Lifranum : Coconstruire un corpus et une archive de littérature nativement numérique

Christian Cote

Prairial
Pôle éditorial
Lyon Saint-Étienne

Nouveaux Cahiers de Marge 8 - 2024
Fiction & données
ISSN : 2607-4427
Éditeur : université Jean Moulin Lyon 3

Lifranum : Coconstruire un corpus et une archive de littérature nativement numérique

Christian Cote

DOI :

Lifranum : Coconstruire un corpus et une archive de littérature nativement numérique

Christian Cote
Maître de conférences en sciences de l'information et de la communication, HDR, université Jean-Moulin Lyon 3, laboratoire Marge

Résumé : Nous présentons la première phase du projet ANR Lifranum, qui consiste à élaborer un corpus et une archive du web littéraire francophone. Nous présentons une méthodologie qui permet d'identifier sur le web les URL pertinentes, d'enregistrer la partie du réseau social dans lequel chacune de ces URL se trouve insérée, et enfin de récupérer ces URL par un *crawl*. Cette analyse est d'abord manuelle et repose sur les réseaux sociaux d'écrivains. La particularité de notre travail réside dans le fait que nous couplons cette méthode par une procédure de validation inverse, qui part d'une analyse automatique pour ensuite opérer un tri manuel et enfin réaliser un *crawl* complémentaire, à visée patrimoniale et réalisé par la BNF. Enfin, nous nous interrogerons sur la façon dont on peut caractériser et utiliser ces objets numériques, corpus et archives, dans le cadre d'investigations diverses. Nous pourrons alors proposer les contours de ce corpus comprenant plus de 1000 écrivains publiant sur le web, et nous envisagerons les implications qu'il peut avoir sur l'approche du fait littéraire, quelles que soient les entrées scientifiques choisies.
Mots clés : littérature web, archives du web, corpus, réseaux

Abstract: We present the first phase of the ANR Lifranum project, which involves building a corpus and archive of the French-language literary web. We present a methodology for identifying relevant URL on the web, recording the part of the social network in which each of these URL is inserted, and finally retrieving these URLs through a crawl. This analysis is initially manual, and is based on writers' social networks. The particularity of our work consists in the fact that we couple this method with a reverse validation procedure, which starts with an automatic analysis, then performs a manual sorting and finally carries

out a complementary crawl, with a patrimonial aim and carried out by the BNF. Finally, we'll look at how we can characterize and use these digital artifacts, corpora and archives, in a variety of investigations. We will then be able to propose the contours of this corpus, comprising over 1,000 writers publishing on the web, and we will consider the implications it may have on the approach to the literary fact, whatever the scientific entries chosen.
Keyword: web literature, web archive, corpus, networks

Introduction

La création web littéraire est une production massive de textes qui est le fait d'auteurs amateurs, mais liés entre eux par des réseaux d'interconnaissance. La reconnaissance mutuelle est un phénomène marquant de cette pratique littéraire qui se traduit par la participation à des ateliers, associations, etc., mais aussi par l'échange de liens hypertexte. Il s'agit d'une forme de sociabilité littéraire.

L'expression web, notamment du fait du média, renouvèle les pratiques d'écriture et les formes mêmes de la production littéraire. Notamment, le rapport entre le texte et l'image est modifié, parce que l'image produit plus de commentaires et d'interactivité du fait de la pluralité des interprétations qu'elle propose.

Par ailleurs, la longueur des textes diminue avec le web et s'accompagne d'une plus forte fréquence de publication. Si les circuits de la publication littéraire écrite ne sont plus pertinents, c'est également la nature des projets littéraires (associés à un site ou un blog) et le processus de création ponctuelle d'un texte qui sont transformés.

Ainsi, la littérature web réintroduit une pratique amateure de la littérature, qui est aussi une pratique massive (et non plus le fait de quelques personnes) et cela renouvèle l'approche des phénomènes littéraires.

Toutes ces intuitions sont à l'origine de notre projet, qui vise d'abord à fournir un cadre d'étude ayant les dimensions quantitatives de cette production littéraire.

Dans cet article, nous présenterons la démarche que nous avons adoptée de façon à proposer un corpus de cette littérature qui réponde aux critères de la représentativité et qui enregistre le plus grand nombre d'auteurs possible. Pour nous, la constitution du corpus est un élément fondamental de la constitution d'un objet scientifique destiné à être appréhendé par des chercheurs portant des problématiques différentes. En ce sens, c'est la méthodologie de constitution du corpus qui constitue le fondement de la scientificité des approches qui suivront. Nous positionnons ce travail en considérant que le corpus réalisé constitue une entrée

pour l'observation de phénomènes littéraires, linguistiques et sociologiques notamment.

Hormis certains présupposés que nous explicitons systématiquement, telle la dimension collective et interpersonnelle de la production littéraire, nous nous limitons à la caractérisation du corpus, laissant en cela les travaux ultérieurs l'explorer.

Nous présentons cette élaboration de corpus en trois temps : l'identification des productions littéraires à partir des phénomènes de reconnaissance mutuelle, la récupération des données, et l'analyse inverse, reposant sur une analyse systématique des liens, réalisée par un autre acteur du projet.

Cadre de travail

L'ensemble des travaux d'élaboration d'un corpus présentés ici est issu du projet Lifranum ANR 2019-2024[1]. L'idée centrale est de constituer un ensemble de ressources scientifiques à partir de la production littéraire web : un corpus et une archive.

Au départ, c'est un projet mené par des littéraires cherchant à donner une visibilité à la littérature web. Des compétences en sciences de l'information se sont adjointes et dès lors l'idée de constituer un véritable corpus et une archive a émergé[2]. Dans les répertoires existants, les œuvres sont enregistrées sur le principe

1. Le projet dirigé par Gilles Bonnet comprend trois partenaires : UR Marge (Lyon 3), UR ERIC (Lyon 2), BNF. Deux IGR, Lorraine Feugères (Marge) et Kévin Locoh-Donou (BNF) ont particulièrement intensément travaillé sur cette partie du projet. Les équipes de Marge comprennent Belen Hernandez-Marzal, Alice Pantel-Cassagnaud, Fanny Mézard, Lucien Perticoz, de la BNF : Christine Génin, Alexandre Faye, Julien Starck, Clara Wiatrowski, ERIC : Julien Velcin, Enzo Terreau, Javier Espinoza, Jérôme Darmont, Sabine Loudcher. Pour en savoir plus sur le projet, voir le site de l'Agence nationale de la recherche, URL : https://anr.fr/Projet-ANR-19-CE38-0007 [consulté le 8 avril 2024].

