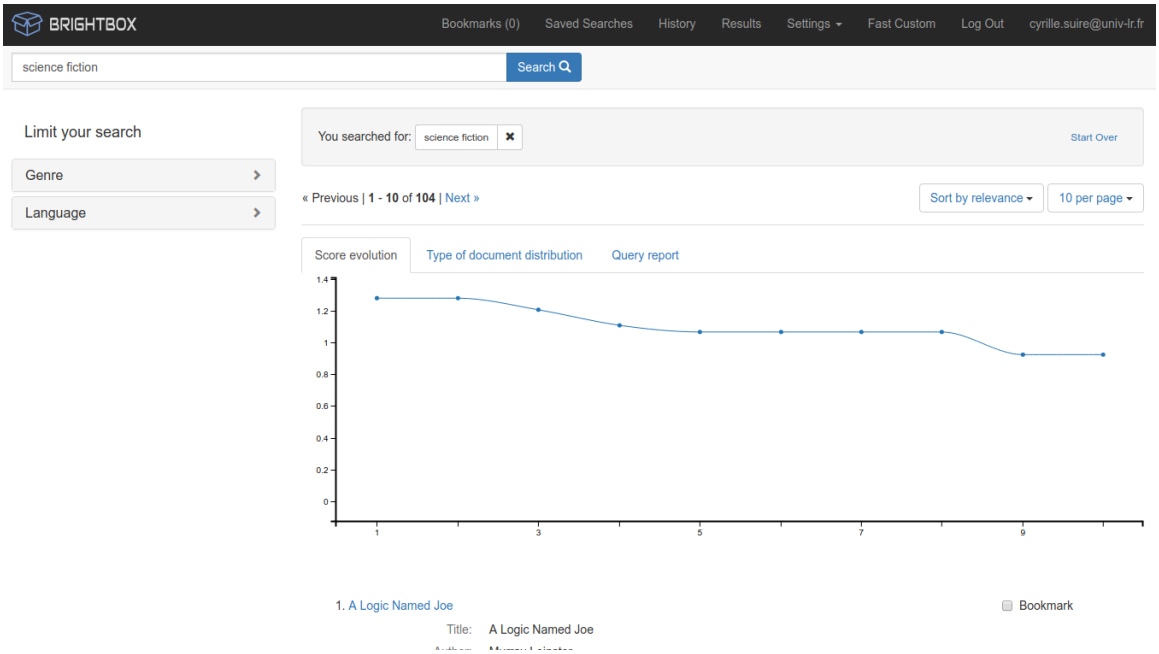
6.2.2 Stratégie pédagogique et mode de fonctionnement

Pour atteindre ces objectifs, nous avons conçu et mis en place une démarche d'enseignement ou de formation à la recherche d'information en ligne. Cette démarche est de portée générale, même si elle est illustrée dans le contexte particulier adapté aux travaux d'un historien. Nos réflexions préliminaires et les échanges que nous avons pu mener avec des collègues nous ont conduit à développer une approche qui repose essentiellement sur une stratégie pédagogique fondée sur l'expérimentation. Il nous a semblé que sur un tel sujet, il était important de limiter au maximum les discours pédagogiques purement transmissifs dans la mesure ou le discours théorique est assez éloigné de la formation de base des étudiants visés. Par ailleurs, les objectifs de la formation et son influence potentielle se situent sur le terrain de la pratique. Nous avons donc choisi de faire reposer la formation sur un dispositif expérimental, autrement dit de mettre au maximum les étudiants en situation de travaux pratiques. De cette manière, que l'on travaille sur les biais du côté du moteur de recherche, dans notre premier mode de fonctionnement, ou sur les biais liés cette fois au comportement de recherche d'information, notre second mode de fonctionnement, la stratégie pédagogique repose, autant qu'il est possible, sur l'expérimentation.

Ainsi, même lorsque nos objectifs requièrent des phases d'explication plus théoriques, nous fondons toujours notre stratégie sur l'expérimentation et le constat du fonctionnement par les étudiants eux-mêmes. Dans ce mode de fonctionnement, le formateur est libre de modifier le fonctionnement du système autant que nécessaire, les requêtes des étudiants sont soumises à ces variations mais leur comportement de recherche, bien qu'il soit observé et enregistré, importe peu. L'enjeu est bien qu'ils puissent constater les effets des divers paramétrages possibles. Ce mode offre donc des représentations dédiées à ces objectifs et en particulier une interface de comparaison des résultats de requêtes (voir figure 6.1 et 6.2). Lorsque les étudiants ont pu constater les changements, il est possible de poser des hypothèses sur le fonctionnement du ou des paramètres que

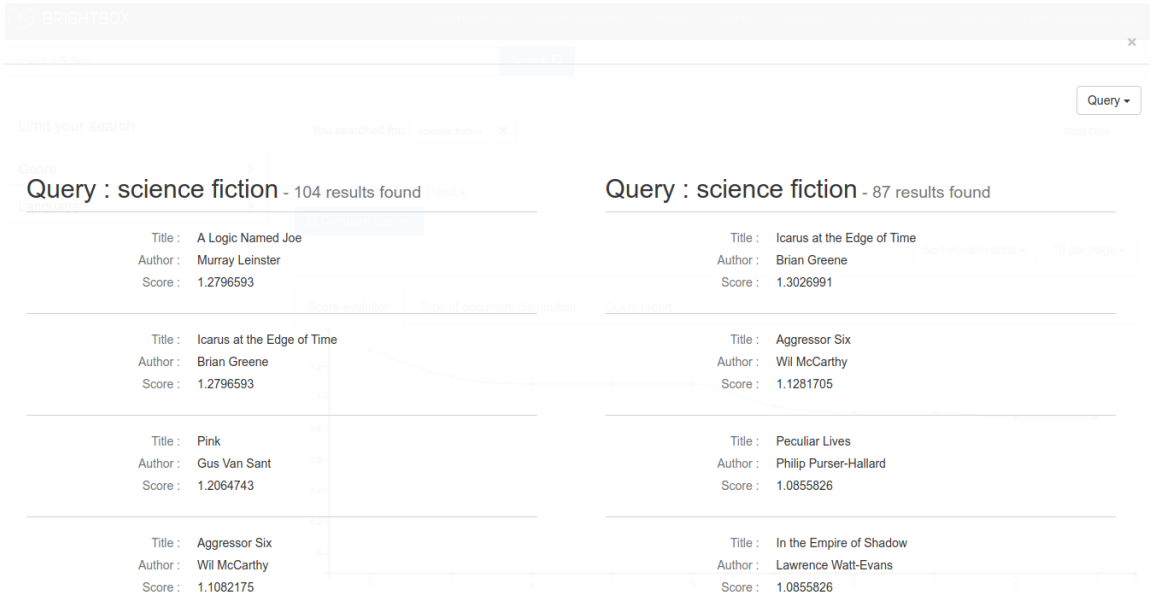
le formateur a choisi de faire varier et d'en débattre. Enfin, une phase d'explication plus théorique vient conclure la discussion.

FIGURE 6.1 – Interface de visualisation des résultats de recherche



Sur cette figure se trouve un exemple de courbe représentant l'évolution du score de pertinence pour les résultats visibles sur la page.

FIGURE 6.2 – Interface de comparaison des résultats de recherche



Cette interface présente deux listes de résultats d'une même requête dans deux contextes différents. Elle permet de comparer aisément l'impact du contexte sur les résultats de la requête.

Le second mode de fonctionnement de notre outil est quant à lui plus naturellement expérimental puisqu'il s'agit d'observer, de mesurer et d'expliquer le comportement des étudiants face à un besoin d'information. Cet objectif de formation requiert évidemment que les étudiants manipulent l'outil et accomplissent une ou plusieurs recherches pour que l'on puisse observer les résultats et les réinvestir dans la discussion. Dans ce mode de fonctionnement de l'outil, des tâches de recherche d'information sont affectées aux étudiants. Ils disposent alors du temps nécessaire pour les accomplir. Ici, le formateur ne modifie pas le fonctionnement du moteur de recherche, le paramétrage reste le même du début à la fin de la tâche[3]. Le comportement des participants est observé et enregistré et des indicateurs de plus haut niveau sont générés par le système. Cette première phase est bien entendu expérimentale, dans la mesure où les étudiants manipulent l'outil. Cependant, elle est complétée par une seconde phase d'analyse des résultats ou les visualisations et les indicateurs, individuels et collectifs peuvent être observés et commentés. Les explications provenant de l'enseignant sont ainsi directement connectées à l'expérience de recherche d'information que viennent d'accomplir les participants.

6.2.3 Implémentation technologique

L'outil logiciel que nous avons développé et utilisé lors de nos expérimentations, que nous avons nommé Brightbox, repose sur l'architecture logicielle décrite dans notre chapitre 3. Il dispose ainsi de toutes les fonctionnalités de notre plateforme expérimentale, sur les aspects bibliothèques numériques et sur les aspects observation. Cependant, les objectifs pédagogiques et techniques que nous avons déterminés ont imposé d'en modifier l'architecture sur les points suivants.

Gestion des rôles

Afin de permettre à l'enseignant de modifier les nombreux paramètres disponibles dans Brightbox tout en ne permettant pas aux étudiants d'y avoir accès, nous avons mis en place une gestion des rôles. Le rôle enseignant donne accès aux diverses fonctions de paramétrage. Il donne également accès à l'interface servant à déclarer de nouvelles tâches à effectuer pour les étudiants et permet de visualiser les résultats des sessions en cours et des sessions passées. Le rôle étudiant interdit l'accès à ces fonctionnalités mais autorise celui aux résultats des sessions courantes et aux multiples visualisations disponibles. Ces rôles sont gérés à l'inscription par l'administrateur de l'outil.

Modification et sauvegarde des contextes

D'ordinaire, notre système n'est pas configurable en temps réel et repose sur des fichiers de configuration qui doivent être modifiés par un administrateur du système. La

3. Sur ce point, le système n'interdit pas une telle modification du paramétrage en cours de tâche. Si le formateur le souhaite, pour une raison ou un autre, il peut changer le contexte algorithmique. Cependant, un changement de ce type biaise évidemment les résultats de l'observation stricte du comportement des participants.

stratégie pédagogique que nous avons définie, fondée sur l'expérimentation, nous a imposé de pouvoir changer dynamiquement ce mode de fonctionnement. Nous avons donc développé une couche logicielle permettant cette fonctionnalité. Cette couche logicielle est fondée sur l'enregistrement du contexte courant, autrement dit du jeu de paramètre courant en base de données. Ces paramètres sont modifiables via une interface dédiée au formateur.

Lors de chaque requête, le système lit la configuration courante dans la base de données et exécute la requête avec ce jeu de paramètres. À chaque changement de paramètre, un nouveau contexte est créé et s'applique automatiquement. Les anciens contextes sont toutefois sauvegardés, le formateur peut y revenir autant qu'il le veut [4]. Par ailleurs, les requêtes des étudiants sont associées au contexte dans lequel elles ont été exécutées de telle sorte que l'on puisse les rejouer par la suite. Par ailleurs, cette contextualisation a un impact sur le modèle de trace que nous utilisons. Chaque observation est désormais liée à un contexte de recherche. Il est ainsi possible d'extraire les traces en fonction de tel ou tel paramètre du contexte.

Gestion des tâches

Pour la gestion des tâches de recherche d'information. L'implémentation de Brightbox diffère assez peu de l'implémentation générale de notre bibliothèque numérique telle qu'elle a été présentée durant les chapitres précédents. Brightbox reprend les modèles de tâches déjà évalués. La différence se situe dans l'interface, puisque le formateur a là encore toute latitude pour prévoir, gérer et supprimer les tâches à accomplir par les étudiants. Le formulaire de déclaration de tâche de recherche (voir annexe A.2) lui permet de donner des instructions, le résultat attendu ainsi que les paramètres indispensables que sont la catégorie de la tâche et sa durée prévue. Depuis la liste des tâches, il peut par ailleurs avoir un accès direct aux résultats et aux différents indicateurs fournis par Brightbox.

Calcul des indicateurs

Si les indicateurs fournis par Brightbox sont en partie issus de nos travaux décrits précédemment, leur mode de calcul est différent. Lors de nos expérimentations précédentes à l'aide de notre bibliothèque numérique, il n'y avait pas de nécessité que ces indicateurs soient calculés rapidement. Dans un contexte pédagogique, ils doivent être très vite disponibles afin de permettre à l'enseignant de les expliquer aux participants. Nous avons donc revu notre processus de calcul des indicateurs pour qu'ils puissent être générés en temps réel, du moins dès que possible. La grande majorité des indicateurs sont ainsi calculés au plus tôt, dès la fin d'une tâche par exemple. Nous avons par ailleurs choisi d'implémenter ces calculs dans une API séparée du corps de notre outil. Cette implémentation nous permet d'utiliser ces calculs aussi bien dans le contexte de Brightbox, en temps réel donc, que dans un autre contexte avec des résultats anciens. Un simple appel à cette API avec des données correctement formatées permet d'obtenir

4. Plus précisément, avant de créer un nouveau contexte, le système vérifie qu'un contexte identique n'existe pas déjà, auquel cas, ce contexte préexistant est ré-utilisé.

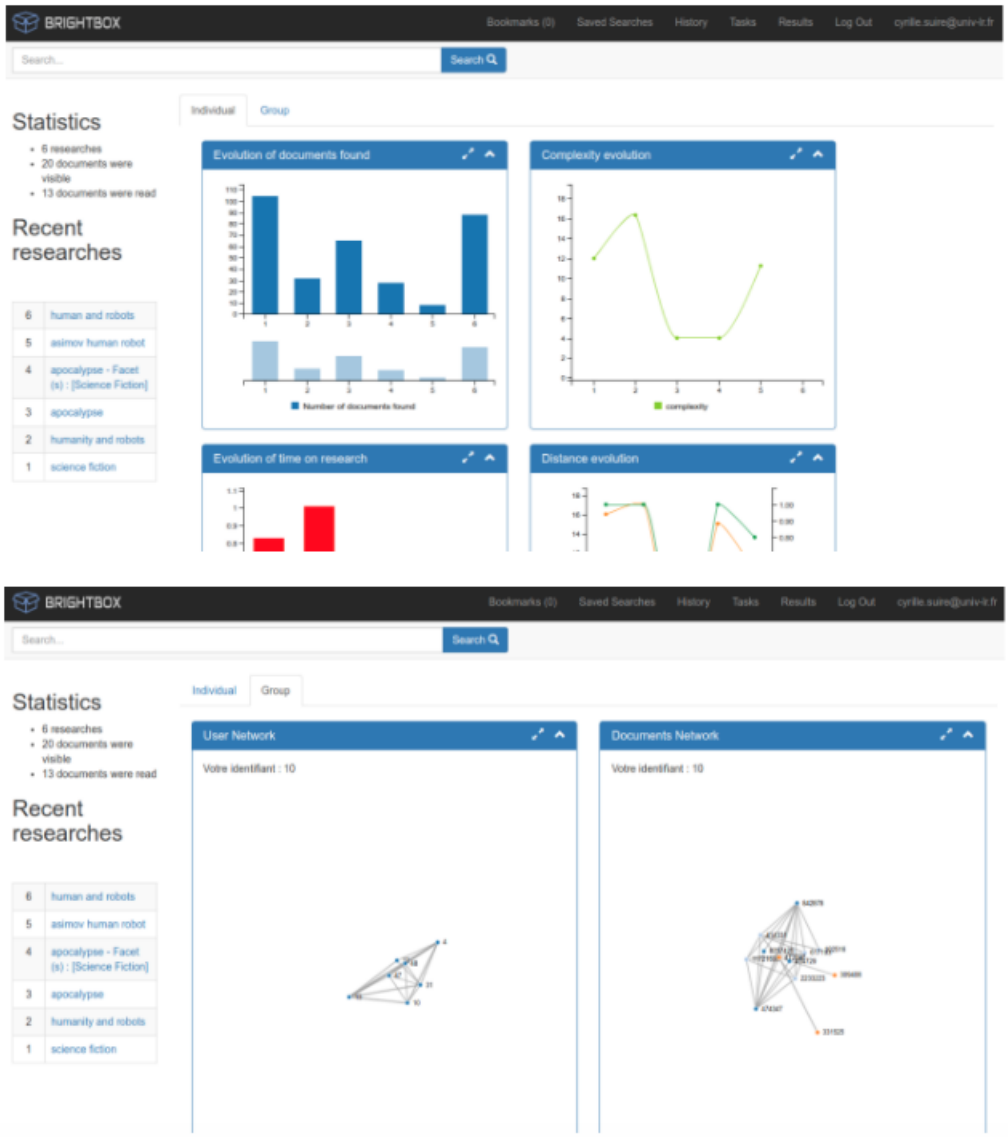
le résultat, il n'y a pas d'obstacle à utiliser des données issues d'expérimentations plus anciennes, voire d'un autre outil.

Interface et visualisations

Par rapport à l'implémentation générale de notre prototype, nous avons ajouté pour Brightbox plusieurs interfaces dédiées à la visualisation des résultats. Les rôles étudiant et enseignant ont ainsi accès aux valeurs calculées de nos indicateurs aussi bien pour leurs sessions de recherche libres que pour les tâches accomplies. Ces interfaces présentent diverses visualisations, dont certaines sont visibles à la figure 6.3. Ces visualisations reposent sur une bibliothèque logicielle libre [5].

5. En l'occurrence d3.js (https://d3js.org/).

FIGURE 6.3 – Interface de consultation des statistiques individuelles et de groupe



L'interface de consultation des statistiques individuelles, ici sur une tâche, est visible en haut. L'interface visible en bas permet de visualiser des résultats sur le groupe entier, ici sous forme de graphe.

