

# Cartographie du Web Littéraire Francophone

Crédit image : « Méandres », Fabrice CLAPIES, 1997 ; <https://geo-graphique.tumblr.com>

## Table des matières

<b>Introduction.....</b>	3
<b>I. Cartographier le web littéraire francophone .....</b>	3
A. Ancrages méthodologiques .....	3
1. Collecte des données et constitution du corpus .....	3
2. « L'écriture au contact du monde » .....	6
B. Interopérabilité des formats : les outils et technologies mobilisés .....	6
C. Nos missions dans le cadre du projet professionnel du master .....	8
1. Missions et premières pistes de travail .....	8
2. Qu'est-ce que cartographier le web littéraire .....	9
<b>II. L'évolution du projet .....</b>	10
A. Divers questionnements .....	10
1. Autour de l'enrichissement des données .....	10
2. L'extraction des entités nommées .....	10
B. Retour sur les premiers scénarios .....	11
1. Le premier schéma.....	11
2. Le deuxième schéma .....	12
C. Le problème spécifique de la collecte des données <i>Facebook</i> .....	13
<b>III. Présentation du scénario.....</b>	14
A. Des sources de données hétérogènes.....	15
1. Méthode de <i>scrapping</i> .....	15
2. Le cas particulier des notices biographiques .....	16
3. De l'exploration contextuelle à l'extraction de connaissances.....	16
B. Utilisation de l'XML-TEI .....	17
1. Les avantages.....	17
2. Le modèle .....	18
C. Remplissage des fiches et sorties .....	24
<b>IV. Les résultats .....</b>	26
A. Éléments de statistiques descriptives .....	26
B. Exploration textuelle des notices biographiques .....	28
1. Présentation des logiciels utilisés et importation des données .....	28
2. Description du corpus textuel.....	30
3. Premiers éléments.....	34
C. Pistes de visualisations en réseaux .....	38
1. Les entités auteurs et les individus qui rédigent leurs notices biographiques .....	38
2. Les entités auteurs et leurs relations de collaborations et de co-autorat relevées au sein des notices biographiques et des répertoires d'autorité .....	39
3. Les entités auteurs et les noms de domaine qui leurs sont associés .....	40
<b>Conclusion .....</b>	41
<b>Bibliographie .....</b>	41
<b>Table des figures .....</b>	44

## Introduction

Notre travail, encadré par Christian Cote, Jean-Pierre Fewou Ngouloure et Julien Velcin, s'inscrit au sein du projet interdisciplinaire IDEX *Cartoweb*. Porté par les laboratoires MARGE et ÉRIC, il vise à construire une cartographie des productions de littérature numérique francophone. Il sert notamment de « premier étage » à un projet ANR plus vaste intitulé *Lifranum* (Littératures Francophones Numériques), initié par Gilles Bonnet, dont la finalité serait la mise en place d'une plateforme web permettant une « interrogation épistémologique sur la littérarité des contenus repérés et la dynamique des sociabilités identifiées ».<sup>1</sup> Ce projet ambitionne donc l'identification et la stabilisation du corpus des productions littéraires francophones nativement numériques accessibles sur Internet (sites, blogs, réseaux sociaux). Pour commencer, « il s'agit donc d'effectuer un repérage massif des ressources, d'étudier leur répartition, d'établir des modalités standardisées de description afin de constituer un répertoire. »<sup>2</sup> Si notre mission principale est la visualisation des données du corpus de *Cartoweb*, il nous a fallu rapidement nous confronter à la réalité d'une méthodologie de collecte encore embryonnaire pour laquelle il a été nécessaire de proposer des ajustements afin de répondre à notre objectif.

Dans l'étude qui suit, nous nous proposons tout d'abord de recontextualiser la méthodologie et les outils mobilisés par le projet CartoWeb dans le cadre de nos missions. Nous reviendrons également sur les questionnements de parcours qui nous ont accompagnés. Nous traiterons ainsi de la question de l'enrichissement des données, du problème spécifique lié à la collecte des données *Facebook* ainsi que de l'extraction des entités nommées pour lesquels nous avons été confrontés lors de l'établissement des scénarios de réponses. Ceux-ci nous ont finalement permis d'enrichir la méthodologie de collecte existante et d'aboutir à un modèle de regroupement et de structuration des données en XML-TEI. Il sera enfin question des résultats, qui seront abordés par le biais de statistiques descriptives, l'exploration des données textuelles et de visualisations en réseaux.

### I. Cartographier le web littéraire francophone

#### **A. Ancrages méthodologiques**

##### **1. Collecte des données et constitution du corpus**

Le point de départ du projet *Cartoweb* a été d'envisager une manière généralisable et systématique de répertorier les ressources. Cette étape devant faire abstraction d'un jugement qualitatif sur les littéralités rencontrées afin de ne pas porter *d'a priori* sur le traitement futur des contenus. La première approche considérée a été celle de lancer des requêtes sur les moteurs de recherche, principalement Google, à l'aide de mots-clés thématiques. Même en exploitant les possibilités de la recherche avancée, les résultats sont apparus peu satisfaisants.<sup>3</sup>

---

<sup>1</sup> <https://www.univ-lyon3.fr/projet-lifranum>

<sup>2</sup> D'après le document cadre du projet HN fourni le 16 septembre, p.1.

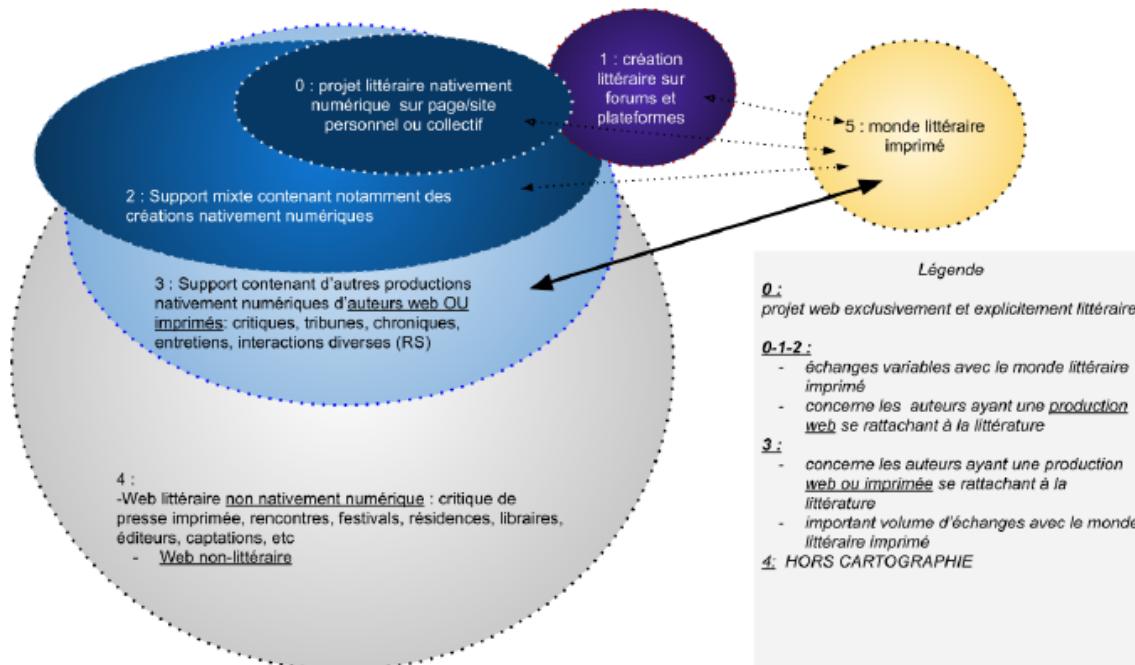
<sup>3</sup> MONAT Jean-Baptiste, *A la recherche du Web littéraire*, mémoire de master 2 Humanités Numériques soutenu à l'Université Lumière Lyon 2 sous la direction de Jérôme DARMONT et de Gilles BONNET, 2019, [93 p.] p. 13.

Cette méthode étant en effet beaucoup trop chronophage pour être systématisée. Ces requêtes ont néanmoins permis de mettre en évidence « l’existence de réseaux locaux serrés » et la récurrence de sites « de référence » tels que *île en île*<sup>4</sup> pour la zone Caraïbe. Ces sites ont néanmoins tendance à répertorier des auteurs déjà « installés » ou institutionnalisés, du moins ayant à leur actif des publications imprimées et valorisées.<sup>5</sup> Qu’est-ce qu’un auteur aujourd’hui ? Dans notre société l’édition papier ne suffit pas à définir le métier d’auteur, car le web est un support d’expression pour les auteurs non édités. Sur quelles bases devons-nous considérer une personne comme étant un auteur et une autre non ?

« Tout comme le Web et ses usages littéraires défont la traditionnelle “chaîne du livre” au profit d’une “chaîne auteur” en laquelle le livre n’est plus qu’une activité parmi d’autres, le site détisse la “chaîne du texte”, organisant une circulation complexe des articles les uns par rapport aux autres, brouillant les relations entre un texte censé recteur et ses marges paratextuelles. »<sup>6</sup>

Le schéma ci-dessous a été élaboré par l’équipe MARGE pour prendre acte du caractère hybride des ressources trouvées (sites personnels n’ayant pas d’objectifs clairs, mais publiant parfois des textes littéraires, blogs communautaires, forums et plateforme d’édition numérique) et du caractère problématique de l’idée de contenus qui seraient « nativement numériques ».

Figure 1 Schéma lié à la construction du corpus élaboré par l’équipe MARGE<sup>7</sup>



Le cercle 0 intègre les projets les plus « centraux », à savoir ceux qui se présentent à la fois comme littéraires et exclusivement créés par des outils numériques. Le cercle 1 considère les contenus à la fois littéraires, francophones et nativement numériques sur des sites collectifs objectivement dévolus à tous ces critères.

<sup>4</sup> <http://ile-en-ile.org/haiti/>

<sup>5</sup> MONAT, *op.cit.*, p. 14.

<sup>6</sup> BONNET Gilles, « L’hypertexte. Poétique de la relecture dans l’œuvre numérique de François Bon », in *Poétique*, n° 175, Paris, 2014, pp. 21-34 [p. 33]

<sup>7</sup> MONAT, *op.cit.*, p. 14.

Le cercle 2 concerne les auteurs ayant une production web se rattachant à la littérature. Le cercle 3 considère également les productions littéraires imprimées (ex. critiques, tribunes, chroniques, entretiens). Le cercle 4 est hors corpus. Il traite en effet du Web littéraire non nativement numérique (ex. critique de presse imprimée, rencontres diverses, festivals) et du Web non-littéraire. Enfin, le cercle 5 indique les échanges avec le monde littéraire imprimé.

Cette représentation permet notamment de mettre en évidence la prépondérance du monde littéraire imprimé dans les échanges ainsi que l’hybridation des contenus, une caractéristique essentielle du « web littéraire ». Milad Doueihi note d’ailleurs que « le numérique consacre l’hybride comme l’aboutissement de notre culture » et qu’une promesse technologique telle celle d’un « web sémantique » ne tient pas, devant la variété d’outils et de formes observées.<sup>8</sup> La cartographie intègre ainsi des éléments du contexte des œuvres imprimés, dès lors qu’ils relèvent de sources nativement numériques et, émanant de l’auteur, contribuent à enrichir ou prolonger l’œuvre imprimée.<sup>9</sup>

Une requête Sparql destinée à extraire les noms d'auteurs francophones vivants a ensuite été élaborée sur le site de data.bnf.fr. Néanmoins, si le genre des documents publiés n'est pas décrit par les métadonnées, il est possible d'isoler des documents textuels et d'extraire les noms des auteurs suivant leur nationalité.<sup>10</sup> L'interrogation des moteurs de recherche est ensuite systématisée sur les noms d'auteurs correspondant aux critères établis.

« Les listes obtenues n'étant que des points d'entrée *par l'imprimé*, il conviendra d'explorer les liens qui peuvent mener d'un site d'auteurs répertorié ici à un auteur non répertorié par la BNF, en misant sur le fait que des réseaux soient suffisamment constitués pour glisser des uns aux autres en produisant un relevé représentatif. Il faudra enfin déterminer précisément *quels contenus* relèvent de notre cartographie et *quels contenus* doivent en être écartés : ici c'est la notion de *littéraire* et de *littérature* qui doivent permettre ce partage. »<sup>11</sup>

Le corpus a ainsi été constitué à partir d'une requête des notices de la BNF, à travers l'exploration de liens, de manière intuitive et suivant la détection de la présence sur le web par l'interrogation de moteurs de recherche. Toute information étant conditionnée par le contexte social, économique et politique spécifique dans laquelle elle a été produite,<sup>12</sup> il est essentiel de pouvoir contextualiser les différents sous-corpus observés au risque de perdre les aspects fondamentaux liés à leur construction scientifique.<sup>13</sup> La méthodologie de collecte ainsi décrite peut donc être envisagée de manière exploratoire à condition de garder à l'esprit que l'apparente masse de données ne suffit *de facto* pas à atteindre une objectivité automatique.

---

<sup>8</sup> DOUEIHI Milad, *Pour un humanisme numérique*, Paris, 2011. Cité par MONAT, p. 15.

<sup>9</sup> *Ibid.*, p. 15.

<sup>10</sup> *Ibid.*, p. 16.

<sup>11</sup> *Ibid.*, p. 17.

<sup>12</sup> BELISLE Claire, « Du papier à l'écran : lire se transforme », in *Lire dans un monde numérique*, Villeurbanne, 2011, p. 143.

<sup>13</sup> « La notion de construction de corpus est centrale : tous les résultats obtenus sont relatifs à sa définition, donc dépendent d'un choix éclairé du chercheur. Une recherche philologique s'impose avant le traitement du corpus, afin de choisir entre plusieurs versions d'un même texte ou d'identifier précisément le(s) auteur(s) » in LEMERCIER Claire, ZALC Claire, *Méthodes quantitatives pour l'historien*, Paris : La Découverte, 2008, p. 51.

## 2. « *L'écriture au contact du monde* »<sup>14</sup>

Comment créer du lien ? Il s'agit d'explorer par exemple les commentaires, les liens connexes et l'évocation des auteurs dans d'autres pages. Un premier problème se pose lorsqu'on souhaite étudier le lien entre deux documents. Pouvons-nous procéder par mots-clés ? D'après François Rastier, la question n'est pas résolue.

« Aucun mot ni aucun passage ne peut prétendre résumer un texte. Certes, définir une saillance, comme on le fait en faisant figurer en tête d'un article une liste de mots-clés, c'est donner une "instruction" interprétative : la clé n'est cependant pas une clé qui ouvre la serrure du sens, car il reste à construire dans l'interprétation. Aussi, les métadonnées doivent-elles garder trace du texte et du contexte et permettre d'y accéder — sans jamais pouvoir s'y substituer. »<sup>15</sup>

Les forums et les plateformes d'écritures littéraires en ligne sont par ailleurs autonomes dans la constitution de communautés d'auteurs. Parfois accessibles sur identification, elles effacent *de facto* l'individualité des auteurs derrière le pseudonymat et la logique collective. Sur Facebook et Twitter, certains auteurs sont déclarés « amis » avec d'autres principalement pour suivre des actualités liées à leur activité. Il est alors possible de suivre un réseau professionnel.

« Le passage d'une simple littératie à cette *translittératie* définie par Sue Thomas comme "l'habileté à lire, écrire et interagir par le biais d'une variété de plateformes, d'outils et de moyens de communication".<sup>16</sup> Diffuse sur plusieurs canaux, l'*identité numérique* devient alors le "centre" paradoxal, car pluriel et mobile, d'une nouvelle modalité de discours sur soi, *l'autobiographie*. »<sup>17</sup>

Il est également possible de parler de liens si deux auteurs partagent un même pays, une même région ou encore qu'ils utilisent les mêmes médias. Le réseau tend ainsi à étudier des relations sociales réciproques. La centralité devant dès lors être pensée par rapport à son objet d'étude.<sup>18</sup>

## B. Interopérabilité des formats : les outils et technologies mobilisés

L'outil provisoirement choisi par l'équipe MARGE pour stocker les URLs et concentrer les métadonnées recueillies par zones géographiques est le logiciel de gestion des références bibliographiques *Zotero*.<sup>19</sup> Il permet de recueillir, sauvegarder les références dans une dynamique de travail collaboratif, avant de les exporter au format RDF pour les transférer vers un outil d'indexation des contenus. Le projet *Lifranum* s'appuyant ensuite sur cet outil pour récupérer les contenus et en défricher d'autres grâce à une exploration automatisée des liens hypertextes récupérés à partir des premières URLs.<sup>20</sup>

<sup>14</sup> LEJEUNE Philippe, *Signes de vie. Le pacte autobiographique* 2, Paris, 2005, p. 58.

<sup>15</sup> RASTIER François, « Sémantique du Web vs. Sémantic Web. Le problème de la pertinence », in *Syntaxe et sémantique*, n° 9, Caen, 2008, p. 15-36. [p. 25]

<sup>16</sup> THOMAS Sue et al., « Transliteracy : Crossing Divides », *First Monday*, vol. 12, n° 12, décembre 2007, (<http://journals.uic.edu/ojs/index.php/fm/article/view/2060/1908>). Ce passage est cité dans la traduction de GUITÉ François (<http://www.francoisguite.com/2007/12/la-translitteratie/>).

<sup>17</sup> BONNET Gilles, « L'*autobiographie*. Écritures numériques de soi », in *Poétique*, n° 177, Paris, 2015, pp. 131-143. [p. 143]

<sup>18</sup> Pierre MERCKLE, *Sociologie des réseaux sociaux*, Paris, 2004, 2011, 2016.

<sup>19</sup> <https://www.zotero.org/>

<sup>20</sup> MONAT, *op.cit.*, p. 18.

L'outil envisagé est *Heritrix*, un robot d'indexation et utilisé par *Internet Archive* pour l'archivage du web. Il s'agit d'un logiciel libre programmé en Java. Il dispose d'une interface principale accessible depuis un navigateur web, mais aussi d'un outil en interpréteur de commandes, ce qui facilite sa prise en mains.<sup>21</sup>

La technologie RDF est utilisée afin d'indexer les contenus. Le RDF (Resource Description Framework) est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre un traitement automatique des données recueillies.<sup>22</sup> À chaque sujet, ici des URLs, est alors associé un prédicat (une propriété) et un objet (sa valeur). Ce modèle de description est destiné à décrire de manière formelle les ressources Web et leurs métadonnées, de façon à permettre un traitement automatique des données recueillies. Ce formalisme peu normalisé propose son propre langage de requêtage : le SPARQL. C'est d'ailleurs par ce biais que l'accès aux données de la BNF est rendu possible. L'une de ses principales syntaxes est le langage RDF/XML qui associe une autre technologie, l'XML-TEI.<sup>23</sup>

« Si l'adoption de standards comme XML ou RDF permet une interopérativité de principe, elle ne résout pas le problème de la production, de l'identification et de l'évolution des connaissances. Les débats portent en amont sur le problème de la réification des connaissances hors des contextes d'utilisation, ou complémentairement sur l'adéquation de leurs modes de représentation à leur utilisation effective. »<sup>24</sup>

Aussi, pour parer à ce problème des ontologies RDF complémentaires sont envisagées telles que FOAF<sup>25</sup> utilisée pour la description des personnes et les relations qu'elles entretiennent entre elles et BIBO<sup>26</sup> pour la gestion des références bibliographiques. Cette dernière a d'ailleurs été récemment importée dans Zotero et Omeka<sup>27</sup> une plateforme de base de données en ligne.