2. Nous nous distinguons des projets autour de ELO (*Electronic Literature Directory* [en ligne], URL : https://directory.eliterature.org/ [consulté le 8 avril 2024]), notamment ELMCIP (*Electronic Literature Knowledge Base* [en ligne], URL : https://elmcip.net/ [en ligne]), le NEXT The NEXT (*The NEXT* [en ligne], URL : https://the-next.eliterature.org/ [consulté le 8 avril 2024]), les répertoires du NT2 (*ALN/NT2* [en ligne], URL : https://nt2.uqam.ca/ [consulté le 8 avril 2024]) ou PO.EX (Arquivo Digital da PO.EX – Poesia Experimental Portuguesa [en ligne], URL : https://po-ex.net/ [consulté le 8 avril 2024]), ou encore le projet brésilien SOBRE (Observatório da literatura digital bresileira [en ligne], URL : https://www.observatorioldigital.ufscar.br/ [consulté le 8 avril 2024]), parce qu'ils suivent un principe déclaratif et non un recueil de données.

de la déclaration, elles possèdent un statut d'œuvre à part entière, notamment parce qu'elles sont référencées comme telles sont isolées de leur contexte de production comme de réception et enfin sont classées en fonction de leurs caractéristiques physiques et intellectuelles. Dans notre projet, c'est nous-mêmes qui identifions les œuvres, qui les répertorions, les indexons et les mettons à disposition dans leur contexte social. Nous ne partons pas d'un modèle de type collection, qui discrétise les données, mais d'un modèle de linguistique de corpus, marqué par la continuité et les relations entre les productions. Nous privilégions donc le contexte sur la caractérisation de l'œuvre comme objet. Au-delà de la question de la mise en relation des œuvres dans le cadre du corpus, mise en relation permise par les liens hypertextuels, on rend possible l'étude du lien entre les productions littéraires et leurs cadres sociaux, interpersonnels, linguistiques notamment.

Des problématiques en informatique (bases de données et traitement automatique) se sont ajoutées, avec un double challenge pour les informaticiens : arriver à structurer un corpus à partir d'une archive et mettre en œuvre un système automatique de détection de traits permettant d'indexer le texte littéraire.

Deux points essentiels en matière d'innovation scientifique seront développés ici : la méthodologie de recueil des œuvres littéraires web et la production de données. On abordera rapidement l'indexation qui constitue le travail en cours.

La méthodologie que nous présentons est un *workflow*[3] fondé sur une méthode d'identification, un recueil et une structuration des ressources identifiées pour alimenter un *crawl*[4] et enfin une évaluation, articulée à la production de l'archive (et effectuée par la BNF). Elle se compose donc de deux phases :

- L'une vise à acquérir des URL par une identification préalable de façon à opérer un premier *crawl* qui constituera le corpus ;

3. « Un *workflow* est un processus industriel ou administratif au cours duquel des tâches, des documents et des informations sont traités successivement, selon des règles prédéfinies, en vue de réaliser un produit ou de fournir un service. » Source : https://www.culture.fr/franceterme/terme/INFO620 [consulté le 8 avril 2024].

4. Un *crawl* est une opération d'archivage (ou de « capture ») sur le web, effectuée par un agent automatisé, appelé « crawler », « robot » ou « araignée ». Les *crawls* identifient les documents sur le web sur la base de votre choix d'URL de départ et d'étendue. Le *crawl* peut également référencer le contenu archivé associé à l'action. Source : https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [consulté le 8 avril 2024].

- L'autre à acquérir des données complémentaires par une identification automatique des liens associés aux URL identifiées lors de la première phase à l'aide de Hyphe[5] pour une analyse pour enfin la constitution de l'archive par un deuxième *crawl*.

Les résultats de cette deuxième analyse complètent ainsi ceux de la première, en intégrant des URL en marge des réseaux d'écrivains préalablement identifiés. Cette seconde analyse nous aura montré que la première avait surtout identifié des écrivains dont les liens de reconnaissance mutuelle sont ancrés dans les pratiques numériques, alors que le second aura mis en évidence des auteurs qui se réfèrent à ces premiers réseaux sans en faire véritablement partie et sans chercher une reconnaissance particulière de leur part.

Structuration générale du projet

L'objectif du projet est l'élaboration d'une plateforme permettant de consulter et analyser ce corpus et l'ensemble des données produites par l'équipe et associées à l'élaboration et au traitement des données. La dualité du corpus et de l'archive est liée à des usages différents entre le corpus (comme « donnée primaire de la science ») et l'archive comme patrimoine accessible sur site à la BNF.

Concernant l'identification des productions littéraires « nativement web », nous avons élaboré une méthodologie, fondée sur la reconnaissance mutuelle des auteurs par le partage de liens. Cette méthodologie s'inspire à la fois de la linguistique de corpus appliquée au web et à la sociologie des réseaux : la démarche d'acquisition de données à partir du web a largement été élaborée en linguistique de corpus (Teubert, 2005), en considérant le web comme un réservoir infini de réalisations linguistiques écrites. Dans ce cadre, toute production est par définition pertinente, et c'est la perspective linguistique qui construit l'objet de recherche.

Le premier intérêt d'un corpus est la capacité d'une grande quantité de données à observer des phénomènes réguliers derrière des variations de surface. Le corpus n'est pas construit pour valider une hypothèse préalablement formulée, mais pour observer des régularités qui seront ensuite utilisées pour construire des hypothèses. L'objectif de ce type d'étude est d'abord la détection d'évolutions et l'identification de règles et de régularités qui n'apparaissent pas sans données massives et outils quantitatifs. Les perspectives proposées par le laboratoire ERIC, reposant

5. *Hyphe* [en ligne], URL : https://hyphe.medialab.sciences-po.fr/ [consulté le 8 avril 2024].

sur les modèles vectoriels BERT (Jawahar, Sagot, Seddah, 2019) d'analyse des données textuelles, sont essentielles pour mener ce projet[6]. Néanmoins, à la différence des autres corpus disponibles d'œuvres littéraires numérisées, notamment Gutenberg[7], notre corpus est composé de textes inédits et fortement différents par rapport aux œuvres classiques, tant par leur forme, leurs règles d'écriture, que par leur statut social. Ainsi, il ouvre des perspectives nouvelles d'analyse pour les outils du traitement vectoriel des textes.

Notre stratégie a permis la réalisation d'un *crawl* massif de façon à disposer d'une somme de données considérable (600 Go), consultables à partir d'un point d'accès unique comme une interface de recherche d'information. Nous aboutissons à une somme de 1000 écrivains environ (avant *crawl*), dont la plupart ne publient pas de livres édités, donc ne sont pas répertoriés dans les catalogues des bibliothèques. Il en découle une faible pertinence des approches fondées sur les auteurs (stylométrie par exemple, ou sur les genres classiques – poésie, nouvelle, roman), mais une très forte potentialité d'identification de nouvelles formes d'expression.

