De la littérature numérique à la création d'un corpus de littérature web

Christian Cote, Gilles Bonnet, Fanny Mezard, Enzo Terreau, Julien Velcin, Alice Pantel, Belén Hernandez Marzal, Lucien Perticoz

De la littérature numérique à la création d'un corpus de littérature web.

COTE Christian[1], BONNET Gilles[1], MEZARD Fanny[1], TERREAU Enzo[2], VELCIN Julien[2], PANTEL Alice[1], HERNANDEZ-MARZAL Belen[1], PERTICOZ Lucien[1].

1 : MARGE, Université Jean Moulin Lyon3

2 : ERIC, Université Lumière Lyon2

Introduction.

Nous présentons dans cet article l'état de nos travaux relativement à la constitution d'un corpus de de la littérature web francophone[1]. Dans un premier temps, nous définirons cette littérature et nous situerons notre travail par rapport à d'autres réalisations dans ce domaine. Puis nous présenterons rapidement le corpus, sa méthode d'identification et de collecte, avant de nous intéresser à la caractérisation et indexation des contenus en vue de leur usage dans un cadre à la fois de recherche et d'enseignement.

La littérature numérique comme objet d'étude.

La littérature numérique pose de nombreuses questions à la pratique littéraire et à son analyse : questions de l'autorialité, du rôle des dispositifs techniques dans l'activité créatrice, nouvelles sociabilités et reconnaissance, mais aussi interrogation et renouvellement des outils d'analyse et des appareils critiques. La littérature numérique constitue un objet hétérogène, comprenant autant des œuvres closes que des formes continues ou temporaires. Enfin, puisque les contraintes de l'édition classique disparaissent, la panoplie des auteurs et de leur mode d'engagement dans un projet littéraire augmente et se diversifie considérablement. Comment une telle production, massive et diverse pourra être observée, et sous quelles modalités ?

- Une première approche a permis la création de fonds et de répertoires fondés sur le principe de la contribution volontaire à des bases de données qui enregistrent la littérature électronique œuvre par œuvre, considérant toute création de littérature électronique comme indépendante et autonome. ELMCIP[i] et d'autres projets comme ELO[ii], le NEXT[iii], les répertoires du NT2[iv] ou PO.EX[v] suivent le même principe déclaratif.
- Une seconde approche consiste à identifier les publications littéraires sur le web et à les enregistrer dans un ensemble de données unique permettant une exploration similaire à celle des corpus ou des archives.

ELMCIP, comme le NEXT, proposent une base de données relationnelle qui indexe les œuvres en tant qu'unités. Les relations, notamment entre les œuvres et les écrits critiques, sont construites après l'indexation. Notre propos[2] est tout autre : nous considérons que la production littéraire électronique est fondamentalement une relation entre un projet littéraire et une reconnaissance collective entre pairs qui n'est pas nécessairement une forme de critique. Ces distinctions ont un autre fondement : ELMCIP enregistre et décrit des œuvres et des critiques électroniques alors que nous proposons une

[1] Cette recherche bénéficie du soutien de l'ANR https://anr.fr/Projet-ANR-19-CE38-0007 .

[2] https://anr.fr/Projet-ANR-19-CE38-0007

représentation de la création littéraire web ou littérature nativement numérique où la notion d'œuvre est disséminée dans la production sociale de discours[3]. NT2 et PO.EX représentent des œuvres de littérature électronique : les œuvres sont décrites par leurs formats, mécanismes, procédures et types[vi] . Ces travaux reposent sur un thésaurus fondé sur les différentes dimensions des œuvres, et ne proposent pas d'éléments intégrant les spécificités de la littérature web[vii].

Nous proposons donc une approche par données massives, regroupées dans un corpus, et qui permet de saisir la production littéraire web dans son ensemble. A la différence des approches précédentes, fondées sur le concept de « littérature électronique », nous rejoignons ainsi l'approche développée par Christine Genin[viii] à la BNF, qui archive de façon pérenne la production littéraire nativement numérique, celle qui est conçue et éditée d'abord sur le web. Par ailleurs Valérie Beaudoin[ix] pose l'importance des réseaux dans la constitution du web littéraire en insistant sur le rôle des liens dans la constitution d'une notoriété et donc d'une structuration de l'espace littéraire web. L'identification des œuvres se fait par leur intégration dans les réseaux de leurs collègues et partenaires.

De la création d'un répertoire vers la constitution d'un corpus.

L'étude de cette littérature repose donc avant tout sur la constitution d'un ensemble de données constituant un corpus de référence. La création d'un répertoire a été le projet le plus simple à mettre en œuvre. Nous avons conçu le répertoire WEBLITT de la Littérature web francophone : https://weblitt.msh-lse.fr/. Ce répertoire reste tributaire de la volatilité des sites et blogs, ce qui en fait un outil sans mémoire et dont la structuration est difficile, parce qu'il est très difficile d'exploiter des URLs pour l'indexation, même si celles-ci sont dotées de sens : on ne dispose pas pour le moment d'outils permettant d'analyser ces expressions. Dès lors, la conception d'un corpus pérenne est apparue comme une réponse en même temps qu'il permettait une approche diachronique des URLs.