Par ailleurs, il existe plusieurs librairies python relatives au *datamining* qui pourraient être mobilisées afin d'optimiser l'extraction des métadonnées et des contenus : *Beautiful Soup*<sup>28</sup> afin de relever des tags sur les pages ; *Newspaper*<sup>29</sup> concernant l'extraction du contenu des blogs ; *SpaCy*<sup>30</sup> pour l'étiquetage morphosyntaxique ou encore *Lefff*<sup>31</sup> (Lexique des Formes Fléchies du Français) une surcouche du précédent adapté au français offrant une lemmatisation efficace.

---

<sup>21</sup> <https://github.com/internetarchive/heritrix3/wiki>

<sup>22</sup> <https://www.w3.org/TR/rdf-concepts/>

<sup>23</sup> « L'Initiative d'Encodage de Textes est un *consortium* qui élabore et maintient collectivement une norme pour la représentation des textes sous forme numérique. Son principal produit livrable est un ensemble de lignes directrices qui précisent les méthodes de codage des textes lisibles par machine, principalement dans les sciences humaines, les sciences sociales et la linguistique. » in <https://tei-c.org/>

<sup>24</sup> RASTIER F., *op.cit.*, p. 20.

<sup>25</sup> <http://www.foaf-project.org/>

<sup>26</sup> <http://bibliontology.com/projects.html>

<sup>27</sup> <https://omeka.org/>

<sup>28</sup> <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

<sup>29</sup> <https://newspaper.readthedocs.io/en/latest/>

<sup>30</sup> <https://spacy.io/>

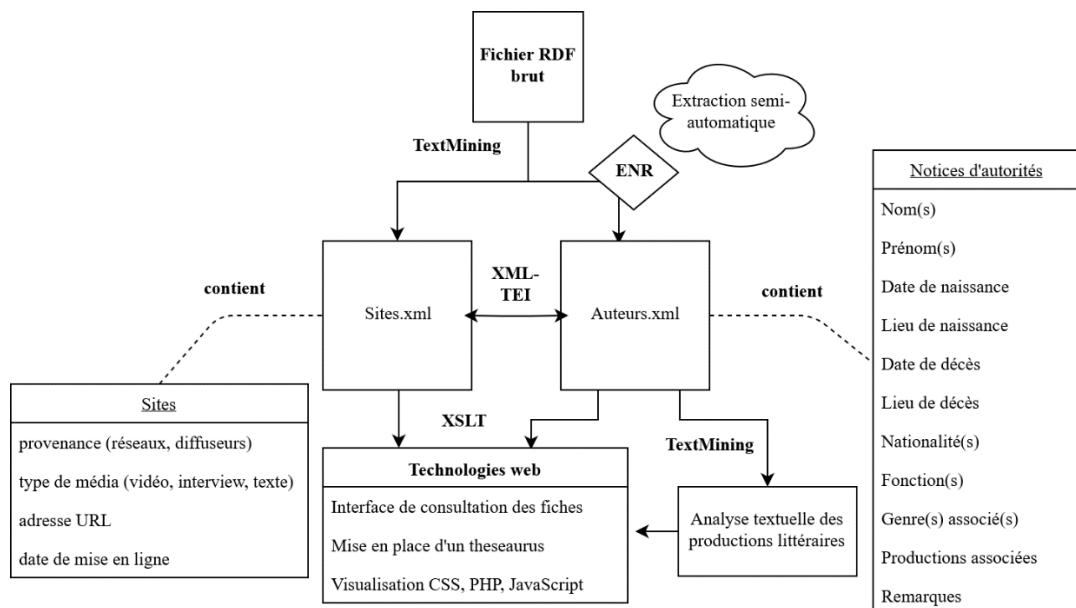
<sup>31</sup> <https://www.labri.fr/perso/clement/lefff/>

## C. Nos missions dans le cadre du projet professionnel du master

### 1. Missions et premières pistes de travail

Les missions qui nous ont été confiées dans le cadre de *Cartoweb* sont « la mise en place d'une base de données et des outils pour verser automatiquement les données issues d'un nombre limité de pages Web au format clairement défini ainsi qu'une réflexion sur les visualisations les plus adaptées à un utilisateur non expert et l'implémentation d'une ou plusieurs de ces visualisations. »<sup>32</sup> Pour effectuer ce travail, une liste d'adresses Web répertoriant des sites et des pages d'écrivains francophones publient en ligne ont été mises à notre disposition au format RDF. Les sites et pages peuvent être associés à un ou plusieurs auteurs et ont été recueillis à partir de métadonnées relativement limitées telles que la localisation, des éléments de temporalités, l'auteur ou encore le type de média utilisé. À partir du fichier RDF, nous avons commencé par élaborer des premières pistes de travail et fait émerger quelques problèmes.

Figure 2 Schéma conceptuel des pistes de travail, réalisé fin septembre 2019, sous Draw.io



Il s'agit, à partir du fichier RDF et d'une extraction semi-automatique des entités nommées, de constituer deux fichiers XML : un dédié aux fiches d'autorités des auteurs et un autre dédié aux URI qui proposeraient des métadonnées liées à la provenance des données, au type de média utilisé ainsi qu'aux dates de mise en ligne et de mise à jour de page. Cette méthode vise un enrichissement des métadonnées existantes, la création de notices d'autorités et la mise en place d'un *thesaurus* par mots-clés. Nous avons d'abord pensé mettre en place un prototypage d'annotations XML-TEI ainsi qu'une analyse textuelle des contenus sur un corpus restreint de productions littéraires à partir desquelles les visualisations viseraient à proposer des parcours de lecture originaux grâce à la mise en place de filtres (par type de média, localisation ou encore par métadonnées auteur).

<sup>32</sup> Document cadre du projet, p. 2.

## 2. Qu'est-ce que cartographier le web littéraire

D'après Bertrand Gervais, la littérature se fait, se conserve, se présente et se reçoit principalement en fonction des dispositifs informatiques.

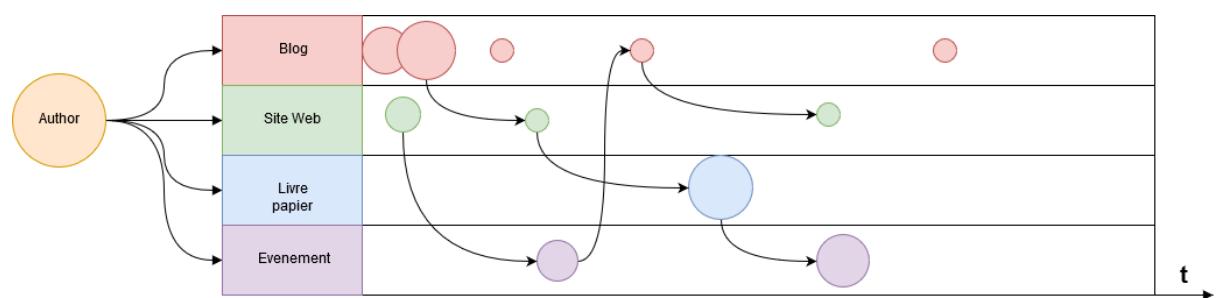
« Il convient [alors] de ne plus distinguer les pratiques littéraires en fonction des médias et des dispositifs utilisés pour les produire, les transmettre et les recevoir. »<sup>33</sup>

Il s'agirait en effet d'aboutir à la fin de la segmentation théorique entre l'oralité et l'écrit. À forcer le continuum, on en vient à diluer un certain nombre de pratiques. L'important serait donc de revenir sur le dynamisme de l'expérience littéraire, à travers une multiplicité des plateformes, où le langage ne serait plus seulement une expérience symbolique.<sup>34</sup>

Marcello Vitali-Rosati<sup>35</sup> adopte quant à lui une position plus radicale en affirmant que la « littérature numérique » n'existe pas. La notion se trouve en effet diluée au sein de pratiques et d'outils très hétérogènes. Il est ainsi peut-être préférable dans cette « omniprésence du “numérique” » de parler plutôt d'une littérature à « l'époque du numérique ». Et finalement, qu'est-ce que la littérature aujourd'hui ? Qu'est-ce que le Web nous dit actuellement de la redéfinition du champ littéraire ? Il est certain que le changement induit par le numérique est à la fois culturel, global et universel.

Il s'agit plutôt d'appréhender le rapport stigmergique entre la technique et la littérature, et le champ littéraire par rapport au numérique. Il faut rechercher la littérature là où elle s'est cachée, dans diverses plateformes aux formats homéotiques, redéfinir les limites de la littérature et ses fonctions de légitimation. De plus, peut-on encore parler d'écrivain ? La figure change et le vocabulaire se doit de l'être aussi. Ainsi, ce n'est pas tant la qualité ni la quantité qui comptent, mais ce qui est possible de visualiser : un processus d'acquisition d'un savoir-faire et de définition d'un statut.<sup>36</sup> Une visualisation chronologique des parcours sociaux des auteurs permettrait, par exemple, d'identifier les processus d'accomplissement des auteurs et aider à leur définition.

Figure 3 Exemple de visualisation permettant d'explorer des parcours inter-littéraires



<sup>33</sup> GERVAIS Bertrand, « Imaginer un partenariat ancré en culture numérique », Colloque international Cartographie du Web littéraire francophone, Lyon, mercredi 22 Janvier 2020.

<sup>34</sup> *Ibid.*

<sup>35</sup> VITALI-ROSATI Marcello, « Je ne suis pas un littéraire. Plaidoyer pour des frontières disciplinaires poreuses », Colloque international Cartographie du Web littéraire francophone, Lyon, mercredi 22 Janvier 2020.

<sup>36</sup> *Ibid.*

## **II. L'évolution du projet**

À ce stade nous avons eu de nombreuses réflexions générales relatives aux données de la littérature du web francophone, notamment concernant l'importance de leur représentativité, les technologies pouvant être mises à profit et la légitimité des auteurs. Voyons à présent la limite de l'automatisation de la collecte des métadonnées ainsi qu'une définition de ce que pourrait être un enrichissement « raisonné » des données. Enfin, les questions relatives à l'extraction des entités nommées et à l'exploitation des données Facebook seront abordées.

### **A. Divers questionnements**

#### ***1. Autour de l'enrichissement des données***

Il est intéressant de réfléchir sur la limite de l'automatisation. Son interruption ne doit pas être décidée lorsque les possibilités techniques nous y contraignent, mais lorsque la qualité scientifique des données en dépend. Au-delà d'un certain seuil, les informations continueront d'être collectées par la machine, mais les résultats obtenus seront de moins en moins pertinents. Il en est de la responsabilité des chercheurs d'effectuer un contrôle qualité : l'idéal étant de trouver un compromis entre la performance de la machine sur la collecte des métadonnées et le temps nécessaire à l'homme pour les valider, les corriger et les compléter. Nous avons donc réfléchi de manière à faciliter le travail du chercheur sans occulter l'importance de l'intellect humain qui ne peut pas et ne doit pas être remplacé par la machine. François Rastier insiste d'ailleurs sur la nécessité de l'« implication » disciplinaire pour une « application véritable » :

« [qui] se doit d'intervenir, même de façon auxiliaire, à diverses étapes : [lors de la] création des logiciels, [de la] constitution des corpus, [des] balisages, [des] expérimentations avec outils sur corpus (balisés), [leur] interprétation et [la] discussion des résultats » ; et où « toutes ces étapes de la chaîne de traitement (...) se révèlent indispensables ».<sup>37</sup>

Un enrichissement automatique, bien qu'il ne puisse pas être contrôlé, vient appuyer l'importance d'un élément dans un texte. L'exploitation des données par la machine implique immanquablement une validation par l'homme. Ce n'est qu'après cela que l'on peut parler d'une forme enrichie des données.

#### ***2. L'extraction des entités nommées***

Nous avons extrait les entités nommées grâce à la librairie *SpaCy*, qui nous permet de recueillir des noms propres, des entités numériques et géographiques, de même que des noms et sigles d'entreprises ou d'institutions publiques.<sup>38</sup> Il convient de considérer les localisations et les personnalités qui seraient repérées sur différentes pages nous renvoyant ainsi les liens partagés par plusieurs auteurs. Ces entités provenant d'une procédure automatique ne seront alors pas confondues avec celles que nous aurons pu extraire des notices d'autorités, puisque celles-ci émaneront du virtuel et non du vérifié.

---

<sup>37</sup> RASTIER, *op.cit.*, pp. 22-23.

<sup>38</sup> <https://spacy.io/api/annotation#named-entities>

Elles pourront néanmoins permettre de créer des liens potentiels. Toutefois, cette procédure étiquetage morphosyntaxique repose sur l'application de systèmes entraînés dans d'autres contextes et peuvent donc être source d'erreurs. Il est par ailleurs utile de souligner la diversité des langues parlées dans la zone Caraïbes telles que le créole et ses patois, l'espagnol, l'anglais et le français qui interrogent sur une éventuelle normalisation des sorties.

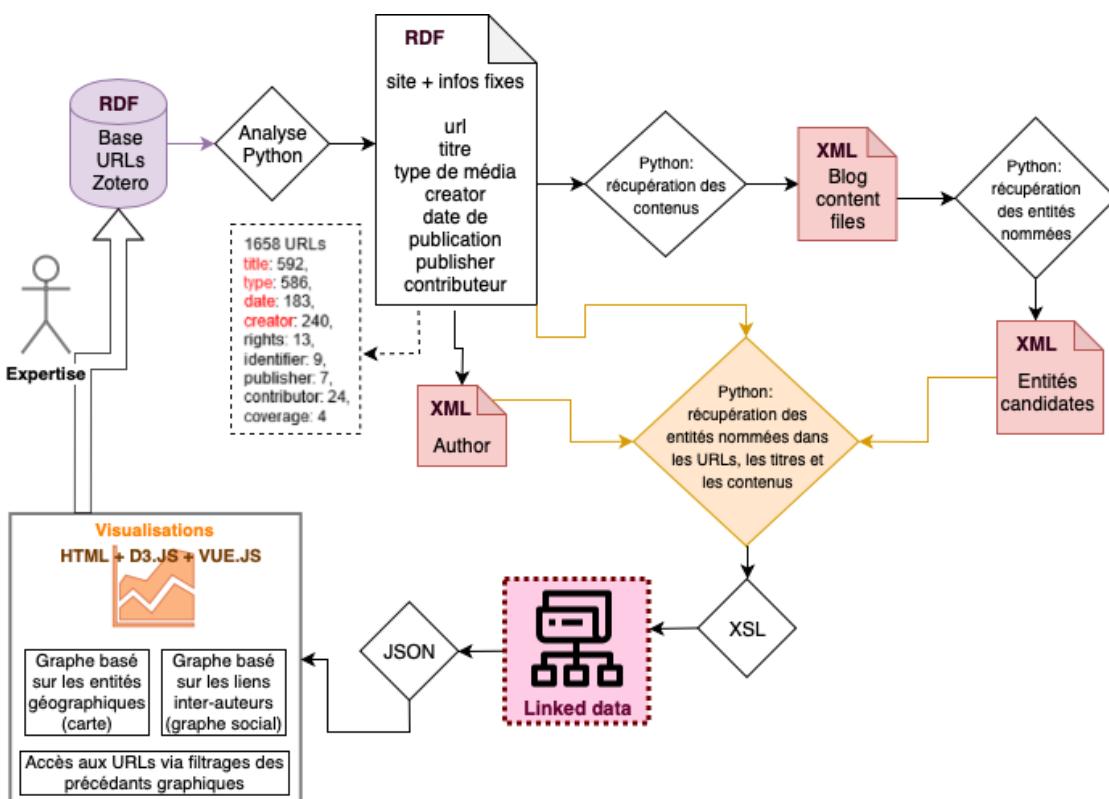
Aussi, que faire des auteurs dont nous n'avons peu ou pas de métadonnées ? Il s'agit par exemple de ceux qui ne mentionnent pas, leur profession, voire même qui n'utilisent pas leur vrai nom, mais un pseudo, du fait de la censure de leur pays d'origine. Dans ce cadre, le manque de données est significatif et doit être considéré. Il en est de même pour les blogs qui ont été fermés. La cartographie s'attache à des ressources accessibles au long du déroulé du projet :

« un blog non alimenté depuis plusieurs années pourra être cartographié, en revanche aucune recherche ne sera faite dessus ».<sup>39</sup>

## B. Retour sur les premiers scénarios

### 1. Le premier schéma

Figure 4 Schéma conceptuel du premier scénario, réalisé mi-octobre 2019 sous Draw.io



Notre premier scénario, réalisé mi-octobre 2019, utilise comme support une base d'URLs, sous format RDF, issue d'une bibliothèque *Zotéro*. Le contenu des liens peut être extrait pour relever automatiquement des éléments explicites. Une première limite peut dès à présent être soulignée, car nous perdons tout ce qui relève de l'implicite et qui ne peut pas être automatisé.

<sup>39</sup> MONAT, *op.cit.*, p. 15.

En croisant les fiches auteurs et les entités nommées, il est possible de générer un fichier unique qui contiendrait les liens entre les différents auteurs. Une exploitation JSON, format le plus adapté à la manipulation en JavaScript, permettra la réalisation de visualisations.

Le scénario a rapidement posé problème en raison du manque de métadonnées disponibles dans le RDF dû au mauvais référencement de la bibliothèque *Zotéro*. À l'issue de notre premier échange avec les commanditaires, il nous a été demandé de nous focaliser sur l'extraction des métadonnées et ainsi de proposer des visualisations élémentaires sur l'ensemble du corpus.

**1658 URLs**

**title:** 592,  
**type:** 586,  
**date:** 183,  
**creator:** 240,  
**rights:** 13,  
**identifier:** 9,  
**publisher:** 7,  
**contributor:** 24,  
**coverage:** 4

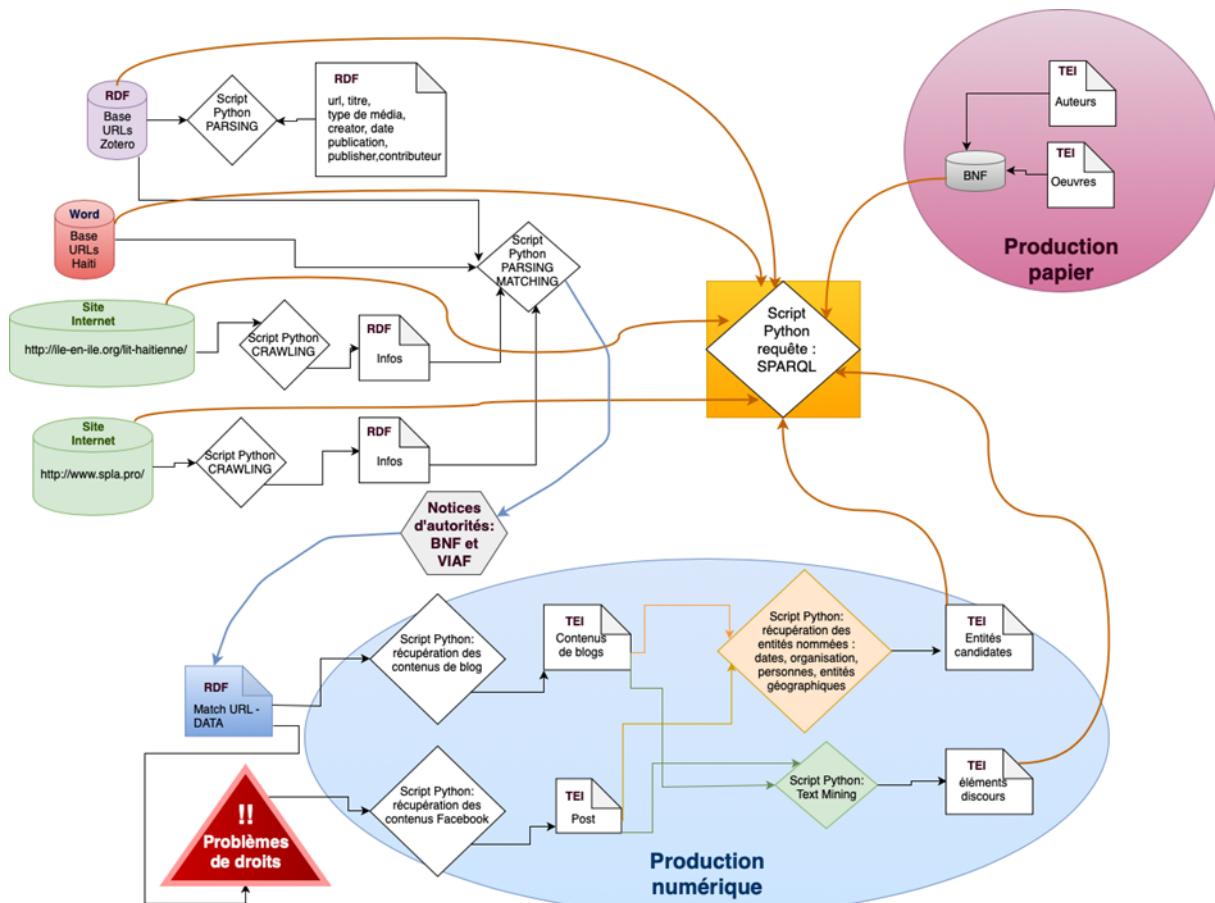
*Figure 5 Extrait du résultatat d'interrogation des métadonnées*

Sur les 1658 URLs contenues dans le RDF, seules 592 (35,7 %) disposaient au moins d'un titre, 586 (35,3 %) d'un type, 183 (11 %) étaient datées et 240 (14,5 %) possédaient une balise « creator ». Cependant, il peut aussi bien s'agir de l'auteur de la page que de l'instigateur du site et donc ne pas posséder les contenus. Face à ces problèmes, le corpus a alors été réduit à la zone Caraïbes.