Par conséquent, l'indexation sera fondée sur des marques de similarité soit de types de discours (narratif, descriptif, etc.), soit d'identité affichée de discours (psychologique, métadiscursif, communicationnel, etc.), soit encore de structure formelle.

Fondement social : la reconnaissance mutuelle

Les archives du web ont un fondement historique et sociologique, en premier lieu pour les sciences politiques. À partir du moment où il est apparu que l'archive permettait de recueillir un ensemble massif d'informations sur un thème donné, alors l'archive s'est inscrite dans le cadre de la massification des données en sciences humaines et sociales (Schroeder, Brügger, 2017).

Idéalement, l'accès au web ne requiert pas d'intermédiaires et permet tout autant d'accéder à la production médiatique, institutionnelle que la publication quotidienne. Par ailleurs, les dispositifs d'archivage du web se sont développés très rapidement

6. Pour une présentation simple de ces modèles, voir : https://eduscol.education.fr/sti/sites/eduscol.education.fr.sti/files/ressources/pedagogiques/14960/14960-word-embedding-les-mots-et-le-machine-learning-ensps.pdf [consulté le 8 avril 2024].

7. *Project Gutenberg* [en ligne], URL : https://www.gutenberg.org/ [consulté le 8 avril 2024].

(notamment l'Internet Archive[8]) à partir de là, tout comme les initiatives nationales d'archivage du web.

Or, le problème est celui de l'accès et de la structuration de l'archive. L'archive massive est difficile à exploiter. Seuls des *crawls* restreints, donc des archives limitées, sont utilisables aisément avec toutes les questions théoriques et méthodologiques que cela pose, notamment le problème de la pertinence d'une archive limitée à priori.

Depuis le lancement de l'outil de recherche SolrWayback (Egense, Myrvoll, 2018) et des outils autour du projet AUT[9], les difficultés d'usage des archives du web tendent à s'amenuiser : néanmoins, les questions posées relèvent à la fois de l'exploration et de l'organisation des collections, de l'analyse du matériau et enfin de la disponibilité des données pour un large public (Maemura, 2019).

Parallèlement à une conceptualisation des activités d'archivage, un champ de recherche s'est construit progressivement autour des liens URL, associant la théorie des réseaux sociaux (Zachary, 1977) et de la modélisation mathématique de ceux-ci (Boyd, Ellison, 2007), aidés en cela par les visualisations de Hyphe. Il consiste à construire une conceptualisation autour des possibilités d'exploration et de visualisation des liens URL (Severo, Venturini, 2016).

Notre recherche est fondée sur l'archive du web, à savoir l'identification d'œuvres et d'auteurs publiant sur le web. En ce sens on se distingue d'une littérature pensée autour de l'usage du numérique comme constituant de la démarche créatrice. Ici, la dimension numérique est essentiellement considérée comme un outil de publication.

En ce sens, les œuvres peuvent renvoyer au champ de la création littéraire numérique, mais notre corpus est plus large et repose sur la caractérisation d'un domaine au travers des liens numériques (URL) opérant entre les auteurs de ces œuvres sans par ailleurs statuer sur ces dernières. Ainsi, le lien marqué par un échange d'URL permet d'explorer un univers social qui ne constitue probablement pas un champ (Sapiro, 2021), mais une constellation de réseaux sociaux.

Notre objectif a été d'éviter les deux pièges que constitue d'une part une limitation très forte du *crawl*, et, d'autre part, une masse de données contenant des informations totalement inutiles au propos et donc difficiles à caractériser comme corpus.

---

8. *Internet Archive* [en ligne], URL : https://archive.org/ [consulté le 8 avril 2024].
9. *Archives Unleashed* [en ligne], URL : https://archivesunleashed.org/ [consulté le 8 avril 2024].

Des approches sociologiques ont déjà montré l'intérêt des réseaux de sociabilité pour structurer le domaine de la littérature numérique (Beaudouin, 2012). On reprend cette donnée pour bâtir notre découverte d'œuvres en se basant sur la reconnaissance entre pairs comme marqueur de littérarité. Une production textuelle devient littéraire dans le cadre du web par la reconnaissance des pairs, et cette reconnaissance se marque par des liens de type « partage d'URL ».

Cette reconnaissance se marque par le partage d'URL, mais également par des types de fonctions différents associés aux projets littéraires manifestés par ces URL : des auteurs, des communautés, des espaces collectifs, des revues. On peut identifier aussi beaucoup de « relais » : certaines œuvres numériques sont reprises par d'autres, sur leur support, avec le plus souvent un référencement minimal. Par conséquent, on observe une dispersion des œuvres et des productions d'auteurs dans différents lieux de publication numérique.

Néanmoins, ce domaine est relativement structuré si l'on considère les réseaux sociaux de reconnaissance mutuelle, comme le montre la visualisation issue de l'analyse HYPHE réalisée par Kévin Locoh-Donou (IGR-BNF) :

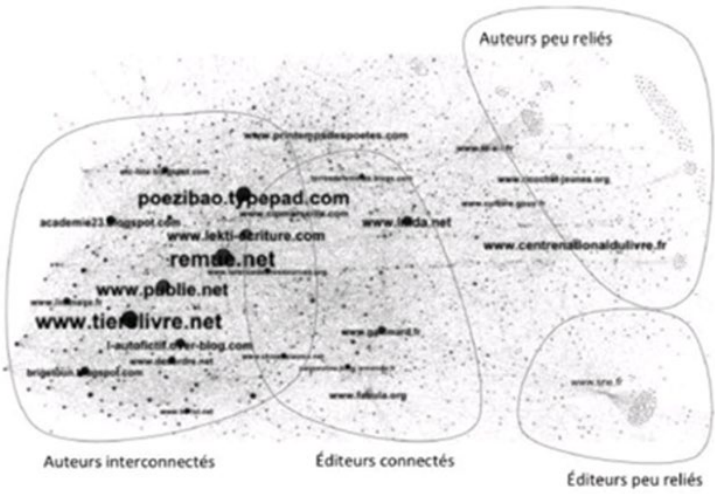

Figure 1. Figure 1. Visualisation des réseaux sociaux d'écrivains du web
© Kévin Locoh-Donou

Sur cette visualisation, on voit l'existence d'une sphère, qui caractérise des communautés de pratiques, à partir de laquelle les autres sphères s'organisent. Elle montre aussi l'importance

des espaces numériques de sociabilité, notamment le « tiers livre » de François Bon et « Remue.net ».

Par ailleurs, l'intitulé des sites et blogs montre que l'unité d'indexation auteur/œuvre, classique des bibliothèques, fonctionne mal ici. En effet, les œuvres dont il est question ici ne sont pas référencées par des règles bibliographiques. Comme nous le verrons, les auteurs comme les titres et les limites des œuvres peuvent être aléatoires ou floues. En ce sens, il est difficile d'extraire des indications d'auteur et d'œuvre de façon sûre, antérieurement à une analyse.

