Inès Burri (Université Jean Moulin Lyon 3)
Anaïs Chambat (ENS de Lyon)
Célian Ringwald (Université Lumière Lyon 2)

Master Humanités numériques
Livrable 2 – UE projet
13 novembre 2019

Cartographie du web littéraire francophone

⁂

Table des matières

Introduction

Notre travail s'inscrit au sein du projet interdisciplinaire IDEX *Cartoweb,* porté par les laboratoires MARGE et ERIC, visant à construire une cartographie des productions de littérature numérique francophone. Il sert de « premier étage » à un projet ANR plus vaste intitulé *Lifranum* (Littératures Francophones Numériques) 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 ».[1] 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. »[2] 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 retracer les différents ancrages méthodologiques portés par le projet, aux outils et technologies mobilisés ainsi qu'à un exemple de visualisation particulier. Nous reviendrons également sur nos missions ainsi que sur nos premières pistes de travail lesquelles ont fait émerger des problèmes et des questionnements de parcours. Ainsi, nous traiterons ensuite de la question de l'enrichissement des données, du problème spécifique lié à la collecte des données *Facebook* ainsi qu'à l'extraction des entités nommées et à la qualification des évènements. Enfin, nous exposerons nos deux propositions socles permettant d'enrichir la méthodologie de collecte existante ainsi que des pistes d'analyses complémentaires.

I. Présentation du projet

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.[3] 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*[4] pour la zone Caraïbes.

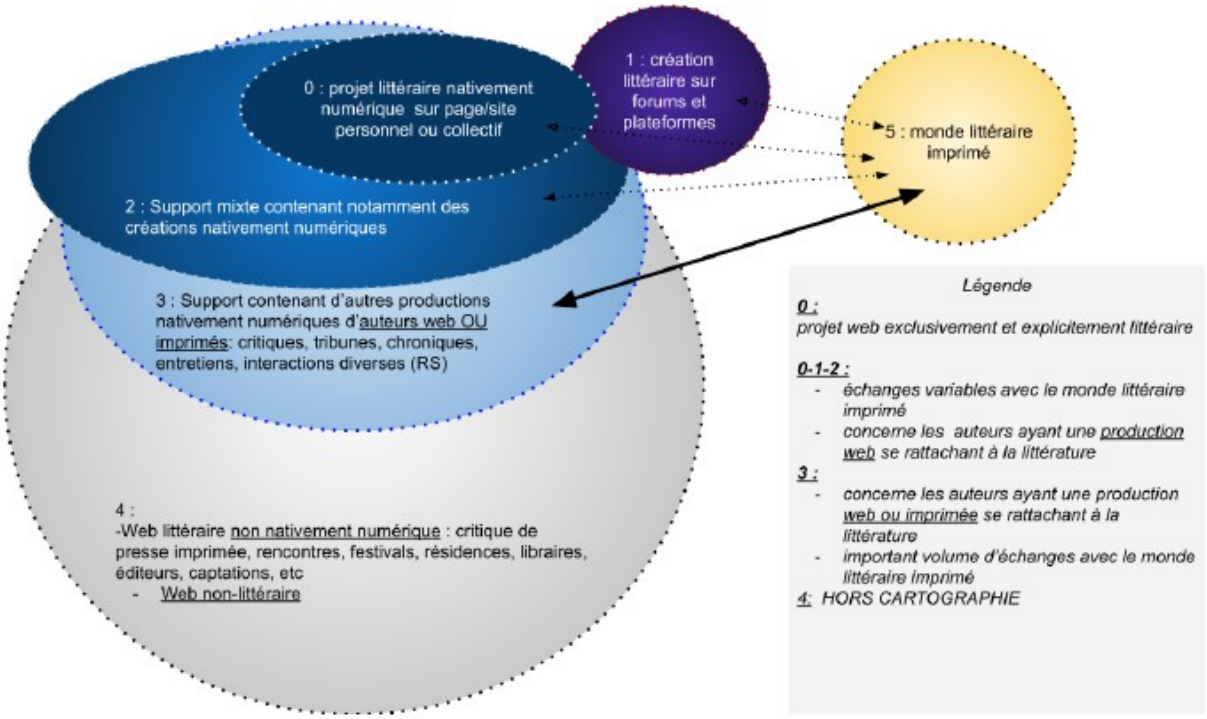
[1] https://www.univ-lyon3.fr/projet-lifranum
[2] D'après le document cadre du projet HN fourni le 16 septembre, p.1.
[3] MONAT Jean-Baptiste, *A la recherche du Web littéraire*, mémoire de master 2 Humanités Numériques soutenu à l'Université Lumière Lyon 2 sous la direction de Jérôme DARMONT et de Gilles BONNET, 2019, [93 p.] p. 13.
[4] http://ile-en-ile.org/haiti/