Malgré l'impossibilité actuelle de réaliser ce scénario, nous vous le présentons toutefois, car il pourrait être utilisé ultérieurement.

## 2. Le deuxième schéma

*Figure 6 Schéma conceptuel du deuxième scénario, réalisé début novembre 2019 sous Draw.io*



Ce scénario a été effectué plus tard, début novembre 2019, lorsque nous nous sommes rendu compte qu'il y avait trop peu de données exploitables pour aboutir à des visualisations intéressantes et interprétables. La même problématique que celle du premier scénario concernant le manque de métadonnées s'est rapidement posée pour le sous-corpus d'Haïti.

Nous avons donc été obligés de compléter les métadonnées existantes et de collecter de nouvelles informations afin de constituer une base de données augmentée. Nous avons décidé de prendre comme point d'entrée deux sites de références d'auteurs : *île-en-île*<sup>40</sup> un corpus déjà constitué d'auteurs littéraires haïtiens et *spla.pro*<sup>41</sup> qui constitue une source d'informations pour la mise en réseau des acteurs culturels du Sud. Afin de suivre le processus d'accomplissement des auteurs du web à la publication papier, nous aimerais également y ajouter des informations issues de répertoires d'autorité tels que la BNF et le VIAF.

Le fichier RDF bleu sera une fiche unique contenant la normalisation des quatre corpus : les liens, les auteurs, les pays et la langue. Elle aura été enrichie des informations renseignées par la BNF et des notices VIAF. On sera alors en mesure de constituer un fichier TEI d'entités nommées, mots-clés pour la constitution des liens en réseau. La finalité du modèle serait un fichier Python SPARQL pour pouvoir concrétiser les visualisations des corpus en réseau.

### C. Le problème spécifique de la collecte des données *Facebook*

Une bonne partie des liens fournis dans le RDF étaient des pages et des profils Facebook. Prenons l'exemple du sous-corpus haïtien, qui décrivait 52 URLs mises en ligne par 33 auteurs. Une matière hétérogène, composée de blogs personnels, de plateformes éditoriales et 16 pages Facebook, ces dernières avaient attrait au style de vie des auteurs, à la promotion de leurs écrits, mais aussi les évènements auxquels ils participaient.

Néanmoins, la collecte des données sur ce type de support est difficile sans accréditation. Il nous a fallu d'abord créer un profil Facebook spécifique « *Facebook for developers* », fournir une preuve de notre identité (passeport, carte d'identité...) et créer une page d'entreprise. Ce compte permet de récupérer de nombreuses informations (*posts*, localisations, *likes*, listes d'amis, etc.). Ce point d'accès est néanmoins limité en termes du nombre de requêtes possibles environ 15 tous les quarts d'heure. Chaque demande d'informations devra ensuite être validée ou non par les administrateurs du système à l'appui d'une vidéo et d'un texte.

Notre demande de création de compte n'a par ailleurs jamais été acceptée.<sup>42</sup> De plus, certaines de ces pages n'existent plus, aussi ne faudrait-il pas les retrouver à travers *Internet Archive*. Les données issues de ce travail ne pourront être intégrées au projet, celles-ci relevant de la responsabilité personnelle des membres de notre groupe.

---

<sup>40</sup> <http://ile-en-ile.org/lit-haitienne/>

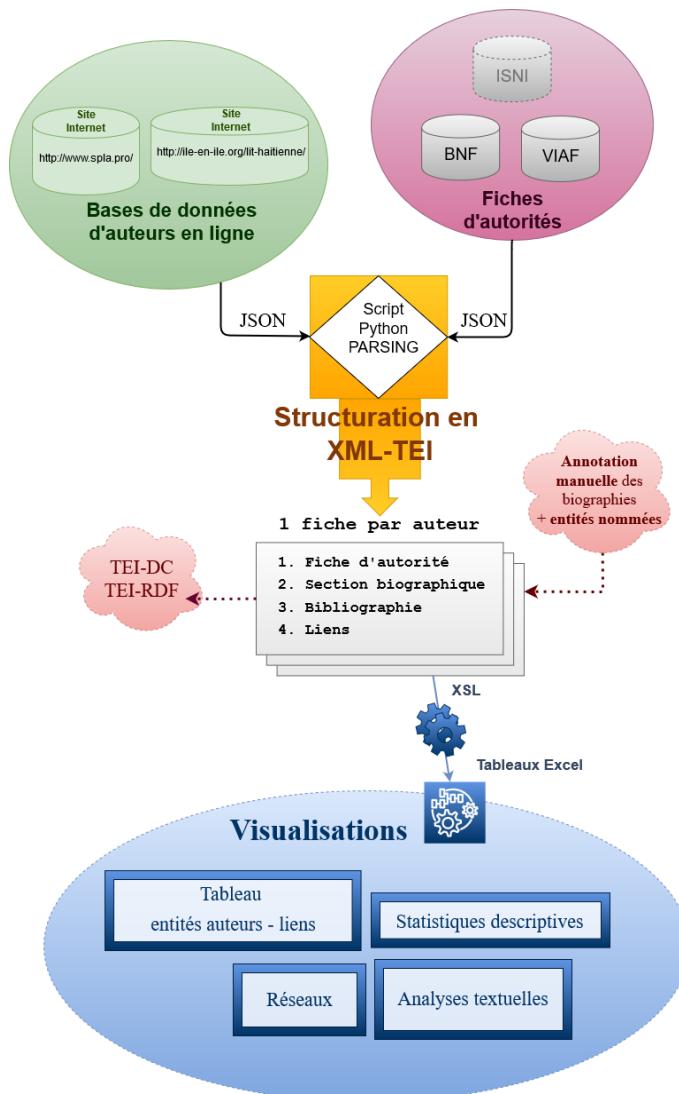
<sup>41</sup> <http://www.spla.pro/spla-presentation.html>

<sup>42</sup> La demande a été formulée le 30 octobre 2019.

### III. Présentation du scénario

Les problèmes rencontrés, à l'issue de ces deux scénarios, ont provoqué un changement de paradigme dans notre manière de traiter la demande. En effet, il ne s'agit plus d'étudier des contenus littéraires, mais plutôt de revenir aux auteurs, qu'ils soient issus du web ou non, de lister leur production web et papier, et les évènements liés. De surcroît, il était primordial de revenir sur une base de données, non seulement structurée, mais surtout ayant un contenu visant un maximum d'exhaustivité. Ainsi, pour le corpus d'Haïti nous nous en remettons aux sites *ile-en-ile.org* et *spla.pro*. Il est également essentiel de se référer à des fiches d'autorité internationales afin d'enrichir le corpus de données certifiées. C'est donc à la suite de ce processus que nous vous proposons l'explicitation de notre méthodologie.

*Figure 7 Schéma conceptuel du scénario final, réalisé début fin décembre 2019 sous Draw.io*



Notre schéma utilise des sources dont les données, accessibles en ligne, sont récupérées et redistribuées sous la forme d'un modèle XML-TEI de fiches entités auteurs. Ces dernières seront utilisées pour réaliser de nouvelles sorties.

## A. Des sources de données hétérogènes

Notre corpus est centré sur des auteurs haïtiens, qu'ils soient établis ou non, et sur leur communauté. Nous avons pu identifier trois types de sources d'enrichissement potentielles : des sites de références d'auteurs (*île-en-île.org* et *spla*), des données issues du web social (Twitter, Facebook, Wattpad, Scribd) et des sources du web sémantique (BNF, VIAF). Bien que le traitement des « données du web social » n'ait pas abouti, il est toujours possible de déverrouiller cette piste dans la suite du projet. De cette manière, ils pourront effectuer par la suite un approfondissement des contenus ainsi retrouvés. Chacune de ces sources ont des particularismes. Nous vous proposons un tableau (voir ci-dessous) pour mieux les cerner.

*Figure 8 Tableau comparatif des sources potentielles d'enrichissement auteur*

	Scraping de sites de « référence »	Données du web social	Données du web sémantique
Qualité de formalisation	- très peu formatées - souvent textuelles	Formatées	- formatées + vérifiées
Informations disponibles	- Profils auteurs - Bibliographies - Références web	- Profils auteurs - Productions - Evénements + liens	- Profils auteurs - Bibliographies - Liens sémantiques
Accès à l'information	- Moteur de recherche	- API + M.R.*	- API + M.R.*
Garant éditorial	- Garant « amateur »	- L'auteur et son réseau	- Garant « expert »

\*MR : moteur de recherche

### 1. Méthode de scrapping

Afin de récupérer les informations disponibles sur les différents sites, nous avons construit des scripts de *scrapping* mettant à profit *Selenium* et *Beautiful Soup*, deux librairies Python. Nous avons pu simuler un comportement utilisateur afin d'accéder aux données des répertoires d'auteurs. C'est au fil de cette « navigation préprogrammée » que nous avons récupéré le contenu des documents HTML. Nous avons enregistré ces informations dans des fichiers JSON formés de données non-structurées et adaptés à nos besoins puisque nous ne connaissons pas *a priori* la structure de ces pages. De cette manière, des biographies très peu structurées ont été récupérées, ainsi que d'autres variables plus normalisées telles que la langue maternelle de l'auteur, sa profession, son pays, etc.

Sur les 139 auteurs récupérés sur île-en-île seulement 69 étaient présents parmi les 193 référencés sur *spla*, c'est-à-dire 263 auteurs au total dont 60 % possèdent un identifiant BNF et/ou VIAF. Nous avons effectué des regroupements dû à la déduction des différentes combinaisons onomastiques possibles afin d'éviter les doublons. Ainsi nous avons obtenu une base de référence de 220 entités auteurs uniques.

## 2. Le cas particulier des notices biographiques

Nous avons extrait les noms de personnes et les lieux évoqués dans les notices biographiques. Nous avons donc pu obtenir, grâce à cette démarche, des réseaux potentiels d'acteurs, lesquels ne sont pas forcément auteurs, et une suite de lieux qui peut permettre de retracer des parcours de vie. Des erreurs sont toutefois à relever. Prenons par exemple la ville de « Port-au-Prince ». Le système identifie « Prince » comme étant un chanteur connu, et donc une entité personne, alors que nous sommes bien face à un nom de lieu. De même, l'auteur « Jean D'Amérique » subit la situation inverse puisque le système repère que « Jean » est un nom de personne et « Amérique » un nom de lieu.

Figure 9 Exemple d'extraction d'entités nommées

Bonel Auguste, né à Port-au-Prince, Haïti, le 6 mars 1973, est poète, bibliothécaire et animateur culturel. Il a étudié la linguistique à l'Université d'Etat d'Haïti et a reçu une formation en histoire de l'art à l'Institut français d'Haïti. Fondateur de l'atelier de poésie « Dimanche en poésie » à la bibliothèque Étoile filante, et animateur des « Vendredis littéraires » de l'université Caraïbe, il écrit en français et en créole.

Né en Haïti en 1994, Jean D'Amérique est écrivain, auteur de Petite fleur du ghetto, recueil qui lui vaut une mention spéciale du Prix René Philoctète 2015 et une sélection au Prix Révélation de Poésie 2016 de la Société des Gens de Lettres.

Afin de corriger ces erreurs, plusieurs pistes sont à envisager. Nous pouvons ré-entrainer notre modèle sur un corpus restreint à une zone géographique donnée. Il devra être accompagné d'une liste d'entités potentiellement employée au sein de celui-ci.

## 3. De l'exploration contextuelle à l'extraction de connaissances

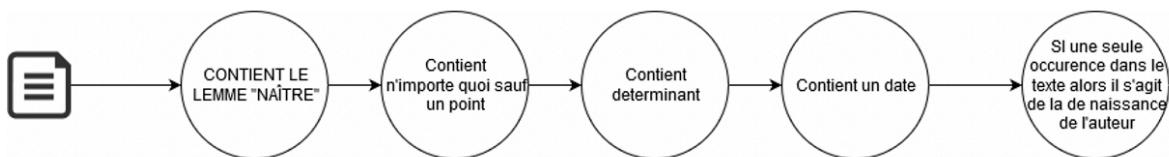
Nous avons eu recours à l'exploration contextuelle<sup>43</sup> pour détecter les entités liées aux dates d'état civil. Cette méthode a notamment été utilisée par Célian Ringwald dans le cadre de ces projets au sein de *DataObserve*. Basée sur les étiquettes morphosyntaxiques établies par *SpaCy*, la méthode utilise un package python nommé *DateExtractor*, qui permet d'extraire automatiquement les date de nos contenus. Ces algorithmes parcourront alors les notices biographiques à la recherche d'un passage répondant à une suite de règles d'énonciation d'une date de naissance. Ainsi, si un nom se trouve à proximité du lemme « naître », d'un déterminant et d'une date, il s'agit alors de sa date de naissance. Voir ci-dessous.

<sup>43</sup> DESCLES Jean-Pierre, CARTIER Emmanuel, JACKIEWICZ Agata et MINEL Jean-Luc, « Textual Processing and Contextual Exploration Method », Colloque Context' 97, Rio de Janeiro, Février 1997.

Figure 10 Exemple de la détection d'informations liées à la naissance

Cleante Desgraves naît le 13 janvier 1891 à Port-au-Prince (Haïti) d'une mère américaine, née Alice Cunningham, et d'un père haïtien, Hector Desgraves, qui était pharmacien et également pianiste.

Bonel Auguste, né à Port-au-Prince, Haïti, le 6 mars 1973, est poète, bibliothécaire et animateur culturel. Il a étudié la linguistique à l'Université d'Etat d'Haïti et a reçu une formation en histoire de l'art à l'Institut français d'Haïti. Fondateur de l'atelier de poésie « Dimanche en poésie » à la bibliothèque Etoile filante, et animateur des « Vendredis littéraires » de l'université Caraïbe, il écrit en français et en créole.



De cette façon, d'autres informations peuvent ainsi être extraites : telles que les lieux de décès, de formation ou encore de profession.

## B. Utilisation de l'XML-TEI

### 1. Les avantages

Nous avons choisi d'utiliser l'XML-TEI, qui possède de multiples avantages, pour réunir les différentes données recueillies. Lou Burnard nous les résume ainsi :

« (...) pourquoi, tout d'abord, prendre la peine d'utiliser la TEI ? (...) La première [réponse] est que le XML-TEI s'intéresse au sens du texte plutôt qu'à son apparence. La deuxième est que le XML-TEI est indépendant de tout environnement logiciel particulier. La troisième est que le XML-TEI a été conçu par la communauté scientifique, qui est aussi en charge de son développement continu. »<sup>44</sup>

L'enrichissement des données s'effectue grâce à un système d'annotation de document à même le contenu sans le modifier. Il aide à la conception d'une structuration généralisée et flexible de tous les types de données, tout en permettant l'accès à un large écosystème de sorties, allant du Dublin Core, au RDF, au HTML ou encore au TXT.

« Le XML-TEI nous offre un cadre pour représenter tout ce que nous considérons important à propos d'un texte, et non pas uniquement son apparence, afin que les logiciels puissent agir sur les différentes caractéristiques identifiées, générant de nouvelles visualisations et de nouvelles perspectives. En conséquence, prendre des décisions quant à l'encodage XML-TEI à utiliser est toujours une activité scientifique, impliquant systématiquement un choix conscient. (...) Cette combinaison de flexibilité et d'un intérêt centré sur les besoins scientifiques d'une communauté particulière fait de la TEI un cas particulier parmi les autres efforts de standardisation, et est également l'un des facteurs expliquant sa longévité. »<sup>45</sup>

<sup>44</sup> BURNARD LOU, *Qu'est-ce que la Text Encoding Initiative ?*, en ligne : <https://books.openedition.org/oep/1297>, Marseille, 2015.

<sup>45</sup> *Idem.*

Notre modèle XML-TEI a été créé d'après une base générique en vue de pouvoir être évolutif face aux besoins encore indéfinis du *Lifranum*. Un autre de ces avantages concerne l'aide à la compréhension du langage grâce à l'implication d'une communauté d'utilisateurs.

« (...) la connaissance et l'expertise de la TEI peuvent aujourd'hui facilement se rencontrer dans la plupart des disciplines des sciences humaines, et ne sont pas recluses dans une zone obscure des sciences de l'information. »<sup>46</sup>

Il existe d'ailleurs un manuel de référence<sup>47</sup> qui explicite et codifie la pratique de la TEI.

## 2. *Le modèle*

### a) La structure générale

Figure 11 Capture d'écran du modèle général XML-TEI, dans le logiciel Oxygen

```
TEI
1 <?xml version="1.0" encoding="UTF-8"?> <!-- déclaration du fichier TEI et de son encodage -->
2 <TEI xmlns:xi="http://www.w3.org/2001/XInclude" xml:lang="fr" xml:id="Marie-Celie_Agnant">
3
4   <!-- 1 fichier TEI par entité auteur -->
5   <!-- xml:id= correspondant à celui du fichier master.-->
6   <!-- Normalisation de l'id: prenom_nom de entité auteur -->
7
8   <teiHeader> [21 lines]
30
31   <div xml:id="Marie-Celie_Agnant_fiche" n="1"> [65 lines]
97
98   <div xml:id="Marie-Celie_Agnant_bio" n="2"> [73 lines]
172
173   <div xml:id="Marie-Celie_Agnant_biblio" n="3"> [81 lines]
255
256   <div xml:id="Marie-Celie_Agnant_web" n="4"> [11 lines]
268
</TEI>
```

Notre corpus est constitué de fichiers XML-TEI qui correspondent pour chacun à une entité auteur. Nous avons privilégié le terme « entité-auteur » et non « auteur » pour distinguer les auteurs seuls des associations ou des collectifs. Pour chacune de ces fiches, un *xml:id* unique et normalisé est donné (voir l'encadré violet de la figure ci-dessus). La fiche est organisée en deux ensembles : le TEI Header et des balises *<div>*.