Dans notre cadre, l'auteur est construit à partir de sa reconnaissance : dès lors, il est auteur de par sa capacité à être identifié comme tel par d'autres, ce qui rend sa définition non pas liée à une analyse littéraire, mais à un jugement social. C'est d'ailleurs le propre des corpus entendus au sens de la linguistique de corpus de ne pas prédire le statut des objets avant l'analyse du corpus. Une œuvre littéraire sur le web se définit donc par une visibilité dans un réseau. La contrainte de la reconnaissance par les pairs, qui constitue un microcosme du fonctionnement général du web, amène à considérer l'auteur comme une entité sociale, inscrite dans un réseau. En ce sens, il ne peut y avoir d'auteur isolé : ce sont les liens sociaux qui créent la reconnaissance d'auteur dans le corpus.

Méthode quotidienne de travail

L'objectif de cette présentation n'est pas seulement de présenter un corpus, mais l'ensemble de la méthodologie qui a permis de le réaliser. Si nous insistons sur la dimension intellectuelle de la méthodologie, cette dernière repose sur des documents de recherche, qui eux-mêmes pourront être exploités au-delà de leur usage dans la menée du projet.

Rappelons que le projet aboutit à deux ensembles de données, le corpus et l'archive (cette dernière étant stockée, gérée et mise à disposition par la BNF). La chaîne de production de ressources repose sur une succession et un enrichissement entre les deux. L'archive a été élaborée à la suite du *crawl* du corpus et est enrichie d'une analyse automatique de liens fondée sur Hyphe.

Par ailleurs, comme une analyse automatique par traitement automatique des langues à partir des WARC[10] est impossible pour

10. Les WARCs sont un format *open source* développé par l'Internet Archive et norme ISO (CD 28500) pour les archives web, composé d'enregistrements WARC désagrégés (provenant de différents hôtes). Source https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml [consulté le 8 avril 2024] ; pour consulter un glossaire des termes spécifiques à l'archivage

le moment, un corpus de travail a été élaboré par le laboratoire ERIC pour expérimenter l'indexation. Nous travaillons essentiellement sur la base de fichiers JSON, CSV et Excel, échangés entre les équipes et permettant d'exploiter les données obtenues de façon automatique (Hyphe, BERT pour les analyses textuelles). Ces fichiers sont tous stockés dans un espace de gestion de données Nuxeo.

Nous présentons ci-dessous la totalité des données produites et récoltées dans cette première partie du projet :

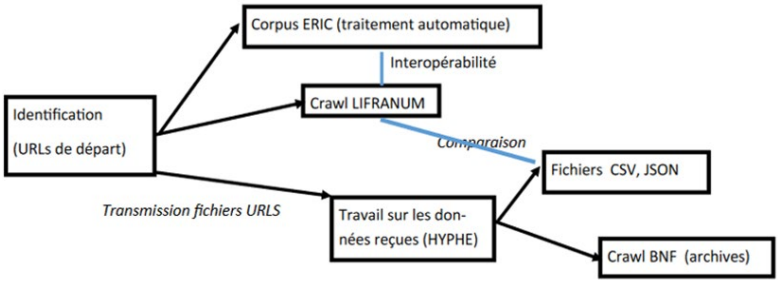

Figure 2. Représentation de la production et de l'usage des produits associé à la méthode de travail

Concernant le corpus, l'objectif est à la fois de fournir un objet d'étude structuré pour les chercheurs du domaine (avec accès réservé) et de constituer un espace d'exploration pour tous ceux qui s'intéressent aux textualités numériques. L'archive a vocation de constituer un patrimoine de cette littérature dans son contexte social.

Notre *workflow* fait alterner systématiquement les outils automatiques et les méthodes manuelles, ce qui permet soit de paramétrer les outils, afin qu'ils répondent à des objectifs circonscrits, soit d'interpréter des données produites automatiquement :

	Méthode manuelle	Méthodes automatiques	Méthode manuelle
Corpus LIFRANUM	Identification des ressources intéressantes et répertoires XML	Crawling HERITRIX sans contrôle dans le déroulement du processus	
Archive BNF		Analyse des liens découverts et crawling supervisé	Suivi de lien en utilisant HYPHE

Figure 3. Représentation de l'articulation des méthodes manuelles et outils automatiques dans le cours du travail

web, voir : https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [consulté le 8 avril 2024].

Procédure d'identification de la littérature web

Le problème fondamental de l'acquisition de ce corpus consiste en la difficulté à identifier la production littéraire web dans toute sa diversité. En effet, la littérature web n'est pas directement identifiable, parce que l'on ne dispose ni de mots-clés ni d'indices spécifiques ou récurrents, à la différence d'archives covid (Messens, Lieber, Chamber, Geeraert, 2022) pour lesquelles certains mots-clés permettent d'identifier les sources. Donc, les réseaux de sociabilité, qui sont à la fois un fait social et technologique (échange de liens URL), constituent le point d'entrée sur le web[11].

Pour l'identification, nous avons mis en place une chaîne permettant de recueillir et structurer les URL et de constituer un répertoire pour le *crawl*. L'identification utilise la recherche avancée Google en structurant la requête de façon à ce qu'à la fois on obtienne les URL pertinentes et sans bruit.

Nous utilisons le cadre formel de la théorie des situations (Barwise, Perry, 1983), qui est un modèle de sémantique formelle, par laquelle les prédicats signifient relativement à trois situations : une situation individuelle dans le monde (correspondants aux déictiques), une situation mentale (ou image mentale de la situation individuelle) et une situation de référence, à partir de laquelle les deux précédentes peuvent être interprétées et donner lieu à des inférences. Ces trois situations fondent la façon dont nous structurons la requête, en considérant que la situation individuelle doit être strictement identifiée, la situation décrite caractériser l'objet qui nous concerne (la production littéraire) et enfin la situation ressource, à partir de laquelle on limite le champ d'interprétation.

On formule ainsi les requêtes obtenues : « Pour l'entité nommée suivante, caractérisée comme l'entité du monde à propos de laquelle on veut obtenir de l'information (un nom propre ou entité nommée), je veux connaître les URL liées à ses propriétés d'être écrivain ou poète et cela en spécifiant les médias au travers desquels ce lien entre une propriété et un objet sont attestés ». Ainsi, on distingue un individu du monde (l'entité nommée), les univers mentaux (la propriété) enfin le contexte dans lequel on souhaite que cette relation opère (les supports web de publication). En plaçant l'individu au centre de la requête, nous limitons les résultats à ceux qui le concernent. En spécifiant certaines propriétés et le contexte, on élimine les homonymes et surtout on identifie tous les contextes dans lesquels il apparaît comme

11. Ce travail a nécessité le recrutement d'une ingénieure (Lorraine Feugères) sur un an.

écrivain et dans le cadre ou en référence à une URL (blogs, sites et réseaux sociaux). Les URL obtenues font explicitement référence à cette « personne en tant qu'écrivain » et au travers du web des blogs, réseaux sociaux et sites. La référence à un contenu web évite les librairies et autres bibliothèques. Ainsi, chaque URL pointant vers cet auteur pourra avoir comme auteur une personne candidate à une nouvelle requête.