Ces sites ont néanmoins tendance à répertorier des auteurs déjà « installés » ou institutionnalisés, du moins ayant à leur actif des publications imprimées et valorisées.[5] 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. »[6]

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[7]



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. 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).

[5] MONAT, *op.cit.*, p. 14.
[6] BONNET Gilles, « L'hypéritexte. Poétique de la relecture dans l'œuvre numérique de François Bon », in *Poétique*, n° 175, Paris, 2014, pp. 21-34 [p. 33]
[7] MONAT, *op.cit.*, p. 14.

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.[8] 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.[9]

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.[10] 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. »[11]

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,[12] il est essentiel de pouvoir contextualiser les différents sous-corpus observés au risque de perdre les aspects fondamentaux liés à leur construction scientifique.[13] 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.

---

[8] DOUEIHI Milad, *Pour un humanisme numérique*, Paris, 2011. Cité par MONAT, p. 15.
[9] *Ibid*., p. 15.
[10] *Ibid*., p. 16.
[11] *Ibid*., p. 17.
[12] BELISLE Claire, « Du papier à l'écran : lire se transforme », in *Lire dans un monde numérique*, Villeurbanne, 2011, p. 143.
[13] « 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 »[14]

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. »[15]

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".[16] 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'autoblographie*. »[17]

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.[18]

B. Interopérabilité des formats et projets similaires

1. Outils et technologies mobilisés

L'outil provisoirement choisi par l'équipe MARGE pour stocker les URL et concentrer les métadonnées recueillies par zones géographiques est le logiciel de gestion des références bibliographiques *Zotero*.[19] 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.

---

[14] LEJEUNE Philippe, *Signes de vie. Le pacte autobiographique 2*, Paris, 2005, p. 58.
[15] 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]
[16] 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/).
[17] BONNET Gilles, « L'*autoblographie*. Écritures numériques de soi », in *Poétique*, n° 177, Paris, 2015, pp. 131-143. [p. 143]
[18] Pierre MERCKLE, *Sociologie des réseaux sociaux*, Paris, 2004, 2011, 2016.
[19] https://www.zotero.org/

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 URL.[20] 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.[21]

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. [22] L'une de ses principales syntaxes est le langage RDF/XML qui associe une autre technologie, l'XML-TEI.[23]

> « 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. »[24]

Aussi, pour parer à ce problème des ontologies RDF complémentaires sont envisagées telles que FOAF[25] utilisée pour la description des personnes et les relations qu'elles entretiennent entre elles et BIBO[26] pour la gestion des références bibliographiques. Cette dernière a d'ailleurs été récemment importée dans Zotéro et Omeka[27] 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*[28] afin de relever des tags sur les pages ; *Newspaper*[29] concernant l'extraction du contenu des blogs ; *SpaCy*[30] pour l'étiquetage morphosyntaxique ou encore *Lefff*[31] (Lexique des Formes Fléchies du Français) une surcouche du précédent adapté au français offrant une lemmatisation efficace.

---

[20] MONAT, *op.cit.*, p. 18.
[21] https://github.com/internetarchive/heritrix3/wiki
[22] https://www.w3.org/TR/rdf-concepts/
[23] « 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/
[24] RASTIER F., *op.cit.*, p. 20.
[25] http://www.foaf-project.org/
[26] http://bibliontology.com/projects.html
[27] https://omeka.org/
[28] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[29] https://newspaper.readthedocs.io/en/latest/
[30] https://spacy.io/
[31] https://www.labri.fr/perso/clement/lefff/

