Méthodologie pour l'élaboration d'un corpus et d'une archive du web littéraire francophone.

Christian Cote

Méthodologie pour l'élaboration d'un corpus et d'une archive du web littéraire francophone.

Methodology for building a corpus and archive of the French-language literary web.

COTE Christian, MCF- HDR

(UR-MARGE, Université Jean-Moulin Lyon3)

18 rue Chevreul
69362 Lyon Cedex 07

christian.cote@univ-lyon3.fr

Résumé

Cet article propose une méthodologie pour la constitution d'un corpus et d'une archive du web littéraire francophone. Elle emprunte à la fois aux méthodes de la linguistique de corpus et à l'archivage du web et propose une méthode originale pour acquérir des données précises à partir du web. En effet, le problème fondamental de l'acquisition de ce corpus consiste en la difficulté à identifier la production littéraire web dans toute sa diversité : la littérature web n'est pas directement repérable parce que l'on ne dispose ni de mots-clés ni d'indices spécifiques ou récurrents.

Nous utilisons donc dans ce corpus différentes méthodes et outils, coordonnés et permettant, par la complémentarité des points de vue et toujours sous contrôle manuel, de constituer un corpus sinon exhaustif, du moins représentatif de cette littérature. Nous avons pour cela emprunté des concepts comme celui de réseau de sociabilité et constitué des ensembles de données liées permettant de décrire la structure de ces communautés d'écrivains à partir des différents phénomènes de reconnaissance mutuelle.

Enfin, au-delà de cette méthodologie et de sa validation, nous présentons quelques éléments relativement à la structuration de corpus, et notamment son indexation.

Abstract

This article proposes a methodology for building a corpus and archive of the French-language literary web. It borrows from both corpus linguistics and web archiving, and proposes an original method for acquiring accurate data from the web. Indeed, the fundamental problem in acquiring this corpus lies in the difficulty of identifying web literary production in all its diversity: web literature is not directly identifiable because we have neither keywords nor specific or recurring indices.

In this corpus, we are therefore using a variety of complementary methods and tools, which, through the complementarity of viewpoints and always under manual control, enable us to build up a corpus that is, if not exhaustive, at least representative of this literature. To this end, we have borrowed concepts such as the sociability network, and set up linked data sets to describe the structure of these writers' communities, based on the various phenomena of mutual recognition.

Finally, beyond this methodology and its validation, we present a few elements relating to corpus structuring, and in particular its indexing.

0. Introduction.

Nous présentons une méthodologie pour la constitution d'un corpus et d'une archive du web littéraire francophone.

Cette méthodologie constitue une part du projet ANR-LIFRANUM[1], qui consiste notamment à construire un corpus et une archive de la littérature web francophone. Ce projet aboutit à une ressource composée d'un corpus spécialisé (comme entendu en linguistique de corpus (Teubert, W., 2005)[i] par exemple) et d'une archive web spécialisée (Brügger, 2017). 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 repérable parce que l'on ne dispose ni de mots-clés ni d'indices spécifiques ou récurrents, à la différence d'archives COVID (Messen, F., & alii 2022) pour lesquelles certains mots-clés permettent d'identifier les sources. L'objectif est bien de mettre en place une méthode permettant d'identifier ces types de production textuelles et d'évaluer nos résultats.

La méthodologie que nous présentons est un workflow fondé sur une méthode d'identification, un recueil et une structuration des ressources identifiées pour alimenter un crawling et enfin une évaluation, associée à la production de l'archive (et effectuée par la BNF). Elle se compose de trois phases, (1) une visant à acquérir les données par une identification préalable et un stockage des URLs obtenues de façon à opérer un premier crawl qui constituera le corpus, (2) une autre à acquérir des données complémentaires par une identification automatique des liens associés aux URLs de la première phase à l'aide de HYPHE[2], et enfin (3) une analyse pour constituer l'archive par un deuxième crawl aboutissant à une archive publique, consultable sur les espaces dédiés de la BNF. Les résultats de cette deuxième analyse complètent ainsi ceux de la première, en intégrant des URLs en marge des réseaux d'écrivains préalablement identifiés.

1. Problématique et hypothèses

Les œuvres littéraires web sont difficiles à identifier car définir a priori ce qui est ou n'est pas de la littérature web introduit un système de valeur axiomatique. Marcello Vitali-Rosati (Vitali-Rosati, M., 2015) explique que la littérature numérique - c'est-à-dire "la littérature dans un contexte numérique" - fait référence à de nouveaux modes de production, de circulation et de réception des œuvres[3]. La création littéraire a donc une spécificité web qui dépend du contexte de la diffusion : le corpus concerne à la fois les œuvres et la manière dont ces œuvres sont reconnues comme telles dans le cadre du web. Or, les réseaux d'écrivains sont distendus et largement intégrés à d'autres réseaux, et les œuvres littéraires sont diffusées dans différents

[1] https://anr.fr/Projet-ANR-19-CE38-0007. Projet ANR_2019-2024. Le projet dirigé par Gilles Bonnet comprend trois partenaires : UR MARGE-Lyon3, UR ERIC-Lyon2, 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, BNF : Christine Génin, Alexandre Faye, Julien Starck, Clara Wiatrowski, ERIC : Julien Velcin, Enzo Terreau, Javier Espinoza, Jérôme Darmont, Sabine Loudcher.
L'ensemble des données dont il est question ici sont disponibles sur notre espace de travail.

[2] https://hyphe.medialab.sciences-po.fr/ [consulté le 3 juillet 2024]. HYPHE est un outil de curation de corpus web intégrant un crawleur.

[3] Nous distinguons la littérature numérique qui concerne les œuvres éditées dans des formats numériques de la littérature web qui définit la littérature créée dans le cadre des formats web. Notre corpus concerne ces deux types de littérature.

réseaux, médias et lieux. Enfin, un projet web peut avoir une dimension littéraire parmi d'autres thématiques : culturelle, sportive, etc. Dès lors il nous a fallu mettre en place un cadrage théorique qui peut être résumé de la façon suivante :

- La littérature web est d'abord affaire de réseaux d'écrivains travaillant en communautés.
- La littérature web, notamment du fait du média, renouvèle les pratiques d'écriture et les formes même de la production littéraire.
- La littérature web réintroduit une pratique amateure et massive de la littérature (et non plus le fait de quelques auteurs édités).

L'importance des réseaux, la pratique ordinaire et le renouvellement des formes impliquent des choix méthodologiques fondamentaux pour la menée du travail :

- Concernant l'identification des productions littéraires « nativement web » : utilisation d'une méthodologie fondée sur la reconnaissance mutuelle des auteurs par le partage de liens.
- Choix d'une stratégie de crawl massif de façon à disposer d'une somme de données consultables par un portail unique. Nous aboutissons à une somme de 1000 écrivains (avant crawl), dont la plupart ne sont pas identifiés comme tels par les bibliothèques : d'où ensuite, faible pertinence des approches fondées sur les auteurs (stylométrie par exemple), ou sur les genres classiques (poésie, nouvelle, roman).
- Indexation fondée sur des marques de similarité soit de types de discours (narratif, descriptif, etc.), soit d'identité affichée de discours (psychologique, méta-discursif, communicationnel, etc.), soit encore de structure formelle poétique.

Avant de développer ces aspects, nous aimerions revenir sur les concepts fondamentaux de notre travail et la façon dont on les utilise.

1.1. Corpus et archive

Les distinctions entre corpus et archive ont trait tout d'abord au traitement des données parce que le corpus structure et traite les données pour des usages scientifiques alors que l'archive vise d'abord à sauvegarder des données et à les rendre consultables dans une visée patrimoniale. Par ailleurs, le corpus doit comporter une unité de contenus qui n'est pas requise pour l'archive. A la différence de l'archive, les données du corpus doivent être structurées de façon à faciliter l'usage par les chercheurs, en littérature notamment. Le corpus est une ressource composée d'entités comparables et structurées pour un certain usage, alors que l'archive recueille et stocke le plus grand nombre de pièces associées à une activité dans le monde sans filtrer relativement au contenu de ces entités textuelles. Ainsi, dans notre cas, le corpus est limité à la production littéraire dans le cadre d'un projet web, alors que l'archive intègre le contexte de cette création, avec notamment la dimension critique, éditoriale et de diffusion. Par exemple, une critique d'ouvrage par un écrivain publiant sur son site fait partie du corpus LIFRANUM, alors qu'un site dédié à la critique littéraire fait partie de l'archive BNF.