6.3 Mode 1 : Expérimenter les biais liés au contexte d'exécution

6.3.1 Variabilité des paramètres de recherche et d'indexation

Parmi l'éventail des paramètres que Brightbox permet au formateur de faire varier, par l'intermédiaire des interfaces visibles en annexe A.4 et A.5, il existe deux grandes catégories. La première regroupe les paramètres liés à l'interface utilisateur, qui ont un impact sur les fonctionnalités disponibles pour les étudiants qui peuvent, du seul fait de leur présence ou leur absence dans l'interface, avoir un impact sur le comportement de recherche d'information. Ces paramètres n'ont, par contre, aucun impact sur la

manière dont sont calculés les résultats de recherche. La seconde catégorie rassemble les paramètres liés au fonctionnement du moteur de recherche à proprement parler. L'impact de ces paramètres est limité à la génération des résultats et n'auront pas, en dehors des visualisations spécifiquement conçues pour les démontrer, d'autres effets sur l'interface utilisateur. Le détail partiel [6] des paramètres est le suivant :

Paramètres d'interface utilisateur

— **Facet (*fields*, *prefix*, *minCount*)** Cette série de paramètres régit la disponibilité dans l'interface utilisateur du filtrage par facet. Il est simplement activé ou désactivé. Si il est activé, alors le formateur peut paramétrer le niveau de complexité de ce filtrage disponible pour les étudiants. *fields* et *prefix* identifient les champs de métadonnées que le système doit considérer comme des filtres actifs. *minCount* définit le nombre minimum de documents concernés par chacun des filtres pour que ceux-ci soient affichés sur l'interface utilisateur.

— **Spellcheck (*accuracy*, *onlyMorePopular*)** Ces paramètres permettent d'appliquer aux requêtes une correction orthographique, il est possible de préciser la précision via *accuracy* ou de n'appliquer la règle que si un terme est plus utilisé qu'un autre, sur la base de l'historique d'utilisation (*onlyMorePopular*).

— **Highlight (*fieldList*** Ces options activent la fonctionnalité dite de *highlighting* qui permet de surligner en couleur les termes de la requête dans les champs de métadonnées des résultats. Les champs sur lesquels s'appliquent cette fonction sont paramétrables via *fieldList*. Cette fonction est similaire à celle que l'on trouve sur les moteurs de recherche les plus connus et est visible dans la figure A.1.

Paramètres de manipulation des requêtes et des documents indexés

— **Stopwords** Les *stopwords* ou mots vides en français correspondent à tous les mots « vides » de sens. Ces mots, tels que les articles ou les conjonctions par exemple, peuvent être supprimés d'un texte lors de la phase d'indexation. Supprimer cette catégorie de mots facilite le traitement dans la mesure ou, très souvent, ces mots n'apportent pas de plus value lors d'une recherche d'information. Leur suppression n'est toutefois pas systématique et pas toujours de même ampleur. Il est possible de supprimer plus ou moins de mots vides. La suppression des mots vides, dans le corpus indexé, comme dans les requêtes des utilisateurs peut avoir un impact sur la recherche. Brightbox permet donc d'activer ou désactiver cette opération.

— **Lemmatisation / Stemming** En TALN, la lemmatisation, correspond aux opérations algorithmiques qui visent à réduire un mot à sa forme canonique, le lemme. L'objectif est d'indexer un terme quelle que soit sa forme grammaticale. Imaginons que deux documents d_i et d_j contiennent un mot m au singulier pour l'un et au féminin pluriel pour l'autre. Sans racinisation, l'index contiendrait les deux formes du même mot m. Le document d_i ne pourrait être trouvé qu'en

6. De très nombreux paramètres sont potentiellement modifiables. Du fait de leur grand nombre et de leur impact mineur, voire négligeable, dans un contexte pédagogique, nous décrivons ici seulement les plus importants.

saisissant la forme masculin singulier du mot m et, de la même manière, d_j ne pourrait l'être qu'en saisissant la forme féminin plurielle. À titre d'exemple, les mots « petit » et « petites », respectivement contenus dans d_i et d_j peuvent être réduits au lemme « petit ». Dans ce cas, les deux documents seront trouvés dans l'index, quelle que soit la forme grammaticale de m saisie pour la recherche.

La lemmatisation est souvent complétée par une étape de *stemming* ou racinisation en français, qui suit un principe similaire, mais qui réduit un mot à son radical. À titre d'exemple, les mots « malade, malades, maladie, maladive » seront indistinctement réduit à leur radical « malad ». Ces algorithmes augmentent naturellement le rappel, au détriment toutefois de la précision. Dans ce cas, le sens de « maladie » et « maladive », bien que sensiblement différent, n'impacte pas la recherche d'information. Les algorithmes permettant d'opérer ses transformations sont nombreux et paramétrables, l'impact de ces paramètres se traduit dans l'agressivité de l'algorithme de *stemming*. Une trop forte racinisation peut nuire considérablement à la précision et introduire un important bruit documentaire. Brightbox offre la possibilité de paramétrer ces algorithmes pour démontrer leur rôle et leur impact sur la recherche d'information aux participants.

— **Operators** Parmi les nombreux paramètres qu'il est possible de faire varier, les opérateurs font partie des plus importants. Ils désignent l'opération logique qui est effectuée par le moteur de recherche lorsqu'il traite la requête d'un utilisateur. Le moteur peut ainsi traiter la requête en associant tous les mots (AND) pour chercher les documents contenant tous les mots de la requête ou les dissocier (OR) et chercher les documents qui contiennent l'un ou l'autre des termes de recherche. Il est évident que les règles par défaut ont une importance majeure dans le fonctionnement de la recherche d'information. Notre outil implémente donc cette fonctionnalité et permet d'en observer les conséquences.

— **Boosts** Les algorithmes de recherche et de classement de l'information laissent la possibilité de définir des *boosts*. Ces *boosts* prennent la forme de coefficients qui pondèrent la valeur « naturelle » d'un score. Ils peuvent par exemple être utilisés pour faire en sorte qu'un terme de recherche, lorsqu'il est trouvé dans le titre d'un document ait un poids plus important que lorsqu'il est trouvé dans le corps du texte. En d'autres termes, dans ce cas, les concepteurs du système accordent plus de poids aux mots du titre considérant, assez logiquement, qu'ils décrivent plus précisément le document que les mots des autres champs de métadonnées. Ils sont aussi utiles pour favoriser des documents récents, lorsqu'ils sont appliqués aux dates de publication par exemple.

  Les *boosts* peuvent s'appliquer indistinctement sur des documents, pour en favoriser certains, ou sur des champs de métadonnées, pour en pondérer le poids. Ils également possible de s'en servir pour « booster » négativement, c'est à dire réduire la valeur de tel ou tel score. Naturellement, ces boosts ont un impact considérable sur la manière dont sont produits les résultats des algorithmes de pertinence. Si l'on souhaite interroger la définition de la pertinence dans un moteur, la logique d'application de ces coefficients est primordiale.

— **Phrase Slop (*ps*)** Dans de nombreux cas, les termes d'une recherche d'information sont multiples, le paramètre *ps* permet de définir la fenêtre dans laquelle doivent se trouver ces termes pour considérer le document comme pertinent.

Imaginons que l'on cherche dans un texte l'expression « maison blanche », en référence à la célèbre résidence présidentielle américaine et que le paramètre *ps* a une valeur de 100. Le moteur de recherche trouvera tous les documents dans lesquels les mots « maison » et « blanche » apparaissent dans une fenêtre de 100 mots et pas nécessairement l'un à proximité immédiate de l'autre. Dans ce cas, certains documents apparaitront dans les résultats même s'ils ne concernent pas directement la Maison Blanche.

— **Minimum Should Match (*mm*)** Dans le cas, toujours, où les termes d'une requête sont multiples, il est également nécessaire d'indiquer le nombre de termes de la requête pour lequel le moteur de recherche doit considérer un résultat comme pertinent. Ce paramètre est exprimé par *mm*. Sauf indication contraire (par opérateur par exemple), les termes d'une requêtes sont tous optionnels la valeur de *mm* détermine le nombre de terme minimum requis pour valider la pertinence d'un document. À titre d'exemple, pour une requête « maison blanche washington » et un paramètre *mm* réglé à 2, les documents contenant une combinaison de deux des trois mots composant la requête seront considérés comme pertinents.

  En pratique, fixer une valeur entière et systématique nuirait beaucoup à la recherche d'information, en effet un *mm* fixe, ayant pour valeur 2 générerait beaucoup de bruit documentaire pour des requêtes nettement supérieures en nombre de mots, de plus de 5 termes par exemple. La valeur de *mm* peut ainsi être calculée en fonction des caractéristiques de la requête. L'utilisation d'une fraction exprimée en pourcentage est par exemple possible, tout comme une combinaison de règles. Ainsi, la valeur par défaut de Brightbox pour le paramètre mm est définie ainsi $2 < -1; 5 < -2; 6 < 90\%$. Si la requête ne contient qu'un ou deux termes, alors ils sont tous requis. Entre 3 et 5 termes, *mm* vaut $nb_{termes} - 1$, pour 6 termes *mm* vaut 4 et pour plus de 6 termes, *mm* vaut $nb_{termes} \times 0.9$. Naturellement, l'impact de *mm* est important. La connaissance de sa valeur ou de sa règle de calcul peut permettre à l'utilisateur d'adapter sa stratégie de recherche.

— **Tie breaker** Ce paramètre, exprimé sous la forme d'un nombre flottant compris en 0.0 et 1.0 sert à gérer les cas où un mot est trouvé dans plusieurs champs de métadonnées, par exemple à la fois dans le titre, la description et le texte. Dans un cas de ce type, un score est calculé en fonction de la fréquence d'apparition du mot dans chacun des champs. Le paramètre *tie* détermine la manière dont doivent être pris en compte ces scores. Une valeur $tie = 0.0$ indique que seul la valeur du score du champs ayant obtenu le score le plus élevé doit être utilisé. Une valeur $tie = 1.0$ indique au contraire que tous les champs doivent être pris en compte, en calculant la somme de tous les scores.

6.3.2 Problématique du bruit informationnel

Bien plus qu'un simple paramètre, comme nous avons eu l'occasion de le voir dans notre chapitre 5, la problématique du bruit informationnel impacte de manière importante la pratique de la recherche d'information. La démarche présentée dans le présent chapitre de pouvait pas passer à côté de cette problématique. Il nous est apparu indispensable de mettre en lumière et expliquer cette problématique aux étudiants.

La première étape de notre réflexion reste cependant très liée aux problématiques d'in-

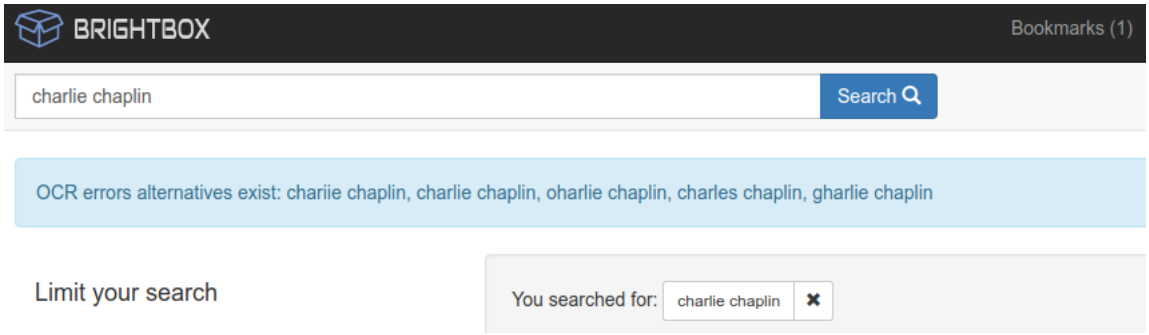
dexation et de paramétrages que nous venons de décrire. Il s'agit d'expliquer l'impact global des variations morphologiques des mots, liées ou non à des erreurs d'OCR sur la recherche d'information. L'idée générale est de fournir à travers l'outil un corpus en deux versions « propre » et « bruitée ». Le formateur peut ainsi choisir la version du corpus et permettre aux étudiants de comparer les sorties du moteur de recherche, mesurant ainsi l'impact du bruit. Comme nous l'avons déjà expliqué dans la section 5.3 de notre chapitre 5 trouver des corpus dans ces deux versions s'avère complexe. Dans le cas qui nous occupe ici, il n'est pas nécessaire que le corpus soit annoté, au sens où nous n'avons pas besoin d'une vérité terrain. Cependant, il existe une contrainte supplémentaire. Le corpus doit être accessible et intéressant pour les étudiants. Ils doivent avoir une connaissance suffisante des documents qu'il contient, afin d'être capable de l'interroger et d'accomplir des tâches de recherche d'information cohérentes, comme nous l'avons vu dans notre chapitre 4.

La solution a ce problème requiert donc une démarche similaire à celle que nous avons utilisée dans notre chapitre 5. Il est nécessaire de partir d'un corpus « propre » connu des étudiants et créer artificiellement du bruit. Une fois ces opérations effectuées, nous pouvons croiser les deux corpus, propre et bruité, pour déterminer un taux d'erreur. Les deux corpus sont alors soumis à l'indexation et deviennent interrogeables dans le moteur de recherche. Le formateur peut ainsi choisir d'interroger le corpus propre ou le corpus bruité, chacun acceptant tout le spectre des paramètres d'indexation et de recherche décrits plus haut. Le bruit informationnel devient ainsi un paramètre parmi les autres.

Malgré tout, cette méthode requiert tout de même de disposer à l'origine d'un corpus de documents « propres », ce qui en matière de documents historiques n'est pas toujours facilement accessible. Au regard de nos précédents développements, il est évident que le choix d'un corpus propre ou bruité n'est pas le seul moyen à notre disposition pour donner à voir et faire comprendre l'impact des variations orthographiques ou des erreurs d'OCR sur le fonctionnement du système. Nous avons également implémenté la méthode décrite dans notre chapitre 5. Cette implémentation rend possible de montrer et d'expliquer l'impact des erreurs d'OCR sans recourir à une dégradation artificielle d'un corpus de documents. Il est en effet possible de sélectionner n'importe quel corpus de documents historiques et de le soumettre à un processus de reconnaissance optique de caractères.

À l'issue de ces manipulations et de l'intégration dans l'outil, le formateur peut activer la fonctionnalité de reconnaissance des entités similaires. Ainsi, à chaque requête saisie, le système vérifie si une variation des mots saisis a été détectée. Cette fonctionnalité permet donc d'afficher toutes les versions alternatives d'un terme, et permet aux étudiants de mesurer les erreurs potentiellement présentes dans le corpus (voir figure 6.4). Naturellement, cette fonctionnalité propose également aux étudiants de chercher toutes les occurrences d'un terme, plutôt que la seule version initialement fournie. Une comparaison entre résultats de requêtes avec et sans cette fonctionnalité est alors possible et témoigne de l'impact des erreurs sur la recherche d'information.

FIGURE 6.4 – Détection et affichage des requêtes alternatives dans Brightbox



Ici, la requête « Charlie Chaplin » provoque l'affichage d'un message indiquant les alternatives trouvées.

6.4 Mode 2 : Comprendre et adapter son comportement face au besoin d'information

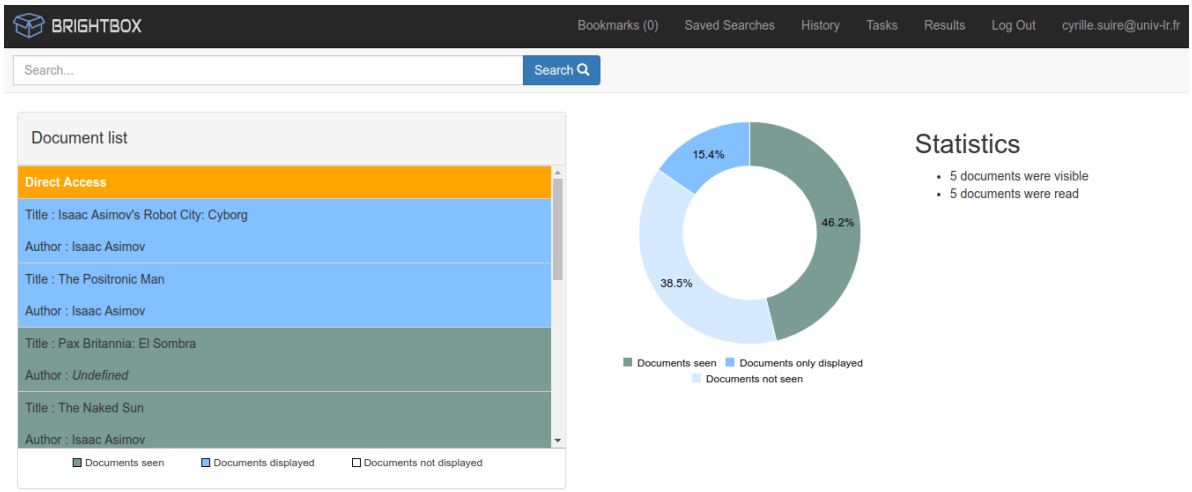
6.4.1 Fonctionnement général

Le second mode de fonctionnement de notre outil est dédié à l'analyse de la pratique des étudiants en situation de recherche d'information. Les fonctionnalités développées dans ce contexte sont essentiellement issues des observations que nous avons menées dans le cadre de notre chapitre 4. Nous avons ainsi pu mettre en place des observateurs et des métriques témoignant de la pratique des étudiants et construire des représentations visuelles leur donnant à voir les biais produits par des pratiques de recherche d'information éventuellement inadaptées à leur tâche. Afin de produire ces indicateurs et ces représentations individuels et collectifs et être capable de les comparer, le système fournit la possibilité au formateur de définir des tâches de recherche d'information aux étudiants. Ces tâches peuvent être classées en fonction des catégories du modèle de Marchionini [MARCHIONINI, 2006] que nous avons également employées lors des expérimentations de notre chapitre 4 et dont l'interface de création est visible à la figure A.2.