2. Un exemple de visualisations : le projet « Mapping the Republic of Letters »

Concernant les visualisations, le projet *Mapping the Republic of Letters*[32] semble particulièrement pertinent, car il développe « des outils de visualisation sophistiqués et intéractifs »[33] ainsi que la « création d'un dépôt de métadonnées » sur leur discipline et « de guidelines pour la saisie future des données ».[34] Le projet étudie les réseaux dans le monde de l'érudition, qui « ont été des lignes de vie de l'apprentissage » et « [qui] ont facilité la diffusion et la critique des idées, la diffusion de l'actualité politique, ainsi que la circulation des personnes et des objets ».[35] Il s'agissait notamment des « réseaux de correspondance qui s'étendaient sur plusieurs pays et continents, les réseaux sociaux créés par les académies scientifiques et les réseaux physiques créés par les voyages ».[36] Les deux projets ont en ce sens des ambitions et objectifs assez communs. Ils tendent à développer une cartographie de la République des Lettres à travers des « études de cas stratégiques par leur portée géographique et temporelle, ainsi que par leur ampleur et leur portée ».[37] Leur corpus s'amplifie au fur et à mesure qu'ils établissent des partenariats, permettant un large éventail d'étude. Ils soulignent que les sources étant toutes singulières, « chacune d'entre elles présente des problèmes différents de traitements des données ainsi que des défis uniques de visualisation de l'information ».[38] Nous pouvons également nous intéresser au traitement de leurs données.

> Il a fallu en effet « créer, nettoyer, compléter et/ou conserver les données historiques [qu'ils] souhaitaient étudier ; les visualisations ont ensuite permis de formuler des hypothèses et de poser des questions sur ces données. L'étape finale du processus de recherche ayant consisté à poursuivre la lecture (des sources primaires et secondaires) et l'analyse (d'ensembles de données identiques ou différents) avant l'interprétation et la rédaction des résultats ».[39]

Les visualisations proposées sont de différents types : interactives ou statiques telles que des schémas de données et des tableaux. Elles ont été réalisées en partenariat avec le *Consortium Gephi,*[40] logiciel *open source* pour la visualisation de données et *Palladio,*[41] laboratoire de recherche en humanités numériques et design de l'Université de Stanford.