<sup>46</sup> *Idem.*

<sup>47</sup> <http://www.tei-c.org/Guidelines>

## b) Un premier ensemble : le TEI Header

Figure 12 Structure du <teiHeader> du modèle XML-TEI, capture d'écran Oxygen

```
<teiHeader>
  <!-- déclaration des métadonnées -->
  <fileDesc>
    <titleStmt>
      <title>Notice de Marie-Célie Agnant - corpus Haïti</title>
      <!-- prénom, nom de l'auteur -->
      <author>Inès Burri</author>
      <author>Anaïs Chambat</author>
      <author>Célian Ringwald</author>
    </titleStmt>
    <publicationStmt>
      <orgName>Université Jean Moulin Lyon 3</orgName>
      <orgName>Université Lumière Lyon 2</orgName>
      <orgName>ENS de Lyon</orgName>
      <pubPlace>Lyon</pubPlace>
      <date>17.01.2020</date>
      <!-- date de l'extraction -->
    </publicationStmt>
    <sourceDesc>Notice et contenus associés à l'entité auteur dont les données sont issues
      de sites de références et de répertoires d'autorités</sourceDesc>
  </fileDesc>
</teiHeader>
```

Le TEI Header est une structure commune basique et obligatoire qui renseigne une description du fichier, des données éditoriales et de la source utilisée.

## c) Un deuxième ensemble : les balises <div>

Figure 13 Description des <div> du modèle XML-TEI, capture d'écran Oxygen

```
TEI
1  <?xml version="1.0" encoding="UTF-8"?> <!-- déclaration du fichier TEI et de son encodage -->
2  <TEI xmlns:xi="http://www.w3.org/2001/XInclude" xml:lang="fr" xml:id="Marie-Celie_Agnant">
3
4    <!-- 1 fichier TEI par entité auteur -->
5    <!-- xml:id= correspondant à celui du fichier master.-->
6    <!-- Normalisation de l'id: prenom_nom de entité auteur -->
7
8  <teiHeader> [21 lines]
30
31  <div xml:id="Marie-Celie_Agnant_fiche" n="1"> [65 lines]
97
98  <div xml:id="Marie-Celie_Agnant_bio" n="2"> [73 lines]
172
173  <div xml:id="Marie-Celie_Agnant_biblio" n="3"> [81 lines]
255
256  <div xml:id="Marie-Celie_Agnant_web" n="4"> [11 lines]
268 </TEI>
```

Le deuxième ensemble regroupe l'intégralité des données, récupérées à l'aide du *scrapping*, sur l'entité auteur. Il s'agit d'une structure en quatre balises <div>, où chacune contient un *xml:id* unique normalisé (voir les soulignés bleus ci-dessus), et est caractérisée à l'aide d'un attribut *n* (voir les encadrés violettes de la figure ci-dessus). Cet attribut *n* est un numéro commun à toutes les autres balises <div> des fiches du corpus. Les balises <div>, dont l'attribut *n* est égal à 1, correspondent toujours aux fiches d'autorité, les numéros deux aux notices biographiques, les numéros trois aux bibliographies. Et, enfin, les numéros quatre renvoient aux listes d'adresses URLs associées.

#### d) Des subdivisions

Figure 14 Description des subdivisions des <div> du modèle, capture d'écran Oxygen

```

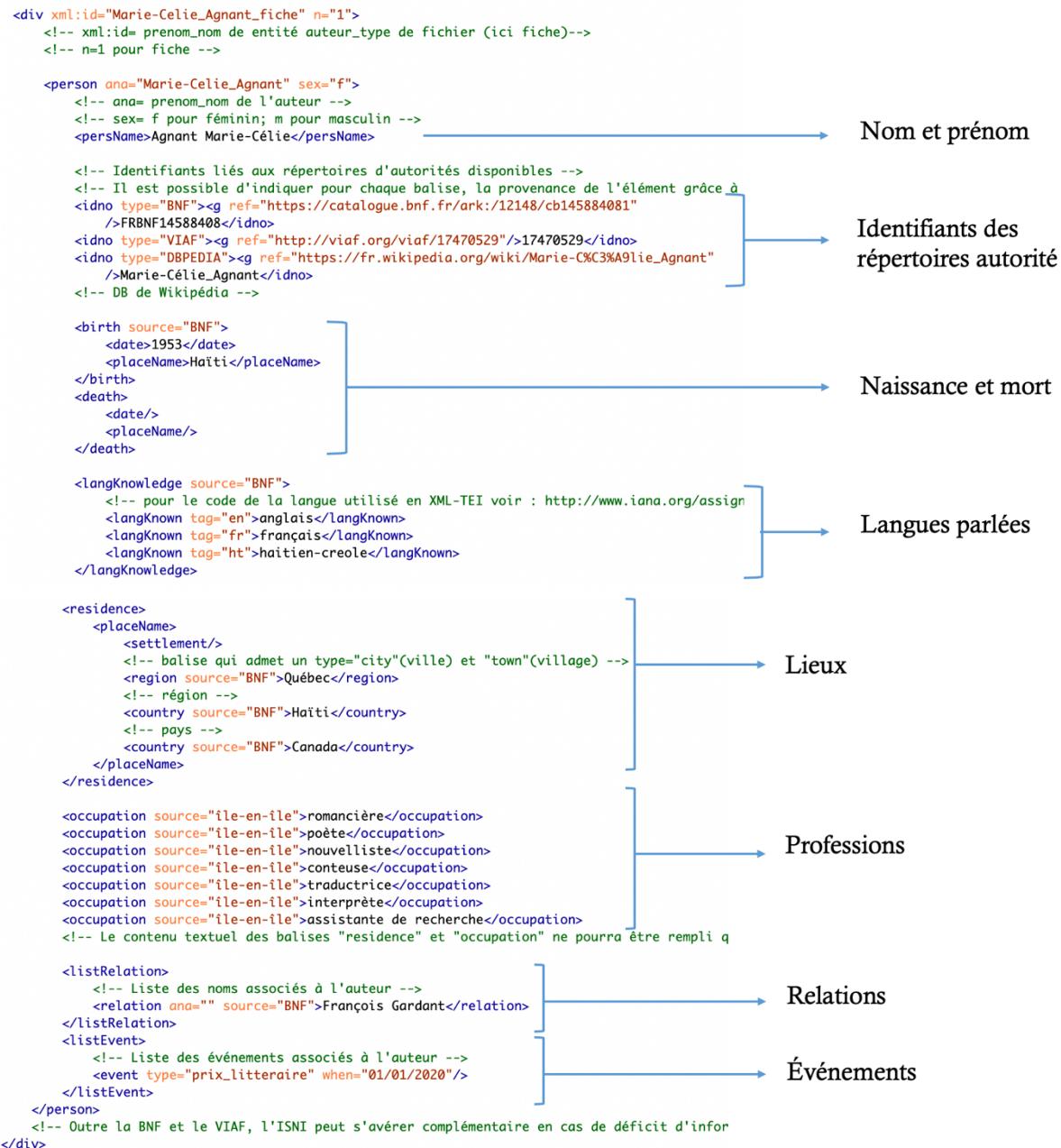
TEI
1  <?xml version="1.0" encoding="UTF-8"?> <!-- déclaration du fichier TEI et de son encodage -->
2  <TEI xmlns:xi="http://www.w3.org/2001/XInclude" xml:lang="fr" xml:id="Marie-Celie_Agnant">
3
4      <!-- 1 fichier TEI par entité auteur -->
5      <!-- xml:id= correspondant à celui du fichier master.-->
6      <!-- Normalisation de l'id: prenom_nom de entité auteur -->
7
8  <teiHeader> [21 lines]
30
31  <div xml:id="Marie-Celie_Agnant_fiche" n="1">
32      <!-- xml:id= prenom_nom de entité auteur_type de fichier (ici fiche)-->
33      <!-- n=1 pour fiche -->
34  <listPerson>
35      <person ana="Marie-Celie_Agnant" sex=""> [54 lines]
90
91      <!-- Outre la BNF et le VIAF, l'ISNI peut s'avérer complémentaire en cas de déficit d'informations -->
92  </div>
93
94  <div xml:id="Marie-Celie_Agnant_bio" n="2">
95      <!-- xml:id= prenom_nom de entité auteur_type de fichier (ici bio)-->
96      <!-- n=2 pour bio -->
97
98      <text ana="Marie-Celie_Agnant" ref="http://ile-en-ile.org/agnant/" hand="Lucie Lequin"> [67 lines]
166
167
168  <div xml:id="Marie-Celie_Agnant_biblio" n="3">
169      <div ana="Marie-Celie_Agnant" source="ile-en-ile" n="3.1" > [42 lines]
212      <div ana="Marie-Celie_Agnant" source="VIAF" n="3.2" > [4 lines]
217  </div>
218
219  <div xml:id="Marie-Celie_Agnant_web" n="4">
220      <!-- xml:id= prenom_nom de entité auteur_type de fichier (ici liens)-->
221      <!-- n=4 pour liens -->
222
223      <list ana="Marie-Celie_Agnant" source="ile-en-ile" subtype="liens"> [4 lines]
228      <list ana="Marie-Celie_Agnant" source="spl" subtype="liens"> [2 lines]
231  </div>
232 </TEI>

```

Ces mêmes <div> sont subdivisées en d'autres balises selon leur fonction. La <div> numéro une contient une balise <listPerson>, qui est elle-même fractionnée en une ou plusieurs balises <person> puisqu'il s'agit de lister la ou les personnes correspondant à l'entité auteur. La <div> numéro deux contient des balises <text>, car elle comporte les textes biographiques. Pour la <div> numéro trois, on retrouve des balises <div>, étant donné que l'on doit hiérarchiser une bibliographie. Et enfin la <div> numéro quatre renferme des balises <list>, puisqu'il est question de lister des URLs. Ces sous-catégories possèdent un attribut *ana* qui fait écho à l'*xml:id* de la fiche. L'important étant ici de pouvoir extraire des balises similaires à toutes les fiches sans perdre le lien avec l'entité auteur.

#### d) I. Le contenu de la balise <div> numéro une

Figure 15 Contenu de la <div> numéro une du modèle XML, capture d'écran Oxygen



La <div> concernant la fiche d'autorité renseigne le nom, les identifiants liés aux répertoires d'autorité, les informations liées à la naissance et à la mort, les langues parlées, les lieux, les professions, une liste des relations et une liste d'événements. Ceux-ci sont respectivement dans des balises <persName>, <idno>, <birth> et <death>, <langKnowledge>, <residence>, <occupation>, <listRelation> et <listEvent>. Grâce à un attribut *source* ou *type*, selon la balise, on peut conserver la source de l'information. Les balises <listEvent> permettront de contextualiser les événements véhiculés par les communautés d'auteurs.

d) 2. Le contenu de la balise <div> numéro deux

Figure 16 Extrait de la <div> numéro deux du modèle XML, capture d'écran Oxygen

```
<div xml:id="Marie-Celie_Agnant_bio" n="2">
  <!-- xml:id= prenom_nom de entité auteur_type de fichier (ici bio)-->
  <!-- n=2 pour bio -->

  <text ana="Marie-Celie_Agnant" source="ile-en-ile" ref="http://ile-en-ile.org/agnant/"
    hand="Lucie Lequin">
    <!-- ana= prenom_nom de l'auteur -->
    <!-- ref= adresse de la bio -->
    <!-- hand = dernier paragraphe de la bio -->

    <p>Poète, nouvelliste, romancière, Marie-Célie Agnant quitte Haïti pour le Canada en
      1970. Traductrice et interprète, elle détient un diplôme en enseignement du français
      et a enseigné plusieurs années tout en travaillant comme assistante de recherche.
      C'est ainsi que son premier roman, La dot de Sara, publié en 1995, sera « construit
      à partir de récits de vie de grands-mères d'origine haïtienne, transmis en créole
      dans le cadre d'une recherche en sociologie ».</p>
    <p>Écrivaine présente et attentive au monde qui l'entoure, son écriture porte à la fois
      le sceau de la poésie et de la violence issue des sociétés postcoloniales qui
      naviguent entre misère criante et opulence indécente. Ses textes, dont certains ont
      été traduits en plusieurs langues, trouvent leur ancrage dans la réalité sociale
      contemporaine ; elle aborde les thèmes tels l'exclusion, la solitude, le racisme,
      l'exil. La condition des femmes, le rapport au passé et à la mémoire font aussi
      partie de son champ d'exploration.</p>
```

La <div> numéro deux comprend une balise <text> pour chaque notice biographique écrite. Ainsi, on renseigne la source, l'URL, et lorsque cela est possible l'auteur de la notice biographique. Respectivement avec les attributs *source*, *ref* et *hand*.

#### d) 3. Le contenu de la balise <div> numéro trois

Figure 17 Contenu de la <div> numéro trois du modèle XML, capture d'écran Oxygen

```
<div xml:id="Marie-Celie_Agnant_biblio" n="3">
  <!-- xml:id= prenom_nom de entité auteur_type de fichier (ici biblio)-->
  <!-- n=3 pour biblio -->

  <div ana="Marie-Celie_Agnant" source="ile-en-ile" ref="http://ile-en-ile.org/agnant/" n="3.1">
    <!-- ana= prenom_nom de l'auteur -->
    <!-- source= nom du site à partir duquel les données ont été extraites -->
    <!-- ref= adresse de la biblio -->
    <!-- n= numéro itératif qui permet de hiérarchiser suivant le nombre de sites cons: -->
    <!-- Chaque item = une référence unique (les romans ont été développés en exemple) -->
    <!-- POUR LES LIST : subtype= titre de la section principale -->
    <!-- POUR LES ITEM : type= titre de la section principale ; subtype= titres des so -->
    <!-- Les références changent d'une entité auteur à l'autre: l'information extraite -->

    <list subtype="romans"> [10 lines]

    <list subtype="nouvelles"> [2 lines]
    <list subtype="poésie"> [2 lines]
    <list subtype="romans pour la jeunesse"> [2 lines]
    <list subtype="contes"> [2 lines]
    <list subtype="textes publiés dans des ouvrages collectifs"> [2 lines]
    <list subtype="distinctions littéraires"> [2 lines]

    <!-- oeuvres traduites -->
    <!-- pour le code de la langue utilisé en XML-TEI voir : http://www.iana.org/assign -->
    <list subtype="catalan" tag="ca"> [2 lines]
    <list subtype="anglais" tag="en"> [2 lines]
    <list subtype="espagnol" tag="es"> [2 lines]
    <list subtype="italien" tag="it"> [2 lines]
    <list subtype="néerlandais" tag="nl"> [2 lines]

    <list subtype="ouvrage collectif"> [2 lines]
    <list subtype="articles sélectionnés"> [2 lines]
    <list subtype="entretiens"> [2 lines]
  </div>

  <div ana="Marie-Celie_Agnant" source="VIAF" n="3.2"> [4 lines]
  </div>
```

La <div> numéro trois possède elle-même une structure en <div> pour chacune des sources dont on a récupéré une bibliographie. On indique, comme précédemment, la source et l'adresse URL avec les attributs *source* et *ref*, mais surtout on hiérarchise ces <div> en affectant un attribut *n* sous-numéroté « 3,1 », « 3,2 », etc. (voir les éléments soulignés dans la figure ci-dessus). À l'intérieur de celles-ci, nous trouvons des balises <list> catégorisées par un attribut *subtype*. Ces dernières contiennent des <item> pour chaque référence bibliographique. Les références changeantes d'une entité à une auteur l'information considérée est ici le type associé à la liste. À ce stade, les catégories littéraires n'ont pas été définies dans le cadre du projet, ainsi nous les avons laissées telles qu'elles.

#### d) 4. Le contenu de la balise <div> numéro quatre

Figure 18 Contenu de la <div> numéro quatre du modèle XML, capture d'écran Oxygen

```
<div xml:id="Marie-Celie_Agnant_web" n="4">
    <!-- xml:id= prenom_nom de entité auteur_type de fichier (ici liens)-->
    <!-- n=4 pour liens -->

    <list ana="Marie-Celie_Agnant" source="ile-en-ile" subtype="liens">
        <!-- ref target= mettre le lien entre les guillements ; entre les bal
            <item><ref target="lien"/>text</item>
        </list>
    <list ana="Marie-Celie_Agnant" source="spla" subtype="liens">
        <item><ref target="lien"/>text</item>
    </list>
</div>
```

La <div> quatre renferme des balises <list> avec, comme toujours, l'indication d'une *source* et d'un *subtype*. Pour chaque lien une balise <item>. Elle donne une description du lien et elle contient une balise <ref> qui renseigne, dans un attribut *target*, l'URL. Comme précédemment, les catégories *subtype* ont été reprises sur *île-en-île.org*, et n'ont pas été redéfinies.

### C. Remplissage des fiches et sorties

Pour remplir les fiches XML, il a fallu construire un script *Python* de *matching* qui allait permettre la réunion de toutes les informations. Puis reproduire le modèle XML en appelant les balises concernées et ainsi remplir les différentes fiches avec les données disponibles.

De ces fiches entités auteurs ainsi créées et remplies, nous avons pu réaliser des sorties fichiers sous format .txt en encodage UTF-8. Pour cela, on utilise des feuilles d'extraction XSL. Afin d'illustrer la démarche nous prendrons en exemple le fichier XSL employé pour réaliser une visualisation réclamée par les commanditaires au cours du projet : un tableau contenant l'entité auteur, le lien, la source du lien et sa description.

Figure 19 Extrait du fichier master, capture d'écran dans le logiciel Oxygen

```
master
1  <?xml version="1.0" encoding="UTF-8"?>
2  <master xml:id="master_auteurs" xml:lang="fr" xmlns:xi="http://www.w3.org/2001/XInclude">
3      <xi:include href="notice_Accilien_Selmy.xml"/>
4      <xi:include href="notice_Alex_Laguerre.xml"/>
5      <xi:include href="notice_Alix_Renaud.xml"/>
6      <xi:include href="notice_Anais_Chavenet.xml"/>
7      <xi:include href="notice_Anderson_Dovilas.xml"/>
8      <xi:include href="notice_Andre_Fouad.xml"/>
9      <xi:include href="notice_Angelucci_Manigat.xml"/>
```

Afin d'appliquer la feuille de transformation XSL à tout le corpus, il faut tout d'abord créer un fichier master. Ce dernier contient tous les *xml:id* des fiches entités auteurs, listés dans des balises <xi : include>, au moyen d'attributs *href*. Ce fichier aura pour fonction d'appeler chacune des fiches entités auteurs répertoriées dans le master.