L'outil est doté des mêmes observateurs que notre plateforme expérimentale décrite à la section 4.3.2. Ces observateurs fonctionnent que l'outil soit ou non en mode « tâche ». Il est donc possible d'observer le comportement d'un utilisateur en dehors de ce mode, la tâche est alors remplacée par une session qui débute lors de la connexion et se termine à la déconnexion. En revanche, en mode « tâche » les points de départ et de fin sont déterminés par les clics sur les boutons de début et de fin de tâche de l'interface.

Les objectifs pédagogiques fixés à l'outil se focalisant principalement sur l'activité de recherche d'information, certains observateurs que nous avions préalablement évalués n'ont pas été implémentés dans Brightbox. Ainsi, les observateurs de consultation en plein texte ou en version originale ne sont pas utiles dans ce contexte. À l'inverse, nous avons augmenté certains observateurs pour fournir des représentations plus précises sur certains points, en particulier sur l'exploitation des résultats de recherche. À titre d'exemple, la liste complète des résultats de chaque recherche est gardée en mémoire,

FIGURE 6.5 – Visualisation du comportement d'exploitation des résultats de recherche



À gauche de la fenêtre, l'interface présente les résultats d'une requête et colore les résultats en fonction de l'exploitation qui en a été faite par le participant. La partie droite de la fenêtre présente les nombres de documents, vus, lus ou ignorés.

afin de reproduire les résultats de cette recherche et visualiser la manière dont ils ont été exploités par les participants (figure 6.5).

6.4.2 Indicateurs individuels et collectifs

À l'issue de la fin d'une tâche, ou d'une session de recherche libre, les utilisateurs de la plateforme ont accès à une page de résultats sur laquelle leur sont présentés différents indicateurs de leur comportement. Cette page permet d'avoir accès à ses statistiques personnels et à des représentations visuelles tirées de l'expérience collective de la salle.

Statistiques générales

L'utilisateur a accès à des indicateurs chiffrés qui lui indiquent le nombre de recherches menées, le nombre de documents vus dans les résultats de recherche ou encore le nombre de documents effectivement consultés. Ces deux derniers indicateurs sont par ailleurs affichés sous forme graphiques, recherche par recherche, l'utilisateur peut ainsi observer leur évolution dans le temps. Par ailleurs, le système l'informe également des indicateurs développés dans le cadre de notre chapitre 4, par exemple les durées de chacune de ses recherches. Ces informations sont utiles pour révéler les recherches et résultats sur lesquels il s'est particulièrement focalisé et qui peuvent avoir été au centre de l'évolution de sa stratégie de recherche.

Expression du besoin

Les indicateurs de l'expression du besoin sont décrits dans notre section 4.3.2. Une représentation dans le temps, recherche par recherche, est également implémentée dans la plateforme sous forme graphique. Cependant, sous cette forme, ces indicateurs sont relativement limités. En effet, deux requêtes peuvent être très proches du point de vue de leur nombre de mots, ou nombre de caractères, mais très éloignées par leur contenu et leur sens. Afin de permettre de mesurer ces différences, nous avons implémenté une représentation fondée sur les distances suivantes :

— **Distance d'édition (caractères) :** La distance d'édition est une comparaison entre une requête et la requête suivante ou précédente, pour la calculer nous utilisons la distance dite de Levenshtein. Elle est exprimée par la somme des opérations d'ajouts, de suppressions ou de remplacements de caractères nécessaires à établir une égalité entre deux chaînes de caractères[7]. La distance entre deux chaînes de caractères a et b est formalisée par l'équation (6.1).

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{6.1}$$

— **Distance de Jaccard :** Afin de mesurer le plus finement possible les différences entre les requêtes produites par les utilisateurs, la distance d'édition seule ne suffit pas. Parmi les métriques existantes, la distance dite de Jaccard est intéressante. Cette métrique compare la similarité des éléments présents dans deux chaînes de caractères. Cet indicateur est calculable en comparant les caractères ou les mots d'une requête. Dans le premier cas, deux requêtes seront considérées comme similaires si elles partagent beaucoup de caractères en commun. Dans le second cas, elle seront similaires si le nombre de mots commun est important. D'un point de vue formel, cette distance est calculable selon la formule (6.2) ou A et B sont des vecteurs de caractères ou de mots.

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{6.2}$$

Ces représentations graphiques sont utiles pour détecter les ruptures dans la stratégie de recherche d'un utilisateur. Lorsque les deux distances dont nous parlons sont significativement élevées, les deux requêtes comparées n'ont rien de commun. Enfin, ces deux métriques se complètent d'une mesure de la similarité sémantique, calculée à l'aide de Wordnet[8], qui vise à établir une distance entre le sens de plusieurs requêtes. À titre

7. Ainsi, pour les mots « voiture » et « toiture », la distance de Levenshtein est égale à 1, il y a une opération de substitution de caractère sur la première lettre du mot. Il n'y a pas d'autres opérations à effectuer.

8. Wordnet est une base de données lexicale et sémantique. Elle est développée principalement par des linguistes, pour l'anglais. Elle existe cependant, dans des versions plus ou moins avancées pour d'autres langues. Cette base offre des données sur les relations sémantiques entre les mots. Il

d'exemple les mots « résident » et « président », bien que très proches du point de vue des distances d'édition seront éloignés du point de vue de cette mesure.

Exploitation des résultats par recherche

L'application est dotée d'une fonctionnalité permettant de ré-exécuter chacune des recherches effectuées par les utilisateurs dans le contexte d'origine et d'afficher l'exploitation qui en a été faite. Outre les indicateurs chiffrés déjà mentionnés, cette fois accessibles dans le contexte d'une seule recherche, cette vue permet de visualiser la proportion des documents retournés par le moteur, des résultats vus et des résultats cliqués (lus). Cette fonctionnalité est visible à la figure 6.5. Elle se fonde sur la mécanique d'observation mise en place pour notre chapitre 4, et les observateurs complémentaires suivants :

— résultats calculés : résultats calculés par le moteur de recherche et extraits des traces « requête » et « filtre ». Cet indicateur est à la fois un témoin de la précision d'une requête (plus une requête est précise, moins elle retourne de résultats) et est indispensable pour rejouer et visualiser les différentes recherches des utilisateurs.

— Nombre de ressources téléchargées : la plateforme laissant la possibilité à l'utilisateur de télécharger les contenus, cet indicateur est pertinent dans la mesure où il témoigne de la volonté de l'utilisateur de télécharger la ressource pour l'inclure dans son propre système d'information ;

— Nombre de ressources mises en favoris : cet indicateur est la somme des traces « ressources mises en favoris ». Il peut être calculé pour chaque recherche ou par session. Il est dans ce dernier cas égal à la somme calculée pour chacune des recherches.

Résultats collectifs

Enfin, l'application fournit deux représentations sous forme de graphes, similaires à celles que nous avons décrites dans notre section 4.5.3 (figure 4.13). Elle présente deux réseaux, d'utilisateurs dans un cas, de documents dans l'autre, qui sont liés entre eux par leurs usage commun dans la tâche. Ces représentations permettent d'observer la pratique à l'échelle du groupe, d'observer si celle-ci a été homogène ou au contraire si différentes stratégies ont été mises en place. Elles donnent également l'occasion de repérer les documents qui ont été collectivement les plus utiles à l'accomplissement d'une tâche.

6.5 Démarche de validation

L'outil que nous venons de décrire a été pensé pour être un outil dédié à l'enseignement, son expérimentation a ainsi nécessité la conception et l'expérimentation de séance de

est par exemple possible d'établir que les mots « voiture » et « automobile » sont très proches du point de vue sémantique. Le site de la version en anglais de Wordnet est accessible à l'adresse : https://wordnet.princeton.edu/.

formation auprès de notre public cible. Nous décrirons dans cette section l'organisation et le contenu de la séance que nous avons imaginée et notre méthodologie de validation, fondée sur un questionnaire. Nous exposerons par ailleurs dans les sections suivantes les conditions du déroulement de cette expérimentation, le public participant et le jeu de données utilisé. Enfin, après une analyse des résultats, nous développerons quelques pistes de développement potentielles en direction d'autres disciplines ou d'autres contextes d'usage.

Pour évaluer l'intérêt pédagogique des fonctionnalités de notre outils, nous avons choisi d'organiser une séance de formation en trois parties. La première vise à explorer avec les participants leur pratique de la recherche d'information dans un corpus de sources, sur une tâche exploratoire. La seconde partie, quant à elle, a pour objectif de confronter cette pratique avec le fonctionnement du moteur de recherche de notre plateforme, avec différents paramètres. Enfin, la dernière partie s'est focalisée sur les problématiques de bruit informationnel, dans l'objectif de démontrer, via la plateforme l'impact du bruit d'OCR sur la pratique de recherche d'information.

6.5.1 Jeu de données

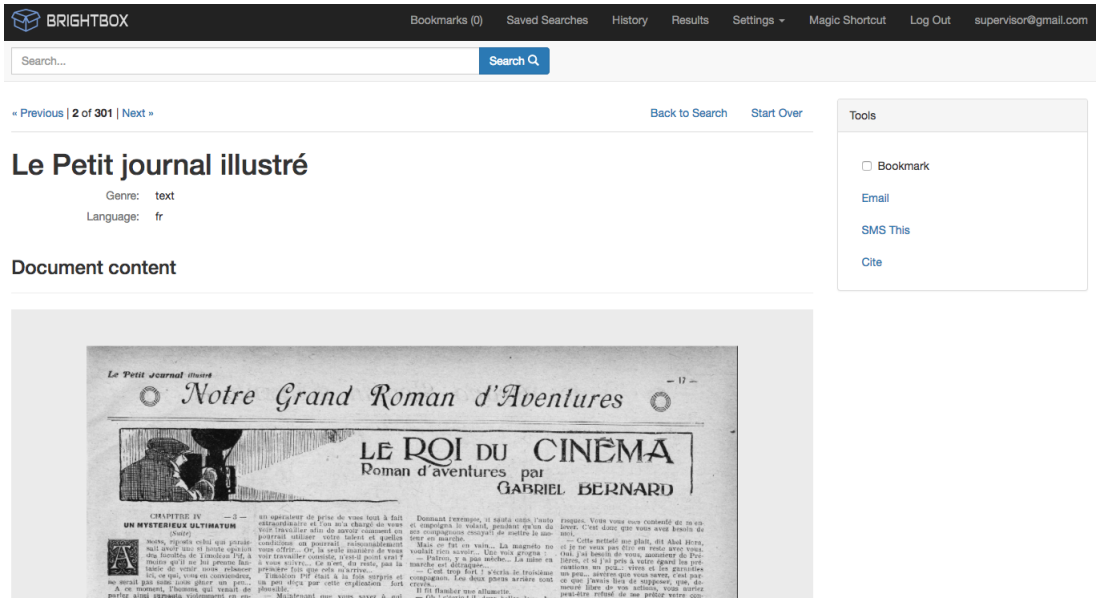
Pour mener nos expérimentations, nous avons du construire un corpus adapté à différentes contraintes. Les premières contraintes sont essentiellement techniques et peuvent être résolues par les moyens détaillés dans la section 6.3.2. Les secondes sont quant à elles liées au niveau de connaissance du domaine que nous avons également évoquées dans notre chapitre 4. Ce niveau de connaissance du domaine est une des variables importantes de la pratique de recherche d'information. Pour mener nos expérimentations il nous est apparu indispensable de faire travailler les étudiants sur des documents d'une période historique qu'ils avaient eu l'occasion de travailler et surtout qui pouvait être réinvestis dans leurs cours d'histoire. En interaction avec les historiens avec lesquels nous avons travaillé, nous avons décidé de travailler sur les jeux de données suivants :

Presse ancienne

La presse écrite est en effet un des types de sources historiques le plus numérisé et le plus facilement disponible en ligne. Elle a par ailleurs, pour les grands titres de la presse nationale et régionale, été largement soumise à la reconnaissance optique de caractères. Le grand volume de document accessible rend par ailleurs intéressant, voire nécessaire, l'usage de la recherche en plein texte. Ainsi, les problématiques que nous souhaitions traiter grâce à nos outils, que ce soit le bruit informationnel, les biais d'indexation et de recherche ou les biais comportementaux pouvaient toutes être démontrées à l'aide d'un corpus de ce type.

En ce qui concerne le choix des documents eux-mêmes, les connaissances des participants à nos expérimentations et les objectifs de réinvestissement des historiens dans leurs cours, nous ont permis de réemployer le corpus B (« Le petit journal illustré ») de notre chapitre 5. Chaque page a été indexée dans notre outils sur la base de ce format textuel et des métadonnées récupérées via les APIs de la BNF. Notre outil disposait donc, pour les expérimentations, de ces documents en version textuelle et

FIGURE 6.6 – Vue d'un document en version originale intégré dans la plateforme



Le lecteur exportable de Gallica est intégré dans la plateforme et affiche la page concernée. Le document est également accessible, plus bas, en plein texte.

en version originale, via le lecteur exportable de Gallica, intégré dans notre outil. Par ailleurs, nous avons également intégré les résultats de notre méthode de détection des variations orthographiques des entités nommées dans la plateforme.

Littérature

Si la presse écrite se prête bien à démontrer les problématiques de bruit informationnel ou de biais algorithmique, elle souffre d'un défaut majeur, le peu de métadonnées disponibles. En effet, dans l'immense majorité des cas, ces métadonnées se limitent au titre du périodique et à la date de parution du numéro. Or, notre plateforme permet également de mener une réflexion sur l'indexation et l'usage des métadonnées pour la recherche d'information. Nous avons donc choisi d'intégrer à la plateforme les ouvrages issus du projet Gutenberg [9]. Ce projet met à disposition des œuvres de littérature libres de droits en plusieurs langues. Ces ouvrages, bien décrits par leurs métadonnées (date de parution, sujet, etc) ouvrent la voie à la démonstration de l'impact de la qualité des métadonnées sur la pratique de recherche avec les participants à l'expérimentation. Nous avons donc téléchargé les contenus en français du Projet Gutenberg ainsi que le catalogue général duquel nous avons exporté les métadonnées de chaque document. À l'issue de leur indexation, la plateforme disposait d'environ 2500 documents en provenance de cette source.

9. Dont le site est accessible à l'adresse : http://www.gutenberg.org/.

6.5.2 Participants et conditions expérimentales

Si le niveau de connaissance du domaine de recherche a un impact important sur les stratégies de recherche d'information déjà clairement démontré, le niveau des étudiants dans leur cursus universitaire sur ce type de problématique n'a pas été étudié. Nous n'avions donc pas d'apriori sur le niveau des étudiants à inclure dans notre expérimentation. Il nous a semblé intéressant de mener des expériences dès l'entrée dans le supérieur, en Licence 1 et en Licence 2, tant les problématiques que nous adressons semblent de plus en plus importantes.

L'expérimentation a donc été organisée avec un groupe d'étudiants en histoire de l'Université de La Rochelle. Ces étudiants ont été recrutés sur la base du volontariat après une présentation rapide des enjeux scientifique et pédagogique de cette expérimentation. Les conditions matérielles de l'organisation de ces expérimentations ont limité le nombre d'étudiants à 12 dont la répartition est donnée dans le tableau 6.1 :

Tableau 6.1 – Expérimentation : participants

Niveau	Nombre	Proportion
L1	5	41,7 %
L2	7	58,3 %

L'expérimentation a été menée en salle de travaux pratiques équipée de postes informatiques individuels [10]. Les salles étaient également dotées d'un vidéo-projecteur pour le formateur. Pour les besoins de l'expérience, seul un navigateur internet est nécessaire sur les postes informatiques de la salle, les participants ont été autorisés à utiliser tout autre outil leur semblant nécessaire afin d'accomplir les tâches demandées, en particulier pour la prise de notes, la recherche d'information sur les documents du corpus devant toutefois se limiter à notre plateforme expérimentale. À l'issue de la séance, il a été demandé aux participants de donner une liste des outils externes, logiciels et sites internet consultés durant la séance. Enfin, le serveur de l'application a été hébergé sur une machine virtuelle de l'Université de La Rochelle.

6.5.3 Déroulement de l'expérimentation

La séance expérimentale de formation, d'une durée de 1h30, s'est déroulée en mêlant l'usage de notre plateforme avec des phases d'explication. Nous avons bien entendu conçu, pour ces dernières, un support pédagogique sous forme d'un diaporama vidéo-projeté [11]. La séance a débuté par une explication générale des problématiques auxquelles nous entendions répondre durant la séance, ainsi que la présentation du déroulement de celle-ci. À l'issue de cette brève introduction d'environ 10 minutes, les étudiants ont été invités à se familiariser quelques minutes avec le fonctionnement de l'outil. Les principales fonctionnalités de recherche d'information, recherche textuelle, facettes, mise en favoris, etc. leur ont été expliquées.

10. Cette séance s'est déroulée le 09 mars 2018 dans les locaux de la Bibliothèque Universitaire de La Rochelle.
11. Des exemples de ces supports sont présentés dans l'annexe A.1.3.