1.2. Réseaux et reconnaissance

Notre projet repose sur l'hypothèse qu'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 association entre un phénomène social et un dispositif technique nous permet de construire le processus d'identification. Cette reconnaissance entre pairs, d'écrivain à écrivain, permet de résoudre la question de la littérarité sans avoir à décider, en fonction de critères externes, de ce qui est littéraire ou pas. Le travail de (Valérie Beaudoin 2012) constitue une première étude reposant sur cette approche par réseaux de reconnaissance.

Par ailleurs, les réseaux que l'on identifie à l'intérieur du web sont une dimension du contexte. Une des particularités de l'hypertexte consiste à expliciter l'ensemble des emprunts (notamment les images, les références), de façon à augmenter la représentation par rapport à une énonciation unique (Atzenbeck, C., & Nürnberg, 2019). Les liens hypertexte sont par définition non hiérarchiques et non structurés. Notre objectif consiste à les explorer de façon à découvrir les réseaux de reconnaissance de la littérature web et à structurer le corpus. Le cheminement de la découverte des œuvres peut être utilisé pour la structuration du corpus, à savoir que les liens découverts peuvent être reversés sur les relations entre les œuvres du corpus, par le biais des métadonnées. Ainsi, la trajectoire de la découverte des œuvres sera enregistrée dans des schémas XML. Les schémas XML que l'on propose sont structurés en distinguant des types de blogs et de sites par leur typicalité.

Comme on pourra le voir à propos des analyses HYPHE, les liens sont par définition multiples et hétérogènes :

- Liens à des plateformes de ressources, comme les images, les vidéos, qui sont des ressources éloignées qui ne constituent que des emprunts. Ils caractérisent une expression augmentée grâce à l'hypertexte.
- Liens associés aux commentaires, qui sont des marques de lecture et de sentiment, mais ne peuvent être considérées comme de la reconnaissance entre pairs. Sur les réseaux sociaux, cette forme de relation sera encore plus présente.
- Liens de sociabilité avec réciprocité, qui constituent des liens de proximité et donc de reconnaissance mutuelle, et qui ne modifient pas l'expressivité des textes par un apport supplémentaire de contenu.

Cette distinction permet également de sauvegarder la distinction entre ce qui relève de la production littéraire, ce qui relève de la communication et de l'audience, et ce qui est constitutif du lien social de l'auteur.

Notre propos consiste à représenter explicitement un contexte, celui qui concerne des liens interpersonnels de reconnaissance parce que ces liens sont fondamentaux pour l'identification des auteurs via les réseaux sociaux. Un travail sur les autres types de liens est possible à partir du corpus et de son exploration par SOLRWAYBACK (NETARCHIVE SUITE 2022). (SOLRWAYBACK est une application web permettant de parcourir les fichiers ARC/WARC, donc des archives web crawlées).

2. Etat de l'art

Notre travail est fondé sur une différenciation par rapport aux perspectives adoptées dans le domaine de la littérature numérique et des corpus littéraires numériques, mais par contre se fonde sur les approches du web en linguistique de corpus et les stratégies de crawling élaborées par les archivistes.

2.1. Les ressources en littérature numérique.

La première difficulté de la littérature numérique est d'abord son identification et la constitution de collections de ces œuvres. Autour de ELO (ELO 2024), association pour l'étude de la littérature numérique, des fonds et des répertoires fondés sur le principe de la contribution volontaire à des collections ont été créés. Ils enregistrent la littérature électronique œuvre par œuvre, considérant toute création de littérature électronique comme indépendante et autonome. ELMCIP et d'autres projets liés à ELO, le NEXT (Electronic Literature Directory, 2022), les répertoires du NT2 (ALNNT2, 2024) ou PO.EX (PO.EX DIGITAL ARCHIVE , 2023) suivent le même principe déclaratif. ELMCIP, comme le NEXT, proposent une base de données relationnelle qui indexe les œuvres en tant qu'unités et propose des relations, notamment entre les œuvres et les écrits critiques, construites après l'indexation. Ces travaux participent à la reconnaissance de la littérature numérique comme objet d'étude. Mais ils ne concernent pas la littérature web, qui est essentiellement publiée sur le web et utilise les langages du web, notamment HTML et les outils

propres au web, à savoir les liens hypertexte entre URLs. En ce sens, nous proposons le premier corpus de cette littérature web, qui peut difficilement être représentée en isolant les œuvres de leur contexte relationnel.

2.2. Corpus en littérature et usages

Les possibilités de traitements massifs des données textuelles, notamment par des outils métriques et statistiques ont entraîné la création de corpus massifs de données littéraires. Dans ce cadre, on peut mentionner le projet Gutenberg, le Stanford Litlab, parmi d'autres initiatives ; au-delà de l'intérêt de la numérisation pour disposer de données accessibles aux méthodes d'analyse quantitative, l'intérêt propre de la démarche consiste à adopter une perspective non linéaire dans l'étude des textes. La dé-linéarisation des textes est un élément essentiel de l'apport de ces travaux, assemblés autour de l'approche dite du « distant reading ». Par ailleurs, l'analyse des cooccurrences permet d'étudier des constructions lexicales et des formes de distribution difficiles à saisir sans des outils de comptage (Bouzereau 2022). Ainsi, les approches de stylométrie ou calcul des particularités stylistiques des auteurs, les analyses thématiques, analyse de sentiments et de changements de registres lexicaux dans les textes (ou entre les textes) deviennent des domaines d'étude en propre, quelles que soient par ailleurs les références disciplinaires convoquées : études littéraires notamment stylistique, psychologie cognitive ou études culturelles (Green, C., 2017).

Ces travaux de construction de corpus visent avant tout à construire des données de façon à étudier les phénomènes d'écriture en accroissant la quantité de données disponibles. Le traitement par « distant reading » vise à globaliser les traitements, notamment afin de mettre en évidence des phénomènes ne pouvant être observables qu'à partir de données massives et donc pertinentes pour un traitement statistique. Nous ne construisons pas les données pour un traitement mais pour l'exploration des œuvres[4].
Pour revenir sur ces corpus, centrés sur des périodes ou des auteurs, quelquefois comparatifs (entre auteurs, genres ou périodes), ils n'interrogent pas les fondements de la caractérisation du texte littéraire et de ses approches. Si notre corpus doit permettre ce type de travaux, nos données se différencient de celles-ci par le fait que les notions traditionnelles d'auteur, de genre et même de texte ne sont pas pour nous acquises mais au contraire des interrogations. La différence tient dans le fait que nous acquérons des données par essence numériques en visant à caractériser leur littérarité, démarche inverse à celle de corpus numérisés qui traitent de données appartenant en propre au domaine littéraire.