Figure 20 Contenu du fichier master du tableau des liens du corpus, capture d'écran Oxygen

```

<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:xi="http://www.w3.org/2001/XInclude"
  exclude-result-prefixes="xs"
  version="2.0"
  xmlns:tei="http://www.tei-c.org/ns/1.0">

  <xsl:output method="text"/> <!-- pour faire une sortie en csv -->

  <xsl:template match="/">
    <xsl:text>Nom_Prénom$source$lien$description du lien &#xA;</xsl:text>
    <xsl:for-each select="//div[@n='4']">
      <xsl:for-each select="list">
        <xsl:for-each select="item">
          <xsl:value-of select="concat(parent::list/@ana,'$',parent::list/@source,'$',ref/@target,'$',..,$',&#xA;')"/>
        </xsl:for-each>
        </xsl:for-each>
      </xsl:for-each>
    </xsl:template>

</xsl:stylesheet>

```

On informe dans un premier temps la sortie souhaité par la balise `<xsl:output>` attribut `method`. Ici nous voulons un fichier .txt, puisque l'on veut obtenir un fichier .csv. Il s'agit, en effet, d'un texte avec un délimiteur pour chaque colonne et un retour à la ligne pour chaque nouvelle ligne. Pour ce tableau, le délimiteur choisi a été le symbole « \$ », car il n'est jamais utilisé dans les URLs. Le titre des colonnes est créé dès le début dans une balise `<xsl :text>`. Puis, pour chaque balise `<item>`, contenue dans une balise `<list>`, elle-même contenue dans une balise `<div>` dont l'attribut `n` est égal à 4, on produit une ligne énumérant l'entité auteur, la source, le lien et sa description. On assigne le fichier XSL au master, on enregistre la sortie en .csv qu'on ouvre ensuite dans *Excel* afin d'obtenir une visualisation de notre tableau.

Figure 21 Résultat de la feuille XSL enregistré sous format .csv et ouvert dans Excel

A	B	C	D
1   Nom_Prénom	source	lien	description du lien
2   Accilien_Selmy	lifranum_db	https://www.facebook.com/www.revedemuse.livre/	Selmy Accilien poète
3   Alix_Renaud	ile_en_ile	http://lie-en-ile.org/alix-renaud-poemes/	Ménage à trois
4   Alix_Renaud	ile_en_ile	http://www3.sympatico.ca/alix.renaud/	Alix Renaud, ou le plaisir du texte
5   Alix_Renaud	ile_en_ile	https://youtu.be/j002arXgZ5g	« Le Mot et la chose », poème de Lattaingnant
6   Alix_Renaud	ile_en_ile	https://youtu.be/C1P65X3rhZU	Des genoux au vertige
7   Alix_Renaud	ile_en_ile	http://www3.sympatico.ca/diamantbleu/accueil.html	À la recherche du Petit Prince
8   Alix_Renaud	ile_en_ile	http://www.litterature.org/recherche/erivains/renaud-alix-389/	Présentation sur L'île
9   Alix_Renaud	ile_en_ile	http://www.litterature.org/recherche/erivains/renaud-alix-389/	Présentation sur L'île
10   Alix_Renaud	ile_en_ile	http://www.potomitan.info/ayiti/renaud.php	Présentation sur Potomitan
11   Alix_Renaud	ile_en_ile	http://www.uefs.br/sitientibus/sitientibus/_/entrevida.pdf	Entrevista
12   Alix_Renaud	ile_en_ile	http://newsgroups.derkeller.com/Archive/Soc/soc.culture.haiti/2005-11/msg00159.html	Le don de jouer avec les mots
13   Alix_Renaud	ile_en_ile	http://www.mondokarnaval.com/alix_chronique_2016.html	Les Chroniques d'Alix Renaud
14   Alix_Renaud	ile_en_ile	http://www3.sympatico.ca/alix.renaud/grac_esp.html	Ofrenda de Gracias
15   Alix_Renaud	ile_en_ile	http://fr.wiktionary.org/wiki/pompion	Pompion
16   Alix_Renaud	ile_en_ile	https://www.thecaribbeancurrent.com/quebec-awards-alix-renaud-for-excellence-in-arts-and-cult	Quebec Awards Alix Renaud for Excellence in Arts and Culture
17   Alix_Renaud	ile_en_ile	http://www.voir.ca/	Voir
18   Alix_Renaud	ile_en_ile	https://voir.ca/livres/2000/02/10/alix-renaud-ovation/	Ovation
19   Anderson_Dovilas	ile_en_ile	http://lie-en-ile.org/anderson-dovilas-pwezi-pefomas/	Anderson Dovilas, Pwezi pèfòmans
20   Anderson_Dovilas	ile_en_ile	http://lie-en-ile.org/anderson-dovilas-la-nuit-nest-pas-le-premier-secret-des-baisers-bleus/	La nuit n'est pas le premier secret des baisers bleus
21   Anderson_Dovilas	ile_en_ile	http://lie-en-ile.org/anderson-dovilas-trois-poemes/	Trois poèmes
22   Anderson_Dovilas	ile_en_ile	http://dovilerson.over-blog.com/	Le blog de Dovilas Anderson
23   Anderson_Dovilas	ile_en_ile	http://www.harmattan.fr/index.asp?navig=catalogue&obj=livre&no=43605	Mémoire d'Outre-Monde
24   Anderson_Dovilas	ile_en_ile	http://www.franccopolis.net/revues/Dovilas-mai2012.html	Notes sur « Vingt poèmes pour traverser la nuit »
25   Anderson_Dovilas	ile_en_ile	http://spcrelophone.blogvие.com/	Sosyete Powet Kreyolofón
26   Anderson_Dovilas	ile_en_ile	http://www.lemanoirdespoetes.fr/poemes-anderson-dovilas.php	« Je vole », « Je voyage » et « Point »
27   Anderson_Dovilas	ile_en_ile	http://www.paroleenarchipel.org/	Parole en Archipel
28   Anderson_Dovilas	ile_en_ile	http://www.paroleenarchipel.com/article-vingt-poemes-pour-traverser-la-nuit-104458858.html	Vingt poèmes pour traverser la nuit
29   Anderson_Dovilas	ile_en_ile	http://www.paroleenarchipel.com/article-en-hollande-la-france-devient-fran-oise-104871758.htm	En Hollande, la France devient Françoise
30   Anderson_Dovilas	ile_en_ile	http://www.paroleenarchipel.com/article-les-iles-en-accent-aigu-105328874.html	Les îles en accent aigu
31   Anderson_Dovilas	ile_en_ile	http://www.paroleenarchipel.com/article-le-sang-de-l-oublie-le-sang-de-l-avenir-105976958.html	Le sang de l'oubli, le sang de l'avenir
32   Anderson_Dovilas	ile_en_ile	http://www.paroleenarchipel.com/article-liminasyon-parcours-d-un-voyageur-106118692.html	Liminasyon: Parcours d'un voyageur
33   Anderson_Dovilas	ile_en_ile	http://issuu.com/pallakaroly/docs/vents_aliz_s_-komansman?mode>window&backgroundColor:	Trois poèmes
34   Anderson_Dovilas	ile_en_ile	http://www.potomitan.info/ayiti/dovilas.php	Pour en finir avec dérogation littéraire en Haïti

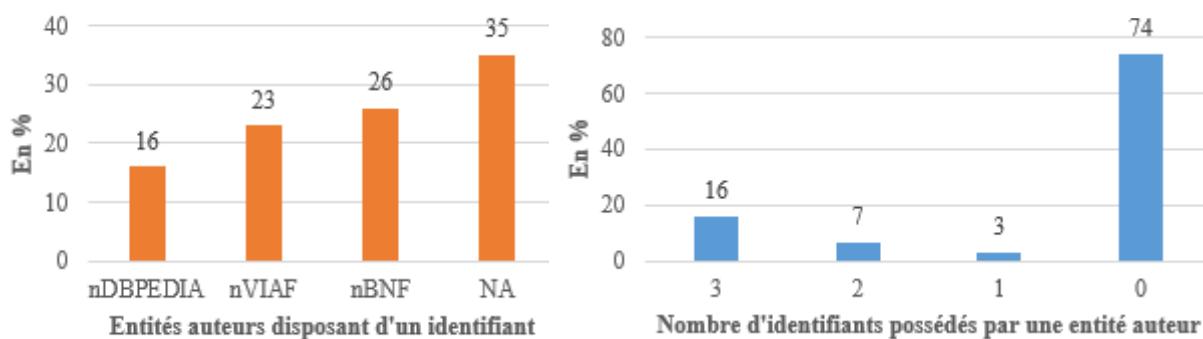
## IV. Les résultats

Nous avons ainsi réalisé une multitude de tableaux en vue de statistiques descriptives, d'analyses textuelles, et de visualisations en réseaux.

### A. Éléments de statistiques descriptives

Grâce au balisage, il a été possible d'observer les possibilités et les limites de l'extraction. D'un point de vue concret, concernant les informations extraites à partir des répertoires d'autorités, nous pouvons établir un premier constat. Si la méthodologie de filtrage et de collecte de données initiée par le projet s'était concentrée exclusivement sur les entrées de la BNF, elle demeure insuffisante.

*Figure 22 Identifiants possédés par les entités auteurs*



Sur 220, seules 7 entités auteurs ne possèdent qu'un identifiant à savoir celui de la BNF. 15 possèdent deux identifiants à savoir ceux de la BNF et du VIAF. 35 possèdent à la fois des identifiants au sein des répertoires d'autorité, mais aussi une entrée au sein de la base de données Wikipédia. 163 demeurent ne pas être directement renseignés soit 74 % du corpus. On peut ici supposer que ce sont précisément ces auteurs, encore méconnus que cherchent à mettre au jour les projets CartoWeb et Lifranum. Les auteurs « traditionnels » francophones sont en effet rendus visibles par leur notice BNF. Les auteurs bénéficiant une certaine visibilité européenne, voire internationale, possèdent également un identifiant VIAF. Enfin, les auteurs prolifiques une entrée au sein de la base de données de Wikipédia. Il s'agit donc de combiner ces diverses sources de données si l'on souhaite d'identifier de nouveaux auteurs potentiels. L'ISNI<sup>48</sup> pourrait ainsi constituer un point d'entrée supplémentaire. Néanmoins, tout comme nous l'avait très justement fait remarquer Pierre Halen à l'occasion de notre communication,<sup>49</sup> si des entités auteurs demeurent après cela sous les radars, c'est qu'ils devraient peut-être le rester.

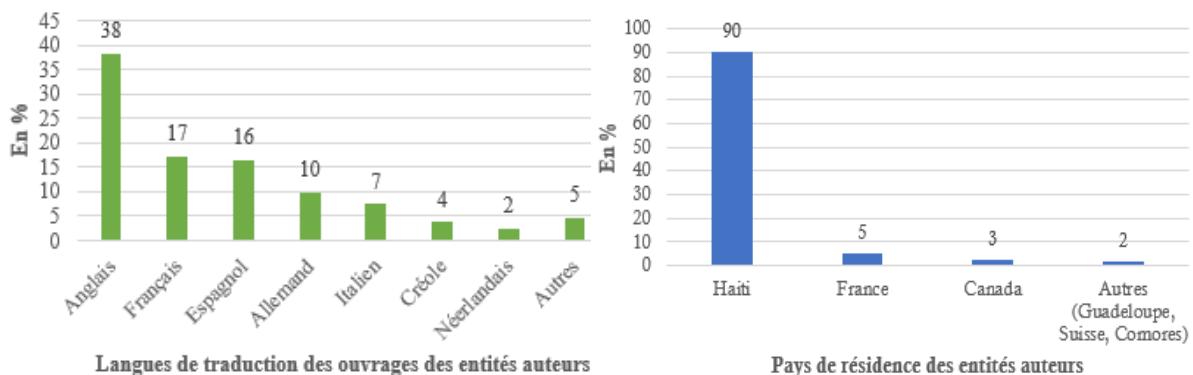
Les notices existantes ne sont d'ailleurs pas nécessairement complètes. Prenons l'exemple du sexe qui est renseigné pour seulement 67 notices soit 30,7 % des entités auteurs.

<sup>48</sup> International Standard Name Identifier, code international normalisé des noms. Il renvoie un identifiant unique pour des entités ayant contribué à des médias comme les livres.

<sup>49</sup> BURRI Inès, CHAMBAT Anaïs, RINGWALD Célian, « Ébauche de visualisation des production littéraires francophones d'Haïti et retour d'expérience », Colloque international Cartographie du Web littéraire francophone, Lyon, mercredi 22 Janvier 2020.

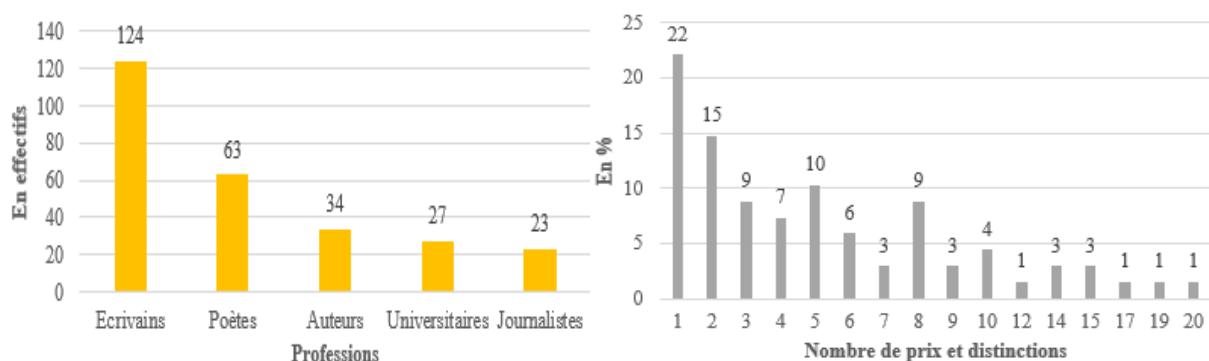
Concernant les données liées à la langue et à la profession, le site de référence d'auteurs *spla* a été d'une aide précieuse. 75 % des données extraites automatiquement proviennent de cette source, le quart restant de la BNF. Concernant la langue maternelle, 98 % des notices mentionnent le français et seulement 2 % le créole. Les pays de résidence des entités auteurs extraites à partir du *spla* sont francophones.

*Figure 23 Langues de traduction (à gauche) et pays de résidence (à droite) des entités auteurs*



Ainsi, Haïti est représenté à hauteur de 90 %, la France 5 % et le Canada 3 %. On remarque ainsi une faible mobilité des auteurs originaires d'Haïti. Ceci est à mettre en regard de la langue de traduction des ouvrages des entités auteurs présents sur *île-en-île*. On observe ici que si l'anglais domine à 38 %, le français et l'espagnol ne sont pas loin derrière avec 17 et 16 %. 4 % des ouvrages ont même été traduits en créole.

*Figure 24 Top 5 des professions exercées (à gauche) et nombre de distinctions et prix attribués aux entités auteurs du corpus (à droite)*



Par le biais de la traduction, la littérature francophone semble donc ainsi gagner en mobilité. Nous avons également pu extraire les professions et dresser un top 5 en effectifs des professions extraites depuis le *spla*. La première place revient ainsi aux écrivains, la deuxième aux poètes, la troisième aux auteurs, la quatrième aux universitaires et la cinquième aux journalistes. On remarque ici le paradigme sémantique entre l'auteur et l'écrivain. Pourquoi revendiquer l'un et pas l'autre ? Un individu peut par ailleurs avoir exercé plusieurs activités ou professions au cours de sa vie. L'extraction ne nous permet pas d'adopter une approche diachronique et successive des entités auteurs et ne peut qu'en rendre compte quantitativement.

Une entrée conceptuelle pourrait ici se situer autour de la légitimité des auteurs : les auteurs prolifiques et légitimés par le milieu académique se sont en effet souvent vus attribué une ou plusieurs distinctions (prix, bourse ou encore invitation en résidence). À partir de quel seuil de distinctions peut-on considérer qu'un auteur est légitime ?

## B. Exploration textuelle des notices biographiques

Traditionnellement utilisée en sociologie ou en linguistique comme préalable à une recherche sur un corpus, cette technique participe à la mise à jour d'une structure du langage. L'analyse textuelle vise à permettre une compréhension statistique des mots dans un corpus. Elle peut être définie comme un « ensemble de méthodes permettant d'opérer des réorganisations formelles des textes et des analyses statistiques portant sur le vocabulaire d'un corpus. [...] Il ne s'agit non pas de chercher le sens d'un texte, mais de déterminer comment sont organisés les éléments qui le constituent. »<sup>50</sup> Cette définition précise un peu davantage ce que l'on pourrait qualifier de lexicométrie « à la française ». L'analyse textuelle est en effet héritée des pratiques factorielles avec lesquelles il est possible d'étudier la structure de répartition des textes et de projeter par-dessus des variables catégorielles.

Si cette technique est envisagée ici comme préalable à l'exploration puis à l'analyse, pour que les résultats statistiques aient un sens et soient recevables, il est essentiel que le corpus et ses éventuels sous-corpus soient contextualisés et ses règles de composition rendues visibles.<sup>51</sup> Ces précautions ont notamment été mises en évidence par Antoine Prost. Il dégage trois règles essentielles selon lesquelles le corpus doit être « homogène, diachronique et contrasté ».<sup>52</sup> Les textes doivent donc être environ de même longueur, concerner le même public et porter sur le même thème. Ils ne peuvent avoir été écrits le même jour et doivent recéler des différences internes. Dans ses travaux de linguistique, Bénédicte Pincemin précise cette typologie en admettant des conditions de signification, d'acceptabilité et d'exploitabilité.<sup>53</sup>

### 1. Présentation des logiciels utilisés et importation des données

Afin d'effectuer l'exploration des notices biographiques, nous mobiliserons deux logiciels TXM<sup>54</sup> et IRaMuteQ<sup>55</sup>.

---

<sup>50</sup> LEBART Ludovic, SALEM André, *Analyse statistique des données textuelles*, Paris : Dunod, 1988 (1994), p. 183.

<sup>51</sup> « La notion de construction de corpus est centrale : tous les résultats obtenus sont relatifs à sa définition, donc dépendent d'un choix éclairé du chercheur. Une recherche philologique s'impose avant le traitement du corpus, afin de choisir entre plusieurs versions d'un même texte ou d'identifier précisément le(s) auteur(s). » in LEMERCIER Claire, ZALC Claire, *Méthodes quantitatives pour l'historien*, Paris : La Découverte, 2008, p. 51.

<sup>52</sup> In PROST Antoine, *Vocabulaire des proclamations électorales de 1881, 1885 et 1889*, Paris : Presses Universitaires de France, 1974, p. 14.

<sup>53</sup> BOMMIER-PINCEMIN Bénédicte, *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne, 6 avril 1999, chapitre VII, A « Définir un corpus », pp. 415-427.

<sup>54</sup> HEIDEN Serge, MAGUE Jean-Philippe, PINCEMIN Bénédicte, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement » in BOLASCO Sergio (dir.), *Proceedings of 10<sup>th</sup> International Conference on the Statistical Analysis of Textual Data*, JADT 2010, Rome, vol. 2, pp. 1021-1032.

<sup>55</sup> <http://www.iramuteq.org/>

### a) TXM

La plateforme open-source TXM a été développée dans le cadre du projet *Textométrie* porté par l'ENS de Lyon. Cet outil s'est rapidement imposé comme une référence dans la construction et l'analyse des corpus annotés et structurés. Il permet notamment d'adopter une approche philologique et sémantique des textes étudiés. Il mobilise les propriétés de *TreeTagger* un outil avancé d'étiquetage morphosyntaxique et de lemmatisation du langage.<sup>56</sup> Les détections des « catégories grammaticales » (POS) permettent d'effectuer d'explorer rapidement des pistes conceptuelles et linguistiques avancées.

« Il sert [plus spécifiquement] à calculer le vocabulaire d'ensemble d'un corpus ou la liste des valeurs d'une propriété particulière ; de construire des tables lexicales ; de rechercher des motifs lexicaux complexes construits à partir des propriétés des unités lexicales et produit des concordances [en contexte] à partir des résultats. [Mais aussi], à calculer le modèle des spécificités de mots ou d'étiquettes situés à l'intérieur d'un sous-corpus ainsi que l'analyse factorielle des correspondances de propriétés des mots sur une partition et peut renvoyer la classification associée. »<sup>57</sup>

Concernant l'importation d'un corpus XML-TEI, elle est relativement aisée. Il suffit en effet d'utiliser l'option « import XML/w + CSV ». Le balisage est conservé et permet une recherche par unités de structures. Si l'on souhaite ajouter des métadonnées supplémentaires, il convient alors de joindre un fichier.csv intitulé « metadata » qui contiendrait les noms des fichiers (Id, identifiants) et les informations souhaitées.