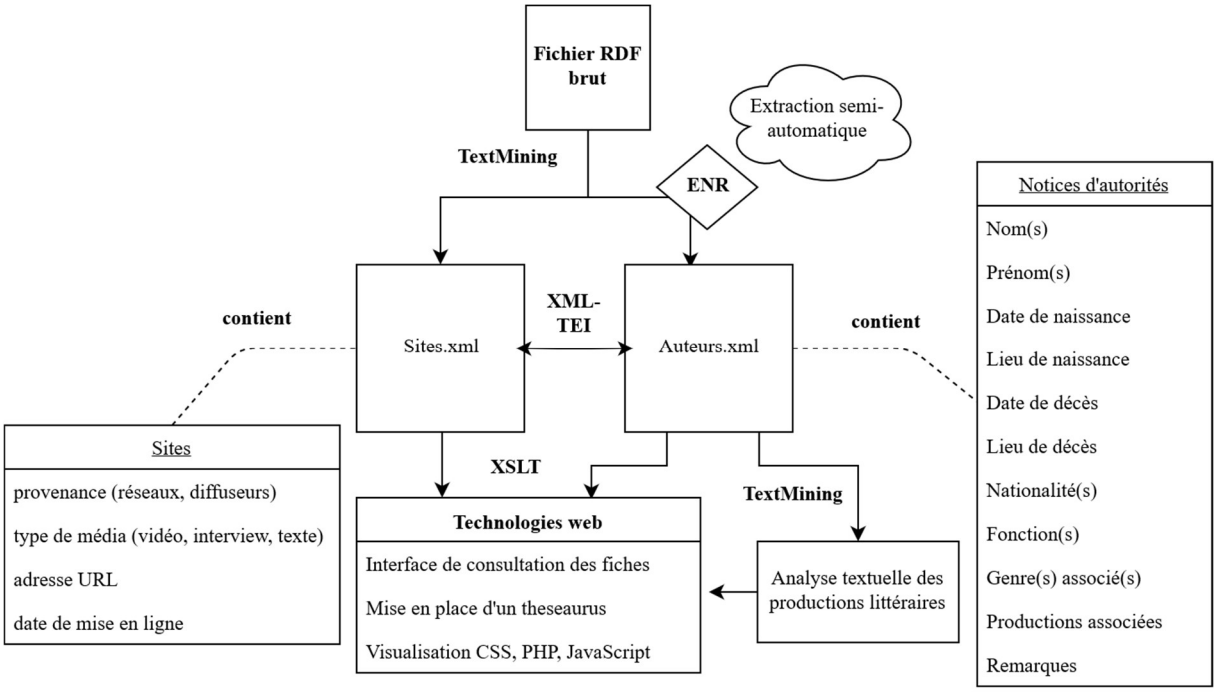
---

[32] http://republicofletters.stanford.edu/index.html
[33] « through the development of sophisticated, interactive visualization tools », *Idem*.
[34] « to create a repository for metadata (. . .), and guidelines for future data capture », *Idem*.
[35] « These networks were the lifelines of learning, from the age of Erasmus to the age of Franklin. They facilitated the dissemination and the criticism of ideas, the spread of political news, as well as the circulation of people and objects. » *Idem*.
[36] « networks of correspondence that stretched across countries and continents; the social networks created by scientific academies; and the physical networks brought about by travel », *Idem*/casestudies.
[37] « The case studies are strategic in geographic range and in time period, and in their breadth and scope », *Idem*.
[38] « The wide range of case studies gives us multiple points of intersection with the Republic of Letters. Most every case study is based on a different information source and so each one presents different problems of data handling as well as unique information visualization challenges », *Idem.*
[39] « The dynamic visualizations of historical data that readers can explore here belong to the middle phase of our work. To reach this stage, we first needed to create, clean, augment, and/or curate the historical data that we wished to study; the visualizations then allowed us to formulate hypotheses and ask questions about these data. The final stage of our research process involved further reading (of primary and secondary sources) and analysis (of the same or different datasets) before composing the articles themselves. » Idem/publications.
[40] https://gephi.org/
[41] http://hdlab.stanford.edu/?fbclid=IwAR0CwGl_v1h-JZLN4qg7_WgNJtAGDfup2sFDCXT4ug47-Hxsdbl7tRTVFx4

C. 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. »[42] Pour effectuer ce travail, une liste d'adresses Web répertoriant des sites et des pages d'écrivains francophones publiant 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 (localisation, date, auteur, 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 sous Draw.io

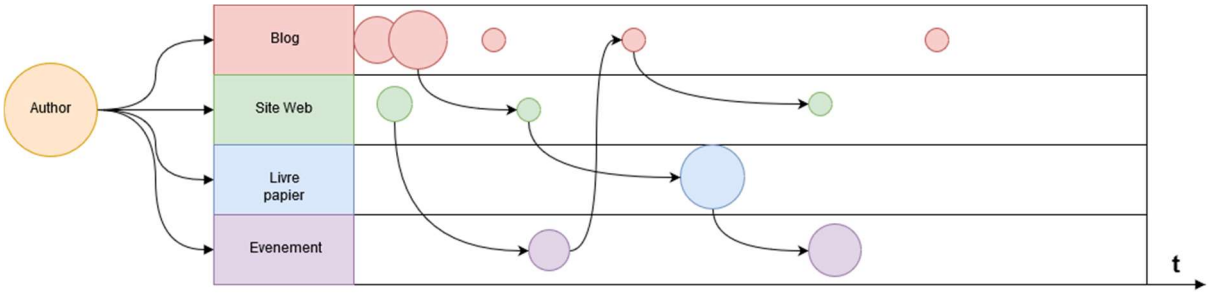


Il s'agit à partir du fichier RDF et d'une extraction semi-automatique des entités nommées de constituer deux fichiers XML : un dédié aux fiches d'autorités des auteurs et un autre dédié aux URI qui proposerait des métadonnées liées à la provenance des données, au type de média utilisé ainsi qu'aux dates 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 et métadonnées auteur).

[42] Document cadre du projet, p. 2.

Les différentes temporalités des contenus pourraient ainsi être mis en valeur de même que les parcours sociaux de leurs auteurs.

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



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. Néanmoins, ceci a rapidement posé problème à la vue du manque de métadonnées disponibles dans le RDF.

Figure 4 Extrait du résultat d'interrogation des métadonnées

1658 URLs
title: 592,
type: 586,
date: 183,
creator: 240,
rights: 13,
identifier: 9,
publisher: 7,
contributor: 24,
coverage: 4

Sur les 1658 URL 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.