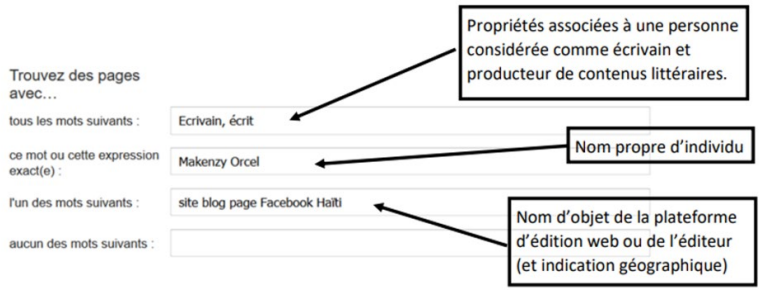

Figure 4. Exemple d'une requête formulée à l'aide des principes de la théorie des situations

Les résultats des requêtes à propos des auteurs ou entités nommées sont reportés dans des schémas XML. Chaque schéma XML reproduit le résultat d'une requête et enregistre la totalité des liens apparus lors de cette requête : il représente les différents liens d'une URL du site/blog d'un écrivain.

Les sites identifiés par la recherche qui pointent vers l'URL de cet auteur seront alors considérés comme des liens. Ensuite, l'analyse des liens sortant de cette URL, marqués par son auteur, identifie une reconnaissance mutuelle (ce qui n'empêche pas d'identifier des liens sortant – ou entrant – sans réciprocité).

Néanmoins, les liens et les formes de reconnaissance sont différents en fonction du projet littéraire. Nous avons donc typé les projets par sortes de liens échangés : chaque type de projet littéraire ou éditorial est spécifié par les différents types de liens. Nous avons caractérisé quatre types de projets : individuel, collectif, éditorial et communautaire :

- Auteurs individuels : production créative signée par un nom de personne. Ce nom doit être identifié comme celui du créateur du média : URL, métadonnées du créateur, propriétaire de la page.
- Auteurs collectifs (différentes personnes identifiées avec une signature unique) : production créative signée par un nom collectif. Même identification que le précédent.

- Communautés : nom collectif associé à un lieu et à des initiatives (cours, conférences, prix, événements) en relation avec différents auteurs individuels et/ou collectifs.
- Supports : nom associé à une initiative contenant l'édition de créations originales d'auteurs individuels ou collectifs. En général, un support peut être assimilé à une revue.

Catégories de projets littéraires
Auteur individuel
Auteur collectif
Communauté d'écrivains
Support de publication (revue)

Figure 5. Table les différentes catégories de projets littéraires

Ensuite, nous structurons les schémas en spécifiant ces différentes formes de liens :

- Lien 0 : liens sortant de la page renvoyant à d'autres réalisations du projet de l'auteur. Il s'agit notamment de liens internes vers le site web de l'auteur ou d'autres pages.
- Lien 1 : liens sortant de la page (lien direct) qui indiquent une relation avec le projet littéraire d'une autre personne.
- Lien 2 : liens entrants mis en évidence par la recherche d'informations (citation directe). L'URL marquée met en évidence un lien URL avec l'objet de la recherche (la page identifiée pointe vers celle qui a été explorée sans que ce lien soit avéré dans l'URL d'origine). C'est l'URL trouvée qui pointe vers la recherche originale. C'est la relation inverse du lien 1.
- Lien 3 : liens mis en évidence par la recherche d'information (citation indirecte). L'URL identifiée se rapporte à l'objet de la recherche, mais sans lien direct. Il s'agit par exemple de la simple citation de l'auteur, de la mention de son existence et de son œuvre. Ce type de lien renvoie à la propriété d'indexation des pages web par les moteurs de recherche. Il s'agit d'un référencement interne des textes qui ne mobilise pas le système des URL.

Caractérisation des liens
Liens d'origine de la mention ou du référencement de l'auteur
Liens sortants à partir du site auteur
Liens entrant par référencement du site auteur sur un autre site
Liens par citation et mention (sans référencement)

Figure 6. Table des types de liens entre les auteurs

Pour le moment, cette modélisation de la reconnaissance reste empirique : l'objectif est qu'elle soit modélisée de façon à être reproductible et permette de représenter des flux d'information. Nous avons utilisé ces méthodes pour identifier des URL sans en inférer des conclusions formelles, qu'il s'agisse d'un outil automatique à concevoir ou d'un modèlve logique qui permette de représenter les échanges informationnels à partir des URL partagées.

Nous présentons ci-dessous un exemple de ces schémas XML qui enregistrent l'ensemble des liens associés à chaque URL dans le cadre de la reconnaissance mutuelle :


```
<schema elementFormDefault="qualified" targetNamespace="http://www.example.org/LIFRANUMidentification">
 <element name="collection" type="string"/>
 <complexType name="network">
  <attribute name="description">
   <simpleType>
    <restriction base="string">
     <enumeration value="personal/communautary"/>
     <enumeration value="personal/personal"/>
     <enumeration value="communautary/communautary"/>
     <enumeration value="communautary/personnel"/>
     <minLength value="0"/>
     <maxLength value="1"/>
     <enumeration value="value"/>
    </restriction>
   </simpleType>
  </attribute>
 </complexType>
 <complexType name="facet">
  <complexContent>
   <extension base="tns:network">
    <sequence>
     <element name="webunit" type="string"/>
    </sequence>
    <attribute name="provenance" type="string"/>
    <attribute name="link0" type="string"/>
    <attribute name="link1" type="string"/>
    <attribute name="link2" type="string"/>
    <attribute name="link3" type="string"/>
    <attribute name="authorproject" type="hexBinary"/>
    <attribute name="communityproject" type="hexBinary"/>
   </extension>
  </complexContent>
 </complexType>
</schema>
```

Figure 7. Exemple de schéma XML utilisé pour enregistrer les résultats des requêtes

Méthodologie du *crawl*

Les listes d'adresses obtenues sont utilisées comme racines pour un *crawl* fortement contraint, de façon à éviter des ramifications hors champ (sachant que le *crawl* sera complété ensuite par la BNF en ce qui concerne le contexte). Notre stratégie a consisté à « crawler » par hôte ou par domaine pour les blogs dépendant d'une API particulière. La limite de deux sauts à partir de la racine a permis d'éviter de recueillir des sites non souhaités, et la profondeur maximale des *crawls* a permis de restituer la totalité des œuvres (dès lors que l'on considère le projet littéraire web comme une œuvre dans la durée).