Première partie : observation du comportement des participants

Nous avons ensuite introduit la partie de l'expérience dédiée à l'observation comportementale, d'une durée de 35 minutes. Une tâche de recherche, paramétrée dans l'outil a ensuite été proposée aux étudiants. Cette tâche de nature exploratoire était présentée dans l'outil de la manière suivante : « Vous préparez une présentation sur le sujet suivant : le monde animal et ses représentations dans la littérature moderne et contemporaine. Vous débutez vos recherches et cherchez des sources que vous pourriez exploiter (analyser et présenter) ». Le format de l'expérimentation ne permettant naturellement pas de répondre de manière exhaustive, il a été demandé aux étudiants de sélectionner quelques sources. La durée de la tâche a été limitée à 12 minutes. Cette durée était évidement trop courte pour permettre d'utiliser les résultats pour faire une analyse précise du comportement des participants. Cependant, là n'était pas l'objectif. Il s'agissait de disposer d'assez de données pour donner à voir aux participants les caractéristiques principales de leur comportement et permettre de les mettre en perspective avec le fonctionnement du moteur.

Les étudiants ont donc pris le temps nécessaire pour débuter leur recherche sur cette tâche. À la fin du temps imparti, ils ont été invités à observer individuellement et discuter collectivement les résultats calculés. Sans détailler ici les résultats individuels, quelques grandes tendances ont pu être constatées par les étudiants. D'abord, bien que la plateforme soit une bibliothèque numérique, et non un moteur de recherche général, ils ont tous formulés leurs premières requêtes comme si ils cherchaient une réponse à la question posée. Des contenus de requêtes tel que « animaux », « animaux domestiques » ou « représentation animalière » ont été constatés. Toutefois, les participants ont rapidement pris conscience de l'inefficacité de cette méthode. Ils ont alors mieux considéré le contenu de la plateforme et formulé des requêtes plus adaptées, cherchant à identifier des sources qui mentionnaient tel ou tel animal, réel ou imaginaire. Ils ont par la suite cherché, pour certains, à limiter les nombreux résultats retournés par le moteur de recherche de la bibliothèque numérique.

En plus de repérer les stratégies élaborées, les étudiants ont constaté qu'ils n'utilisaient que très peu les résultats de recherche. Pour l'essentiel ils ont observé qu'ils ne regardaient rarement plus que les 4 ou 5 premiers résultats de recherche dans la liste des résultats. Autrement dit, ils ont très peu « descendu » dans la page de résultats. Ils ont également très peu cliqué sur des résultats pour consulter les métadonnées et le texte intégral des documents. Ce comportement s'explique toutefois par la nature de la tâche, où les métadonnées de la liste de résultats pouvaient effectivement suffire à inclure ou exclure des documents de sa réflexion. Enfin, nous avons expliqué les différentes métriques et représentations visuelles aux étudiants afin qu'ils puissent les comprendre. Ils ont ainsi été en mesure de retrouver sur ces représentations les moments où ils ont significativement changé de stratégie de recherche.

Seconde partie : fonctionnement de la bibliothèque numérique

Cette tâche a été effectuée avec une série de paramètres qui n'ont bien entendu pas varié durant l'exercice [12]. La seconde partie de l'expérimentation, d'une durée de 30 minutes, visait à confronter les grands traits du comportement de recherche relevés par les étudiants avec le fonctionnement général de la bibliothèque numérique. Cette partie de la séance s'est d'abord appuyée sur un diaporama présentant un exemple de chaîne de traitement qui conduit un document dans sa version originale à être accessible dans une bibliothèque numérique.

Nous avons ainsi présenté les étapes suivantes, mettant systématiquement en avant les enjeux aussi bien pour l'informaticien et le spécialiste de l'information que pour la pratique de l'historien :

— **Numérisation :** explication des notions de taille et de résolution, présentation d'exemple de numérisation à différentes résolution et des enjeux de conservation de la lisibilité pour les hommes et les machines ;

— **OCR :** présentation d'un exemple d'OCR d'un document et explication des enjeux relatifs à la qualité de la reconnaissance ;

— **Prétraitement :** démonstration par l'exemple, sur une phrase, des effets des grandes étapes de prétraitement des données, suppression de la casse, de la ponctuation, des mots-vides, racinisation, etc. et des enjeux pour l'indexation et l'efficacité du système de recherche d'information ;

— **Métadonnées :** présentation des grandes catégories de métadonnées, de leur importance pour la recherche d'information et de leur rôle important pour la description des données ou pour l'interopérabilité ;

— **Indexation :** explication du fonctionnement et des enjeux et de l'indexation et des calculs de pertinence.

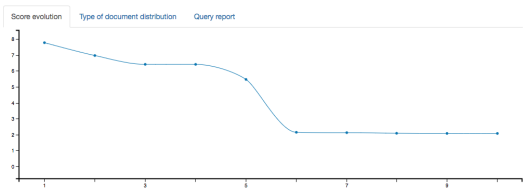
Cette étape de présentation a permis de mettre en lumière les éventuels effets de boîtes noires que produisent chacune des étapes que nous venons de décrire. Sur cette base, nous avons proposé aux étudiants de constater, par eux-mêmes, ces effets. Pour ce faire, nous avons activé dans l'outil les représentations visuelles complémentaires et les détails du traitement des requêtes. Par l'intermédiaire de l'interface formateur, permettant de faire varier le paramétrage de la bibliothèque numérique, nous avons fait varier certains de ces paramètres [13]. À titre d'exemple, nous avons retiré certains champs de métadonnées, le résumé par exemple, qui était un champ de métadonnées très lacunaire du jeu de données. Nous avons également fait varier des paramètres plus complexes, en changeant les règles de prétraitement par exemple, le calcul de la pertinence, ou des règles de *matching* (*PS* ou *QS* par exemple, décrits dans la section 6.3).

À chaque modification des paramètres, les étudiants ont pu exécuter quelques requêtes et constater les variations induites par le nouveau paramétrage. Afin de faciliter cette

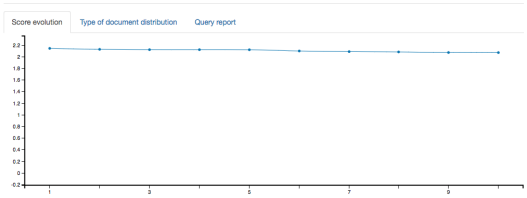
12. Les paramètres essentiels étaient : recherche sur toutes métadonnées et plein texte, texte et requête avec *preprocessing*, pertinence par algorithme BM25, visualisations spécifiques désactivées. Le reste des paramètres a été établi en référence aux standards de Europeana.

13. Des vues de cette interface sont disponibles en annexe A.4 et annexe A.5.

FIGURE 6.7 – Exemple de courbes des scores de pertinence calculées pour une même requête.



(a) paramétrage A.



(b) paramétrage B.

observation ils ont pu utilisé la fonctionnalité de comparaison de résultats de requêtes qui met en parallèle, sur l'écran, deux listes de résultats. Ils ont par ailleurs pu mesurer l'impact des différents paramétrages sur le calcul de la pertinence. À titre d'exemple, la figure 6.7 présente les deux courbes générées par la plateforme pour une même requête, avec dans un cas toutes les métadonnées disponibles et le prétraitement activé (paramétrage A), et dans l'autre le seul plein texte (paramétrage B). Les participants ont pu constater que les documents qui possèdent un résumé sont largement sur-évalués par l'algorithme de pertinence. Il a également été intéressant de faire remarquer et réagir les étudiants sur leur pratique, observée dans la première partie, au regard de ces résultats. Ils ont pu prendre conscience que leur tendance à ne se limiter qu'aux premiers résultats de recherche était la plupart du temps problématique. Un changement mineur du paramétrage pouvant bouleverser le classement produit par le moteur.

Troisième partie : problématique spécifique de l'OCR et des entités nommées

Dans la dernière partie de la séance, plus courte que les précédentes, d'une durée de 15 minutes, nous avons souhaité mettre brièvement l'accent sur les problématiques de bruit informationnel, d'OCR et d'entités nommées que nous avons déjà largement développées dans notre chapitre 5. À l'aide, là encore, d'un court diaporama vidéo-projeté, nous avons expliqué la méthodologie de calcul des indicateurs de qualité de l'OCR que les étudiants peuvent trouver sur certaines plateformes.

Par la suite, nous avons brièvement résumé le fonctionnement de notre méthode de détection des variations orthographiques et des erreurs et montré et expliqué des représentations qui témoignent de l'importance de ces erreurs sur un corpus de presse ancienne. Nous avons par la suite proposé aux étudiants de chercher quelques documents contenant de telles entités, de visualiser les alternatives fournies par la plateforme et de rechercher de nouveau en incluant ces alternatives afin de mesurer l'impact sur les résultats.

6.6 Résultats et pistes de développement

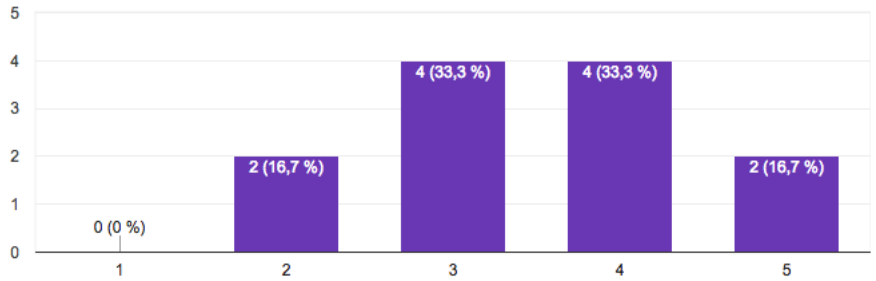
6.6.1 Questionnaire d'évaluation

À l'issue de la séance, les étudiants ont été invités à répondre à un questionnaire disponible en ligne. Il avait pour principal objectif de recueillir les impressions et le niveau de satisfaction des participants sur la séance qu'ils venaient de suivre. Le questionnaire comportaient 16 questions organisées en plusieurs sections : questions générales (niveau de formation des participants), questions relatives à leurs habitudes de recherche d'information et enfin questions sur la séance expérimentale à proprement parler.

Les questions portant sur leurs habitudes en matière de recherche d'informations témoignent, sans surprise, du poids des moteurs de recherche généraux (Google, etc.) dans la pratique des étudiants. Que ce soit pour débuter une recherche, l'approfondir, ou rechercher des faits précis, il sont mentionnés comme une source majeure par 9 étudiants sur 12. À la seconde place, le catalogue numérique de la Bibliothèque Universitaire reste une source très importante devant les ouvrages généraux et leurs bibliographies. Les bibliothèques numériques plus spécialisées, telles que Cairn ou Jstor sont mentionnées par un tiers des étudiants, surtout pour l'approfondissement de la recherche d'information. Ces résultats confirment ce que nous avons déjà eu l'occasion de remarquer par l'analyse de la littérature dans notre chapitre 2. Les moteurs de recherche et autres plateformes numériques constituent les ressources dominantes de l'activité de recherche d'information.

À propos de cette activité, les étudiants se déclarent plutôt débutants qu'experts [14], reconnaissant probablement que des compétences sont nécessaires pour la maîtriser parfaitement. Cette impression se confirme par les réponses portant sur la question du besoin de formation, ou la majorité des étudiants estime avoir des besoin moyens à importants sur ces problématiques (figure 6.8).

FIGURE 6.8 – Évaluation : besoin de formation



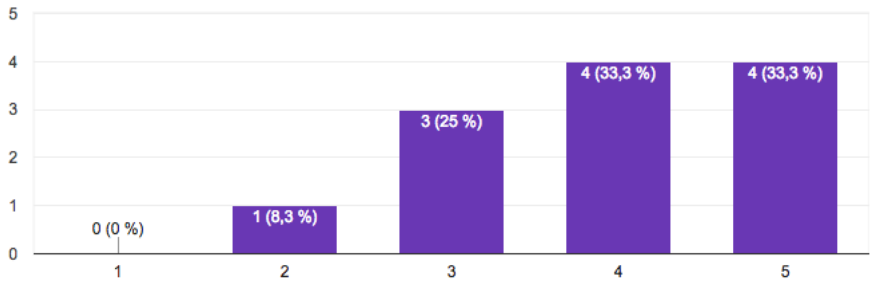
Réponses à la question : « Comment évalueriez-vous vos besoins en matière de formation à la recherche d'information numérique ? », échelle de 1 à 5.

Au regard de ce besoin de formation estimé, la séance est apparue globalement pertinente pour la majeure partie des participants, 11 d'entre eux situant en effet leurs

14. À la question « en matière de recherche d'information, vous décririez-vous comme : débutant (1), expert (5) », les étudiants répondent 2 pour 5 d'entre eux (41,7 %) et 3 pour 7 d'entre eux (58,3 %).

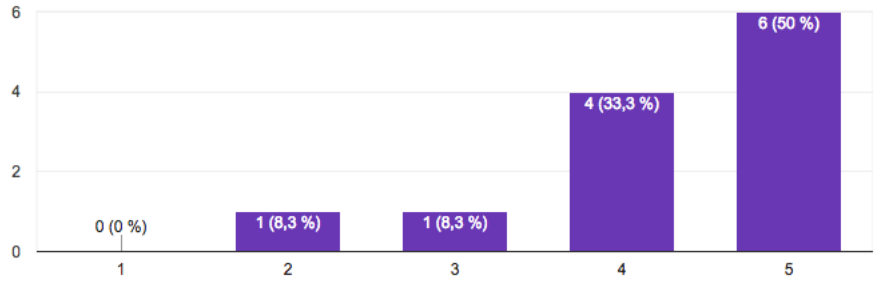
réponses entre 3 et 5 à cette question (figure 6.9). Par ailleurs les participants ont dans une large majorité jugé utile que ce type de formation soit intégré dans leur formation initiale. Ces réponses témoignent de leur compréhension de l'importance de la problématique, à laquelle ils avaient été sensibilisés auparavant durant leur formation, comme de leur satisfaction de la séance elle même.

FIGURE 6.9 – Évaluation : pertinence de la séance



Réponses à la question : « La séance suivie aujourd'hui vous a-t-elle semblée pertinente au regard de vos besoins ? », échelle de 1 à 5.

FIGURE 6.10 – Évaluation : pertinence dans la formation



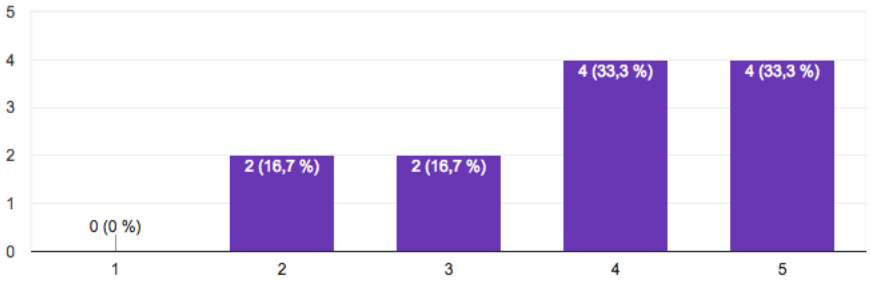
Réponses à la question : « Trouveriez-vous pertinent d'intégrer ce type de questionnement et de séance dans votre formation ? », échelle de 1 à 5.

Par ailleurs, nous souhaitions recueillir le sentiment des participants sur l'intérêt de se trouver pour une formation de ce type en situation de travaux pratiques. Autrement dit, la possibilité de manipuler l'outil et observer directement les effets algorithmiques ou une représentation de sa pratique est-elle de nature à mieux comprendre les problématiques développés dans le discours et les supports pédagogiques ? À cette question, les étudiants ont là encore répondu plutôt favorablement (figure 6.11). L'articulation entre support et discours pédagogique et manipulation de l'outil semble donc avoir fonctionné correctement et a suscité l'intérêt des participants.

Enfin, sur l'ensemble du questionnaire, il ne semble pas exister de différences significatives entre les réponses des étudiants de première année et ceux de deuxième année. Les deux étudiants évaluant le moins favorablement la pertinence de la séance et l'intérêt de l'outil sont aussi ceux qui jugeaient avoir le moins besoin de formation. Difficile de dir, pour autant si cela montre un bon niveau de maîtrise technique des outils ou au contraire un désintérêt pour ces problématiques. Le cursus suivi avant l'entrée dans le supérieur, durant le lycée, ne semble pas non plus discriminant. Le champ d'expression

libre n'a pas donné lieu à beaucoup de commentaires, exceptée la suggestion de développer ce type de formation en lieu et place des formations générales en informatique, essentiellement bureautique, dispensées dans le cadre du C2i.

FIGURE 6.11 – Évaluation : intérêt des manipulations dans l'outil pour la compréhension



Réponses à la question : « La possibilité de manipuler l'outil de recherche d'information vous a-t-elle aidée à mieux comprendre les problématiques développées ? », échelle de 1 à 5.

6.6.2 Limites de l'expérimentation

À la suite des résultats que nous venons de présenter il convient de souligner les limites de cette expérimentation. La principale est bien entendu liée à la taille réduite de l'expérimentation, menée avec un seul groupe d'étudiants. Les conditions matérielles n'ont malheureusement pas permis d'envisager de mener d'autres séances avec d'autres étudiants avec la version finale de l'outil. Les expérimentations intermédiaires que nous avions menées dans des versions précédentes, afin de tester les fonctionnalités et affiner notre cahier des charges de développement n'avaient quant à elles pas vocation à démontrer la pertinence globale de l'outil et de la démarche de formation associée. L'évaluation présentée ici ne peut donc s'entendre que comme une preuve de concept dont l'expérimentation mérite d'être poursuivie.

Au delà du nombre d'étudiants impliqués, leur niveau est lui aussi un point d'interrogation. Si l'expérimentation montre qu'il y a un intérêt à la proposer relativement tôt (en première ou seconde année de Licence), difficile de dire si elle ne serait pas également pertinente plus tard, en troisième année de Licence ou en Master où il est évident que les besoins des étudiants évoluent. Un compromis pourrait sans doute être trouvée en traitant des stratégies de recherche d'information (notre partie 1) tôt dans la formation, et développer les aspects plus algorithmiques un peu plus tard. Ces questions peuvent en effet avoir un intérêt en lien avec l'enseignement d'outil désormais classiques des humanités numériques, comme les outils de lecture distante ou d'analyse de réseaux pour lesquels certains des biais que nous avons mis en avant peuvent aussi prêter à conséquence.