Nous nous étions engagés à proposer une cartographie du web littéraire francophone dans son intégralité. Cependant, le manque de métadonnées et leur mauvais référencement dans la bibliothèque Zotéro n'ont pas permis d'aboutir à un traitement pertinent des données. Le corpus a ainsi été réduit à la zone Caraïbes. Cette même problématique 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. De ce fait, nous n'avons pas encore débuté notre travail sur les visualisations.

II. Sélection et regroupement des problèmes : les questionnements de parcours

À ce stade nous avons eu de nombreuses réflexions générales relatives aux données, notamment concernant l'importance de leur représentativité et la définition d'un auteur. 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 problèmes spécifiques à la collecte des données sur Facebook et la question relative à l'extraction des entités nommées et des évènements seront abordés.

A. Autour de l'enrichissement des données

Il est intéressant de réfléchir sur la limite de l'automatisation. L'interruption de l'automatisation 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 ».[43]

Nous avons alors tenté d'approfondir ce qu'était qu'un enrichissement des données. Quelle est la place de la machine et celle de l'homme dans cette définition ? Sur ce point, nous sommes partagés. Nous pouvons en effet distinguer l'enrichissement, qui serait le propre de l'homme, et la collecte, qui serait le propre de la machine. Lorsque la machine extrait des éléments de manière automatique afin d'y ajouter une valeur ajoutée n'est-ce pas en cela un enrichissement ? Ne faudrait-il pas nuancer le terme afin de distinguer l'enrichissement issu d'une réflexion humaine basée sur des compétences et la légitimité d'un profil, à celle d'une machine qui nécessite d'être contrôlée ? Par exemple, dans le cadre de l'établissement automatique d'un *thesaurus*, François Rastier souligne

> des « inconvénients notoires : une généralité qui ne (…) permet pas de s'adapter aux points de vue sélectifs exigés par les tâches et un manque d'évolutivité qui exige une maintenance manuelle [,] réduis [ant] la langue à une nomenclature, qui ne rend compte ni des structures textuelles, ni des variations considérables de genres et de discours ».[44]

On pourrait alors voir la machine comme effectuant ce que l'on appelle plus simplement une extraction de données :

> « On extrait l'information, puis on la communique, la seule condition mise à la communication se limitant à *l'information packaging*, conçue comme simple emballage des connaissances. »[45]

La création de liens automatisés et basés sur l'extraction de données ne pourrait-elle pas être également perçue comme une forme d'enrichissement ?

---

[43] RASTIER, *op.cit*., pp. 22-23.
[44] *Ibid*., p. 20.
[45] *Ibid*., p. 21.

En effet, un enrichissement automatique, bien qu'il ne puisse pas être contrôlé, vient appuyer l'importance d'un élément dans un texte. Ainsi, nous pourrions y voir plutôt une exploitation des données par la machine en vue d'être validées par l'homme et qui se voudra ensuite comme une forme enrichie des données.

B. 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. Néanmoins, la collecte des données sur ce type de support est extrêmement complexe. Il faut d'abord créer un profil Facebook spécifique « *Facebook for developers* », fournir une preuve de son identité (passeport, carte d'identité…) et de créer une page d'entreprise. Ensuite, pour chaque information collectée, il est nécessaire d'effectuer une demande longue et complexe composée d'une vidéo et d'un texte, laquelle sera ensuite validée ou non par les administrateurs du système. Notre demande n'a toujours pas été acceptée.[46] Par ailleurs, certaines de ces pages n'existent plus. Les données issues de notre travail sur ce site ne pourront pas être intégrées au projet, celles-ci relevant de la responsabilité personnelle des membres de notre groupe.

C. L'extraction des entités nommées

Nous envisageons d'extraire 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.[47] 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. Elles pourront néanmoins permettre de créer des liens potentiels. 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, le français ou le hollandais 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 signifiant 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 ».[48]

Il nous faut également réfléchir à un moyen de contextualiser les évènements véhiculés par les communautés d'auteurs.