Ainsi, les possibilités et les enjeux d'un traitement automatique sont limités et contraints par les formats mêmes des textes. Le traitement automatique aura comme objectif d'une part de caractériser le propre de cette littérature, ses tendances en quelque sorte, et d'autre part de construire des critères descriptifs qui permettraient de saisir automatiquement des traits des textes permettant leur indexation.
Notre point de vue est celui de la préservation et de la mise à disposition des documents. En effet, les documents que l'on enregistre n'ont pas de sauvegarde pérenne hors de notre corpus (éventuellement un crawl de l'archivage du web). Par ailleurs, il nous intéresse de préserver non seulement des textes, mais également les structures éditoriales du web dans lesquelles ils s'inscrivent, et qui par ailleurs définissent le projet littéraire web. Enfin, nous conservons le cadre d'échanges dans lequel se situe le projet littéraire lui-même, donc les réseaux d'interconnaissance dans lesquels s'insère l'auteur.
Le corpus insère les données textuelles dans l'ensemble des processus de diffusion et de communication associés au web. Une part essentielle de la description (par le biais de métadonnées) consistera à préserver la visibilité de ce contexte.
Comme nous le verrons, les plateformes des réseaux sociaux structurent ces processus mais rendent plus difficile leur appréhension, notamment dans le cadre de la reconnaissance.
Nous ne partons pas uniquement des archives du web comme références mais également de la linguistique de corpus utilisant le web parce que l'un de nos objectifs majeurs est la caractérisation de la littérature web comme phénomène poétique spécifique. Nous reprenons donc une propriété des corpus web issus de la linguistique de corpus : on obtient des données massives et comparables sans présupposer à priori des contenus mais en étant certain d'y trouver des textes comportant certaines caractéristiques prédéfinies. On

[4] Nous reviendrons sur cette question plus loin, parce que pour le traitement automatique qui sous-tend l'indexation, il a fallu constituer un corpus de travail.

offre ainsi aux chercheurs le matériau le plus approprié pour explorer la littérature web sans en avoir au préalable caractérisé les règles (puisque ce sont des critères interactionnistes qui nous ont guidé).
Par ailleurs, l'archive du web constitue un domaine de recherche dans lequel nous nous inscrivons pleinement. Néanmoins, concernant l'élaboration de la méthodologie pour identifier et collecter les données en fonction d'un objectif d'élaboration de corpus, les ressources qu'il nous semble essentiel de mobiliser sont celles de la linguistique de corpus appliquée au web.

2.3. Fondements de l'approche par corpus en linguistique

La démarche d'acquisition de données à partir du web a largement été élaborée en linguistique de corpus, en considérant le web comme un réservoir infini de réalisations linguistiques écrites. Toute production est par définition pertinente, et c'est la perspective linguistique qui construit l'objet de recherche. Les règles sont celles de la complétude et de l'équilibre, et laissent au linguiste le choix de la dimension du corpus. Enfin, une telle démarche ne présuppose pas de dimensions particulières aux unités textuelles, le discours lui-même étant une construction (Teubert, W., 2005)..

En général, un corpus est fondé sur le principe de continuité, c'est-à-dire que les différents contenus sont considérés comme des ressources linguistiques documentées. Le terme 'représentatif' signifie que l'étude d'un corpus (ou d'une combinaison de corpus) peut remplacer celle d'une langue entière ou d'une variété de langue. Cela signifie que toute personne effectuant une étude sur un corpus représentatif (considéré comme un échantillon d'une population plus large, son univers textuel) peut extrapoler à partir du corpus (Leech, G., 2007).

Les principes fondateurs de la constitution de corpus ont été définis par (D. Biber 2006) : 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. Ce principe est fondé sur les capacités techniques de l'informatique et s'insère dans le cadre de la Ground Theory (Glaser, B.G., & Strauss, A.L. (1967), qui postule le fondement de la recherche sur les données empiriques. Mais cela a aussi une autre conséquence : 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.

2.4. Stratégies d'acquisition de données à partir du web.

Si l'on considère les corpus linguistiques pour l'étude d'une langue et non les corpus spécialisés, l'intérêt linguistique pour les corpus web est motivé par la diversité des réalisations linguistiques et non sur l'intégralité d'une production textuelle (comme une revue ou un auteur). La particularité d'un corpus web est sa capacité à caractériser les changements et les variations dans l'utilisation de la langue. A la différence des corpus fondés sur des journaux ou des sites institutionnels, les corpus web intégrant les médias sociaux permettent d'enregistrer la pratique quotidienne d'une langue.

La première tâche pour l'élaboration d'un corpus ou d'une archive à partir du web est la caractérisation des racines, c'est-à-dire des URLs qui engagent le crawl. La première méthode utilisée pour construire des corpus hors ligne est BootCat (Baroni, M., & Bernardini, S. 2004), où une double étape a été formalisée : d'abord une recherche Google à partir de lexiques spécialisés pour obtenir les URLs et un crawl pour enregistrer le contenu, en intégrant de nouvelles URLs par la propriété "frontières". Seules les URLs retournées d'une requête sont considérées comme pertinentes. BootCat est néanmoins la première méthodologie qui utilise les requêtes Google pour acquérir des termes et des racines pour un crawl. La première grande différence avec notre méthodologie est que BootCat travaille sur des termes alors que nous voulons identifier les URLs. La deuxième distinction est que nous considérons la recherche d'informations à l'aide de Google comme une première étape pour acquérir des racines pour un crawl ultérieur et non comme un outil d'acquisition de données. Par ailleurs, Google contient beaucoup de biais et c'est pourquoi nous considérons cette étape

comme préliminaire à l'identification des racines. Or, l'intérêt des outils d'archivage apparaît à ce moment-là car il permet d'enregistrer toutes les URLs acquises à partir d'une liste définie de racines. La première qualité des outils de crawl est en effet leur indépendance vis-à-vis des différents biais des moteurs de recherche (Schäfer, R. & alii, 2013).

D'autres stratégies de crawl, dans d'autres contextes que l'acquisition d'un très grand corpus ont été élaborées. Nous présentons trois types de stratégies emblématiques parmi une multitude de méthodologies intermédiaires, liées à des usages différents :

- une quantité de données considérée comme une dimension correcte pour l'analyse linguistique. Le crawl sera alors arrêté lorsque la quantité de données sera atteinte. Par exemple, pour le corpus WaCky Wide Web (Schäfer, R., 2016), "la définition du moment du crawling est la contrainte unique pour la stratégie de crawling : s'assurer que toutes les pages récupérées représentent l'anglais britannique. Les crawls sont réalisés à l'aide du crawler Heritrix, avec une stratégie de crawling « breadth-first multi-threads » : ils sont arrêtés après 10 jours de fonctionnement continu (Baroni, M 2009) "

- un crawl considéré comme un ensemble de données pour un post-traitement très important (qui permet l'élaboration du corpus) associé à une utilisation précise et à un domaine initialement défini. La stratégie consiste à construire le corpus à partir d'un ensemble de données crawlées et est développée par (Hybermal, I. & alii, 2016) pour proposer un large corpus de données textuelles. Le post-traitement peut être assimilé à une véritable construction de corpus où les données sont transformées pour une analyse automatique (que ne permet pas le crawl lui-même puisqu'il ne recueille que les données HTML).

- un crawl considéré comme un outil d'acquisition de données limitées strictement aux racines. Les corpus thématiques sont construits dans ce sens. Le problème reste la maîtrise de la thématique dans le cadre d'un crawl reposant sur les seuls liens d'URLs et donc le contrôle permanent du crawl.

La construction de corpus à partir du web à des fins linguistiques a été abandonnée au profit des données textuelles massives associées à BERT (Devlin, J., 2028) et BLOOM (BLOOM)[ii], mais la chaîne d'outils que nous venons de présenter fournit néanmoins une rationalité du processus d'acquisition de données pour les corpus spécialisés massifs. Enfin, ces corpus restent pertinents pour l'étude de la production écrite spécifique au web.

3. La méthode d'identification : recherche d'information.

Avant de présenter de façon précise la méthode d'identification, nous présentons rapidement la façon dont nous avons travaillé. La méthodologie d'identification consiste à partir des partages de liens directs et indirects. Cette méthode d'abord manuelle permet de contrôler les URLs que l'on utilise comme racines lors du crawl[5]. Dans la deuxième partie du travail, la méthode inverse, qui débute à partir de ces adresses et de récupérer l'ensemble des liens, a été adoptée par la BNF en utilisant l'outil HYPHE. Alors, le tri des URLs parmi l'ensemble de ceux obtenus par l'outil HYPHE s'est fait manuellement.