Pour ce qui est de l'outil lui même, si tout n'est pas parfait et qu'il mériterait d'être amélioré, en terme d'interface par exemple [15], nous avons pu mesurer que son intérêt majeur réside dans l'interactivité qu'il rend possible durant la séance. La nécessaire

15. La présentation des résultats bruts des requêtes mériterait par exemple d'être améliorée pour être plus facilement lisible par les utilisateurs.

mise en relation de l'outil avec un discours pédagogique rend toutefois assez difficile l'évaluation de ce point précis. À ce stade de notre réflexion et à travers la méthodologie d'évaluation et le questionnaire mis en place, il n'est pas possible de dire si les qualités et les défauts relevés son dus à l'outil lui même, ou au discours qui l'accompagne. Enfin, une démarche de formation de ce type ne peut sans doute pas être évaluée en dehors d'un parcours pédagogique plus global. Dans un premier temps sans doute serait-il nécessaire de mettre en relation notre dispositif avec un cours d'histoire. Ce lien permettrait de réinvestir les cas d'étude et les compétences développés dans l'une et l'autre des formations. Cela autoriserait ainsi, à n'en pas douter, une évaluation beaucoup plus fine de l'intérêt de notre démarche au sein d'un écosystème pédagogique réaliste.

6.7 Conclusion et perspectives

La démarche présentée dans ce chapitre constitue tout à la fois un aboutissement et un nouveau champ de perspectives. Un aboutissement d'abord parce qu'elle est fondée sur un démonstrateur logiciel qui n'aurait probablement pas pu être développé sans les contributions de nos chapitres précédents. Les fondements théoriques et techniques mis en place dans le chapitre 3 nous ont assuré une plateforme expérimentale stable et représentative des systèmes existants régulièrement manipulés par les étudiants. Les observateurs et indicateurs que nous avons étudiés dans le chapitre 4 nous ont permis de mettre à disposition des participants les moyens de visualiser et comprendre leur pratique de la recherche d'information. Enfin, les travaux du chapitre 5, implémentés dans notre outil, donnent à voir de manière très claire certains des biais que peuvent induire les divers processus menant un document de sa forme physique à sa forme numérique.

Plus qu'un démonstrateur, toutefois, l'outil et la démarche ici évalués ont montré, malgré les limites que nous avons identifiées, leur potentiel en matière de formation. Les participants semblent sensibles aux problématiques développées et curieux de réfléchir à leur pratique de recherche d'information et plus largement à leur usage du numérique dans le contexte de leurs travaux historiques. De ce point vue, l'outil que nous imaginions au début de son développement remplit son rôle. Il facilite la compréhension de processus informatique parfois complexes pour des étudiants naturellement éloignés de ces problématiques et est à même de nous aider à leur faire développer des stratégies d'usage des ressources numériques pertinentes pour leur activité universitaire. À l'heure ou l'usage du numérique est très largement répandu, il est un moyen de se doter d'un regard critique sur l'usage de ces ressources, ce qui apparaît pour la formation de tout historien, comme une nécessité.

Cette démarche ouvre par ailleurs, comme nous l'écrivions, d'autres perspectives. Lors des présentations de l'outil que nous avons eu l'occasion de faire dans des publications et communications [16], les échanges avec l'auditoire ont soulevé d'intéressantes problématiques et d'autres cas d'usage. La plus récurrente des suggestions a été, sans surprise, d'étendre cette démarche à d'autres disciplines qui pourraient elles aussi trou-

16. En particulier dans les publications et communications suivantes : [Suire et Estraillier, 2017; Suire et collab., 2017a,b].

ver un intérêt à une telle expérimentation. Techniquement, cette extension ne pose pas véritablement de problème et pourrait être menée. Néanmoins, sur le plan scientifique, elle nécessite de se pencher sur les comportements de recherche qui pourraient être spécifiques à ces disciplines.

Pour l'histoire, la recherche dans des sources primaires constituent par exemple une telle spécificité que l'on ne retrouvera pas dans d'autres contextes. Il serait ainsi important pour la validation de l'expérience de se pencher en détail sur l'état de l'art [17] et de valider au préalable les métriques implémentées dans l'outil. Une problématique très similaire se pose pour une autre suggestion récurrente, celle d'expérimenter la démarche avec d'autres publics, plus jeunes en l'occurrence, par exemple avec des élèves du secondaire. Là encore, des recherches doivent être menées pour généraliser l'approche et les moyens technologiques pour l'adapter à cette population.

Parmi les extensions possibles, nous avons également prévu d'expérimenter l'intérêt de l'outil pour l'enseignement en informatique, en particulier pour l'enseignement des algorithmes de traitement du texte, d'indexation et de pertinence. Ces méthodes sont enseignées dès la 2e année pour certaines d'entre elles jusqu'au Master pour les plus complexes. Notre outil pourrait avoir un intérêt pour révéler les effets de ces algorithmes et serait dans ce contexte un moyen de test intéressant pour les apprentis informaticiens. Nous y voyons par ailleurs un moyen pertinent pour leur faire comprendre que les choix techniques qu'ils sont amenés à faire régulièrement ne peuvent être déterminés qu'en fonction des seules considérations techniques, de stabilité ou de rapidité d'exécution. Une séance avec notre outil pourrait être un levier pour leur faire comprendre que ces choix ont des conséquences sur l'usage des utilisateurs finaux qui parfois les dépassent et qu'ils doivent être capables d'intégrer.

Enfin, parmi les suggestions qui nous sont parvenus, certaines proposent de détourner largement l'usage initial sur lequel nous avons travaillé. Il nous a ainsi été suggéré d'utiliser cet outil pour évaluer la qualité du fonctionnement de la bibliothèque numérique. Dans un contexte où l'on connaît les modèles de pratique optimaux et des indicateurs pertinents, il s'agirait de tester avec des utilisateurs un jeu de paramètres particulier et de le confronter avec les modèles connus, pour évaluer s'ils améliorent l'expérience ou si au contraire ils dégradent la qualité du service.

Nous avons prévu d'adresser prochainement certaines de ces problématiques. Néanmoins, elles nécessitent toutes un travail important pour adapter l'outil à de nouvelles conditions expérimentales. Quelques améliorations techniques sont d'ores et déjà planifiées pour faciliter l'usage de cet outil pour d'autres problématiques ou par d'autres équipes. Nous souhaitons en particulier simplifier la procédure d'ingestion des documents dans la plateforme. Quoi qu'il en soit, ces adaptations ne sont pas que d'ordre technique et ne pourront être opérées sans un dialogue interdisciplinaire avec les disciplines concernées, dans une démarche de co-construction similaire à celle que nous avons mise en place pour le cas présenté dans ce manuscrit.

17. Des éléments de l'état de l'art existent pour certaines disciplines. En droit, par exemple, les travaux de [Makri et collab., 2008] offrent déjà un cadre scientifique tout à fait pertinent.

6.8 Références

Bourgatte, M. 2017, «Pour un humanisme numérique en éducation», *Revue française des sciences de l'information et de la communication*, , n° 10. 163

Cardon, D. 2013, «Présentation», *Réseaux*, , n° 177, p. 9–21. 164

Duke, L. M. et A. D. Asher. 2012, *College Libraries and Student Culture : What We Now Know*, American Library Association. 163, 164

East, J. W. 2005, «Information literacy for the humanities researcher : A syllabus based on information habits research», *The Journal of academic librarianship*, vol. 31, n° 2, p. 134–142. 164

Kolowich, S. 2011, «What Students Don't Know», URL https://www.insidehighered.com/news/2011/08/22/erial_study_of_student_research_habits_at_illinois_university_libraries_reveals_alarmingly_poor_information_literacy_and_skills. 164

Makri, S., A. Blandford et A. L. Cox. 2008, «Investigating the information-seeking behaviour of academic lawyers : From Ellis's model to design», *Information Processing & Management*, vol. 44, n° 2, p. 613–634. 190

Marchionini, G. 2006, «Exploratory Search : From Finding to Understanding», *Commun. ACM*, vol. 49, n° 4, doi :10.1145/1121949.1121979, p. 41–46. 176

Moisson, P. 2011, «Les "digital natives" ont encore besoin des bibliothécaires | Enssib», URL http://www.enssib.fr/breves/2011/09/01/les-digital-natives-ont-encore-besoin-des-bibliothecaires. 164

Suire, C. et P. Estraillier. 2017, «Outils et méthodes informatiques pour l'enseignement d'une culture numérique : le cas de la recherche d'information», dans *Journée d'études : Recherche et développement dans le numérique scolaire : des projets aux données de recherche.*, Montpellier. 189

Suire, C., A. Jean-Caurant, R. Champagnat et M. Rabah. 2017a, «Atelier : Outil et démarche pédagogique pour une approche critique de la recherche d'information en ligne.», dans *Colloque Jeux et Enjeux*, La Rochelle. 189

Suire, C., A. Jean-Caurant et C. Illouz. 2017b, «Ouvrir les boîtes noires : un outil pédagogique pour une approche critique de la recherche d'information en ligne», dans *Digital Humanities (DH)*, Montreal. 189

Williams, S. 2010, «New Tools for Online Information Literacy Instruction», *The Reference Librarian*, vol. 51, n° 2, doi :10.1080/02763870903579802, p. 148–162. 164

Chapitre 7

Conclusion et perspectives

Sommaire

7.1 Approche et positionnement de la thèse

À l'origine de ce travail, nous avions pour objectif d'étudier l'impact des outils du numérique sur la pratique de la recherche en histoire. Pour réduire cette trop vaste ambition, nous avons mené un état de la pratique qui nous a permis d'observer que les outils numériques avaient désormais leur place dans toutes les phases et tous les champs d'activité de l'historien, avec beaucoup de nuances toutefois. Certains de ces outils, dont l'usage est attesté par la littérature, sont encore relativement confidentiels à l'échelle de la recherche parce que dépendant d'approches scientifiques spécifiques, par nature favorables à l'utilisation du numérique, comme l'analyse de réseaux par exemple. D'autres outils, en revanche, sont bien installés dans les pratiques. C'est le cas des outils liés à l'accès, à la capture ou à la gestion des sources historiques. L'usage de cette famille d'outils est cette fois bien attesté, largement pour l'appareil photo numérique, plus diversement pour les bases de données et moteurs de recherche de sources en ligne.

Ce sont ces derniers qui ont été au cœur de notre questionnement. L'analyse de la littérature scientifique a montré que leur fonctionnement s'apparentait parfois à une boîte noire qui conduit à des effets de bord importants sur la recherche. Dans l'optique de révéler, expliquer et permettre aux utilisateurs de prendre en compte ces biais, nous avions posé les trois questions de recherche suivantes :

— Sur le plan conceptuel, comment croiser les exigences de la démarche scientifique de l'historien avec les effets des choix de conception et des compromis techniques opérés par l'informatique des moteurs de recherche de sources ?

— Du côté du système, bibliothèque numérique et moteur de recherche, par quels moyens mettre au jour et expliquer à l'historien les biais induits par les processus informatiques impliqués depuis la numérisation jusqu'à la recherche d'information ?

— Sur la base de ces moyens, par quelle démarche favoriser une meilleure maîtrise du système et une meilleure appropriation méthodologique des outils de recherche de sources en ligne ?

Ces questions ont structuré ce manuscrit. Les chapitres 2 et 3 ont ainsi positionné notre cas d'étude au sein d'un écosystème plus vaste, dans lequel nous trouvons de nombreux outils utiles à la recherche, depuis les plateformes d'accès aux sources jusqu'aux outils d'écriture collaborative en passant par les nombreux logiciels d'analyses des données. Dans ces chapitres, nous avons défini un cadre conceptuel, autour de la notion de ressource numérique, à même d'autoriser le croisement des usages avec les logiques de conception qui régissent le fonctionnement des outils. Ce cadre conceptuel a pointé la nécessité de disposer d'un outil d'observation des usages suffisamment représentatif de l'existant. Nous avons donc évalué de manière comparative un échantillon de plateformes existantes, pour garantir la représentativité de notre outil sur le plan fonctionnel.

Dans ce cadre conceptuel, nous avons pu nous intéresser aux différentes facettes impliquées dans le fonctionnement des bibliothèques numériques et moteur de recherche de sources. Notre chapitre 4 s'est focalisé sur le contexte d'usage. Nous nous sommes attachés à y décrire les modèles conceptuels et modèles d'interactions de l'état de l'art

et à y définir une mécanique d'observation et une méthode de traitement de l'usage des utilisateurs en situation de recherche d'information.

En parallèle, nous nous sommes intéressés à étudier les biais issus des autres versants du système, les contextes de production et d'exécution. Pour ce faire, nous avons choisi, dans notre chapitre 5, de prendre pour cas d'étude la problématique de la reconnaissance optique de caractères sur les entités nommées, qui introduit des biais liés aux erreurs d'interprétation du processus. Nous avons mis en place une méthode permettant de révéler les effets du bruit de l'OCR sur cette catégorie de mot que nous avons évaluée en nous concentrant sur leurs conséquences pour l'activité de recherche d'information.

Sur la base des résultats et des outils logiciels développés dans ces deux chapitres, étudiant indépendamment les différents contextes de la ressource numérique, nous avons proposé une approche et un outil permettant de les réinvestir dans une logique de formation. Cette démarche vise à répondre à notre dernière question de recherche et donc à favoriser une meilleure maîtrise et une meilleure appropriation méthodologique des systèmes de consultation de sources en ligne.

7.2 Contributions

En dehors du positionnement et de l'approche globale que nous venons de rappeler, les travaux décrits dans ce manuscrit ont plus précisément permis de contribuer et proposer des pistes d'amélioration à l'étude des pratiques de recherche d'information dans le contexte de la recherche en histoire autour des thématiques suivantes.

7.2.1 Mécanique d'observation et indicateurs de l'usage de la recherche d'information

Comme l'a montré notre analyse de l'état de l'art, l'activité de recherche d'information dans une bibliothèque numérique ou un moteur de recherche de documents est une activité complexe, aussi bien sur les plans conceptuels qu'opérationnels. Les modèles conceptuels, des plus généraux aux plus spécifiques, ont précisé les enjeux des activités de recherche d'information. Étant donné qu'ils sont avant tout des constructions théoriques qui ne se focalisent que sur les actions menées par les utilisateurs, nous avons complété l'état de l'art par des modèles d'interactions qui prennent en compte le système.

Ces modèles ne nous renseignant pas sur les modalités pratiques de l'observation ; nous avons passé en revue les méthodes déjà employées dans la littérature. Aucune d'entre elles ne permettait d'aboutir à notre objectif global, confronter la pratique utilisateur avec le fonctionnement technique du système. Nous avons donc pu conclure à la nécessité de disposer de notre propre outil. Cette approche présentait l'avantage de nous permettre d'observer les interactions des deux côtés, utilisateur et système, et faire varier à notre convenance le contexte d'exécution et les corpus hébergés.

Nous avons donc pu mettre en place différents observateurs tenant compte de l'expression des utilisateurs d'une part et des réponses du système d'autre part (section 4.3).

Ces traces brutes de l'activité ont ensuite été transformées en indicateurs de plus haut niveau. Si certains de ces indicateurs avaient déjà été évalués par la littérature dans certains contextes de recherche d'information, ils ne l'avaient pas été pour des tâches, des corpus et des utilisateurs des sciences humaines et sociales. Nous avons donc cherché à en évaluer la pertinence en adoptant une démarche statistique. Nous avons pu montrer que ces indicateurs sont des témoins intéressants de la pratique des utilisateurs, dans leur grande majorité si l'on tient compte de l'intégralité de la tâche et avec quelques nuances si l'on restreint l'analyse à la première recherche (section 4.5).

Ceci dit, ces résultats sont issus d'une démarche expérimentale et doivent donc être entendus avec un certain recul. Bien que nous ayons cherché à contrôler bon nombre de facteurs influençant les résultats (section 4.4), ceux-ci sont très dépendants des participants et des corpus proposés à l'expérimentation. Renouveler ces expériences avec un public élargi et différents ensembles documentaires permettrait toutefois de dépasser cette limite du travail. Cette limite, due à la nature de la validation expérimentale que nous avons mise en place, n'impacte toutefois pas le réinvestissement des résultats que nous avons été en mesure de proposer, puisque les utilisateurs ainsi que les corpus impliqués sont similaires.

La mécanique d'observation et la démarche de validation associée nous ont donc permis de développer le calcul de ces indicateurs dès la fin d'une tâche ou d'une session de recherche. Ces développements ont favorisé le réinvestissement de ces indicateurs dans la version dédiée à la formation de notre outil. Les participants aux séances de formation peuvent ainsi prendre connaissance de leur pratique immédiatement après leurs sessions de recherche d'information. Il est ainsi très aisé de comparer les résultats à l'échelle individuelle, entre différentes tâches et à l'échelle du groupe (section 6.4). La production rapide de ces indicateurs et leur pertinence favorisent la visualisation et la compréhension par les participants des éventuels biais de recherche produits par des stratégies de recherche souvent inadaptées aux tâches demandées et aux contextes d'exécution et de production du système.

7.2.2 Impact des erreurs d'OCR sur les entités nommées et pratique de la recherche d'information

En ce qui concerne le contexte de production de la ressource numérique, les éléments d'état de l'art présentés dans notre chapitre 3 ont montré qu'il pouvait être la source de nombreux biais pour le chercheur, utilisateur du système de recherche d'information. Depuis les problématiques de qualité de la numérisation et de l'extraction d'information jusqu'aux logiques de classement et de construction des métadonnées, les biais potentiels que nous aurions pu étudier sont nombreux.