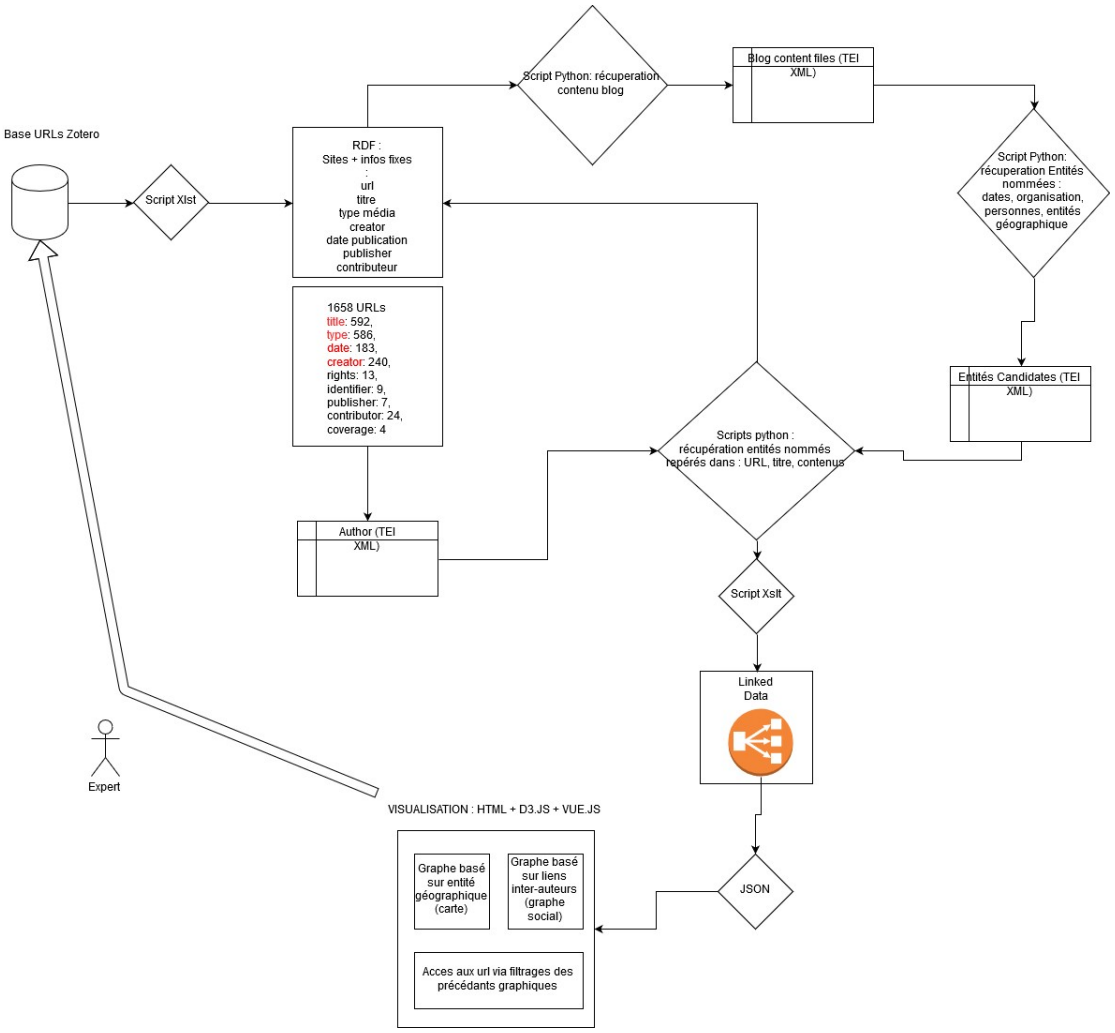
[46] La demande a été formulée le 30 octobre 2019.
[47] https://spacy.io/api/annotation#named-entities
[48] MONAT, *op.cit.*, p. 15.

III. Construction des scénarios de réponse

Il est temps à présent d'exposer nos deux propositions socles permettant d'enrichir la méthodologie de collecte existante ainsi que des pistes d'analyses complémentaires.