Par ailleurs, la liste d'URLs obtenue à l'issue de la démarche d'identification a servi de base aux deux crawls. La seconde démarche, utilisant HYPHE, a essentiellement servi à compléter les listes obtenues.
Nous résumons la chaîne de traitement de la façon suivante :

[5] La partie identification et crawl ont été réalisées entre 2020 et janvier 2021, ce dernier ayant été réalisé en février 2021 par l'équipe MARGE (L. Feugères, C. Cote). La partie vérification a été réalisée par la BNF à partir de novembre 2021 jusqu'à l'été 2022 (K. Locoh-Donou, A. Faye).

	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 1. Organisation du travail et échanges durant le projet.

Dans une première partie du travail, l'identification consiste à explorer le web de façon à éviter la multitude des données non pertinentes. Elle se réalise au travers d'un modèle de recherche d'information web basé sur des principes de sémantique formelle. Nous avons mis en place une chaîne de l'information permettant de recueillir et structurer les URLs et de constituer un répertoire pour le crawl. L'identification utilise la recherche avancée GOOGLE en structurant la requête de façon à ce qu'on obtienne les URLs à la fois pertinentes et sans bruit. Ainsi, nous utilisons le cadre formel de la théorie des situations (Barwise, J., & Perry, J., 1981).

La théorie des situations repose sur l'idée que la prédication, à savoir la capacité d'un énoncé à porter de la signification, repose sur des contraintes. Les prédicats qui composent donc la structure informationnelle sont envisagés comme des limites contraignant l'information, et par là-même permettent de caractériser sa portée. Si l'on considère une requête sur le moteur de recherche comme une suite de contraintes, on peut raisonnablement considérer que la précision et la pertinence du résultat seront liées au paramétrage de la structure informationnelle (ou prédicative). On formule donc la requête comme une sorte de structure de phrase ; elle constitue un type de structure informationnelle et la réponse correspond aux valeurs que l'on peut associer à cette structure. Nous reformulons de la façon suivante :

« 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 connaitre les URLs 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 d'être écrivain ou poète) 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 apparait comme écrivain et dans le cadre ou en référence à une URL (blogs, sites et réseaux sociaux). Les URLs 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 les bibliothèques. Ainsi, chaque URL pointant vers cet auteur pourra avoir comme auteur une personne candidate à une nouvelle requête.

Nous illustrons cette structuration des requêtes ainsi :

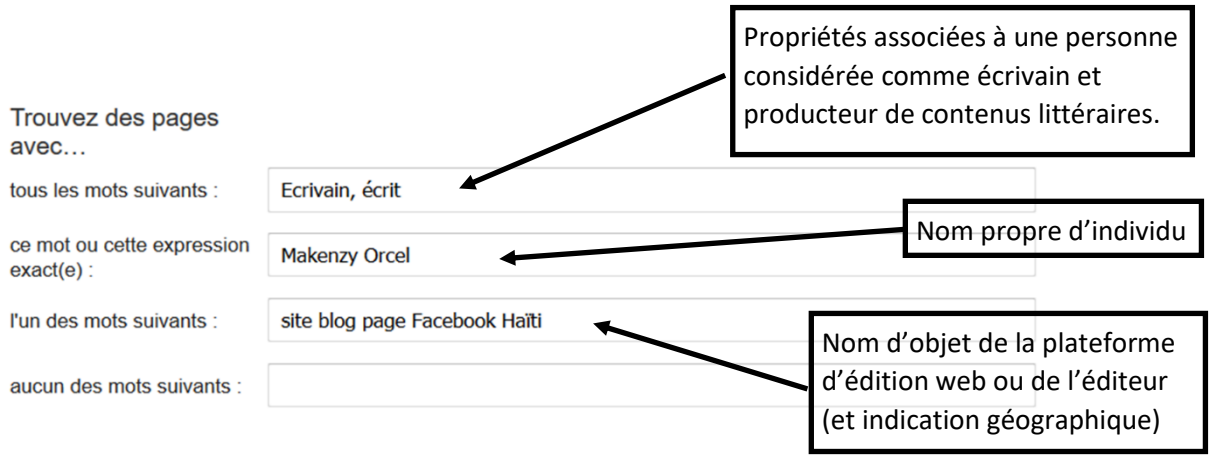


Figure 2. Exemple d'une requête formulée à l'aide des principes de la théorie des situations.

On obtient ainsi le discours sur cet auteur tenu par d'autres, dans le cadre de sa reconnaissance comme poète ou écrivain. Les termes des requêtes sont issus d'une analyse proxémique simple, reprenant l'ensemble des termes utilisables pour désigner l'activité de créateur littéraire pour la propriété, et l'ensemble des noms de supports et plateformes d'édition de contenu et de diffusion pour limiter la recherche aux médias du web.

En tant que corpus représentatif et massif, notre objet ne pouvait se limiter à quelques communautés ou thématiques. En retenant chaque marque de reconnaissance produite par un auteur à propos d'un autre, et les reconnaissances que cet autre produit sur d'autres, on n'a guère de limite pour passer d'un réseau à un autre dès lors que les réseaux sont ouverts (à la différence des forums et de certains groupes Facebook privés).

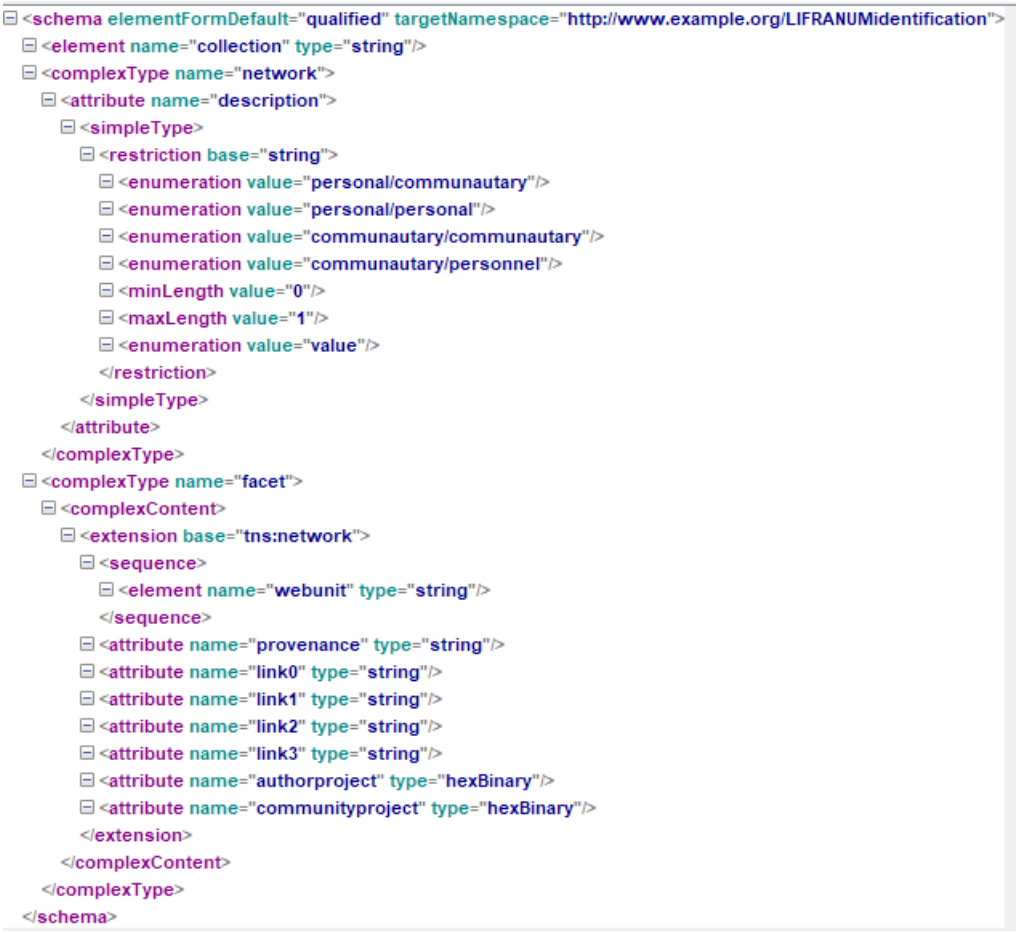
Enfin, nous aimerions présenter quelques éléments afin de justifier notre choix d'un moteur de recherche intégrant l'indexation des contenus :