Comme nous l'avons expliqué dans notre chapitre 5, nous avons choisi d'étudier spécifiquement le problème de la reconnaissance optique de caractères parce qu'il nous a semblé, sur la base de l'état de l'art, particulièrement sensible pour l'histoire. Nous avons par ailleurs focalisé notre attention sur les entités nommées pour étudier le contexte de production en gardant un lien étroit avec l'usage. De ce point de vue, les entités nommées sont pour les utilisateurs une des portes d'entrée principales dans les sources. Elles sont également particulièrement soumises aux erreurs de reconnaissance.

À la suite des travaux de l'état de l'art, nous savions que ces erreurs de reconnaissance pouvaient avoir un impact sur la recherche d'information.

Pour révéler ces biais aux utilisateurs et leur permettre de les intégrer, il était nécessaire de mettre en place une méthode permettant d'identifier au mieux ces entités nommées dans des corpus de documents soumis à du bruit OCR. La méthode évaluée, qui fonctionne en regroupant des mots en fonction d'une combinaison des similarités de contexte et d'édition, présente l'avantage d'identifier ces entités nommées bruitées sans connaissance a priori, c'est à dire sans annotation manuelle préalable. Elle autorise donc de présenter des alternatives sans correction systématique, laissant le choix d'inclure ou non les termes découverts dans la recherche à l'utilisateur.

Cette méthode a malgré tout ses limites. Elle est d'abord dépendante des systèmes de détection des entités nommées, qui bien qu'ayant été entraînés avec un modèle adapté, ne détectent pas toutes les entités. Ensuite, les paramètres de cette méthode doivent être choisis avec soin et déterminés empiriquement pour trouver le meilleur compromis. Avec un jeu de paramètre inadapté, la méthode peut générer beaucoup de bruit, amoindrissant d'autant son intérêt. Enfin, c'est une méthode coûteuse en temps de calcul qui la rend en pratique difficile à exécuter pour de grands corpus sur des machines grand public.

Malgré ces limites, cette méthode produit des résultats intéressants. Au regard de nos objectifs, son évaluation se devait de tenir compte de la finalité de notre étude. Nous avons donc mise en place une méthode d'évaluation de l'impact de la méthode sur l'activité de recherche d'information en prenant le parti de comparer les réponses du moteur de recherche pour des requêtes incluant ou non les alternatives. Suite à cette évaluation, la méthode a montré un impact significatif des erreurs d'OCR sur la recherche d'information, qui varie toutefois en fonction des contextes d'exécution et d'usage, algorithmes de pertinence utilisés ou nombre de résultats pris en compte (section 5.4).

7.2.3 Outils de prise en compte des erreurs d'OCR

La méthode que nous avons mise en place a donné lieu à la production de deux outils décrits à la section 5.3. Le premier prend la forme d'une fonctionnalité intégrée à notre bibliothèque numérique. Lorsque l'utilisateur saisit un ensemble de mots-clés dans le moteur de recherche, le programme vérifie s'il existe au moins une alternative de chacun des termes saisis par l'utilisateur. Dans le cas où de telles alternatives existent, le système les affiche à l'écran sous forme de suggestions et propose à l'utilisateur de les inclure dans sa recherche d'information.

Nous n'avons pas été en mesure d'évaluer spécifiquement cette fonctionnalité dans le contexte strict de l'usage d'une bibliothèque numérique, cela constitue une des limites de notre travail. En revanche, elle a pu l'être dans le contexte de notre démarche de formation pour laquelle la problématique de la reconnaissance de caractères constitue un point important. À l'aide de cette fonctionnalité les étudiants sont en capacité d'observer directement l'impact de ces erreurs. Ils peuvent comparer les résultats produits par la plateforme avec et sans l'inclusion des alternatives. Cette observation peut être mis en parallèle par le formateur avec les visualisations de résultats à plus grande échelle,

représentation globale de l'accessibilité des documents ou distribution des documents en fonction des différents scores calculables.

La seconde implémentation disponible est une version autonome de la méthode. Sous la forme d'une application web, elle permet aux utilisateurs de soumettre leurs propres corpus à l'analyse et dispose d'une interface de configuration des paramètres les plus importants de la méthode. À l'issue de la procédure, l'outil rend disponible les résultats et donne la possibilité de visualiser les différentes formes des entités détectées dans leur différents contextes. Cet outil est évidemment utile dans pour les documents bruités par un processus d'OCR, il a également un intérêt potentiel pour d'autres tâches impliquant des mécanismes similaires, la transcription des sources par exemple. Il faut toutefois noter que cet outil autonome n'a pas été évalué pour ces autres tâches et n'a pour l'heure pas fait l'objet d'une évaluation auprès d'utilisateurs.

7.2.4 Démarche et outil de formation

L'approche et l'outil de formation que nous avons présentés dans notre chapitre 6 est l'aboutissement de notre démarche de recherche. Comme nous venons de le décrire, il est issu du réinvestissement de nos résultats précédents aussi bien pour l'observation du contexte d'usage que pour la révélation des biais du contexte de production de la ressource. Nous avons également doté l'outil de moyens de simulation du contexte d'exécution décrits à la section 6.3. Il permet donc de démontrer des biais méthodologiques provenant des trois contextes définis par notre cadre conceptuel.

Nous aurions sans doute pu réinvestir les résultats de nos différents chapitres d'une autre manière, en proposant et en évaluant des fonctionnalités logicielles d'accompagnement à la recherche de sources. Cependant, nous avons préféré développer une démarche de formation en réponse à une limite de l'état de l'art. En effet, les enquêtes sur l'usage du numérique ont montré que la formation à ces thématiques était utile pour l'historien, faisant semble-t-il partie d'un socle de connaissance devenu fondamental avec l'émergence du numérique. Par ailleurs, notre lecture de l'état de l'art a montré que la thématique de l'*information literacy* était cruciale (voir section 6.1) pour permettre aux étudiants de développer des stratégies de recherche adaptées face à l'explosion du volume de documentation disponible. Malgré l'importance de la méthodologie de recherche d'information pour les étudiants en histoire, comme dans d'autres disciplines d'ailleurs, il n'existait pas à notre connaissance d'outil autorisant une confrontation entre la technique de recherche, autrement dit l'usage, et la logique de fonctionnement des moteurs de recherche.

L'expérimentation de cette démarche a montré sont intérêt pour les étudiants en histoire en début de cycle universitaire, malgré certaines limites. En premier lieu, cette expérimentation est à considérer comme une preuve de concept plus qu'une véritable évaluation, dans la mesure où elle a été menée avec un groupe restreint d'étudiants. En dehors de cet aspect et dans l'optique de généraliser une telle approche, il est important de relever qu'il existe d'autres contraintes.

Si les contraintes matérielles d'un enseignement en présentiel devant des machines sont aisément surmontables, celle de l'hébergement de l'outil dans sa forme actuelle sont plus problématiques. L'outil nécessite en effet une pile technologique complexe

dont l'installation doit être faite sur des serveurs, par des spécialistes. Sur le plan pédagogique, de notre point de vue, cette démarche a un plus grand intérêt si elle est directement liée à la formation historique, c'est à dire si les corpus et les tâches de recherche offerts aux étudiants sont en cohérence avec leur enseignement disciplinaire du moment. Dans notre contexte local, nous avons été en mesure de faire ce lien, mais cela nécessite un important travail d'échanges avec les collègues impliqués, surtout pour la constitution des corpus.

7.3 Perspectives

Pour conclure ce manuscrit, nous présentons ici les pistes de développement que nous souhaitons suivre à l'issue de ces travaux. La première d'entre elles est le développement de l'approche conceptuelle que nous avons mise en place et son extension par l'introduction de nouveaux cas d'étude et d'utilisation. La seconde concerne les outils qui ont été élaborés dans le cadre de ces travaux, nous évoquerons la poursuite de leur développement et leur mise à disposition pour la communauté.

7.3.1 Extension de l'approche de recherche

L'approche de recherche que nous avons présentée dans notre introduction (section 1), représentée graphiquement à la figure 1.4 est une proposition que nous avons souhaitée la plus générique possible. Dans le cadre d'un travail de thèse, nous avons bien entendu été forcé de faire des choix et avons été conduit à limiter l'approche en nous focalisant sur des cas d'étude. Cependant, le cœur de la proposition demeure adaptable à d'autres cas.

Pour ce qui est du contexte de production, par exemple, nous nous sommes focalisé sur la problématique de la reconnaissance de caractères. Ceci dit, nous aurions tout aussi bien pu prendre en compte d'autres traitements, générateurs de biais différents. Il aurait sans doute été possible de nous intéresser aux problèmes liés aux méthodes d'indexation et de description des documents, modélisation et lacunes des métadonnées ou biais de l'indexation par sujet par exemple. Dans ce cas, le réinvestissement possible, aussi bien en terme de nouveaux services offerts qu'en terme d'outil de formation aurait été différent. La focale aurait probablement été déplacée vers les problématiques de modélisation des contenus au détriment des logiques de recherche d'information, mais l'approche globale serait restée la même et les résultats obtenus seraient complémentaires.

Il aurait toujours été question de confronter des usages spécifiques, ceux de l'historien, avec des choix de conception et des logiques d'exécution, pour produire des moyens d'explication et d'appropriation de leurs éventuels effets. Suivant la même logique, nous pourrions envisager de modifier les besoins de l'utilisateur, c'est à dire étudier d'autres cas que celui de l'historien ou d'autres types de corpus. Cela nécessiterait de faire l'état de l'art des stratégies de recherche d'information pour le public considéré, mais cela n'impacterait pas le cœur de l'approche.

À titre d'exemple, lors d'une de nos récentes présentations, une collègue du domaine

des sciences de l'information et de la communication nous suggérait d'utiliser notre approche pour étudier la problématique des *fake news* ou plus généralement des sources d'information douteuses du web. Il s'agirait, sur la base d'un corpus contenant des sources de bonne et de mauvaise qualité, d'observer comment un public s'empare des sources pour différentes tâches et produire des recommandations et des moyens d'explication pour leur permettre de choisir les meilleures sources d'information.

Enfin, il est également envisageable de changer la nature de la boucle de pertinence que nous présentions dans notre introduction et qui a pris, pour notre travail, la forme d'un accompagnement méthodologique sous les traits d'une démarche et d'un outil de formation. Il serait par exemple possible de réfléchir à des mécanismes d'adaptation. En fonction du contexte d'usage détecté, sur la base des indicateurs très tôt discriminants de la pratique (voir section 4.5.2), il serait possible de changer le comportement du moteur pour adapter son fonctionnement à la tâche en cours. Cette adaptation pourrait également fonctionner dans le sens inverse, c'est à dire adapter le fonctionnement du moteur sur la base du type de données indexées. Si les sources ont subi tel ou tel traitement dont on connaît l'impact sur la pratique de l'historien, alors il est envisageable d'adapter le fonctionnement de la bibliothèque numérique pour en minimiser les biais, sans imposer à l'utilisateur de tenir compte des contraintes du contexte de production.

7.3.2 Évolution des outils

Le second axe de perspective auquel nous souhaitons nous attacher concerne la suite à donner aux outils développés dans le cadre de ces travaux. En ce qui concerne l'outil autonome de détection et visualisation des différentes formes d'entités bruitées, son développement est terminé et il sera prochainement déployé et mis à disposition de la communauté pour une première phase de test. Nous prévoyons toutefois déjà d'y ajouter des fonctionnalités. Nous planifions en particulier de le doter d'une fonction d'import de documents plus puissante, qui puisse importer des documents avec leur métadonnées, plutôt que des textes seuls. Cela permettra aux utilisateurs de retrouver plus facilement le contexte du document et expliquer l'origine des différentes formes trouvées pour un même mot, erreur d'OCR, écriture différente, choix de transcription, *etc.*

Pour notre outil de formation, les pistes de développement sont nombreuses. Dans un premier temps, nous souhaitons étendre son expérimentation à un plus grand nombre d'étudiants et une plus grande hétérogénéité de niveaux. Nous prévoyons donc de mener des séances de formation avec des groupes d'étudiants de la première année de Licence au Master et affiner la démarche en fonction des besoins pédagogiques de ces différents niveaux. Par ailleurs, une expérimentation à destination d'autres disciplines est également prévue, avec des objectifs de formation similaires. Enfin, avec des objectifs cette fois différents, nous souhaitons tester l'intérêt de l'outil pour la formation d'étudiants en informatique, dans le contexte de l'enseignement de l'indexation et du traitement du texte.

Dans un second temps, nous souhaitons changer profondément son architecture et sa pile technologique pour faciliter son déploiement. Nous avons pour projet de le rendre autonome et installable directement sur tout poste informatique et limiter l'usage du serveur à l'indexation des corpus. Ce changement technique sera l'occasion de faire évo-

luer sa logique de fonctionnement. Jusqu'à présent, seul le formateur pouvait manipuler les paramètres du moteur de recherche, les actions des étudiants se limitant à constater les effets des configurations choisies. Si nous faisons évoluer l'outil dans cette direction, les utilisateurs auront la possibilité de modifier eux-mêmes certains des paramètres. Nous pourrons ainsi vérifier l'intérêt de cette fonctionnalité sur le plan pédagogique.

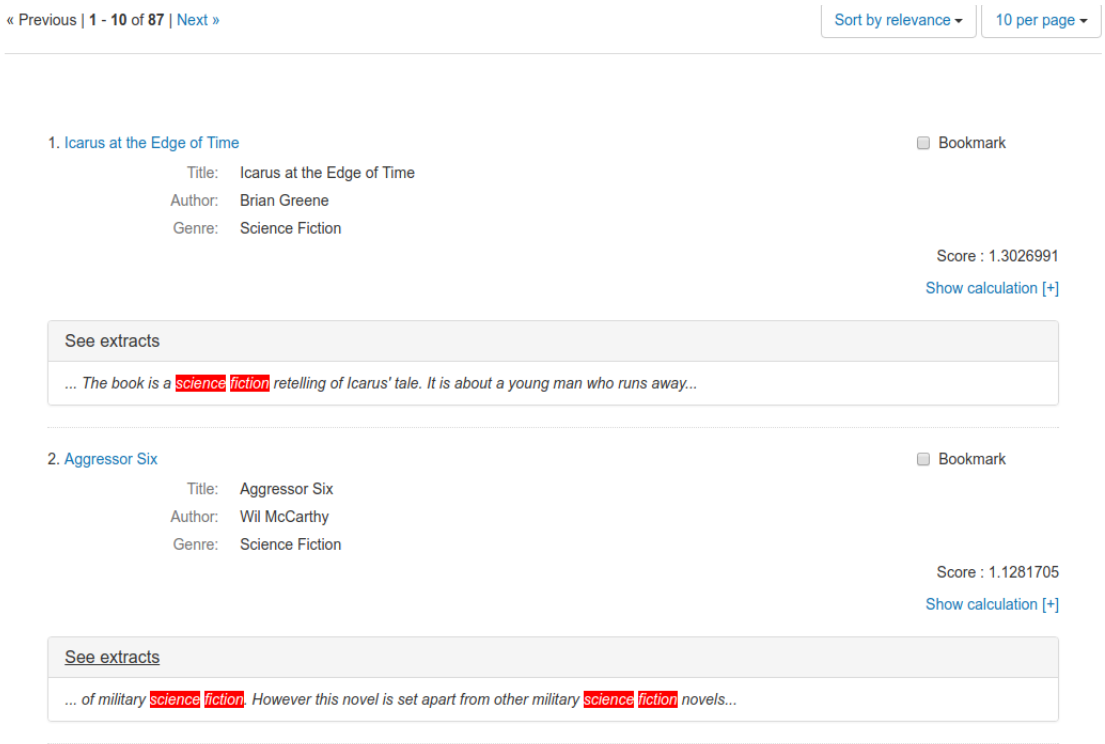
Annexe A

Annexes

A.1 Figures annexes

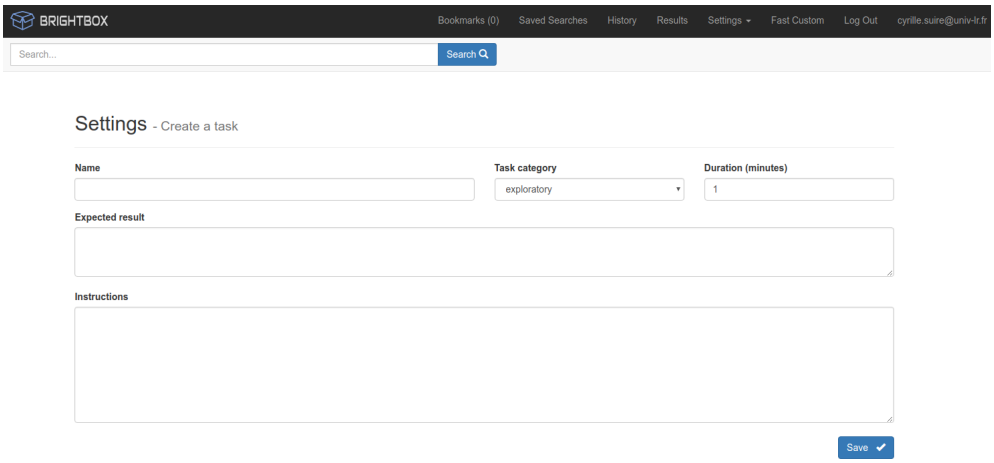
A.1.1 Interfaces des outils expérimentaux

FIGURE A.1 – Exemple de résultats avec *highlighting* activé



Les termes de la requête saisie par l'utilisateur sont ici surlignés en rouge par le système.