A. Le premier scénario

Figure 5 Schéma conceptuel du premier scénario réalisé sous Draw.io

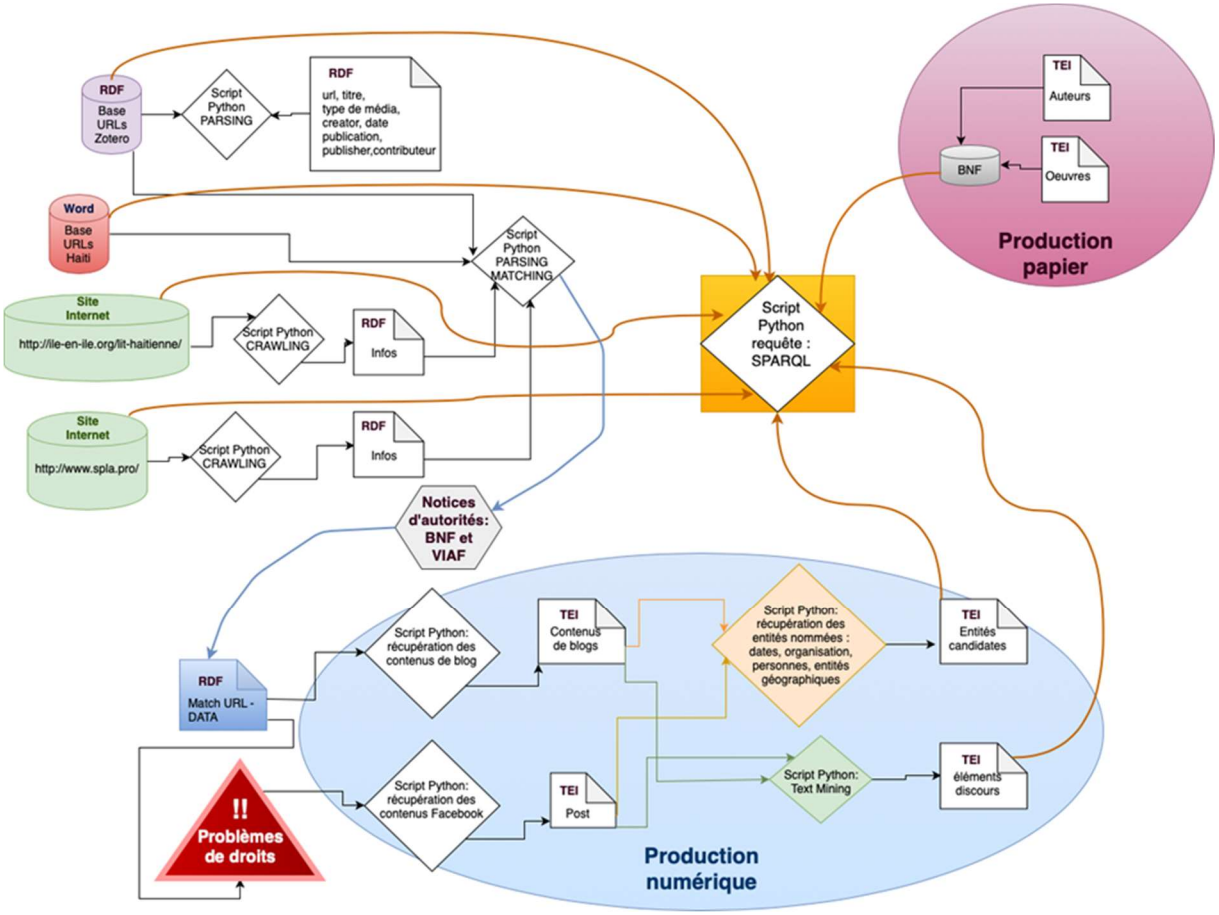


Notre premier scénario débute avec l'importation d'une bibliothèque Zotéro, annotée grâce à BIBO sous la forme d'un fichier RDF avec des métadonnées complètes. De ce fichier, il est alors possible d'extraire le contenu du blog et des fiches d'auteurs en XML. Ensuite, à partir du contenu du blog nous pouvons extraire des entités nommées, et tous autres 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. La main humaine est indispensable pour cette partie. Dans une deuxième étape, nous croisons les fichiers XML (les fiches d'auteurs et les entités nommées) avec le fichier RDF pour créer les liens entre auteurs qui seraient contenus dans un fichier unique. Il s'agirait ensuite de l'exploiter à l'aide de JSON, format le plus adapté à la manipulation en JavaScript. La dernière étape sera la création d'une page web contenant les visualisations.