- Le résultat de la requête identifie les auteurs à partir de leur œuvre et tout discours à propos d'elle ou une partie d'elle. De cette façon, nous trouvons le réseau social de l'auteur considéré non pas seulement par les liens entre URLs mais par les discours puisque l'on interroge des textes indexés par Google et on relève les URLs qui leur sont associées. L'interrogation par la textualité permet de fonder les liens entre URLs sur des contenus indexés, à la différence des outils comme HYPHE qui ne prennent en compte que les URLs.
- Le fondement de l'identification reste la réciprocité de la reconnaissance, qui peut porter sur des objets différents : l'œuvre, le site ou un texte particulier. Le nom de l'auteur doit donc apparaître ainsi que le type de support pour qu'il y ait une reconnaissance effective de l'individu (qui peut aussi être collectif) comme auteur. Cette forme de reconnaissance de l'auteur est distincte d'une appréciation du lecteur qui peut prendre la forme d'un commentaire ou d'un "like", mais cette appréciation du texte ne concerne pas nécessairement l'auteur. Un commentaire ou un "like" n'est pas une reconnaissance mais une émotion par rapport à une œuvre.

3.1. Typage des URLs.

Les résultats obtenus nous amèneront à typer les ressources et à structurer des schémas XML qui reportent les résultats des requêtes à propos des auteurs ou entités nommées. Chaque schéma XML reprend le résultat d'une requête et enregistre la totalité des liens apparus lors de cette requête. Le schéma représente les différents liens d'une URL du site/blog d'un écrivain. Les sites identifiés par la recherche et qui pointent vers l'URL de cet auteur seront alors considérés comme des liens de reconnaissance mutuelle. Enfin, l'analyse des liens sortant de cette URL, marqués par son auteur, indiquent une reconnaissance qui peut être sans réciprocité.

Nous présentons ici le schéma XML pour les sites ou blogs individuels, indiquant la trajectoire de découverte de l'URL :



```
<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 3. Schéma XML caractérisant les sites ou blogs individuels.

Les liens et les formes de reconnaissance sont différents en fonction du projet littéraire et donc il est nécessaire de typer les projets par sortes de liens échangés. Chaque type de projet littéraire ou éditorial est ainsi spécifié par les différents types de liens. Nous avons caractérisé quatre types de projet : 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.

Des relations "sorte de" sont fondamentalement liées à un typage et non à une structure hiérarchique. Cette remarque permet de préciser ce que l'on entend par reconnaissance : entre sites individuels, la reconnaissance est entre auteurs. En ce qui concerne les supports, elle concerne l'intégration dans une sociabilité liée à un projet d'édition collective, et enfin pour les sites communautaires à une finalité qui consiste à faire exister dans le monde social l'activité littéraire. La différence entre les types de structure relève en premier lieu de la configuration éditoriale et la façon dont les liens peuvent apparaitre. Le lien d'appartenance à une communauté est distinct d'un lien marquant la reconnaissance de l'intérêt de cette communauté. Dès lors,

l'identification contient des éléments permettant de caractériser la façon dont une activité éditoriale web organise son identité et ses relations.

On peut ainsi différencier les entêtes des schémas « auteur collectif » et « communauté » de la façon suivante :

Figure 4. Entêtes des schémas « communauté » et « auteur collectif »

Chaque résultat de la requête est reporté dans le schéma XML d'enregistrement : il intègre le fait que les URLs obtenues ne présentent pas nécessairement le même type de liens : certains liens sont des hyperliens, d'autres, qui sont des citations ou des commentaires sur l'auteur et son œuvre, font partie du résultat des requêtes mais seulement par des marques textuelles. Nous structurons donc les schémas en spécifiant ces différentes formes de liens :

- Lien 0 : liens sortants 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 sortants 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.

On obtient ainsi non seulement le réseau de liens, mais également la façon dont il est structuré. On peut ainsi distinguer des liens de reconnaissance associés à un auteur, à des œuvres, à un projet littéraire.

La méthode, adaptée aux blogs et sites, comporte quelques limites : certains créateurs considèrent leur blog ou site comme un répertoire de liens n'ayant pas systématiquement vocation à former un réseau. Nous les avons marqués comme « sites à liens prolifiques » et sélectionné manuellement les liens de reconnaissance en les différenciant des liens hors contexte par rapport au projet.

Enfin, des liens entre schémas peuvent être établis en considérant les URL communes et en utilisant des balises de "provenance" qui marquent le point d'entrée que nous avons retenu pour identifier par quelle URL on a obtenu celle que l'on explore.

Nous pouvons prendre comme exemple les liens caractérisés à partir du blog de Jean-Marc Flick :

```
<complexType name="facet">
 <complexContent>
     <extension base="tns:network">
          <sequence>
              <element name="webunit" type="https://gammalphabets.org/"></element>
          </sequence>
          <attribute name="provenance" type="http://www.fut-il.net/"></attribute>

          <attribute name="link0" type="https://rubatopleinetvidedelies.wordpress.com/"></attribute>
          <attribute name="link0" type="https://bassescontinues.wordpress.com/"></attribute>
          <attribute name="link0" type="https://ilpleuvrademain.com/2012/09/06/vase-communicant-avec-jean-yves-fick-jean_yvesf-vasecommunicant/"></attribute>
          <attribute name="link0" type="https://empreintestrajectoires.wordpress.com/"></attribute>
          <attribute name="link0" type="https://jeanyvesfick.files.wordpress.com/2011/01/un-noyau-de-nuit-comme-avancer-loin-aveuglc3a9-dc3a9ploiement-iii1.pdf"></attribute>
          <attribute name="link0" type="https://www.flickr.com/photos/jean_yvesf/"></attribute>

          <attribute name="link1" type="http://abadon.fr/"></attribute>
          <attribute name="link1" type="http://www.arnaudmaisetti.net/spip/spip.php?page=sommaire"></attribute>
          <attribute name="link1" type="http://yzabel2046.blogspot.com/"></attribute>
          <attribute name="link1" type="http://carnet.marcpautrel.net/"></attribute>
          <attribute name="link1" type="http://deboitements.net/"></attribute>
          <attribute name="link1" type="https://www.face-ecran.fr/"></attribute>
          <attribute name="link1" type="http://www.tierslivre.net/"></attribute>
          <attribute name="link1" type="http://l-autofictif.over-blog.com/"></attribute>
          <attribute name="link1" type="http://www.desordre.net/"></attribute>
          <attribute name="link1" type="http://www.lignesdevie.com/"></attribute>
          <attribute name="link1" type="http://www.liminaire.fr/"></attribute>
          <attribute name="link1" type="http://www.martinesonnet.fr/blogwp/"></attribute>
          <attribute name="link1" type="https://brigetoun.blogspot.com/"></attribute>
          <attribute name="link1" type="http://sabinehuynh.com/index.html"></attribute>
          <attribute name="link1" type="http://sebastienrongier.net/spip.php?page=sommaire"></attribute>
          <attribute name="link1" type="http://christinejeanney.net/"></attribute>

          <attribute name="link2" type="http://i-voix.net/tag/poesie%20-%20j-y%20fick/"></attribute>
          <attribute name="link2" type="http://www.ericdubois.net/article-texte-de-jean-yves-fick-les-vases-communicants-de-mars-2013-115663001.html"></attribute>
          <attribute name="link2" type="https://brigetoun.blogspot.com/2010/06/une-plante-peu-pres-aussi-fraiche-que.html"></attribute>
          <attribute name="link2" type="https://www.tierslivre.net/krnk/spip.php?article761#forum5775"></attribute>
          <attribute name="link2" type="https://rvjeanney.wordpress.com/2010/04/02/vases-communicants-jean-yves-fick-de-gammalphabets/"></attribute>