Nous avons utilisé Heritrix[12] pour le *crawl*[13] et SolrWayback[14] comme outil d'exploration. SolrWayback[15] permet un accès aux contenus grâce aux métadonnées WARC. Ces outils ont été choisis pour une raison liée à la notion de corpus comme unité de format, et seuls des outils hors API peuvent constituer cette unité. Un *crawl* par API aurait obligé à segmenter le recueil en se limitant à une API ou à envisager le *crawl* d'une API l'une après l'autre, éliminant ainsi la dynamique des liens entre sites ou blogs. Par ailleurs, Heritrix et SolrWayback sont les seuls outils utilisés à grande échelle dans le cadre de l'archive du web.

Cette stratégie permet de récupérer l'ensemble d'un site ou d'un blog. Elle est liée, au-delà du principe de reconnaissance mutuelle, à l'idée que la littérature web est fortement inscrite dans un contexte construit par l'auteur (individuel ou collectif) et que ce contexte constitue le projet littéraire lui-même. Ainsi, un site d'auteur peut comprendre d'autres publications que des textes purement littéraires (opinions politiques, questions de société, etc.), mais ils font partie du projet littéraire web. En ce sens, le projet web, chez de nombreux auteurs, insère la production littéraire dans d'autres écrits, voire les indifférencie. Notre stratégie permet donc d'intégrer la diversité dans un projet et

12. « Heritrix : internetarchive/heritrix3 », *Github* [en ligne], 27 juillet 2022 URL : https://github.com/internetarchive/heritrix3 [consulté le 8 avril 2024].
13. Heritrix est le nom du projet de moteur de recherche d'Internet Archive, *open source*, extensible, à l'échelle du web et de qualité archivistique. Source : https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [consulté le 8 avril 2024].
14. « Netarchivesuite/Solrwayback » *Github* [en ligne], 5 juillet 2022, URL : https://github.com/netarchivesuite/solrwayback [consulté le 8 avril 2024].
15. SolrWayback est à la fois une plateforme de découverte pour la recherche de pages web historiques et un outil de lecture pour visualiser les pages web. Il est livré avec divers outils, dont l'outil d'exportation du graphe de liens. Source : https://labs.statsbiblioteket.dk/linkgraph/ [consulté le 8 avril 2024].

par l'absence de restriction sur la profondeur, elle permet de montrer l'évolution du projet littéraire d'un auteur.

Méthode inverse et constitution de l'archive

Le *crawl* à partir de l'identification manuelle constitue la première phase de la constitution du corpus et de l'archive. Elle se devait d'être complétée par une autre, qui permettait à la fois d'élargir le corpus, de compléter et de valider cette première phase. En effet, nous n'avions pas la possibilité de montrer la relative complétude de notre méthode ; seule une méthode inverse, reposant sur des principes totalement différents, permettait d'évaluer la pertinence et les caractéristiques des données recueillies lors de cette première phase.

Nous cherchions ainsi à répondre à deux questions :

- Qu'est-ce que l'on a récolté par rapport au domaine du web littéraire en termes de contenus ?
- Quelles sont les URL qui ne sont pas apparues lors de la première identification, et pourquoi ?

Ainsi, dans un deuxième temps, les URL racines, issues de l'identification, sont réinterprétées par l'équipe de la BNF en produisant des analyses automatiques de liens utilisant Hyphe, de façon à repérer d'autres adresses qui ne seraient pas des liens de reconnaissance. Ainsi, on peut compléter la représentation des reconnaissances par des adresses d'auteurs qui gravitent autour de ces réseaux constitués.

La première difficulté tient d'abord au nombre d'URL dans la liste de départ. Il a fallu travailler sur un échantillon réduit. En prenant en compte uniquement les comptes Wordpress, l'analyse Hyphe a permis de récolter un ensemble impressionnant de liens à partir desquels une sélection manuelle a été opérée de façon à éliminer les sites non pertinents. L'analyse manuelle de cet échantillon a permis de sélectionner entre 10 et 20 % de sites pertinents. Cette dualité d'approche où l'analyse initiale est affinée par une analyse automatique contrôlée *a posteriori* permet de distinguer les liens de reconnaissance entre écrivains formant réseau (identification initiale) et les relations à ces réseaux découvertes par les seuls liens hypertextes.

Cette seconde méthodologie a permis aussi de mettre en évidence le réseau de critiques et de structures accompagnant les auteurs (associations, bibliothèques, éditeurs notamment).

Nous présentons ici les résultats globaux des deux méthodes d'identification, puis la comparaison entre des deux :

Répertoire LIFRANUM (avant crawl)	BC web[7]	Découvert par HYPHE.
937 URLs	152 URLs	646 URLs

URLs Identifiées et crawlées par HERITRIX	URLs seulement identifiées par HERITRIX	Découvertes par BNF et pertinentes	Non pertinentes pour le corpus LIFRANUM
2	46	100	498

Figure 8. Bilan et comparaison des collections

Ainsi, on peut noter qu'aux 937 URL découvertes par la méthode d'identification on peut ajouter une centaine d'adresses découvertes par Hyphe. Par ailleurs, les adresses de BC web reprennent très largement notre identification. Les URL du domaine qui ne contiennent pas de production littéraire (critiques, bibliothèques notamment) sont relativement volumineuses puisqu'elles concernent quasiment 500 URL. Enfin, les contraintes de notre *crawl* ont fait que sur les 100 adresses découvertes par la BNF, 46 étaient déjà identifiées sans être « crawlées », 2 identifiées et « crawlées ». 52 adresses étaient véritablement nouvelles.

Analyse du corpus

Le problème important qui demeure sur le corpus Lifranum pour son analyse est la granularité puisque le *crawl* et donc les WARC renseignent très exactement une URL et non un hôte ou une racine. Pour l'analyse, il faut un travail manuel via SolrWayback pour reconstituer un projet éditorial et son organisation. À terme, un enjeu de la structuration du corpus par l'accrochage des fichiers XML aux racines des URL « crawlées » devrait permettre une structuration du corpus (en plus de l'indexation). L'autre intérêt des schémas XML réside dans le fait qu'ils permettent de mettre en évidence des liens entre les différentes URL, mais également de typer les projets (individuels, collectifs, communautés et supports). En attachant les schémas aux racines (hôte ou domaine) des sites, on obtiendrait aussi leur typage.

Le choix de commencer par une méthode manuelle a permis de limiter un champ qu'aucune méthode automatique ne pouvait circonscrire. Mais du fait de l'outil d'interrogation (Google) et malgré la précision de la formulation des requêtes, de nombreux biais ont demeuré. Par conséquent, une seconde méthode, automatique, a permis de compléter la première recherche et a permis de retrouver des auteurs indirectement associés aux premiers.