Plusieurs corpus sont finalement constitués : un corpus avec l'ensemble des URL, un des sites « points-relais » permettant de créer des liens entre auteurs, un autre contenant la production littéraire en elle-même, les fiches auteurs et la liste des évènements. Ce schéma ne peut être performant que si la bibliothèque Zotéro est correctement complétée. Comme nous l'avons vu précédemment, les difficultés liées au manque de données disponibles ne nous permettent pas actuellement de suivre ce processus. Nous vous le présentons toutefois, car il pourrait devenir le modèle utilisé ultérieurement si la méthodologie enrichie de collecte des données continuait d'être gérée par le biais d'une bibliothèque Zotéro.

B. Le deuxième scénario

Figure 6 Schéma conceptuel du deuxième scénario réalisé sous Draw.io



Ce scénario a été effectué plus tard lorsque nous nous sommes rendu compte qu'il y avait trop peu de données exploitables pour aboutir à des visualisations intéressantes et interprétables. Ainsi, nous avons ajouté la collecte de donnée de deux sites, l'un possédant un corpus déjà constitué d'auteurs littéraires haïtiens,[49] l'autre étant un site très complet d'information et de mise en réseau des acteurs culturels du Sud.[50] Afin de suivre le processus d'accomplissement des auteurs du web à la publication papier, nous aimerions également ajouter un autre corpus des productions littéraires publiées imprimées, qui pourrait nous être fourni par la BNF.

[49] http://ile-en-ile.org/lit-haitienne/
[50] http://www.spla.pro/spla-presentation.html

L'ensemble de ces corpus doivent cependant faire l'état d'un contrôle qualité puisque leur constitution est entièrement automatisée et peut alors contenir des erreurs et des informations inutiles, d'autant qu'ils intègrent des auteurs qui n'ont jamais publié sur le Web.

Le fichier RDF bleu sera une fiche unique contenant la normalisation des quatre corpus : les liens, les auteurs, les pays, la langue, etc. Elle aura été enrichie des informations données par la BNF et des notices VIAF. À partir de cela on pourra alors 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. Pistes complémentaires

Nous avons donc vu qu'elles étaient nos propositions socles, voici maintenant les étapes supplémentaires. Tout d'abord dans l'objectif d'une automatisation constituant une aide au travail des chercheurs, il serait intéressant de s'abonner à des flux RSS afin d'être quotidiennement informé des actualités entourant la littérature du Web francophone. Ce type d'automatisation permet en effet de ne pas fondre les informations au sein du corpus sans être validées par les chercheurs.

Afin d'enrichir le traitement des données concernant les auteurs, nous avons par ailleurs réfléchi à la création d'un corpus indépendant d'auteurs non référencés par la BNF ou par les notices d'autorités. En effet, ces derniers ne doivent pas être occultés puisqu'ici l'absence de donnée est significative.

Conclusion

À cette étape de mi-parcours du projet, nous relevons la défaillance dans l'application de la méthodologie de collecte des données ce qui nous limite *de facto* dans leur exploitation. Il nous semble donc délicat d'utiliser les contenus en raison d'un manque d'expertise. Nous pourrions en revanche exploiter les entités nommées qui auront été extraites automatiquement (les noms propres, les noms de lieux, etc.) afin de mettre en avant certains liens entre les auteurs. Les liens qui feront l'objet d'une visualisation seront issus des métadonnées fournies par le site *spla.org*[51] et des données collectées à partir des fiches d'autorités. Cependant les données que nous avons extraites sont volatiles, et ne sont disponibles que sur nos ordinateurs personnels. Il pourrait ainsi être envisagé pour l'après-projet de construire une base de données afin de les rendre pérennes et ainsi permettre l'élaboration d'une interface d'enrichissement, ce qui constituerait le prolongement de la visualisation ; résolvant ainsi la plupart des problèmes soulignés dans ce rapport puisqu'il s'agirait d'effectuer un contrôle manuel après toute forme d'automatisation.

Nous espérons pouvoir commencer prochainement l'application des scénarios et ainsi contourner les problèmes exposés précédemment.

[51] *Idem*.

Bibliographie

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

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

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

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

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

Pierre MERCKLE, *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.

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

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/).

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://republicofletters.stanford.edu/index.html
http://ile-en-ile.org/lithaitienne/
http://www.spla.pro/spla-presentation.html

Table des figures