          <attribute name="link3" type="https://publications-prairial.fr/marge/index.php?id=333&file=1"></attribute>
          <attribute name="link3" type="https://medium.com/@publienet/poesie-et-livre-electronique-une-question-despaces-10d3055c0ccc"></attribute>
          <attribute name="link3" type="https://furieuxdujeudit.com/verlaine-en-mots-dits/"></attribute>
          <attribute name="link3" type="http://furieuxdujeudit.com/blog/wp-content/uploads/2012/06/Dossier-Lettres-et-le-Savoir-2012-2013-version-longue1.pdf"></attribute>
           <attribute name="link3" type="https://www.parlerfrançais.fr/2011/01/?m=1"></attribute>

          <attribute name="authorproject" type="hexBinary"></attribute>
```

Figure 5. Exemple concernant un auteur individuel

L'exemple montre l'origine de la découverte du site « Gammalphabete », à savoir le site de Christophe Sanchez[6]. Il énumère ensuite l'ensemble des adresses relatives directement à cet auteur (link0), ceux auxquels il est lié (link1), ceux qui pointent sur lui (link2) et enfin ceux qui le mentionnent (link3).

3.2. Usage des schémas.

Les réseaux identifiés ont comme rôle de structurer le corpus en restituant les liens de reconnaissance mutuelle. Ils peuvent donc être associés aux fichiers WARCs. L'idée consiste à enrichir les fichiers WARCs par les schémas construits, de façon à ce qu'effectivement ils permettent, en utilisant SOLRWAYBACK, de constituer des outils de guidage à l'intérieur de l'outil de recherche. Les WARCs sont produites automatiquement à partir de chaque enregistrement. Elles contiennent différentes informations à propos de la page, de son type de média, mais également du moment du crawl et de l'origine de la requête. L'intégration des métadonnées de réseau dans le cadre des WARCs se fait à partir de 6.2. "warcinfo" et 6.6 "metadata". Il s'agit par ailleurs des deux espaces dans lesquels les métadonnées DCMI peuvent également être insérées. Cette opération pourra être étendue en intégrant les traits d'indexation. On intègre les fichiers XML dans le fichier WARC correspondant à l'URL racine crawlée.

Actuellement, SOLRWAYBACK fonctionne essentiellement comme un outil de recherche documentaire mais ne permet pas une exploration du corpus, au sens où il ne permet pas de comparaisons, mises en relations et autres formes de structuration de données. La stratégie d'indexation que nous mettrons en œuvre vise à permettre un usage du corpus non comme ensemble des documents, mais comme ensemble de données liées. Les schémas apparaissent ainsi comme des outils de description et de structuration, en

[6] www.fut-il.net [consulté le 3 juillet 2024].

considérant que des projets littéraires qui se reconnaissent mutuellement ont une parenté. Il ne s'agit pas de classer les objets mais de les mettre en relation par une forme de proximité. Dans la suite du travail, nous verrons que la structuration des textes du corpus privilégiera la dimension relationnelle par rapport à une dimension classificatoire.

Ainsi, nous pourrons non plus simplement découvrir des textes, mais caractériser la façon dont ces textes sont insérés dans un contexte relationnel. L'idée est qu'il soit possible soit d'explorer le contexte de reconnaissance d'un texte en utilisant SOLRWAYBACK, soit de façon plus simple, en construisant une interface permettant d'afficher ces liens (sous forme d'un mur par exemple comme pour SUCHO (Jolicoeur, K.,2023), qui constitue une archive numérique collaborative visant à préserver l'héritage culturel ukrainien et en permettre l'accès le plus large).

3.3. Le cas des réseaux sociaux.

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) et les forums sont fondés sur d'autres modes d'échanges que les mises en commun d'URLs : un lien ou marque d'appréciation ne peut être considérée comme une reconnaissance (Larsson, A. O., 2015). Au problème d'identification s'ajoute celui du crawl lui-même, à savoir l'impossibilité pour HERITRIX (HERITRIX 2022) de crawler les contenus des plateformes. Dès lors, concernant l'identification de ces productions, nous devons mettre en œuvre une procédure d'identification spécifique sauvegardant le principe de la reconnaissance mutuelle. Or, les réseaux sociaux (FACEBOOK, X, etc.) accentuent les dimensions déjà identifiées dans les blogs et les sites : les auteurs écrivant pour être lus, les réseaux, de par leurs fonctionnalités, multiplient les traces de lecture mais également les opérations de dissémination. Dès lors, loin d'être un phénomène hétérogène par rapport au corpus identifié, il s'agit d'une continuité de l'insertion de la production littéraire dans l'ensemble des publications web et probablement d'une modification de la place et du positionnement de la production littéraire dans le cadre de la culture commune[7].

4. Stockage et structuration des ressources

Les listes d'adresses obtenues sont utilisées comme racines pour un crawl fortement contraint utilisant HERITRIX, 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[8] dans la durée). Cette stratégie est également liée à l'idée que la littérature web est fortement inscrite dans un contexte construit par l'auteur (individuel ou collectif). Ainsi, un site d'auteur peut comprendre d'autres publications que des textes purement littéraires (opinions politiques, questions de société, etc.), mais qui font partie de son projet littéraire web.

5. Evaluation et constitution d'une archive.

Dans un deuxième temps, ces URLs racines, issues de l'identification, sont réinterprétées par l'équipe de la BNF en produisant des analyses automatiques de liens utilisant HYPHE[9], de façon à repérer d'autres adresses qui ne seraient pas des liens de reconnaissance. Ainsi, on complète la représentation des

[7] Ce corpus permet entre autres d'étudier ce type de phénomène.

reconnaissances par des adresses d'auteurs qui gravitent autour de ces réseaux constitués. Le contrôle est lié aux biais de l'identification - GOOGLE- et aux limites fixées par le crawl.

La première difficulté tient d'abord au nombre d'URLs 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 20% de sites pertinents au regard des impératifs de la BNF.

Cette dualité d'approche où l'analyse initiale est affinée par une analyse automatique contrôlée a postériori permet de distinguer :

- Des liens de reconnaissance entre écrivains formant réseau (identification initiale)
- Des relations à ces réseaux découvertes par les seuls liens hypertextes et qui concernent des auteurs peu insérés dans les processus de reconnaissance collectifs.

Les URLs qui ont été identifiés par l'équipe de la BNF et qui concernent la production littéraire seule, nous ont permis de compléter la représentativité du corpus. Nous présentons dans la table suivante les URLs découvertes par les différents outils :

Répertoire LIFRANUM (avant crawl)	BC web[10]	Découvert par HYPHE.
937 URLs	152 URLs	646 URLs

Tableau 1. Nombre d'URLS obtenus

Après comparaison avec les données obtenues lors du crawl LIFRANUM, on obtient les résultats suivants :

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

Tableau 2. Comparaison entre les différents crawls.

La complémentarité des deux stratégies permet d'intégrer dans le répertoire des adresses qui ne sont pas systématiquement apparues dans le cadre de l'identification du fait de leur rôle marginal dans le processus de reconnaissance. Par conséquent, les auteurs plus isolés apparaissent pour 1/3 dans le crawl, les deux autres tiers étant à mettre au crédit de l'analyse HYPHE. Ces auteurs ne rentrent pas dans le cadre de la reconnaissance mutuelle. A partir de cette liste complétée, la BNF a mis en place un crawl HERITRIX contrôlé afin de constituer une archive élargie, au niveau hôte, domaine et page.