Corpus et archives.

Deux perspectives sont possibles pour constituer un ensemble de données à partir du web : l'archive telle qu'elle est conçue par la BNF et qui se fonde sur la préservation des données, et le corpus web, qui vise à rendre exploitables comme données de la science un ensemble de faits pertinents, exhaustifs ou représentatifs pour un objet scientifique défini.

Un corpus est pensé avant tout par rapport à un usage et en proposant l'exploitation systématique de données massives et des méthodes d'accès fondées sur des traitements automatiques des données (comme peut l'être le tagging par exemple). Cette perspective, élaborée tout d'abord en linguistique[x], s'est traduite en littérature par les corpus et approches du « distant reading[xi] ».

Trois questions se posent dès lors que l'on constitue ce corpus[xii] :

- La méthode de recueil, à savoir le protocole mis en place pour identifier les données,
- Le format de conservation des données,
- L'accès, à savoir les outils mis en œuvre pour accéder aux données et procéder à des recherches.

[3] Opinions politiques, critiques sportives ou « lifestyles » peuvent voisiner avec la production de textes littéraires, sur de mêmes sites, avec de mêmes auteurs.

Elaboration d'une méthodologie de recueil des données.

Le protocole d'identification des données repose sur des noms propres d'individus et leur implication dans des réseaux sociaux de reconnaissance mutuelle. Une recherche d'information utilisant les modalités de recherche avancées des moteurs permet le suivi des réseaux avec des liens entrants et sortants, des citations et des mentions. Dès ce moment-là, on sort d'une logique d'auteur pour entrer dans des logiques de reconnaissance d'acteurs[4]. Notre méthodologie consiste à reconstituer les réseaux de reconnaissance mutuelle et de les stocker dans des fichiers XML.

Formats de données.

Deux techniques existent pour le format de recueil de données : soit on utilise les APIs des sites et blogs dont Twitter, soit on recueille les contenus par un outil de crawling, sans prendre en compte les APIs. La première solution rend le corpus dépendant des APIs et donc limite les possibilités de caractérisation des liens entre les sites et blogs utilisant des APIs différentes[xiii]. Chaque crawl est limité par l'API choisie et il faut systématiquement reprendre chaque API pour constituer un corpus, qu'il faudra ultérieurement lier aux autres. Par conséquent, l'usage des outils de crawling élaborés pour les archives[5], et qui ne prennent en compte que le HTML, nous a semblé nécessaire afin de disposer d'un corpus homogène, c'est-à-dire pouvant être exploré dans son ensemble, indépendamment des plateformes et de leurs APIs. Le crawling permet également de sauvegarder certains liens de type « réseaux » et la découverte de ressources qui auraient échappé à l'identification. Cela dit, le crawling ne permet pas la sauvegarde des structures de données, ce qui rend impossible pour le moment le traitement automatique et oblige au dédoublement du corpus via un autres recueil, partiel, utiisant les APIs.

L'accès aux données et leur indexation.

Le corpus est exploré avec l'outil de recherche SOLRWAYBACK[xiv], qui fonctionne en utilisant les WARCs[6] et dont le principe de segmentation est l'URL : chaque URL identifie une production textuelle (ou visuelle, graphique) distincte. La notion d'œuvre ne peut s'appliquer à une URL : elle marque une publication en temps et auteur au sein d'un espace web. Les racines d'un site ou d'un blog ne peuvent pas non plus être équivalentes à une œuvre même si elles peuvent être assimilées à un projet littéraire. Mais cette segmentation constitue également une opportunité : l'unité textuelle retenue au travers de l'URL est une unité discursive indépendante de toute structuration et donc permet des mises en relations fondées uniquement sur des marqueurs discursifs et textuels et ainsi proposer une indexation fondée sur des relations ou facettes communes entre des productions, et non sur des principes classificatoires[7]. Ainsi notre approche est transversale : le dispositif de recherche et de navigation

---

[5] Voir notamment HERITRIX : https://github.com/internetarchive/heritrix3

[6] 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/

[7] Néanmoins, notre corpus ne prend en compte que le HTML. Les données sans la mise en forme des APIs sont aujourd'hui encore difficiles à exploiter par les outils de l'analyse automatique. Nous avons dû élaborer un corpus limité, basé sur les APIs de certaines plateformes, notamment BLOGGER et WORDPRESS, pour constituer un corpus de travail connecté au corpus principal.

repose sur des relations entre des textes via des URLs[8]. Cet usage des données, privilégiant des explorations par différentes sortes de similarités (comprenant des découvertes inattendues) plutôt que des recherches d'information structurées (par auteur, date, titre, etc.) est corroboré par les résultats d'une enquête de Fanny Mézard concernant les usages de ce corpus, dans le cadre de notre projet.

