




10 OCTOBRE 2022

# PROJET LIFRANUM

FILIPPO SARRA, ZOE NOEL, BRUNEL TCHEKELI, AMBRINE SOLTANI  
MASTER 2 HUMANITES NUMERIQUES  
Enssib, Lyon 2



## Contexte

Le projet LIFRANUM (<https://projet-lifranum.univ-lyon3.fr/projet>) est un projet visant à identifier et indexer un corpus des productions littéraires francophones et nativement numériques. Ce projet débute en 2020 grâce à un financement de l'Agence Nationale de la Recherche. Il regroupe les laboratoires universitaires MARGE et ERIC. Le laboratoire MARGE est spécialisé en langue et littérature française (ainsi que d'autres ères linguistiques : hispanophone, slave, etc.) et en sciences de l'information et de la communication. Le laboratoire ERIC est quant à lui spécialisé en sciences des données, informatique décisionnelle et humanités numériques. Ils sont localisés au sein des universités Lyon 2 et Lyon 3. À cela s'ajoute une collaboration avec la Bibliothèque nationale de France.

L'objectif du projet LIFRANUM ou "Littérature Francophone Numérique" est d'essayer de répondre à des questions telles que :

- Comment les auteurs produisent et pensent la littérature à l'ère numérique ?
- Comment les auteurs mobilisent-ils les dispositifs numériques, tels que sites, blogs, et réseaux sociaux, pour diffuser leurs productions ?
- De quelle façon peut-on classifier la littérature numérique ?
- Quelles sont les spécificités de la production littéraire francophone numérique ?
- Comment traiter les corpus littérature numérique à l'aide d'outils informatiques ?
- Quels sont les thèmes abordés par la littérature francophone numérique ?

Le but étant de construire des connaissances novatrices sur un type de production littéraire trop souvent inconsiderée par le domaine universitaire littéraire.

Les chercheurs considèrent ces productions originales et spécifiques, au vu des contextes inédit dans lesquelles elles ont vu le jour. Nous constatons également des formes de circulation nouvelles ; et les formes d'échanges et d'interactions entre auteurs sont aussi inédites. Les chercheurs doivent se demander comment définir la littérarité de ces productions, afin de détecter des formes d'écriture, regrouper les auteurs, définir une typologie de la littérature numérique et identifier les thèmes récurrents ainsi que les auteurs et les textes "de référence".

Pour cela, le projet réunit de nombreux chercheurs et chercheuses provenant de disciplines différentes. Un point clé de ce projet est l'interdisciplinarité. Ainsi, la littérature, l'informatique et les sciences humaines et sociales sont amenées à dialoguer ensemble. La démarche scientifique de ce projet est en elle-même un savant mélange de méthodologies, d'approches et de conceptions de ces trois disciplines. Profondément inscrit dans les Humanités Numériques, le projet LIFRANUM a pour ambition de dépasser les frontières disciplinaires afin de construire des protocoles de recherches et méthodologies plus à même de saisir les spécificités des productions littéraires numériques.

Un stage a été réalisé en 2021 par Théo Gady, un étudiant de master, qui a pu réaliser un pré-traitement des données. Il a rédigé un rapport de stage dont nous disposons et à partir duquel nous pourrions commencer à travailler.

Après avoir discuté avec les responsables du projet et notre encadrant, (Alice Pantel-Cassagnaud, Enzo Terreau et Julien Velcin), nous avons compris que jusqu'ici, c'est la partie principalement informatique qui a été mobilisée. Ceci afin de récolter des données, construire le corpus, effectuer un premier traitement sur les données pour les rendre lisibles et assurer la viabilité des données. Il s'agit aujourd'hui

de travailler sur l'analyse en mobilisant les outils informatiques et en portant un regard littéraire tant sur la forme que sur le fonds de ce corpus.

## Reformulation de la demande

### Tâches à réaliser

Suite à notre première réunion avec les commanditaires, nous avons défini quatre étapes dans la réalisation de notre projet :

1. Découvrir le corpus et le nettoyer en identifiant les balises et les points pertinents pour notre travail,
2. Réaliser nos premiers essais et premiers traitements en utilisant les librairies python nous permettant de mettre en lumière des pistes de catégorisations
3. Nommer et classer, en d'autres mots, réaliser notre propre typologie adaptée à notre corpus. Sur cette partie, les commanditaires attendent de nous diverses propositions. Les marqueurs peuvent être thématiques, linguistiques, etc. C'est à nous de trouver des catégories parlantes et cohérentes avec la thématique de la littérature nativement numérique.
4. Une étape d'analyse se découpant en deux sous-étapes :
  - Analyse macro avec des outils permettant une analyse automatique permettant de réaliser des analyses statistiques ou du datamining par exemple.
  - Analyse micro où nous mettrons à profit nos connaissances en SHS. Nous pourrions ainsi proposer des analyses plus fines concernant les interactions entre auteurs, dresser des profils, identifier des spécificités en fonction des zones, analyser les discours (côté SHS) ou réaliser des analyses de textes, des analyses de langage ou de style d'écriture, une identification des références ou inspirations (côté Littérature)