Ainsi, alors que les schémas XML montrent des liens de solidarité et de reconnaissance mutuelle, le corpus révèle d'autres liens, qui n'ont été identifiés que par le crawl. Nous n'avons pas encore réalisé d'étude systématique du corpus. Néanmoins, nous avons comparé des données provenant de la même zone géographique, les Caraïbes, et provenant de sites ayant un volume et une historicité relativement similaires.

[10] BC web désigne la liste des URLs collectées au fil du temps par la BNF, sous la responsabilité de Christine Genin.

Ainsi, "Africultures[11]", considéré comme un espace numérique pour les cultures africaines francophones, est partiellement crawlé en raison de liens vers des sites de publication et non des sites individuels (35 fichiers). En revanche, un site comme "Potomitan[12]", qui couvre la zone Caraïbes et publie des œuvres originales parmi des articles culturels, est beaucoup mieux crawlé (81 fichiers), la plupart des liens étant vers des sites de production littéraire individuelle. Enfin, "Pergolayiti[13]", qui ne contient que des œuvres de création, est très bien couvert par le crawl (499 fichiers). En ce sens, le corpus remplit son objectif d'identification et de recueil d'œuvres et limite les sites ne publiant que peu d'œuvres ou uniquement leur contexte.

Plus précisément, « Pergolayiti » est un site de publication communautaire : lors d'une requête SOLRWAYBACK, il n'est associé à aucun autre domaine. Il a été identifié par les schémas XML, mais pas par des sauts pendant le crawl. En effet, « Pergolayiti » provient bien d'une identification manuelle, à partir de la revue "Francopolis[14]" (revue littéraire numérique autour de la francophonie et publiant des textes, des comptes-rendus et des liens) par un lien sortant. Il s'agit donc d'une reconnaissance par une communauté.

A l'inverse, « Africultures » est lié à six sites différents, certains constituant des relais importants de la littérature francophone (« Larencore[15] » et Catherine Boudet (« catherineboudet.Weebly[16] », et son site de publication « poesiedeslavesbleues[17] »)), d'autres des liens de citation plus fins : « infusionrevue[18] », revue culturelle et artistique sans lien direct à « Africultures » (sinon par référence), « La marche au pages[19] » qui republie des textes courts ou encore « Mondes Francophones[20] » qui publie toutes sortes de ressources relatives à la francophonie, avec un ancrage universitaire.

« Potomitan » est lié à sept domaines : de nouveau « Larencore » et les sites de Catherine Boudet, mais également une revue Openedition « Transcontinentales »[21] , « oeuvresouvertes.net »[22] (le site de Laurent Margantin, qui n'est pas un auteur de la Caraïbe), et enfin le site d'Edouard Glissant[23] et « Pergolayiti » : néanmoins, il s'agit ici essentiellement du mot créole et non du site.

Dans le cadre de l'identification, c'est à partir des activités de publication multiples de Catherine Boudet, autrice réunionnaise, que « Africultures – à partir de la page « Paroles et voix de femmes » - et « Potomitan », où elle a publié quelques poèmes, ont été identifiés. De la même façon, « Larencore », le site de l'écrivaine réunionnaise Patricia Larenco, est apparu très vite dans l'identification (à partir du site du poète Christophe Sanchez[24]), et contient l'ensemble des liens vers « Africultures » et « Potomitan ». Dans ces deux cas, au-delà de la reconnaissance mutuelle, c'est la stratégie de dissémination des œuvres, opérée par ces auteures, qui permet de créer la communauté.

Ainsi, sur un exemple très restreint, on peut illustrer la façon dont l'identification et le crawl fonctionnent en complémentarité : les URLs sont identifiées et leur contexte immédiat est crawlé lors de la seconde opération. Ce sont alors d'autres formes de reconnaissance qui apparaissent lors du crawl : reconnaissance universitaire, visibilité dans un jeu d'emprunts et de références culturelles, mais aussi importance des

---

[11] https://africultures.com/ [consulté le 3 juillet 2024].
[12] https://www.potomitan.info/ [consulté le 3 juillet 2024].
[13] https://www.pergolayiti.com/accueil/ [consulté le 3 juillet 2024].
[14] http://www.francopolis.net/ [consulté le 3 juillet 2024].
[15] http://larencore.blogspot.com/ [consulté le 3 juillet 2024].
[16] https://catherineboudet.weebly.com/ [consulté le 3 juillet 2024].
[17] https://poesiedeslavesbleues.wordpress.com/ [consulté le 3 juillet 2024].
[18] https://infusionrevue.wordpress.com/ [consulté le 3 juillet 2024].
[19] http://la-marche-aux-pages.blogspot.com/ [consulté le 3 juillet 2024].
[20] https://mondesfrancophones.com/ [consulté le 3 juillet 2024].
[21] https://journals.openedition.org/transcontinentales/pdf/397 [consulté le 3 juillet 2024].
[22] https://oeuvresouvertes.net/ [consulté le 3 juillet 2024].
[23] http://www.edouardglissant.fr/ [consulté le 3 juillet 2024].
[24] https://www.fut-il.net/ [consulté le 3 juillet 2024].

interventions ponctuelles d'auteurs (lectures, critiques, recommandations, etc.), qui ne sont pas nécessairement référencées sur leurs propres sites.

6. Promesses et difficultés d'exploitation des résultats du crawling pour la constitution du corpus.

L'usage du crawl MARGE ne pourra être celui d'un corpus que lorsque les données auront pu être indexées de façon à faciliter l'entrée en fonction de problématiques littéraires. Pour l'indexation et le traitement des données, il est impossible de travailler directement avec les résultats du crawl. D'où la constitution d'un corpus complémentaire, élaboré par ERIC[25] à partir des APIs de WORDPRESS et de la BLOGGER de façon à disposer de données « propres » pour une analyse automatique.
L'organisation du travail est actuellement la suivante :

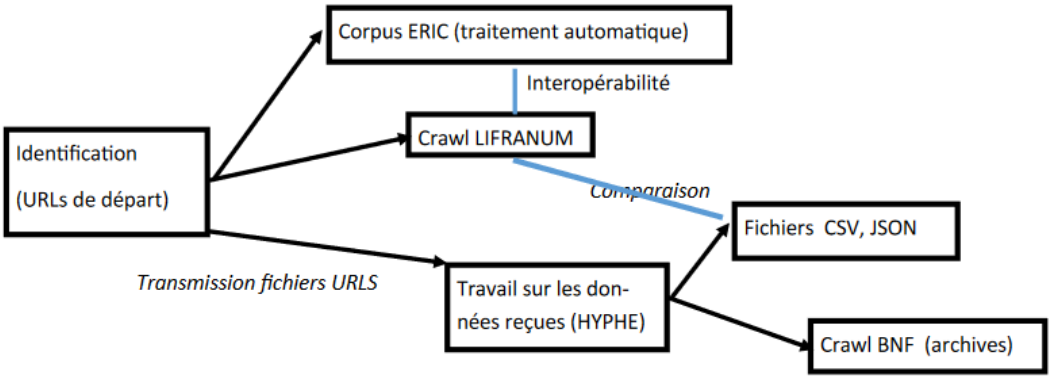


Figure 6. Organisation du travail dans la deuxième partie du projet.

La stratégie d'indexation repose sur deux stratégies complémentaires :

- Une stratégie supervisée, où l'on demande à l'outil de retrouver certaines formes grammaticales (par exemple les connecteurs argumentatifs et les verbes) de façon à caractériser des types discursifs et des positionnements de discours,
- Une stratégie non-supervisée, où on cherche à faire émerger des tendances à partir d'une vectorialisation systématique de l'ensemble des marqueurs lexicaux et grammaticaux.

La stratégie d'indexation repose aussi sur des observations de pratiques de chercheurs pour lesquels la navigation par proximité constitue un aspect essentiel de la consultation de la littérature web. 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.

Conclusion

Le corpus LIFRANUM, dont on vient de présenter la première phase, constitue la première entreprise de constitution d'un corpus massif à partir d'une archive spécialisée du web. A trois reprises dans notre travail, la question des APIs s'est posée : pour l'analyse automatique, relativement à la possibilité d'explorer les données par des outils intégrant la syntaxe des discours, pour la sélection des liens à partir de l'analyse HYPHE et enfin le problème récurrent de l'accès aux réseaux sociaux.

---

[25] Le travail de l'équipe ERIC a consisté à élaborer le sous-corpus avec API et à réaliser les opérations d'analyse vectorielle. Ce travail a commencé au printemps 2021 et s'est achevé à l'été 2023. Il a mobilisé E. Terreau et J. Velcin.