### b) IRaMuTeQ

IRaMuTeQ, Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires, logiciel développé par Pierre Ratinaud au sein du LERASS propose quant à lui différents types d'analyses basées sur les méthodes statistiques (calculs de spécificités, analyses factorielles et Méthode Reinert), la visualisation de données textuelles (nuage de mots) et l'analyse de réseaux de mots (analyses de similitudes).

Il s'agit pour le traitement dans IRaMuTeQ de regrouper l'ensemble des textes d'un corpus dans un fichier texte (.txt) au format UTF-8.<sup>58</sup> Les textes sont séparés par une ligne « étoilée » comportant les variables associées à chaque texte. Les variables qualitatives que nous avons retenues pour nos analyses sont l'identifiant de l'entité auteur et la source de la notice. Dans R, on commence par définir le répertoire de travail et charger le fichier qui correspond au jeu de données contenant les variables que l'on souhaite analyser. Elles doivent avoir au préalable été recodées pour répondre aux objectifs de la recherche textuelle. Les champs vides ont donc été retirés. Enfin, le nom des variables ou des modalités ne doit pas comporter d'accents. La première variable est introduite par une suite de quatre étoiles et les suivantes d'une seule.

---

<sup>56</sup> <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

<sup>57</sup> In PINCEMIN Bénédicte, HEIDEN Serge, « Qu'est-ce que la textométrie ? Présentation », Site du projet TXM, <http://textometrie.ens-lyon.fr/>.

<sup>58</sup> L'encodage reconnu par le logiciel lors de l'importation peut différer. Pour notre cas, nous avons dû utiliser l'encodage par défaut « cp1252 - Western Europe ».

Concernant le texte, il est nécessaire de sauter une ligne entre les variables, les unités d'analyses et après. Le corpus est dès lors au bon format. On termine donc par l'enregistrer en .txt.

Figure 25 Script R - Import IRaMuTeQ<sup>59</sup>

```
#définition du répertoire et chargement des librairies
setwd("D:/Bureau/ADT")
library(readxl)
#importation des données
bio <- read_excel("bio_contenus.xlsx")
names(bio)

#on ne conserve que les colonnes qui sont utiles pour l'analyse
iram_bio <- bio[c(1,2,3)]
head(iram_bio, 3)

#formatage
iram_bio$nom_prenom <- paste("**** *Bio_", iram_bio$nom_prenom, sep = "")
iram_bio$source_bio <- paste("*source_", iram_bio$source_bio, sep = "")
iram_bio$bio<- paste("\n",iram_bio$bio, "\n", sep=" ")
write.table(iram_bio, file="iram_bio.txt", sep=" ", quote=FALSE,
            row.names=FALSE, col.names=FALSE)
#enregistre au format.txt
```

Ce qui donne ceci :

```
**** *Bio_Alix_Renaud *source_ile_en_ile
Alix Renaud naît à Port-au-Prince le 30 août 1945. Il vit au Québec depuis 1968.
```

## 2. Description du corpus textuel

### a) Dimensions

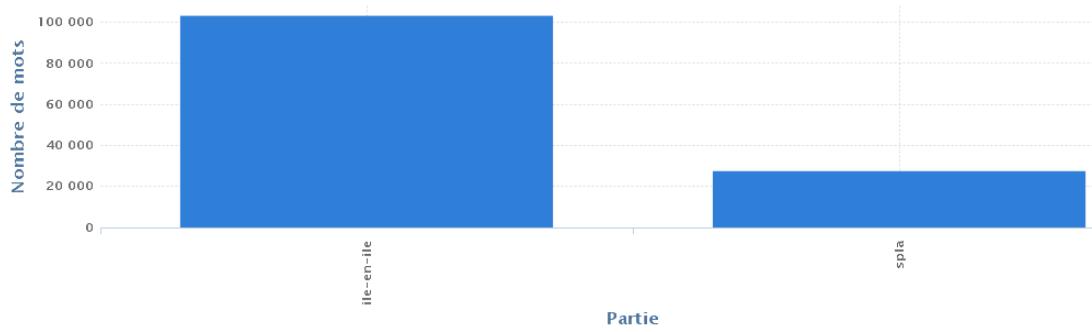
L'outil d'analyse textométrique TXM permet notamment de renvoyer la structure générale du corpus et notamment le nombre de balises complétées. Les notices biographiques générées comprennent un total de 274 850 mots répartis en quatre unités lexicales à savoir les lemmes, les catégories grammaticales, les caractères et les mots<sup>60</sup> ainsi qu'en quatre unités de structure qui reprennent l'annotation XML-TEI vue précédemment : la balises « text » qui renvoie à la totalité de la notice biographique, les balises « p » aux paragraphes des notices biographiques, les balises « persName » et « placeName » à l'ensemble des noms de personnes et des localisations rencontrées.

On dénombre ainsi 272 liens, 898 références à des personnes uniques et 469 noms de lieux différents. Sur 220 entités auteurs, 127 possèdent au moins une notice biographique à l'issue du croisement des deux sites de références et 74 deux notices. Ainsi, seulement 19 entités auteurs soit 9 % du corpus demeurent non renseignées.

<sup>59</sup> D'après le tutoriel de ROQUEBERT Corentin, « Constituer un corpus Europresse utilisable dans R, IRaMuTeQ et TXM », quatrième étape : transformer le tableau aux formats IRaMuTeQ et TXM, in RPubs, 4 décembre 2018.

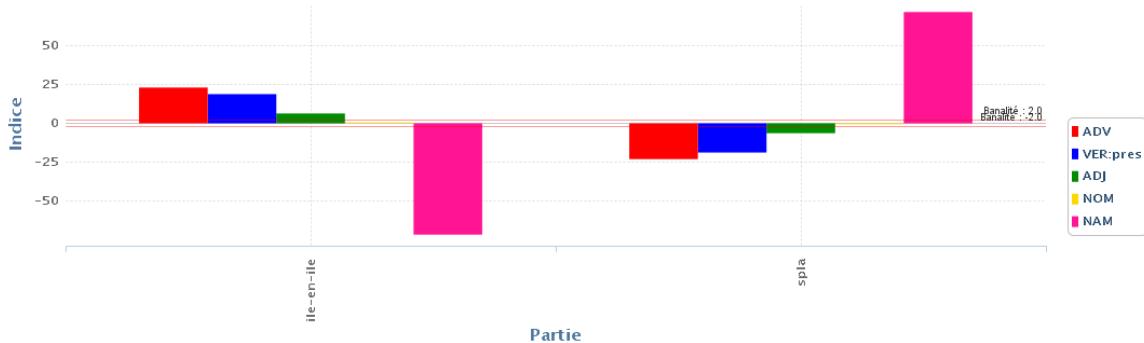
<sup>60</sup> Les unités lexicales sont abrégées dans TXM de la manière suivante : *frlemma*, *frpos*, *n* et *word*.

Figure 26 Dimensions des deux partitions « *ile-en-ile* » et « *spla* »



Nous avons partitionné le corpus en deux sous-parties : « *ile-en-ile* » et « *spla* » suivant la source des notices. La partition *ile-en-ile* est ainsi composée de 137 notices et celle du *spla* de 149 notices. Nous avons tout d’abord souhaité observer l’homogénéité en termes de nombre de mots de ces partitions. La première partition contient 103 225 mots soit un nombre moyen de 753 mots par notice et la deuxième 27 473 mots soit 184 mots en moyenne par notice. Les notices d’*ile-en-ile* sont ainsi en moyenne 4 fois plus longues que celles du *spla*. Si l’on observe à présent les catégories grammaticales signifiantes<sup>61</sup> des deux partitions, on observe que l’usage des noms communs est banal. Les noms propres sont en revanche surreprésentés au sein des notices *spla* tandis que les verbes conjugués au présent, les adjectifs et les adverbes sont sous-représentés. Une telle représentation est cohérente avec la présence de textes courts.

Figure 27 Analyse de spécificités de l’usage des principales catégories grammaticales



Contrairement à TXM, lors de l’indexation du corpus, IRaMuTeQ transforme les unités textuelles en minuscules (tokénisation) et segmente le corpus tous les 40 mots suivant la ponctuation.<sup>62</sup> Il effectue ensuite la lemmatisation du corpus. « Les occurrences sont réduites à leur racine ; les verbes sont ramenés à l’infinitif, les noms au singulier et les adjectifs au masculin singulier. »<sup>63</sup> Des clés permettent de spécifier quelles formes seront analysées.

<sup>61</sup> ADV : adverbes ; VER : verbes au présent ; ADJ : adjectifs ; NOM : noms communs ; NAM : noms propres.

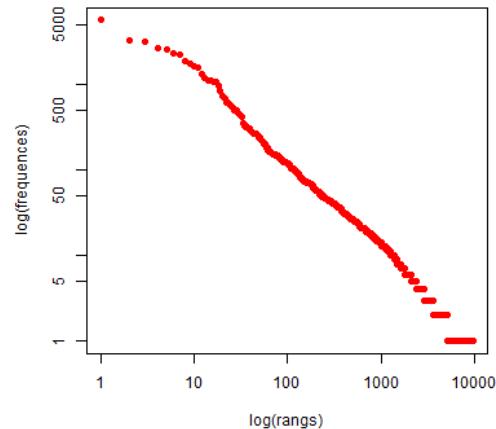
<sup>62</sup> « Les segments de texte sont construits à partir d’un critère de taille et de ponctuation. IRaMuTeQ cherche le meilleur ratio taille/ponctuation (par ordre de priorité, les “.” “?” “!” en premier, puis en second “;” et les “:” en troisième la virgule et en dernier l’espace). L’objectif est d’avoir des segments de tailles homogènes en respectant le plus possible la structure du langage. » in PELISSIER Daniel, « Initiation à la lexicométrie. Approche pédagogique à partir de l’étude d’un corpus avec le logiciel IRaMuTeQ », mars 2017, <https://presnum.org.hypotheses.org/>, p. 8.

<sup>63</sup> *Ibid.*, pp. 9-10.

Par défaut, ce qui est mis en actif (codé 1) sont les adjectifs, les adverbes, les formes non reconnues, les noms communs et les verbes ; ce qui est mis en supplémentaire (codé 2) sont les mots-outils.<sup>64</sup> Tout comme dans TXM, il est possible de générer des sous-corpus à partir des variables catégorielles associées. Aussi, les résultats obtenus à l'échelle du corpus entier et des partitions peuvent légèrement varier.

*Figure 28 Table récapitulative du corpus et de ses*

	<b>Corpus entier</b>	<b>ile-en-ile</b>	<b>Spla</b>
Unités textuelles	1 242	1 100	142
Segments de textes	3 245	2 641	604
Occurrences	108 835	88 334	20 501
Formes différentes	9 648	8 661	3 673
Hapax	4 552	4 183	2 049
% Hapax occurrences	4	5	10
% Hapax formes	47	48	56
Moy. d'occurrences par textes	88	80	144



*partitions*

*Figure 29 Loi de Zipf (corpus entier)*

On observe notamment que le corpus entier est constitué de 1 242 unités d'analyses textuelles, de 3 245 segments de textes et de 108 835 occurrences réparties en 9 648 formes différentes. Il contient 4 552 hapax, occurrences dont la fréquence est de 1, soit 4 % des occurrences et 47 % des formes. La moyenne d'occurrences par textes est de 88.

Le graphique ci-dessus présente en abscisses les logarithmes des rangs et en ordonnées les logarithmes des fréquences des formes. Les mots les plus fréquents sont ceux qui ont le rang le plus élevé. La courbe illustre la « loi » de Zipf qui stipule que le produit rang\*fréquence est une constante.<sup>65</sup> La concentration d'hapax est significative : derrière le cadre en apparence normalisé et structuré des notices biographiques, se dissimule un vocabulaire riche et varié. L'effet « pallier » observé est dû à la prépondérance des mots-outils dans le corpus. Le mot qui occupe le rang 1 du corpus, mais aussi des partitions est « de » avec 5 654 occurrences. C'est pourquoi la suite de nos analyses portera essentiellement sur les formes actives. Nous avons effectué ensuite une requête sur l'ensemble du corpus afin d'identifier les mots signifiants.

### b) Représentation des syntagmes les plus fréquents

*Figure 30 Extrait de la table lexicale du corpus (top 5)*

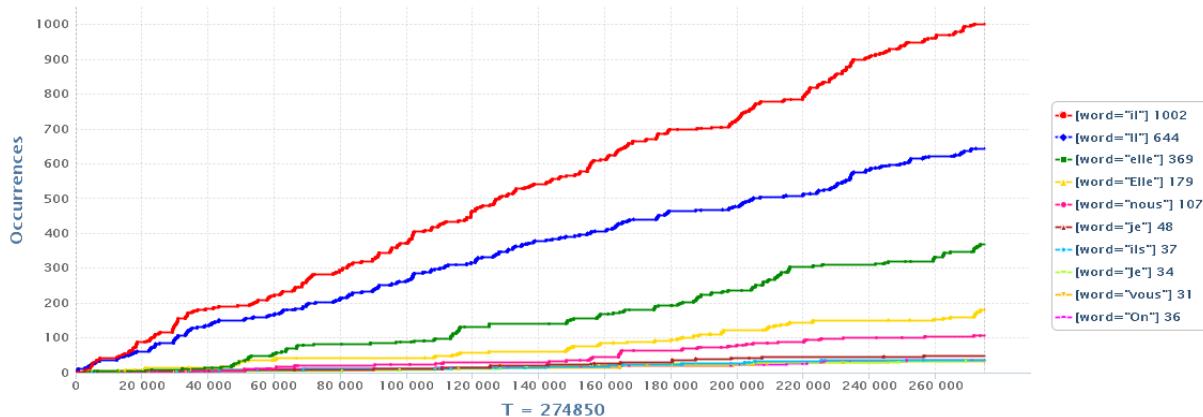
word_frpos	Fréquence T=130698	ile-en-ile t=103225	spla t=27473
est_VER:pres	1293	1071	222
Haïti_NAM	629	472	157
a_VER:pres	626	423	203
Prince_NAM	305	227	78
Port-au_-NAM	285	214	71

<sup>64</sup> Concernant le paramétrage de la lemmatisation, voir BARIL Elodie, GARNIER Bénédicte, « Utilisation d'un outil de statistiques textuelles. IRaMuTeQ 0,7 alpha 2 », avril 2015, <http://www.iramuteq.org/>, pp. 8-10.

<sup>65</sup> LEBART Ludovic, SALEM André, *op.cit.*, p. 34.

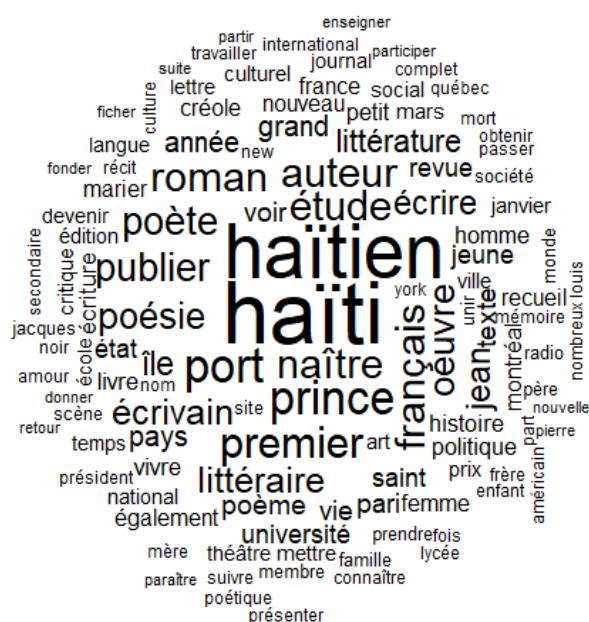
Tous les syntagmes composés de mots, adjectifs, adverbes et verbes du corpus sont au nombre de 61 419, dont 15 609 de formes différentes. Le mot observant la fréquence maximale de 1 293 est la forme conjuguée au présent du verbe « être » à la 3<sup>e</sup> personne du singulier, ceci est propre à la nature des notices. Une entité auteur se décrit en effet rarement elle-même. De même, la forme conjuguée au présent du verbe « avoir » à la 3<sup>e</sup> personne du singulier apparaît en troisième position avec une fréquence de 626, viennent ensuite des références au pays étudié Haïti ou encore Port-au-Prince, sa capitale.

*Figure 31 Progression de l'usage des pronoms au sein du corpus*



Comme peut le constater sur le graphique ci-dessus, l'emploi du pronom personnel masculin « il/Il » prédomine avec 1 646 des occurrences du corpus. Le pronom « elle/Elle » a quant à lui été utilisé à 548 reprises. Si l'usage des pronoms « nous » et « vous » marquent la collectivité et servent à prendre le lecteur à parti, l'usage du « on » est impersonnel. Enfin, le « je/Je » a été utilisé à seulement 82 reprises sur l'ensemble du corpus.

*Figure 32 Nuage de mots, fréquence >= 50 soit 150 formes actives*



A présent, nous choisissons de représenter les formes actives sous la forme d'un nuage de mots. La taille des formes/mots est proportionnelle à leur fréquence. Les mots les plus cités, sont placés au centre du nuage. Le nombre de formes actives représentées a été abaissé à 150, soit toutes les occurrences ayant une fréquence supérieure ou égale à 50 afin d'en faciliter la lecture.