Par ailleurs, l'intérêt de la dualité entre le corpus et l'archive a d'abord été de distinguer des projets littéraires et un domaine.

Vers la structuration des données

Pour l'indexation et le traitement automatique des données, il est impossible de travailler directement avec les résultats du *crawl* Heritrix. D'où la constitution d'un corpus complémentaire, élaboré par ERIC à partir des API de Wordpress et de Blogger de façon à disposer de données « propres » pour une analyse de traitement automatique du langage par le modèle BERT.

La stratégie d'indexation repose sur le constat que les catégories usuelles de la classification littéraire (roman, poésie, nouvelle, etc.) ne sont guère pertinentes ici et plus généralement, qu'il n'existe guère de critères distinctifs sûrs pour qualifier une production littéraire, puisque le propre de la recherche en littérature consiste à la caractériser.

C'est donc une indexation sans langage contrôlé préalable[16]. Ainsi, une part de l'objectif consiste justement à construire ce langage documentaire qui permettra de structurer ces données massives. Plutôt que de partir sur une indexation thématique (nous savons qu'un texte littéraire peut avoir un tout autre sens que les unités référentielles qu'il contient), nous construisons plusieurs points de vue, relatifs à l'identité revendiquée du texte et les choix esthétiques, qui tous constituent des caractéristiques textuelles relevant d'un projet d'expression artistique.

Deux stratégies complémentaires ont donc été suivies :

- Une stratégie supervisée, où l'on demande à l'outil de retrouver certaines formes grammaticales de façon à caractériser des types discursifs et des positionnements de discours. La démarche supervisée s'appuie sur des marqueurs discursifs dont le rôle est avéré dans la construction de types de discours (narratif, descriptif, etc.). Ces marqueurs sont les connecteurs argumentatifs et les verbes. Ainsi, on affiche des proximités entre textes liées à des similarités d'usages de connecteurs et de verbes.
- Une stratégie non supervisée, où l'on cherche à faire émerger des tendances à partir d'une vectorisation systématique de l'ensemble des marqueurs lexicaux et grammaticaux. Ici, les choix dans la langue, mesurés par des outils statistiques (entropie de Shannon, constantes de Yule, indices de Simpson, hapax et dislegomena notamment), font émerger des particularités d'usage de la langue. Ces particularités, analysées au niveau du texte, permettent de caractériser des choix expressifs spécifiques à chaque texte.

16. Un premier usage de ce corpus consiste justement à explorer ces caractéristiques du texte littéraire web.

La stratégie d'indexation repose aussi sur des observations de pratiques de chercheurs pour qui la navigation par proximité constitue un aspect essentiel de la consultation de la littérature web. En effet, le lecteur ne va pas rechercher tel ou tel type de texte, ou au contraire des textes sur un sujet précis, mais bien se laisser guider par des similitudes ou des parentés esthétiques entre les textes.

L'indexation permet des recommandations liées à des traits communs (de types de discours, de portée revendiquée et de formes) ou au contraire à des ruptures entre les textes.

L'indexation permet des recommandations liées à des traits communs (de types de discours, de portée revendiquée et de choix esthétiques) ou au contraire à des ruptures entre les textes. Dans le contexte qui est le nôtre (écrivains non référencés, œuvres disparates et courtes), il nous est apparu pertinent de partir de traits énonciatifs et de choix esthétiques pour proposer des types de textes liés par une communauté de choix et d'affichage[17].

Limites et perspectives

Nous proposons maintenant une mise en perspective de notre travail et l'ouverture de pistes permettant de poursuivre l'analyse proposée. Notre projet propose un corpus de recherche correspondant au *crawl* LIFRANUM et qui contient tous les matériaux de recherche associés, notamment les schémas XML d'identification, visualisations Hyphe, données de comparaison. Ces données font partie intégrante du corpus parce qu'il s'agit de résultats scientifiques demandant à être exploités.

L'objectif de ce *workflow* a consisté à établir une méthodologie robuste pour le recueil de données (corpus et archives) de façon à généraliser le travail sur corpus web. Nous avons voulu également montrer que les méthodes manuelles et automatiques sont complémentaires. Elles ont également permis à la BNF de comparer avec ses collections BCweb, élaborées depuis une dizaine d'années par Christiane Génin.

Les méthodes de collecte traditionnelles du web littéraire, mises en œuvre par la BNF, reposent essentiellement sur une identification manuelle à partir du dépôt légal. Cette méthode permet de suivre les sites déclarés comme étant de littérature, mais ne permet pas de les insérer dans le contexte de leur production et réception (sachant que la reconnaissance est une part

17. Le corpus de travail de ERIC (sur lequel se fonde l'indexation) sera relié au corpus général et permettra à terme d'indexer l'ensemble de la production littéraire web.

de la réception). Leur comparaison et leur exploitation mutuelle ne sont pas encore achevées.

Par ailleurs, l'archive est rejouable[18] et il est possible de traiter à part les URL contextuelles. En effet, la totalité des outils et des données de travail est disponible et réutilisable, voire modifiable. Le budget de la BNF étant seulement d'1Ta, le *crawl* est limité, notamment relativement aux images et à la profondeur. Le suivi du *crawl* a permis d'éviter les dérives de Heritrix sans par ailleurs contraindre les sauts, à la différence du premier crawl. Le corpus a été réalisé en une seule fois, ce qui permet d'avoir une vision générale de l'état de la littérature web en France. Dans l'idéal, le *crawl* doit être réitéré de façon à intégrer les évolutions de la production littéraire web.

Néanmoins, un certain nombre de difficultés subsistent. Le problème de l'utilisation des données WARC pour l'analyse automatique n'a pas été résolu, et nous oblige à travailler avec un corpus textuel d'appoint, constitué à partir des API.

Nous insistons encore ici sur la complémentarité entre les deux phases de la méthodologie : la seconde permet d'améliorer les résultats de la première : sur les 1000 noms d'auteurs, l'analyse Hyphe a permis d'en ajouter une centaine. Par ailleurs, elle a permis d'analyser le crawl, puisqu'une cinquantaine d'adresses avait été répertoriée sans être « crawlée » (deux avaient été identifiées et crawlées). Leur lien aux schémas XML et l'analyse des réseaux est encore à développer : les deux analyses de réseau (schémas XML et Hyphe) sont différentes, mais complémentaires.

Notre corpus contient un certain nombre de restrictions que nous souhaitons développer ici. Le web des blogs et des sites repose sur l'échange d'adresses sans autre forme de contraintes, alors même que les plateformes de réseaux sociaux, de publication (Wattpad[19] notamment) et les forums[20] sont fondés sur d'autres modes d'échanges que les mises en commun d'URL et donc qu'un lien ou marque d'appréciation ne peut être considéré comme une reconnaissance. Au problème d'identification s'ajoute celui du *crawl* lui-même, à savoir l'impossibilité pour Heritrix de « crawler » les contenus des plateformes. Dès lors, concernant l'identification de ces productions, nous devons mettre en

18. Rejouer signifie que les versions archivées peuvent être représentées telles qu'elles étaient lors de leur mise en ligne.