Dans cet article, nous avons insisté sur deux aspects :

- Le contrôle et la meilleure représentativité afin que le corpus soit la représentation la plus fidèle possible du web de production littéraire,
- La proposition d'une méthode pour garantir la construction d'un corpus répondant à des impératifs de cohérence thématique et de similarité de contenu.

Enfin, au-delà du corpus lui-même, notre travail a produit des données qui seront utilisées dans la panoplie des formes de liens entre les œuvres : schémas XML d'identification, visualisations HYPHE, données de comparaison. Ces données font partie intégrante du corpus.

Si au départ la stratégie d'un corpus fondé sur les APIs n'a pas été retenue, c'est parce qu'il devait rester indépendant de la segmentation du web issue de la logique des APIs. Néanmoins, pour une analyse intégrant la dimension syntaxique, nous avons dû y avoir recours. L'indexation, produite à partir des analyses à base syntaxique du corpus avec APIs, peut être intégrée dans le corpus crawlé tout en restant relative aux sites présents dans les deux corpus.

Les choix d'outil de constitution du corpus, au départ techniques, ont ensuite été corrélés aux difficultés à archiver les contenus du fait de la fermeture de l'accès de certaines plateformes à l'archivage. Or, cette impossibilité à pérenniser des oeuvres, associée à la privatisation des contenus par ces plateformes (des contenus qu'elles n'ont pas produits), constitue le cœur des difficultés de la constitution d'un corpus de la littérature web intégrant l'ensemble des pratiques.

Notre travail a permis tout d'abord de valider les propositions d'A. Gefen (Gefen, A. 2022) sur la pratique sociale de la littérature, mais également de les traduire sous forme de métadonnées, inscrites dans le corpus. Ainsi, nous pouvons ainsi rendre compte de la dimension collective de cette création, ce qui constitue un apport par rapport aux jeux habituels de métadonnées, centrées sur des documents isolés.

ALNNT2 : Laboratoire de recherche sur les arts et les littératures numériques. *ALNNT2 : Laboratoire de recherche sur les arts et les littératures numériques* [en ligne]. [consulté le 3 juillet 2024].Disponible à l'adresse : http://nt2.uqam.ca/

Atzenbeck, C., & Nürnberg, P. J. (2019, September). Hypertext as method. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (pp. 29-38). https://doi.org/10.1080/13614568.2021.1942237

Baroni, Marco, et al. "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora." *Language resources and evaluation* 43.3 (2009): 209-226.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In Proceedings of LREC (pp. 1313–1316).

Barwise, J., & Perry, J. (1981). Situations and attitudes. *The Journal of Philosophy*, *78*(11), 668-691.

Beaudoin, V., 2012. Trajectoires et réseau des écrivains sur le Web : Construction de la notoriété et du marché. *Réseaux*. 2012. Vol. 5, no. 175, pp. 107-144. DOI : 10.3917/res.175.0107.

BLOOM : *BigScience Large Open-science Open-access Multilingual Language Model* [en ligne]. [consulté le 3 juillet 2024]. Disponible à l'adresse : https://huggingface.co/bigscience/bloom

Biber, D., Conrad, S., & Reppen, R. (2006). Corpus linguistics - Investigating language structure and use (5th ed.). Cambridge: Cambridge University Press. isbn: 9780521499576

Bouzereau, C., Pajona, C., & Sitbon, C. (2022). L'hétéronymie à l'épreuve de la logométrie: quand Vian rencontre Sullivan. *Corpus*, (23). https://doi.org/10.4000/corpus.6711

Brügger, Niels; Schroeder, Ralph (Hg.): *The Web as History. Using Web Archives to Understand the Past and the Present*.London: UCL Press 2017. DOI: https://doi.org/10.14324/111.9781911307563

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Electronic Literature Directory, 2022. *Electronic Literature Organization* [en ligne]. 2022. [Consulté le 11/01/2023]. Disponible à l'adresse : http://directory.eliterature.org/

Electronic Literature Knowledge Base. *Elmcip*[en ligne]. [consulté le 3 juillet 2024].Disponible à l'adresse : https://elmcip.net/

ELO NEXT. *ELO NEXT* [en ligne]. [consulté le 3 juillet 2024]. Disponible à l'adresse :https://the-next.eliterature.org/
Gefen, A. (2022). 19. La démocratisation de l'écriture. Dans : Olivier Bessard-Banquy éd., *Splendeurs et misères de la littérature: Ou la démocratisation des lettres, de Balzac à Houellebecq* (pp. 421-439). Paris: Armand Colin. https://doi.org/10.3917/arco.bessa.2022.01.0421
Glaser, B.G., & Strauss, A.L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago, IL : Aldine.
Green, C., 2017. Introducing the Corpus of the Canon of Western Literature : A corpus for Culturomics and Stylistics. *Language and Literature : International Journal of Stylistics*. Novembre 2017. Vol.26, no.4, pp. 282-299. https://doi.org/10.1177/0963947017718996
Habernal, Ivan, Omnia Zayed, and Iryna Gurevych. "C4Corpus: Multilingual Web-size corpus with free license." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.
Voir notamment HERITRIX : internetarchive / **heritrix3 ,** 27 Juillet 2022 [consulté le 3 juillet 2024]. Disponible à l'adresse : https://github.com/internetarchive/heritrix3
Jolicoeur, K. (2023). Saving Ukrainian Cultural Heritage Online and the Mission to Preserve Digital Cultural Heritage. *Museum and Society*, *21*(2), 58-64. https://doi.org/10.29311/mas.v21i2.4302
Leech, G (2007), New resources, or just better old ones? The Holy Grail of representativeness. in M Hundt, N Nesselhauf & C Biewer (eds), *Corpus Linguistics and the Web*. Rodopi, Amsterdam, pp. 133-149.
Netarchivesuite, 2022. Solrwayback. Github.com [en ligne]. 5 Juillet 2022. [consulté le 3 juillet 2024].Disponible à l'adresse : https://github.com/netarchivesuite/solrwayback
Larsson, A. O. (2015). Comparing to prepare: Suggesting ways to study social media today—and tomorrow. *Social Media+ Society*, *1*(1), 2056305115578680.
Par exemple : Messens, Fien; Lieber, Sven; Chambers, Sally; Geeraert, Friedel, 2022, "Seed list mini pilot COVID-19collection", Social Sciences and Digital Humanities Archive – SODHA,V1 https://doi.org/10.34934/DVN/SE8NUY
PO.EX DIGITAL ARCHIVE : Portuguese Experimental Poetry. *PO.EX DIGITAL ARCHIVE : Portuguese Experimental Poetry* [en ligne]. Dernière modification le 10 janvier 2023. [consulté le 3 juillet 2024]. Disponible à l'adresse : https://po-ex.net/
Project Gutenberg : *Project Gutenberg is a library of over 60,000 free eBooks* [en ligne]. [consulté le 3 juillet 2024]. Disponible à l'adresse :https://www.gutenberg.org/
Schäfer, R., Barbaresi A., and Bildhauer F. "The good, the bad, and the hazy: Design decisions in web corpus construction." *8th Web as Corpus Workshop*. 2013.
Schäfer, R. (2016). On Bias-free Crawling and Representative Web Corpora. In Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, Berlin, Germany.
Stanford Literary lab [en ligne]. [consulté le 3 juillet 2024].Disponible à l'adresse : https://litlab.stanford.edu/
Teubert, W. (2005). My version of corpus linguistics. *International journal of corpus linguistics*, *10*(1), 1-13. https://doi.org/10.1075/ijcl.10.1.01teu
Les WARCs sont les métadonnées associées à chaque URL et reprenant celles associées à la page et d'autres, spécifiques à l'archivage : https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/ [consulté le 3 juillet 2024].
Vitali-Rosati, M., (2015) "La littérature numérique, existe-t-elle?", *Digital Studies / Le champ numérique* 6(1). doi: https://doi.org/10.16995/dscn.42