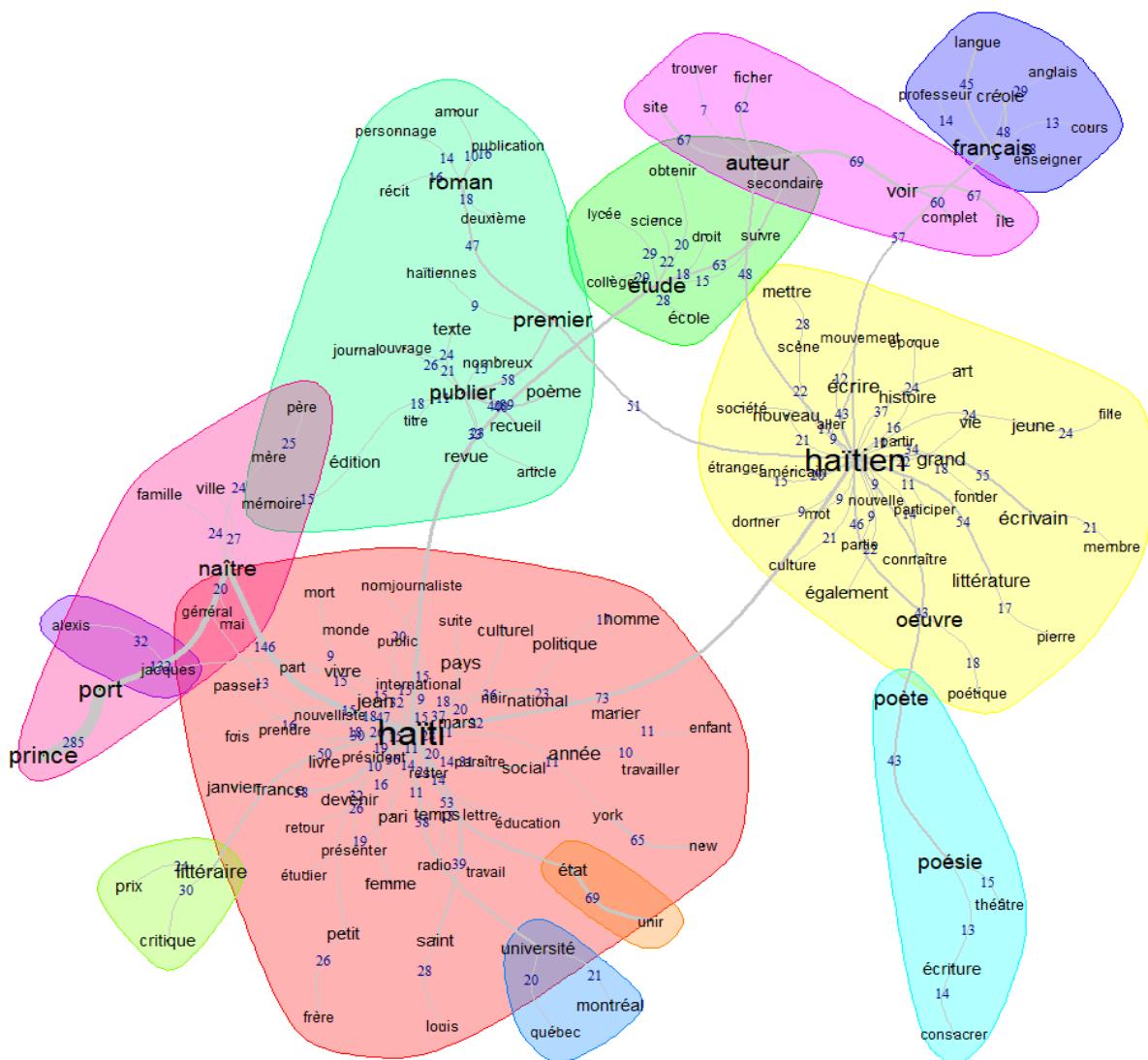
Ces différents outils ne permettent pas de montrer les liens entre les différentes formes contrairement à l'analyse de similitudes.

### 3. Premiers éléments

#### a) ADS

Il s'agit en effet d'une analyse de cooccurrences présentée sous forme de graphiques de mots associés. La cooccurrence renvoie généralement à « l'apparition simultanée de deux ou plusieurs éléments ou classes d'éléments dans le même discours. »<sup>66</sup> La cooccurrence est déterminée au niveau des segments de textes. Ainsi, l'indice correspond au comptage du nombre de segments dans lesquels une forme est associée à une autre. Cet indice est visible sur les arêtes et les modalités textuelles sur les sommets. Plus la taille des mots est grande, plus ils sont fréquents dans le corpus ; plus les liens/arêtes sont épais plus les mots sont cooccurrents.<sup>67</sup>

*Figure 33 Analyse de similitude du corpus entier, fréquence >= 50 soit 150 formes actives, arbre maximum, indice de cooccurrence, algorithme graphopt,<sup>68</sup> halos communautaires*



<sup>66</sup> <http://www.cnrtl.fr/definition/cooccurrence>

<sup>67</sup> BARIL Élodie, GARNIER Bénédicte, *op.cit.*, p. 24.

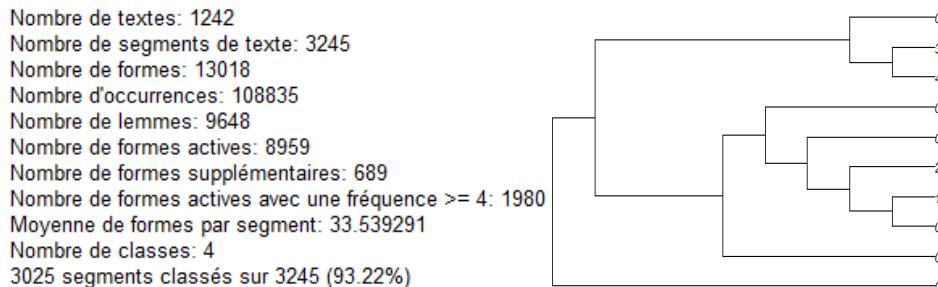
<sup>68</sup> Un algorithme de visualisation peut d'ailleurs être utilisé afin d'optimiser l'affichage du graphe et visualiser les mots les plus « centraux ».

Sans surprise, la principale communauté de mots est structurée autour du pays « Haïti » et de sa nationalité « haïtien ». On constate la présence de plusieurs îlots de cooccurrences. Ainsi, « Haïti » est notamment associé à 146 reprises au verbe « naître » lequel est relié à « Port-au-Prince », sa capitale ou encore à 50 reprises à l'îlot dédié aux « prix, critiques et distinctions littéraires ». Cette représentation exploratoire permet également de rendre compte des spécificités internes du corpus. La nationalité est ainsi spécifiquement liée aux parcours de vie et aux langues parlées : le français ainsi lié au créole lui-même lié à l'anglais.

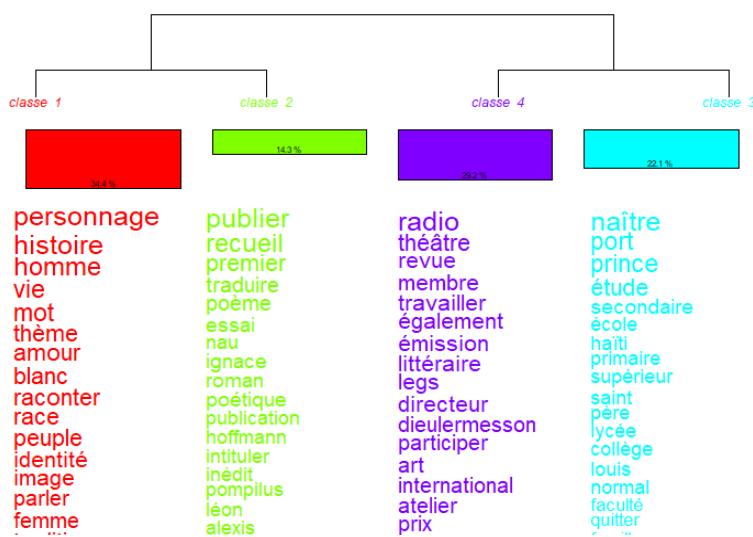
### b) Classification de Max Reinert ou Méthode Alceste

Enfin, la classification de Max Reinert cherche à rendre compte de l'organisation interne d'un discours par la mise en évidence des « mondes lexicaux » investis par le locuteur.<sup>69</sup> Plus précisément, elle vise à classer l'ensemble des formes d'après leur indépendance mesurée par un test du Chi<sup>2</sup>. Nous avons privilégié une « classification simple sur segments » pour visualiser le corpus entier. Le nombre minimal de segments de textes par classe que nous retenons est 0 et le nombre de classes terminales de la phase 1 est 10.<sup>70</sup> Il est ensuite possible de représenter les résultats dans des dendrogrammes de formes et de les projeter sur une AFC.

*Figure 34 Sorties logiciel de la classification hiérarchique descendante du corpus entier*



*Figure 35 Dendrogramme de la CHD et de formes*



<sup>69</sup> REINERT Max, « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », in *Les cahiers de l'analyse de données*, 8/2, 1983, pp. 187-198.

<sup>70</sup> À noter que ces paramètres d'analyses fonctionnent avec les deux scripts proposés « svdR » et « irlba ».

La Méthode Alceste a mis en évidence quatre classes, 93,22 % des segments ont pu être classés, ce qui démontre une excellente qualité de l'analyse. Deux sous-ensembles de classes se distinguent : les classes 1 et 2 d'une part et les classes 3 et 4 d'autre part. Les pourcentages indiqués par le dendrogramme représentent la quantité d'information résumée pour chaque classe soit 34,35 % pour la première classe (1 039 segments de textes sur 3 025), 14,35 % pour la deuxième (434/3025), 22,08 % pour la troisième (668/3025) et 29,22 % pour la quatrième (884/3025). Le dendrogramme de formes présenté plus haut fournit quant à lui la liste des formes les plus associées pour chaque classe : la classe rouge, les contenus textuels des publications, la verte, les événements liés à celles-ci et aux genres littéraires, la violette, les événements professionnels et la bleue, les éléments liés au parcours de vie.

Nous choisissons d'effectuer une AFC d'après la CHD pour interpréter les résultats. Deux axes ont été retenus par le logiciel, le premier représente 45 % de l'inertie et le deuxième 33 %. Rappelons qu'outre les formes de mots placées en variables actives et les mots-outils en variables supplémentaires, l'identifiant des entités auteurs et l'origine des notices ont été indiqués comme variables illustratives.

*Figure 36 AFC sur variables actives, 30 points par classes*

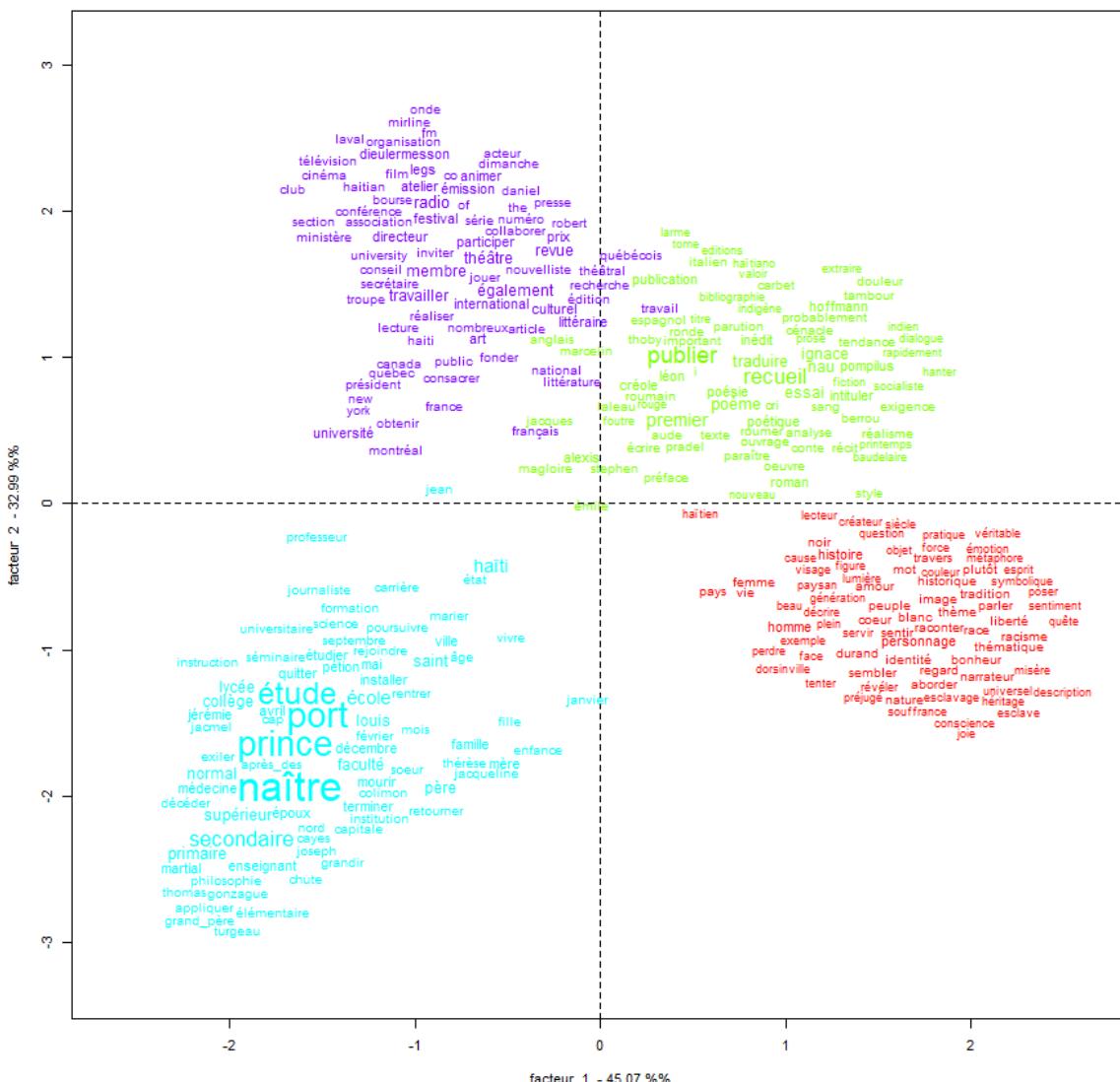
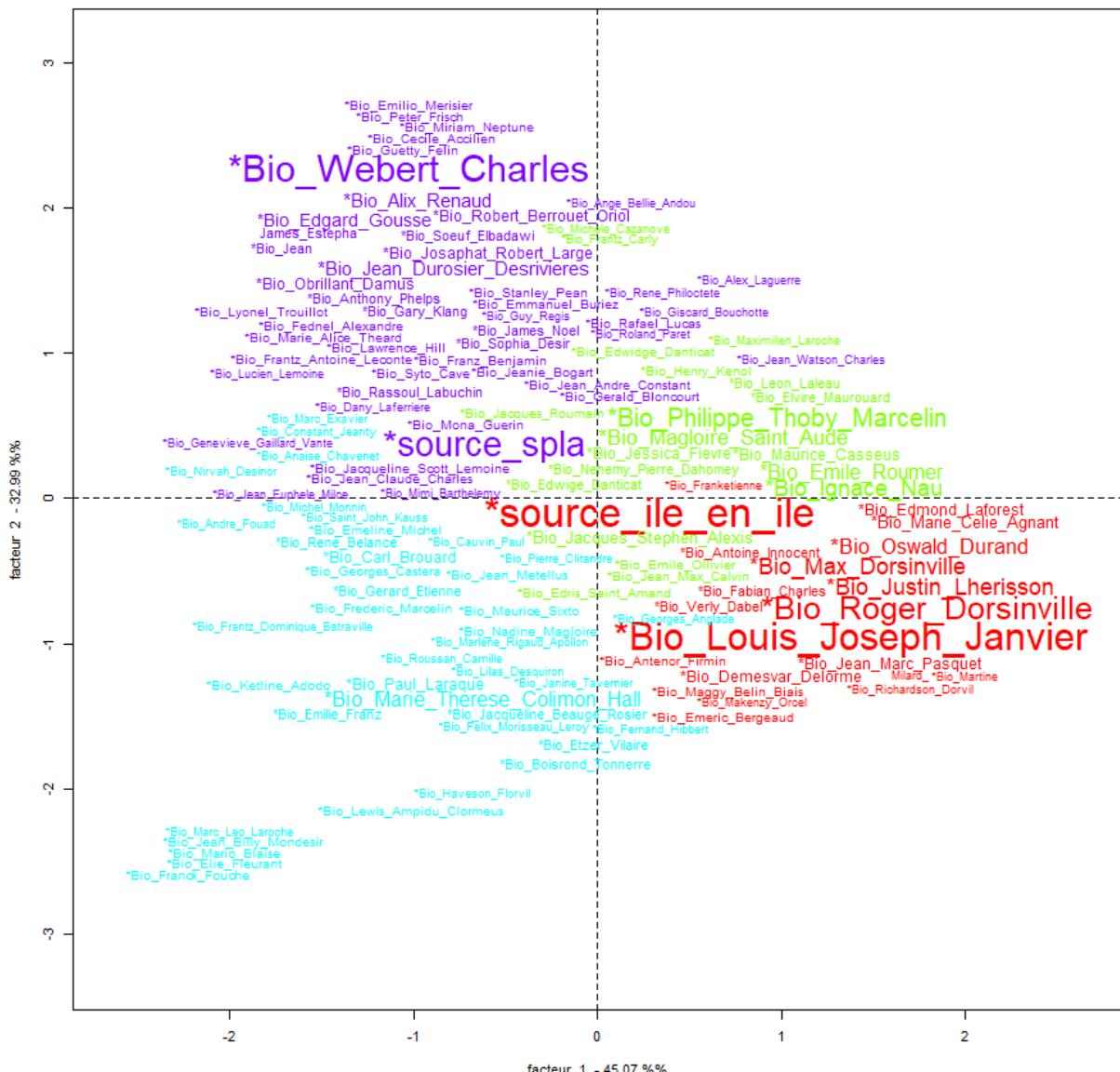


Figure 37 Variables illustratives de l'AFC, 30 points par classes



L'AFC identifie quatre profils qui s'opposent entre eux. Les analyses précédentes concordent avec ces observations : le profil violet est caractéristique des notices *spla*. Il regroupe principalement les événements professionnels. Il semble s'opposer au profil rouge, caractéristique des notices *ile-en-ile*, dont la part accordée aux contenus textuels des publications prédomine. Les profils bleus et verts se mélangent au sein des notices. On remarque cependant que pour chaque profil il est possible d'identifier des notices jugées représentatives.<sup>71</sup> Sans contextualisation de la part des experts du champ littéraire, ces outils et premiers constats ne peuvent être envisagés que de manière exploratoire pour signifier une tendance. Il convient en effet de garder à l'esprit que l'apparente masse de données ne suffit *de facto* pas à atteindre une objectivité automatique.

<sup>71</sup> Profil 1 (violet) : Spla, notice de Charles Webert.

Profil 2 (rouge) : Ile-en-ile, notices de Louis Joseph Janvier, Max et Roger Dorsinville, Justin Lherisson.

Profil 3 (vert) : notices de Philippe Thobé Marcellin, Ignace Nau, Emile Roumier et Magloire Saint-Aude.

Profil 4 (bleu) : notice de Marie-Thérèse Colimon-Hali

## C. Pistes de visualisations en réseaux

Partant du noyau originel de 220 auteurs haïtiens, nous avons pu constater qu'il était possible d'identifier près de 1000 autres auteurs potentiels. Ce phénomène a notamment pu être exploré par le biais de visualisations en réseaux.<sup>72</sup> Grâce à l'enrichissement des données, nous avons été en mesure d'identifier trois types de liens différents.

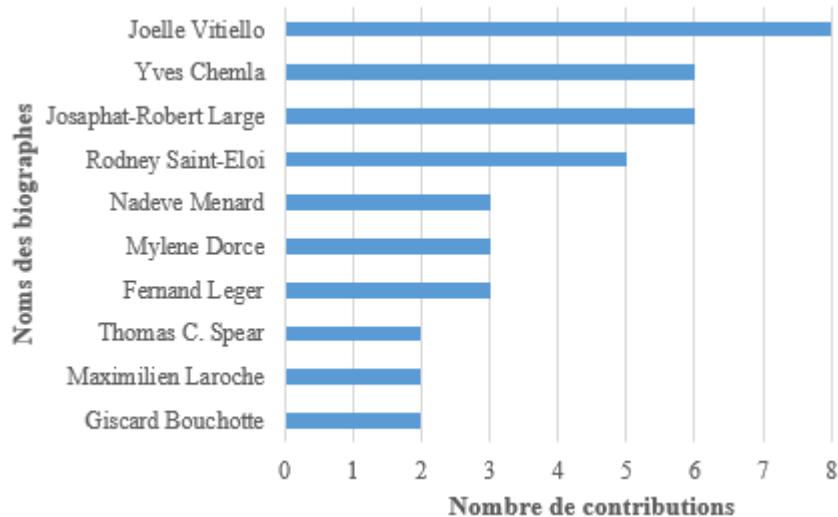
### 1. Les entités auteurs et les individus qui rédigent leurs notices biographiques

*Figure 38 Réseau des liens entretenus entre les entités auteurs et leurs biographies, Gephi, algorithme Fruchterman Reingold, partition des nœuds suivant leur poids*



Sur un ensemble de 201 notices, on en dénombre 38 issues d'*île-en-île* pour lesquelles son auteur est mentionné. Le quart a été rédigé par une dizaine d'individus. La visualisation met en évidence un réseau épars structuré par quelques figures.