FIGURE A.2 – Interface de création de tâche par l'enseignant



Cette interface permet de renseigner le nom, la catégorie, la durée maximale et les consignes de la tâche. Il est aussi possible de renseigner une correction, présentée aux participants à l'issue de la tâche.

FIGURE A.3 – Exemple de visualisation du processus d'accomplissement d'une tâche

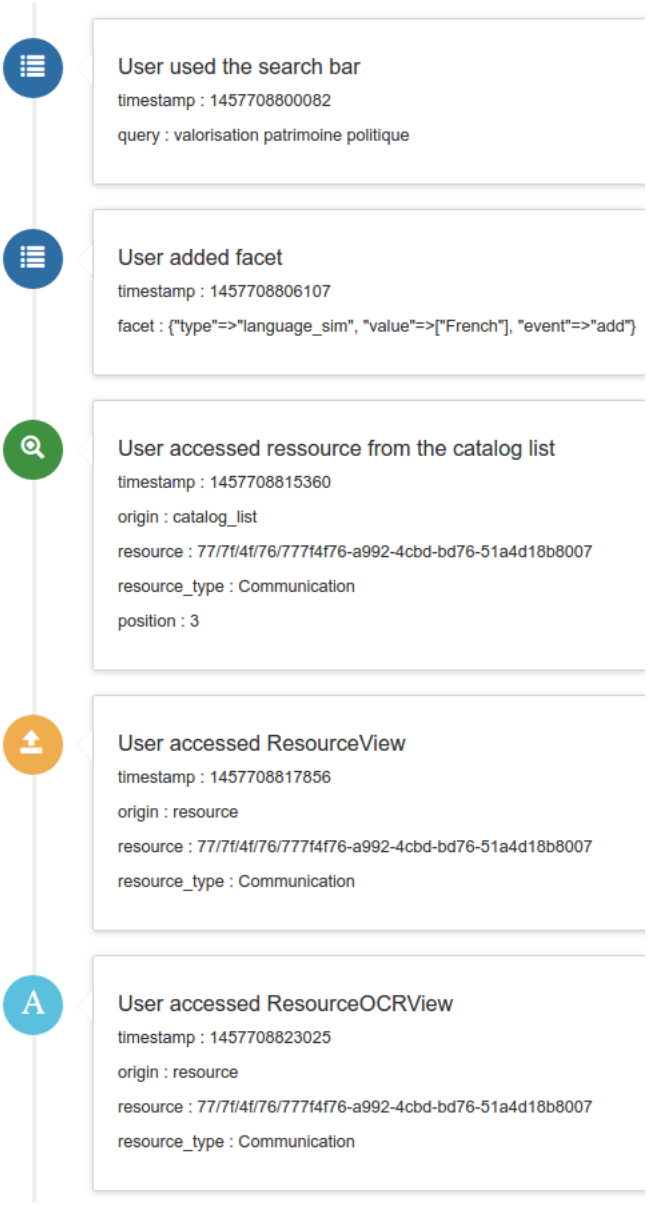
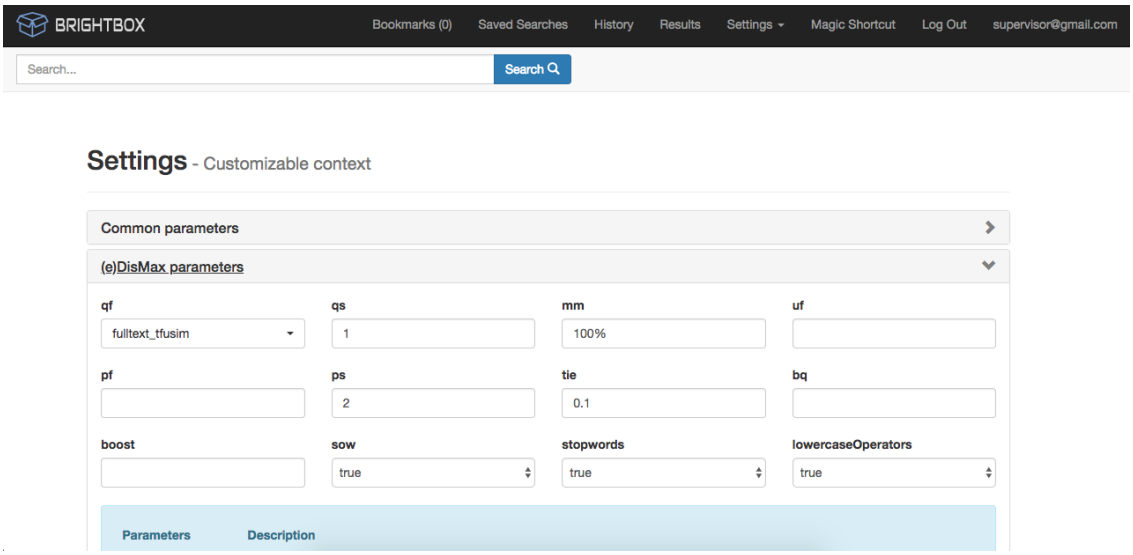
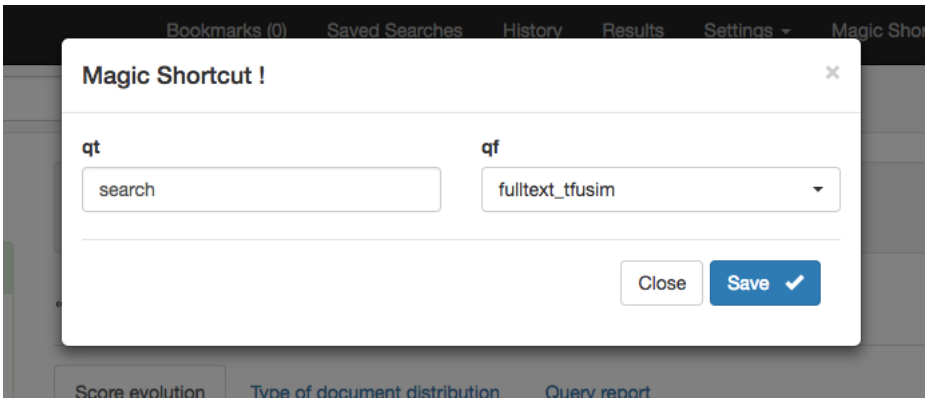


FIGURE A.4 – Vue de l'interface de configuration des paramètres du moteur



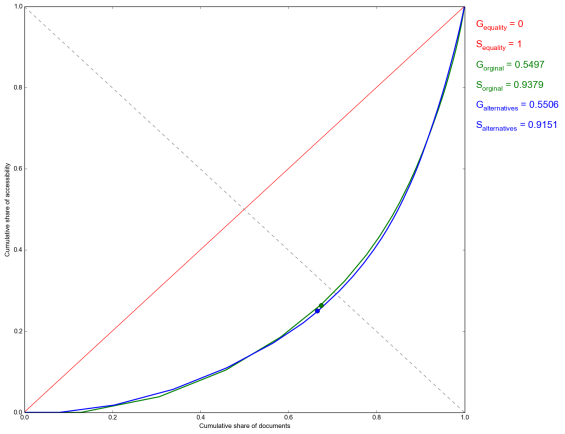
Cette interface permet de configurer tous les paramètres du moteur et du traitement des documents, y compris les plus fins.

FIGURE A.5 – Vue de l'interface de configuration rapide

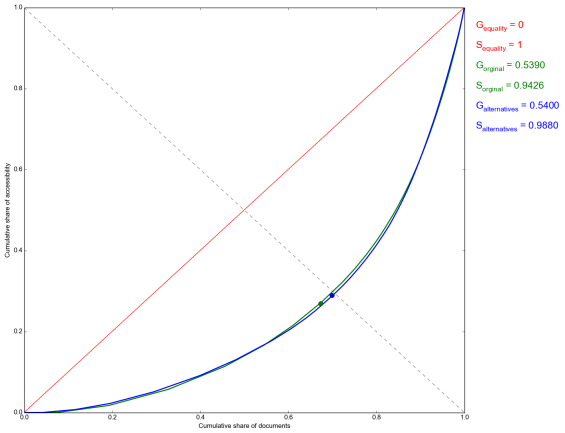


Cette fenêtre, accessible de partout dans l'application, permet uniquement de configurer les champs de métadonnées et de données sur lesquels opère le moteur et de changer de *query handler* (paramètre *qt*).

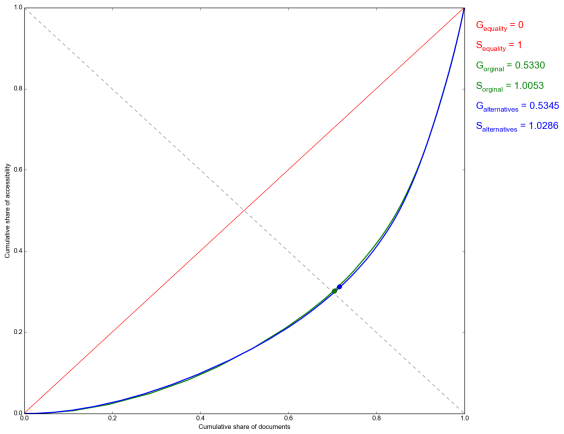
A.1.2 Représentations de l'accessibilité des corpus A et B avec *tf-idf*



(a) c = 10

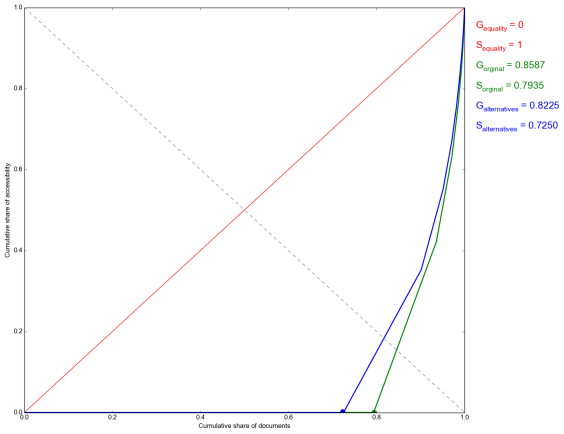


(b) c = 20

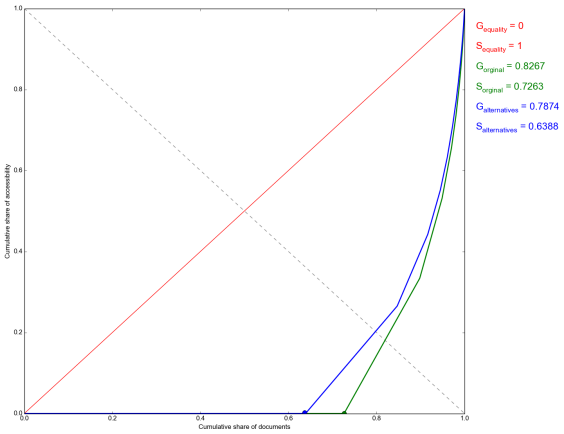


(c) c = 50

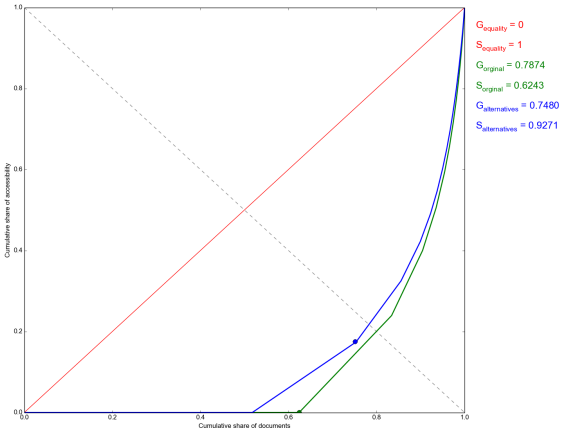
FIGURE A.6 – Représentation de l'accessibilité pour c = 10, c = 20 et c = 50 avec *tfidf*



(a) c = 10



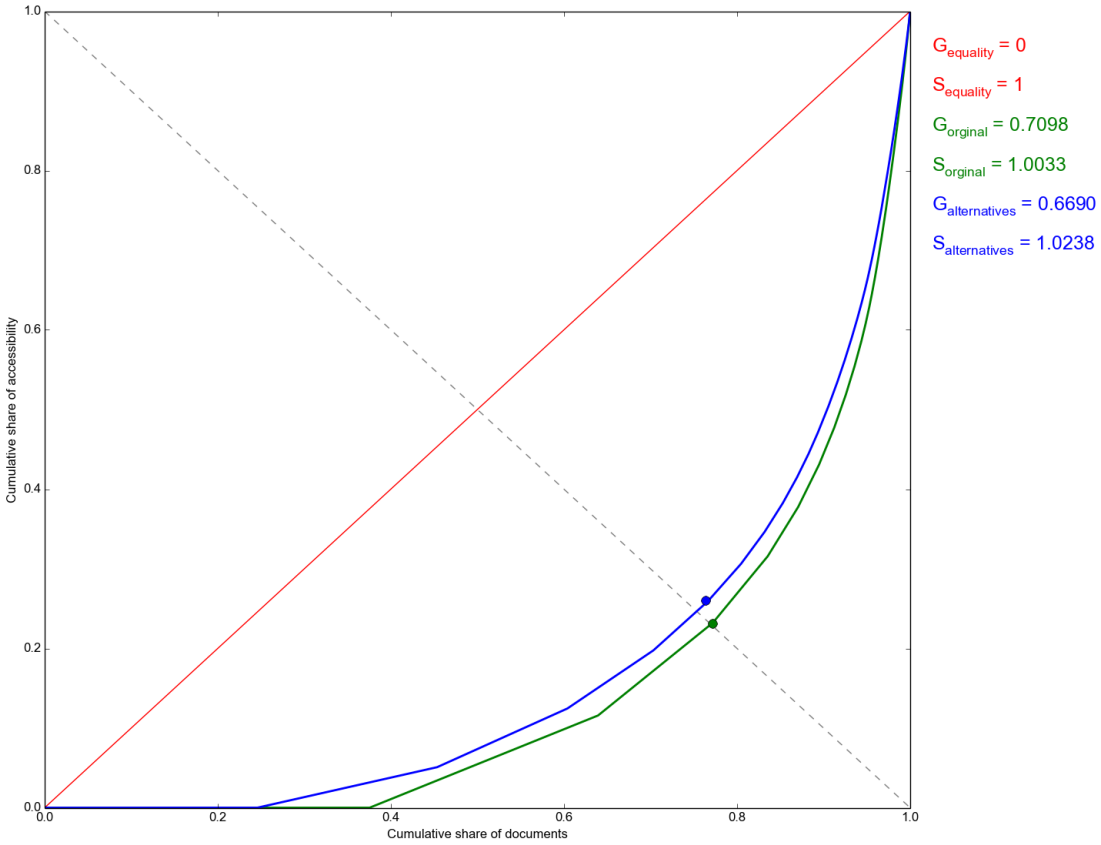
(b) c = 20



(c) c = 50

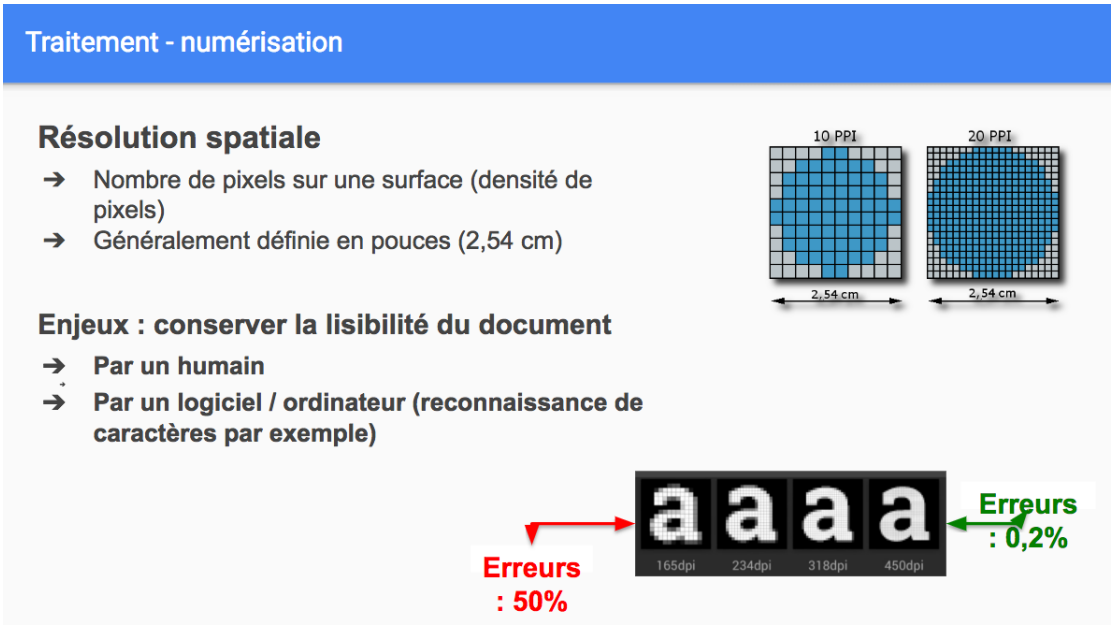
FIGURE A.7 – Représentation de l'accessibilité pour c = 10, c = 20 et c = 50 avec *tfidf*

FIGURE A.8 – Corpus B : Accessibilité globale sans seuil c avec *BM25*



A.1.3 Exemples de support pédagogique pour l'explication des traitements subies par les données de la bibliothèque numérique

FIGURE A.9 – Exemple de support : résolution spatiale et OCR



Cette figure explique le concept de résolution spatiale et montre son impact sur la qualité de la reconnaissance optique de caractères.