L'indexation est considérée non à partir d'un thésaurus mais sur la base de traits de discours pouvant être extraits automatiquement, notamment à l'aide des outils LEXCONN[xv] pour les connecteurs et VERBNET[xvi] pour les verbes. Nous avons distingué trois niveaux de constantes du texte littéraire :

- Forme textuelle ou genre (descriptif, narratif, lyrique, dialogal, émotionnel, argumentatif)
- Marques d'activité (mentale, psychologique, communicationnelle et créatrice)
- Poéticité.

Ces traits permettent de mettre en relation des textes indépendamment de leur auteur, plateforme ou date d'édition et de constituer un maillage du corpus permettant de multiples chemins exploratoires.

Conclusion et perspectives. Question des usages : données scientifiques et recueil à vocation pédagogique.

Les choix d'indexation, mais également les possibilités des outils de recherche, mettent à mal la prédominance de l'auteur dans la caractérisation des œuvres, voire la notion d'oeuvre. En revanche, elle permet de mettre en évidence la dimension collective et communautaire associée à la production littéraire web. Nous élaborons dans cet esprit un sous-corpus, à vocation pédagogique (et à données publiques) et qui reprend ces principes organisationnels de façon à familiariser les élèves à ces formes d'écriture et de partage de lectures.

En définitive, ce travail transforme à la fois la façon dont on peut appréhender la littérature (et envisager son enseignement) et la façon dont on peut la décrire au travers de questionnements sur les principes et outils des Sciences de l'Information, à l'aune des possibilités offertes par le traitement automatique de données massives.

Références bibliographiques :

[i] Electronic Literature Knowledge Base. *Elmcip*[en ligne]. [Consulté le 11/01/2023]. Disponible à l'adresse : https://elmcip.net/

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

[iii] ELO NEXT. *ELO NEXT* [en ligne]. [Consulté le 11/01/2023]. Disponible à l'adresse :https://the-next.eliterature.org/

[8] Cette approche n'est pas exclusive, dans la mesure où l'on peut prendre en charge, les racines des sites pour construire un sous-corpus autour d'un nom d'auteur, associé à cette racine. Voire, il est possible de l'étendre par recherche de ce nom d'auteur dans la totalité du corpus.

[iv] 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 11/01/2023]. Disponible à l'adresse : http://nt2.uqam.ca/

[v] PO.EX DIGITAL ARCHIVE : Portuguese Experimental Poetry. *PO.EX DIGITAL ARCHIVE : Portuguese Experimental Poetry* [en ligne]. Dernière modification le 10 janvier 2023. [Consulté le 11/01/2023]. Disponible à l'adresse : https://po-ex.net/

[vi] Taxonomies definition. *CELL : Consortium on Electronic Literature* [en ligne]. [Consulté le 11/01/2023]. Disponible à l'adresse : https://cellproject.net/taxonomies-definition.

[vii] Le vocabulaire descriptif des œuvres littéraires numériques (VODOLIN). *Opentheso* [en ligne]. Dernière modification le 27 octobre 2022. [Consulté le 11/01/2023].
Disponible à l'adresse : https://opentheso.huma-num.fr/opentheso/?idt=th267

[viii] GENIN, Christine, 2016. Le devenir Web de la littérature. *Revue de la Bibliothèque Nationale*. Avril 2016. Vol. 52, no.1, pp. 152-162. *De quoi le peuple est-il le nom ?* ⟨hal-01315464⟩

[ix] BEAUDOUIN, Valérie, 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. Disponible à l'adresse : https://www.cairn.info/revue-reseaux-2012-5-page-107.htm

[x] TEUBERT, Wolfgang, 2005. My version of corpus linguistics. *International journal of corpus linguistics*. 1 janvier 2005. Vol.10, pp. 1-13.

[xi]GREEN, Clarence, 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.

[xii] Ces questions ont été largement développées par :
Maemura, E. (2018). What's Cached is Prologue: Reviewing Recent Web Archives Research Towards Supporting Scholarly Use. In L. Freund (Ed.), Proceedings of the Association for Information Science and Technology. Hoboken, NJ : Wiley, Février 2019, pp. 327– 336. https://doi.org/10.1002/pra2.2018.14505501036

[xiii] FAHEEM, Muhammad, 2014. Intelligent Content Acquisition in Web Archiving [en ligne]. Lieu : Telecom ParisTech. Thèse de Doctorat. [Consulté le 11/01/2023]. Disponible à l'adresse : https://theses.hal.science/tel-01177622/document

[xiv] Netarchivesuite, 2022. Solrwayback. Github.com [en ligne]. 5 Juillet 2022. [Consulté le 11/01/2023]. Disponible à l'adresse : https://github.com/netarchivesuite/solrwayback

[xv] LEXCONN. *LEXCONN* [en ligne]. [Consulté le 11/01/2023]. Disponible à l'adresse : http://www.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml

[xvi] Lima Publisher, 2016. Verbenet. *Github.com* [en ligne]. 28 Juin 2016. [Consulté le 11/01/2023]. Disponible à l'adresse : https://github.com/aymara/verbenet