*Figure 39 Diagramme des principaux biographes*



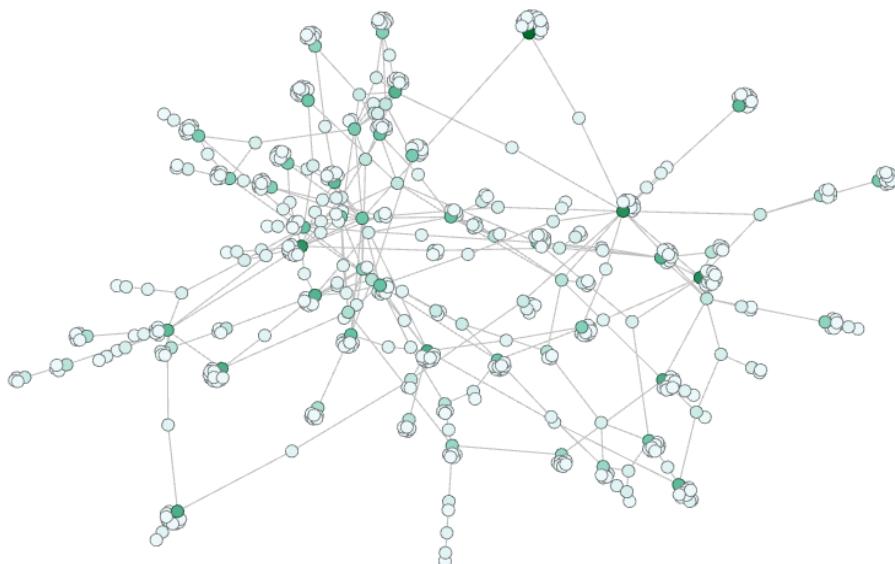
Cette première analyse nous permet d'affirmer l'hypothèse selon laquelle il est possible d'identifier de nouveaux auteurs potentiels à partir d'individus pivots, ici les biographes. C'est ainsi que Joëlle Vitiello, Yves Chemla, Josaphat-Robert Large, mais aussi Rodney Saint-Éloi participent à la structuration du réseau du corpus d'*île-en-île*.

<sup>72</sup> La théorie des graphes considère la force d'attraction entre deux nœuds quelconques, la somme des vecteurs déterminant la direction vers laquelle un nœud doit se déplacer.

## 2. Les entités auteurs et leurs relations de collaborations et de co-auteurat relevées au sein des notices biographiques et des répertoires d'autorité

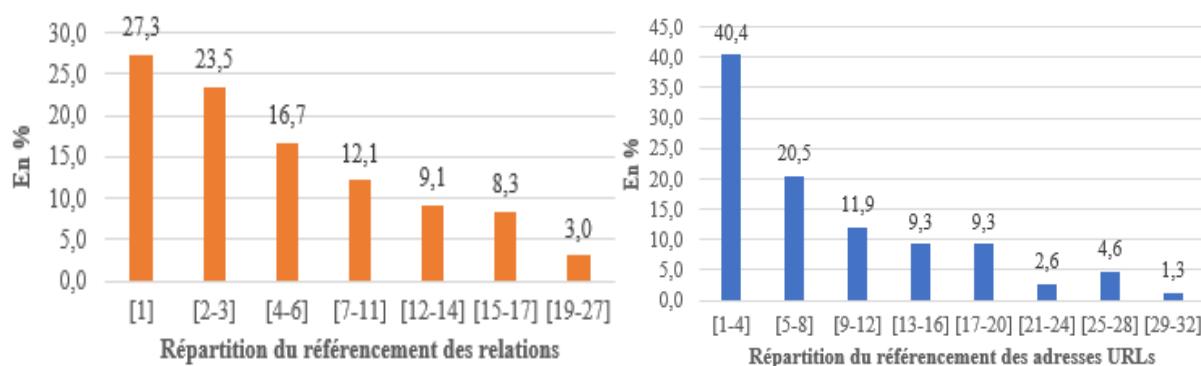
On dénombre 326 entités nommées extraites automatiquement depuis *ile-en-ilé* et près de 506 issues du *VIAF*. Sur ces 832 entités, on note 583 entités uniques soit 86 % du corpus. 65 apparaissent à deux reprises (10 %) et 18 à trois reprises (3 %). Outre ces remarques préliminaires, le réseau est structuré une fois encore autour de personnes pivot entretenant plusieurs relations de collaborations et de co-auteurat. Des figures sont ainsi mises en avant comme Vincent Nedjmhartine cité à quatre reprises, Alexis Stephen ou encore Jacqueline Wiener cités à cinq reprises, Jacques Roumain et Rodney Saint-Éloi cités à six reprises, René Depestre, cité à 8 reprises ou encore Lyonel Trouillot, cité à 13 reprises.

*Figure 40 Réseau des liens de collaborations et de co-auteurat entretenus par les entités auteurs, Gephi, algorithme Force Atlas2, partition des nœuds suivant leur poids*



Si l'on regarde à présent la répartition du référencement des relations et des liens pour chaque entité auteur, on remarque que la majorité du corpus d'étude entretient le plus souvent des relations uniques 27 % du corpus et ne possèdent qu'une à quatre adresses URLs (40 %).

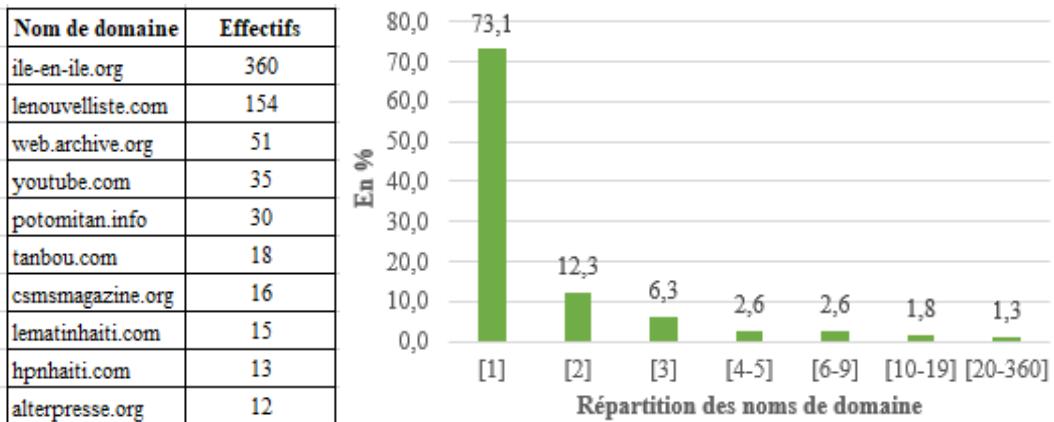
*Figure 41 Répartition des entités suivant le nombre de relations entretenues (à gauche) et le nombre d'adresses URLs identifiées (à droite)*



### 3. Les entités auteurs et les noms de domaine qui leurs sont associés

On dénombre 5 liens issus de la base de données RDF d'origine ; 26 liens issus du *spla* et 1258 d'*ile-en-ile*. Il a été possible d'isoler les noms de domaine des adresses URLs.

Figure 42 Noms de domaines récurrents (à gauche) et répartition par entité auteur (à droite)



Les noms de domaines uniques et sites personnels constituent près de 73 %. Parmi les noms de domaines récurrents, on retrouve sans surprise *ile-en-ile*. *Lenouvelliste* est quant à lui le seul quotidien publié en Haïti à Port-au-Prince depuis la catastrophe de 2010. En troisième position, on retrouve un site d'archives en ligne et juste derrière, la plateforme *Youtube*. Enfin, *potomitan* et *tanbou* sont des sites de promotion des cultures et langues créoles. On remarque également la présence de quotidiens et de magazines numériques.

Figure 43 Réseau des liens de référencement des entités auteurs, Gephi, algorithme Yifan-Hu, partition des nœuds suivant leur poids et par nom de domaine



Cette visualisation et les graphiques précédents nous permettent ainsi de conclure à l'existence de communautés d'auteurs nativement numériques structurées autour de relations de collaborations et de la diffusion de leurs travaux.

## Conclusion

En cette fin de projet, nous pouvons constater l'existence d'une « littérature numérique » en train de se faire. Elle suscite actuellement de nombreux débats et est de par sa nature peu normalisée. Aussi, il est nécessaire de regrouper les informations disponibles, de les structurer et de les analyser avant même d'envisager une étude des contenus, des productions et de leurs auteurs. Par l'interrogation de répertoires d'autorité tels que la BNF, le VIAF et de sites locaux de références d'auteurs, nous sommes parvenus à enrichir les données d'origine et à en structurer leur collecte.

Le modèle XML-TEI que nous avons proposé est flexible, adaptable et ouvert. Il permet d'ajouter de nouvelles sources d'informations au fil des découvertes du projet, telles que les données issues du web social, d'autres répertoires d'autorités comme l'ISNI ou des ressources sémantiques complémentaires à l'exemple de la base de données Wikipédia. De plus, ce processus d'enrichissement offre de multiples occasions d'identifier de nouveaux acteurs littéraires ainsi que des noyaux communautaires. Ceci a notamment été rendu possible grâce à un travail et à une réflexion sur l'extraction des entités nommées et à l'induction de données liées aux parcours de vie des auteurs. Ces pistes permettraient de décrire ceux qui passent sous les radars des références d'autorités. Cette méthodologie nous semble par ailleurs généralisable du local au global, puisque l'exemple du corpus haïtien est tout à fait applicable à d'autres corpus géographiques de la littérature francophone : il existe autant de répertoires d'auteurs locaux que de communautés.

Si nous sommes conscients qu'il ne s'agit ici que de la première étape d'un projet qui a encore tout à définir, nous espérons qu'il aura permis de mettre en avant certaines problématiques, et que certaines solutions trouvées aideront à les surmonter. Il nous semblait également important de relever que cette expérience de recherche fut enrichissante de par sa matière et les débats qu'elle a suscités.

Nous tenions enfin à remercier nos commanditaires et notre tuteur qui ont su nous faire confiance et nous offrir la liberté de poursuivre ces différentes pistes, qui semblent-elles, offrent de belles perspectives de recherches.

## Bibliographie

BARIL Elodie, GARNIER Bénédicte, « Utilisation d'un outil de statistiques textuelles. IRaMuTeQ 0,7 alpha 2 », avril 2015, <http://www.iramuteq.org/>.

BELISLE Claire, « Du papier à l'écran : lire se transforme », *Lire dans un monde numérique*, Villeurbanne, 2011, pp. 112-162.

BOMMIER-PINCEMIN Bénédicte, *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne, 6 avril 1999, chapitre VII, A « Définir un corpus », pp. 415-427.

BONNET Gilles, « L'hypérítex. Poétique de la relecture dans l'œuvre numérique de François Bon », in *Poétique*, n° 175, Paris, 2014, pp. 21-34.

BONNET Gilles, « L'autobiographie. Écritures numériques de soi », in *Poétique*, n° 177, Paris, 2015, pp. 131-143.

BURNARD Lou, *Qu'est-ce que la Text Encoding Initiative ?*, OpenBooksEdition, Marseille, 2015, disponible sur <https://books.openedition.org/oep/1237>.

BURRI Inès, CHAMBAT Anaïs, RINGWALD Célian, « Ébauche de visualisation des production littéraires francophones d'Haïti et retour d'expérience », Colloque international Cartographie du Web littéraire francophone, Lyon, mercredi 22 Janvier 2020.

DESCLES Jean-Pierre, CARTIER Emmanuel, JACKIEWICZ Agata et MINEL Jean-Luc, « Textual Processing and Contextual Exploration Method », Context' 97, Rio de Janeiro, Février 1997.

GERVAIS Bertrand, « Imaginer un partenariat ancré en culture numérique », Colloque international Cartographie du Web littéraire francophone, Lyon, mercredi 22 Janvier 2020.

HEIDEN Serge, PINCEMIN Bénédicte, « Qu'est-ce que la textométrie ? Présentation », Site du projet TXM, <http://textometrie.ens-lyon.fr/>.

HEIDEN Serge, MAGUE Jean-Philippe, PINCEMIN Bénédicte, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement » in BOLASCO Sergio (dir.), *Proceedings of 10<sup>th</sup> International Conference on the Statistical Analysis of Textual Data*, JADT 2010, Rome, vol. 2, pp. 1021-1032.

LEBART Ludovic, SALEM André, *Analyse statistique des données textuelles*, Paris : Dunod, 1988 (1994).

LEJEUNE Philippe, *Signes de vie. Le pacte autobiographique*, 2, Paris, 2005.

LEMERCIER Claire, ZALC Claire, *Méthodes quantitatives pour l'historien*, Paris : La Découverte, 2008.

MERCKLE Pierre, *Sociologie des réseaux sociaux*, Paris : La Découverte, 2004, 2011, 2016.

MONAT Jean-Baptiste, *A la recherche du Web littéraire*, mémoire de master 2 Humanités Numériques soutenu à l'Université Lumière Lyon 2 sous la direction de Jérôme DARMONT et de Gilles BONNET, 2019, 93 p.

PELISSIER Daniel, « Initiation à la lexicométrie. Approche pédagogique à partir de l'étude d'un corpus avec le logiciel IRaMuTeQ », mars 2017, <https://presnumorg.hypotheses.org/>.

PROST Antoine, *Vocabulaire des proclamations électorales de 1881, 1885 et 1889*, Paris : Presses Universitaires de France, 1974.

RASTIER François, « Sémantique du Web vs. *Semantic Web*. Le problème de la pertinence », in *Syntaxe et sémantique*, n° 9, Caen, 2008, pp. 15-36.

REINERT Max, « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », in *Les cahiers de l'analyse de données*, 8, 1983, pp. 187-198.

ROQUEBERT Corentin, « Constituer un corpus Europresse utilisable dans R, IRaMuTeQ et TXM », quatrième étape : transformer le tableau aux formats IRaMuTeQ et TXM, in *RPubs*, 4 décembre 2018.

THOMAS Sue *et al.*, « Transliteracy: Crossing Divides », *First Monday*, vol. 12, n° 12, décembre 2007, (<http://journals.uic.edu/ojs/index.php/fm/article/view/2060/1908>), traduction de GUITÉ François (<http://www.francoisguite.com/2007/12/la-translitteratie/>).

VITALI-ROSATI Marcello, « Je ne suis pas un littéraire. Plaidoyer pour des frontières disciplinaires poreuses », Colloque international Cartographie du Web littéraire francophone, Lyon, mercredi 22 Janvier 2020.

## Sitographie

- <https://www.zotero.org/>
- <https://github.com/internetarchive/heritrix3/wiki>
- <https://www.w3.org/TR/rdf-concepts/>
- <https://tei-c.org/>
- <http://www.foaf-project.org/>
- <http://bibliontology.com/projects.html>
- <https://gephi.org/>
- <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- <https://newspaper.readthedocs.io/en/latest/> <https://spacy.io/>
- <https://www.labri.fr/perso/clement/lefff>
- <http://ile-en-ile.org/lithaitienne/>
- <http://www.spla.pro/spla-presentation.html>
- <http://www.isni.org/>
- <http://viaf.org/>
- <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

## Table des figures

Figure 1 Schéma lié à la construction du corpus élaboré par l'équipe MARGE .....	4
Figure 2 Schéma conceptuel des pistes de travail, réalisé fin septembre 2019, sous Draw.io .....	8
Figure 3 Exemple de visualisation permettant d'explorer des parcours inter-littéraires .....	9
Figure 4 Schéma conceptuel du premier scénario, réalisé mi-octobre 2019 sous Draw.io .....	11
Figure 5 Extrait du résultat d'interrogation des métadonnées .....	12
Figure 6 Schéma conceptuel du deuxième scénario, réalisé début novembre 2019 sous Draw.io.....	12
Figure 7 Schéma conceptuel du scénario final, réalisé début fin décembre 2019 sous Draw.io .....	14
Figure 8 Tableau comparatif des sources potentielles d'enrichissement auteur.....	15
Figure 9 Exemple d'extraction d'entités nommées .....	16
Figure 10 Exemple de la détection d'informations liées à la naissance .....	17
Figure 11 Capture d'écran du modèle général XML-TEI, dans le logiciel Oxygen .....	18
Figure 12 Structure du <teiHeader> du modèle XML-TEI, capture d'écran Oxygen.....	19
Figure 13 Description des <div> du modèle XML-TEI, capture d'écran Oxygen.....	19
Figure 14 Description des subdivisions des <div> du modèle, capture d'écran Oxygen.....	20
Figure 15 Contenu de la <div> numéro une du modèle XML, capture d'écran Oxygen .....	21
Figure 16 Extrait de la <div> numéro deux du modèle XML, capture d'écran Oxygen.....	22
Figure 17 Contenu de la <div> numéro trois du modèle XML, capture d'écran Oxygen.....	23
Figure 18 Contenu de la <div> numéro quatre du modèle XML, capture d'écran Oxygen.....	24
Figure 19 Extrait du fichier master, capture d'écran Oxygen.....	24
Figure 20 Contenu du fichier master du tableau des liens du corpus, capture d'écran Oxygen.....	25
Figure 21 Résultat de la feuille XSL enregistré sous format .csv et ouvert dans Excel .....	25
Figure 22 Identifiants possédés par les entités auteurs .....	26
Figure 23 Langues de traduction (à gauche) et pays de résidence (à droite) des entités auteurs .....	27
Figure 24 Top 5 des professions exercées (à gauche) et nombre de distinctions et prix attribués aux entités auteurs du corpus (à droite) .....	27
Figure 25 Script R - Import IRaMuTeQ .....	30
Figure 26 Dimensions des deux partitions « ile-en-ile » et « spla ».....	31
Figure 27 Analyse de spécificités de l'usage des principales catégories grammaticales .....	31
Figure 28 Table récapitulative du corpus et de ses partitions.....	.....
Figure 29 Loi de Zipf (corpus entier) .....	32
Figure 30 Extrait de la table lexicale du corpus (top 5).....	32
Figure 31 Progression de l'usage des pronoms au sein du corpus.....	33
Figure 32 Nuage de mots, fréquence $\geq 50$ soit 150 formes actives.....	33
Figure 33 Analyse de similitude du corpus entier, fréquence $\geq 50$ soit 150 formes actives, arbre maximum, indice de cooccurrence, algorithme graphopt, halos communautaires .....	34
Figure 34 Sorties logiciel de la classification hiérarchique descendante du corpus entier.....	35
Figure 35 Dendrogramme de la CHD et de formes .....	35
Figure 36 AFC sur variables actives, 30 points par classes.....	36
Figure 37 Variables illustratives de l'AFC, 30 points par classes .....	37
Figure 38 Réseau des liens entretenus entre les entités auteurs et leurs biographies, Gephi, algorithme Fruchterman Reingold, partition des nœuds suivant leur poids .....	38
Figure 39 Diagramme des principaux biographies .....	38
Figure 40 Réseau des liens de collaborations et de co-autorat entretenus par les entités auteurs, Gephi, algorithme Force Atlas2, partition des nœuds suivant leur poids .....	39
Figure 41 Répartition des entités suivant le nombre de relations entretenues (à gauche) et le nombre d'adresses URLs identifiées (à droite) .....	39
Figure 42 Noms de domaines récurrents (à gauche) et répartition par entité auteur (à droite) .....	40
Figure 43 Réseau des liens de référencement des entités auteurs, Gephi, algorithme Yifan-Hu, partition des nœuds suivant leur poids et par nom de domaine .....	40