FIGURE A.10 – Exemple de support : image originale et texte reconnu

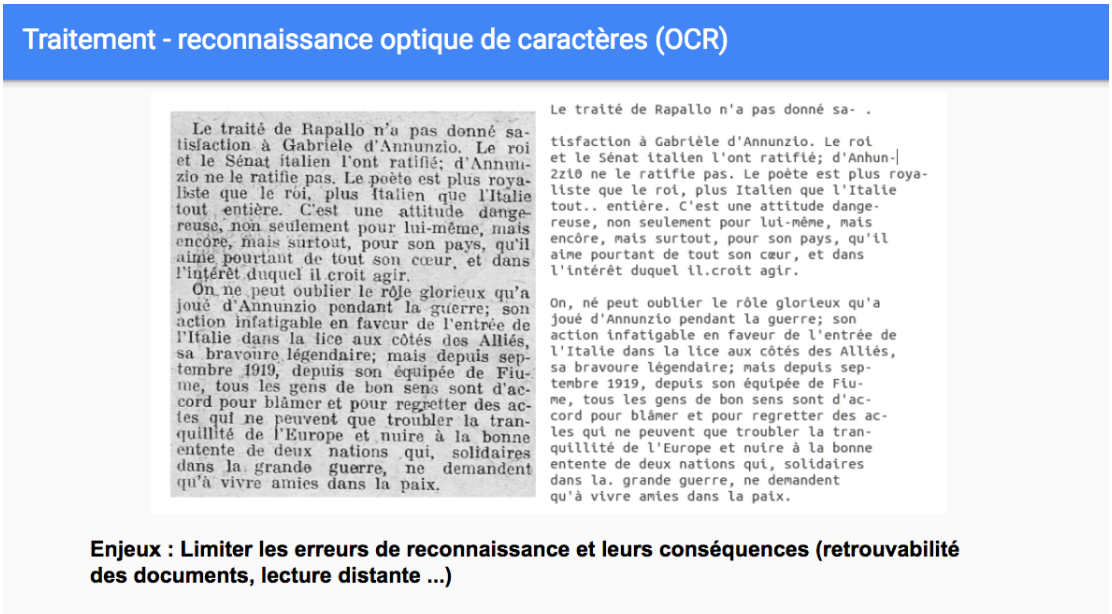
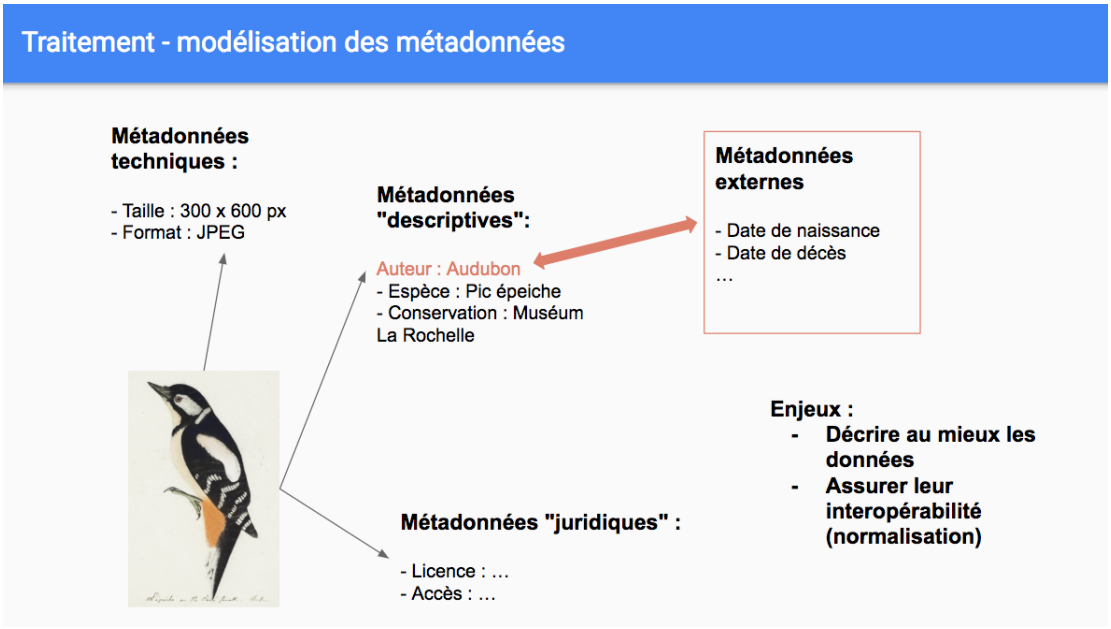
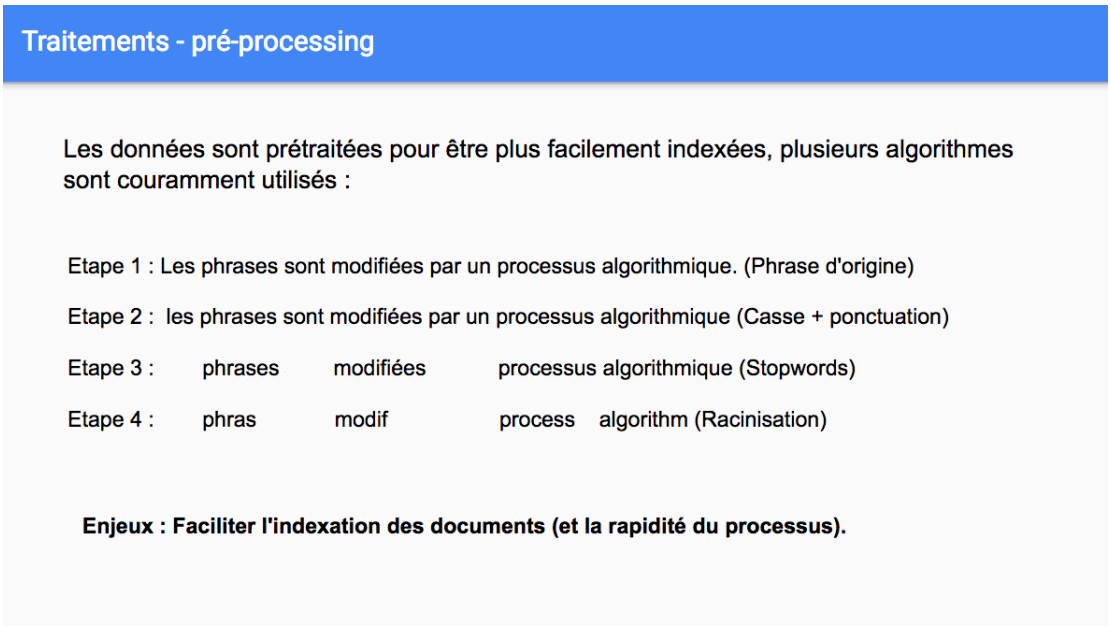


FIGURE A.11 – Exemple de support : modélisation des métadonnées



Ce support présente les différents types de métadonnées et les enjeux de leur modélisation.

FIGURE A.12 – Exemple de support : pré-processing du texte indexé et des requêtes



Ce support permet d'expliquer les traitements que peuvent subir les données textuelles issues des sources ou des requêtes.

A.2 Tableaux annexes

A.2.1 Résultats expérimentaux pour l'évaluation du comportement

Tableau A.1 – Évaluation de la phase de prise en main de la plateforme (comparaison $T1/T2$)

	first query (f')	overall task (f)
	T1/T2	T1/T2
	query length	
p	*	0.97
Z	-2.35	-0.04
	duration	
p	0.12	0.47
Z	-1.55	-0.73
	max scroll value	
p	0.21	0.18
Z	-1.26	-1.34
	number of clicked items	
p	0.39	0.36
Z	-0.87	-0.91
	number of viewed documents	
p	*	*
Z	-2.27	-2.07
	dwelling	
p	*	**
Z	-2.19	-2.87

Différence statistique entre deux tâches de même nature (*fact finding*), exécutées dans l'ordre $T1$ puis $T2$. La partie gauche du tableau expose les résultats durant la première recherche quant la partie droite présente les résultats pour l'intégralité de la tâche. Les résultats sont calculés à l'aide d'un test des rangs signés de Wilcoxon. Les valeurs p-values avec * sont statistiquement signifiantes avec * for $p < 0.05$, ** pour $p < 0.01$ et *** pour $p < 0.001$.

Tableau A.2 – Résultats expérimentaux complémentaires pour la tâche $T4$

Indicateurs	Valeurs
Documents ouverts (toutes sources confondues)	217
Nombre total de requêtes	358
Nombre de requêtes uniques	127
Nombre de requêtes contenant un filtre par facette	33

A.3 Détail des événements observés

A.3.1 Expression du besoin d'information

Requête (*query*) La requête est l'expression du besoin d'information sous forme de texte, principalement des mots-clés saisis par l'utilisateur dans la barre de recherche de l'application. Lorsque l'utilisateur a saisi les termes de sa requête, celle-ci est transmise au serveur. Ce dernier enregistre alors cette chaîne de caractères sous forme d'un événement que l'on appelle "*query*" puis calcule et transmet à l'utilisateur la liste des résultats correspondant à cette requête.

Filtrage par facette (*facet*) Le second moyen utilisable pour exprimer le besoin d'information est le filtrage par facette (*facetting search*). L'utilisateur peut sélectionner une ou plusieurs valeurs de filtrage parmi les catégories disponibles dans l'application, telles que la langue d'une ressource ou son type. La valeur de cette sélection est enregistrée par le serveur de l'application sous la forme d'un événement "*facet*".

A.3.2 Exploitation des résultats de recherche

Résultats de recherche (*display*) Afin de pouvoir analyser la manière dont les utilisateurs ont traité les résultats de recherche, le système garde une trace de la liste des résultats générée par le moteur lors d'une recherche. Cette trace est naturellement associée à la trace de l'événement qui a conduit à la production de cette liste de résultats.

Documents visibles (*visible*) Dans la liste des résultats de recherche, compte tenu des métadonnées affichées en plus du titre ou de la référence de la ressource, tous les résultats ne peuvent pas être visibles à l'écran. Suivant le taille et la résolution de l'écran seuls quelques résultats sont affichés, de l'ordre de 3 à 4. Notre application dispose donc d'un programme capable d'observer l'affichage de résultats supplémentaires. Cette observation résulte d'une opération de *scroll* de la part de l'utilisateur. Lorsqu'un utilisateur « descend » dans la liste de résultat, le programme enregistre une trace des documents qui deviennent visibles, après un court délai (250 ms) permettant d'éviter d'enregistrer des traces issues de *scroll* trop rapide pour que le document ait été réellement visible par l'utilisateur. Cet événement ne doit pas être confondu avec l'événement correspondant à un changement de page dans la liste des résultats. L'application tient toutefois compte de ces changements de page pour l'observation des ressources visibles.

Documents vus (*show*) L'application garde également une trace de l'événement « document vu ». Cet événement est déclenché lorsque un utilisateur clique sur un

résultats de recherche. A la suite de cette action, l'utilisateur peut visualiser l'intégralité des métadonnées disponibles sur ce document et a accès aux différents moyens de consommation du document, lecture ou téléchargement.

Documents vus par recommandation (*recommendation*) Lorsque l'utilisateur clique sur un document, pour en visualiser les détails, à partir d'un lien de recommandation, une variante de l'événement *show* est émise. La trace de cet événement contient les mêmes informations, mais précise simplement l'origine de l'événement. Cela permet de mesurer l'impact des recommandations sur la pratique de l'utilisateur.

Recherches sauvegardées (*saved search*) L'utilisateur peut s'il le souhaite sauvegarder dans l'application les résultats d'une recherche. Cette action peut signifier la pertinence de cette recherche pour l'activité de l'utilisateur et son souhait de pouvoir la retrouver rapidement. A ce titre, il est naturel d'observer cette fonctionnalité et d'en garder une trace. Cette trace est associée au contexte d'usage lors de la sauvegarde de la recherche et aux éventuels contextes de réutilisation de cette sauvegarde.

A.3.3 Exploitation des documents

Documents mis en favoris (*bookmark*) L'application dispose de fonctionnalités de mise en favoris permettant à l'utilisateur de retrouver très rapidement un document, sans faire de nouveau une recherche. Le système génère un événement lorsqu'un document est mis en favoris et lui associe une trace. Naturellement cette trace est contextualisée de sorte qu'il soit possible de connaître le contexte dans lequel l'événement a été généré.

Documents consultés en version originale (*original*) Le module de consultation des contenus de notre application fournit différents modes de lecture. Le premier d'entre eux est évidemment dédié à la consultation des versions originales des documents. Lorsque l'utilisateur sélectionne cette option de visualisation, le serveur génère un événement qui se traduit en une trace dans laquelle sont enregistrés les détails de la consultation. Cette trace étant datée, il est par ailleurs possible de calculer la durée de la consultation.

Documents consultés en plein texte (*fulltext*) L'application dispose par ailleurs d'un mode de consultation des documents en *fulltext*. Cette option n'est naturellement disponible que si le contenu *fulltext* est disponible, c'est-à-dire si le contenu est de nature textuelle et si une extraction du texte a pu être menée. De la même manière que pour le mode original, un événement est levé et la trace correspondante est enregistrée.

Documents téléchargés (*download*) Enfin, outre les modes de consultation directement intégrés à l'application, les utilisateurs peuvent télécharger le document, que ce soit sa version originale ou sa version *full-text*. Cette action indique souvent la volonté de pouvoir conserver et réutiliser un document et donc indique sa pertinence pour la recherche de l'utilisateur. Nous avons donc choisi d'observer cette action et d'en garder une trace contextualisée.

A.4 Représentations vectorielles d'éléments textuels

A.4.1 Sacs de mots

La représentation par « sac de mots » est une représentation classique de l'information textuelle, utile à de nombreuses application de traitement automatique de la langue ou de recherche d'information. Dans cette représentation, la seule information conservée est la multiplicité des mots, le nombre de fois où ces derniers apparaissent dans un texte. Ainsi pour deux documents d_1 et d_2 composés de la manière suivante :

— d_1 : une belle maison
— d_2 : la demeure cossue

le corpus est défini par un vocabulaire V :

$$\overrightarrow{V} = \{une, belle, maison, la, demeure, cossue\} \tag{A.1}$$

de cette manière d_1 et d_2 peuvent être représentés de la manière suivante, ou chaque représentation du document est égale à la taille du vocabulaire V :

$$\begin{aligned} d_1 &= [1, 1, 1, 0, 0, 0] \\ d_2 &= [0, 0, 0, 1, 1, 1] \end{aligned} \tag{A.2}$$

Enfin, chaque mot peut à son tour être représenté par un vecteur représentatif, tel que :

$$\begin{aligned} \overrightarrow{maison} &= [0, 0, 1, 0, 0, 0] \\ \overrightarrow{demeure} &= [0, 0, 0, 0, 1, 0] \end{aligned} \tag{A.3}$$

Le vecteur représentatif d'un document est ainsi égal à la somme des vecteurs représentatif de chacun des mots qui composent ce document.

A.5 Détail du fonctionnement de la méthode de détection des erreurs d'OCR

A.5.1 Construction du graphe

Les sommets V du graphe $\mathbf{G}_n^t$ sont connectés s'ils représentent des entités nommées similaires du point de vue contextuel (sim_{ctx}) et si leur similarité d'édition sim_{edit} dépasse un seuil t paramétrable. Pour chaque entité e_i dans l'ensemble des entités $\mathbf{E}$, on sélectionne le sous ensemble Top_n les n entités les plus proches de e_i en privilégiant, par un opérateur lexicographique, la similarité contextuelle. Par la suite, un lien e_i e_j est créé si $sim_{edit}(e_i, e_j) \geq t$.

L'opérateur lexicographique utilisé est défini de la manière suivante :

$$\begin{gathered}([0,1]\times[0,1])^2 \to [0,1]\times[0,1] \\ ((x,y),(x',y')) \mapsto lex_{max}((x,y),(x',y')) = \\ \begin{cases} (x,y) & if x > x' \vee (x = x' \wedge y > y') \\ (x',y') & otherwise. \end{cases}\end{gathered} \tag{A.4}$$

Cet opérateur fonctionne de la même manière que la recherche d'un mot dans un dictionnaire, en comparant la première lettre du mot recherché avec la première lettre des mots d'une page, nous déduisons en fonction de la position des lettres dans l'alphabet si le mot recherché et avant ou après un autre mot. Dans le cas d'une égalité, la comparaison s'effectue sur la seconde lettre. A titre d'exemple, si l'on compare les mots « le » et « la » pour recherche quel terme est situé en premier dans le dictionnaire, la première lettre étant identique, on passe à la seconde. Dans ce cas on constate que le caractère « a » est situé avant le « e » dans l'alphabet. Il est possible de cette manière d'établir que le mot « la » est situé avant le mot « le » dans le dictionnaire.

A.5.2 Détail du calcul du coefficient d'asymétrie de Lorenz

Le coefficient d'asymétrie Lorenz est défini dans l'équation suivante. Il est le résultat de la fonction de répartition des documents une fois ordonnés en fonction de leur accessibilité F et la fonction de répartition de cette accessibilité L.

$$S = F(\mu) + L(\mu) \tag{A.5}$$

Les équations suivantes présentent les détails du calcul de F et L :

$$\delta = \frac{\mu - r(d_m)}{r(d_{m+1}) - r(d_m)} \tag{A.6}$$

$$F(\mu) = \frac{m + \delta}{n} \tag{A.7}$$

$$L(\mu) = \frac{\Sigma_{i=1}^{m} r(d_i) + \delta_r(d_{m+1})}{\Sigma_{j=1}^{m} r(d_j)} \tag{A.8}$$