19. *Wattpad* [en ligne], URL : https://www.wattpad.com/ [consulté le 8 avril 2024].

20. Comme par exemple : *Forum des jeunes écrivains* [en ligne], URL : https://www.jeunesecrivains.com/ [consulté le 8 avril 2024], *Mauvaises Herbes* [en ligne], URL : https://mauvaisesherbes.forumactif.com/ [consulté le 8 avril 2024], etc.

œuvre une procédure d'identification spécifique sauvegardant le principe de la reconnaissance mutuelle.

Enfin, ce corpus constitue un outil pour caractériser des formes et des expressions littéraires nouvelles, jusque-là étudiées de façon parcellaire et sans l'aide de la systématicité permise par ce corpus massif.

Conclusion

À l'issue de cette construction de corpus, la figure de l'auteur est relativement transformée, au sens où il n'est plus simplement celui qui produit le texte, mais aussi celui qui l'édite et s'assure de son référencement comme de sa mise en réseau. Pour la constitution du corpus, l'auteur se caractérise avant tout comme la personne qui va s'assurer d'une mise en visibilité de ses productions, ce qui passe avant tout par sa socialisation.

Nous avons relaté dans cet article l'élaboration d'un corpus antérieurement aux analyses qui peuvent en être faites, sachant que l'objectif d'un corpus est de rendre accessible cette production à l'analyse littéraire. Son enregistrement repose sur la sociabilité littéraire comme principe d'identification. Nous sommes restés en deçà de l'analyse littéraire – ou autre – de cette production.

L'enjeu social de ce travail est tout d'abord de fournir un objet d'études. Il s'agit du rôle d'un corpus. Il permet, du fait de ses dimensions mêmes, une entrée pour des analyses différentes, relevant de la littérature, mais également de la sociologie et de la linguistique notamment. Ces analyses sont, pour le moment, tributaires des possibilités de SolrWayback, qui permet néanmoins des recherches lexicales tout autant que des approches par facettes.

Si comme nous l'avons déjà énoncé, les analyses de type BERT, nécessitent de dupliquer le corpus pour réaliser des opérations de traitement automatique de la langue, d'autres outils ne nécessitent pas cette duplication, comme ceux proposés par AUT[21] et ARCH[22]. En ce sens, le référencement interne, l'extraction de textes autant que l'exploration des liens hypertextuels peuvent être effectivement mis en œuvre à partir de l'archive elle-même.

Enfin, au-delà de la performance de la réalisation d'un corpus de web littéraire, l'enjeu de ce travail est avant tout de formaliser une première méthodologie pour construire des corpus à la fois massifs et cohérents à partir du web. C'est aussi pour cette

21. *Archives Unleashed Toolkit* [en ligne], URL : https://aut.docs.archivesunleashed.org/ [consulté le 8 avril 2024].

22. *Archives Research Compute Hub* [en ligne], URL : https://webservices.archive.org/pages/arch [consulté le 8 avril 2024].

raison que nous sommes restés en deçà des analyses possibles du corpus : il est un outil, et nous nous sommes contentés d'éclairer certains présupposés (notion d'auteur notamment associée à la construction de l'outil).

Bibliographie

Boyd D. M., Ellison N. B., « Social network sites: Definition, history, and scholarship » *Journal of computer-mediated Communication*, 2007, vol. 13, n° 1, p. 210-230.

Barabási A. L., Albert R., « Emergence of scaling in random networks », *Science*, 1999, vol. 286, n° 5439, p. 509-512.

Barwise J., Perry J., *Situations and Attitudes*, Massachusetts, MIT Press, 1983.

Beaudouin V., « Trajectoires et réseau des écrivains sur le Web : Construction de la notoriété et du marché », *Réseaux* [en ligne], 2012, vol. 5, n° 175, p. 107-144, DOI : 10.3917/res.175.0107.

Egense T., Myrvoll A. K.,, *SolrWayback* [en ligne], 2018,URL : https://netpreserve.org/ga2018/wp-content/uploads/2018/11/IIPC_WAC2018-Thomas-Egense_Anders_Klindt_Myrvoll-SolrWayback.pdf [consulté le 8 avril 2024].

Jawahar G., Sagot B., Seddah D., « What does BERT learn about the structure of language? », *in* : *57th Annual Meeting of the Association for Computational Linguistics* [en ligne], Florence, Italie, Association for Computational Linguistics, juillet 2019, p. 3651–3657, DOI : 10.18653/v1/P19-1356.

Maemura E., « What's cached is prologue: Reviewing recent web archives research towards supporting scholarly use », *Proceedings of the Association for Information Science and Technology*, 2019, vol. 55, n° 1, p. 327-336.

Messens F., Lieber S., Chambers S., Geeraert F., 2022, « Seed list mini pilot COVID-19 collection », *Sodha (Social sciences and digital humanities archive)* [en ligne], V1, 2021, DOI : 10.34934/DVN/SE8NUY.

Sapiro G., « Le champ littéraire. Penser le fait littéraire comme fait social », *Histoire de la recherche contemporaine* [en ligne], 2021, Tome 10, nº 1, p. 45-51, DOI : 10.4000/hrc.5309.

Schroeder R., Brügger N. (dir.), *The Web as History: Using Web Archives to Understand the Past and the Present*, Londres, UCL Press, 2017, p. 296.

Severo M., Venturini T., « Enjeux topologiques et topographiques de la cartographie du web : Le cas du patrimoine culturel immatériel français », *Réseaux* [en ligne], 2016, vol. 1, no 195, p. 85-105, DOI : 10.3917/res.195.0085.

Zachary W. W., « An information flow model for conflict and fission in small groups », *Journal of anthropological research*, 1977, vol. 33, nº 4, p. 452-473.

Sources web :

Archives Unleashed [en ligne], URL : https://archivesunleashed.org/ [consulté le 8 avril 2024].

Hyphe [en ligne], URL : https://hyphe.medialab.sciences-po.fr/ [consulté le 8 avril 2024].

Project Gutenberg [en ligne], URL : https://www.gutenberg.org/ [consulté le 8 avril 2024].

« Heritrix : internetarchive/heritrix3 », *Github* [en ligne], 27 juillet 2022 URL : https://github.com/internetarchive/heritrix3 [consulté le 8 avril 2024].

« Netarchivesuite/Solrwayback » *Github* [en ligne], 5 juillet 2022, URL : https://github.com/netarchivesuite/solrwayback [consulté le 8 avril 2024].