La troisième étape est grandement attendue par notre équipe et par les commanditaires. En effet, comme nous l'avons mentionné précédemment, il nous est ici demandé d'apporter un regard littéraire et SHS sur ce corpus numérique. En cohérence avec nos différentes formations, il sera important de nous approprier une partie des objectifs du projet LIFRANUM afin de se focaliser sur des questions thématiques, ou stylistiques, ou socio anthropologiques qui sont au cœur de ces productions. Cette étude sera centrée sur les possibles façons d'analyser et d'appréhender ce corpus de productions littéraires francophones nativement numériques. Bien évidemment, étant inscrits dans la démarche HN, nous emploierons les outils informatiques des fouilles de texte et de fouilles de données qui font l'objet de nos cours.

Ces différentes tâches s'inscrivent dans un objectif plus large qui est celui de visualiser et de diffuser une quantité très importante de données. Ce corpus doit pouvoir servir à la recherche, mais pas uniquement. En effet, l'ambition est de rendre ces données et les connaissances produites autour de ce corpus accessibles à un large public (enseignants, primaire et secondaire, bibliothécaires et pourquoi pas un public amateur).

### Rendu final

Dans le cadre de ce projet, plusieurs livrables sont possibles. En effet, il n'y a pas de consigne précise sur la forme que doit prendre le rendu final. Cependant, un livrable écrit ainsi qu'une présentation orale sont prévus par la structure de cet enseignement. Il est demandé aux étudiants et aux commanditaires de définir le livrable le plus approprié au projet. Ainsi, les commanditaires nous ont proposés de participer à un colloque portant sur la littérature numérique, au mois de mars 2023. Nous allons donc évaluer cette proposition en identifiant nos capacités et en prenant en compte le temps qui nous est imparti (au mois de mars nous serons en stage). Nous envisageons, pour le moment, que notre

rendu prendra la forme d'un article universitaire, comme proposé par les commanditaires, d'une quinzaine de pages. Cet article reprendra notre méthodologie, ainsi que nos apports réflexifs et nos différentes manipulations sur les données. Si nous décidons de participer au colloque, nous serons amenés à présenter le contenu de cet article lors de ce dernier.

## Problèmes et difficultés

La plus grosse difficulté qui va se poser à nous, est celle de la catégorisation. C'est une démarche scientifique qui n'est pas simple, d'autant plus que nous travaillons sur un corpus de taille très importante. Nous devons réaliser un effort de réflexion important afin de trouver et élaborer les catégorisations les plus pertinentes et exploitables par la suite.

Ce projet de grande envergure est le premier auquel nous sommes confrontés. Nous devons donc être attentifs à nos capacités et évaluer avec attention nos possibilités en prenant en compte nos compétences mais aussi le temps qui nous est imparti. Il est facile face à un projet tel que celui-ci de se retrouver face à une quantité de travail trop importante ou face à des situations nous demandant des compétences techniques dont nous ne disposons pas. Nous porterons la plus grande attention à évaluer la faisabilité de nos missions en nous reportant notamment à nos professeurs et aux commanditaires.

De plus nous sommes tous les quatre issus de disciplines différentes et suivons des formations différentes. Le profil interdisciplinaire de notre groupe est un réel atout cependant, la gestion de l'emploi du temps, les concertations et réflexions de groupe vont nous demander de grandes capacités organisationnelles. Nous savons que l'organisation d'un groupe peut être une considération très importante et nous devons mettre au point des stratégies afin de fluidifier les échanges et faciliter le partage d'information.

Il est clair que nos principales difficultés consistent dans la prise en main des données contenues dans le corpus, et par conséquent leur exploitations. En effet, nous ne maîtrisons pas tous les outils informatiques. Pour la réalisation de notre projet il faudra en tenir compte et ainsi porter une attention particulière à la programmation du temps de travail. Nous n'hésiterons pas à demander des conseils et de l'aide technique à monsieur Terreau.

## Les données

Le corpus est déjà existant, les commanditaires nous ont fourni les données provenant des API de Wordpress et Blogger. Pour le moment, ces données sont très nombreuses et toutes ne vont nous servir. Enzo Terreau nous a donc proposé de sélectionner les différentes balises dont nous aurions besoin pour pouvoir nous envoyer les données dans un format csv et non dans le format json (format dans lequel elles sont actuellement disponibles).

Pour chaque API, nous disposons de différentes catégories d'information, celles sur les blogs, sur les posts et sur les commentaires. Nous n'allons pas traiter la partie concernant les commentaires, cette dernière n'étant pas assez représentative (il y a en réalité, peu de commentaires sur ces blogs).

Un notebook nous a également été fourni par Enzo Terreau grâce auquel nous pourrions plus facilement explorer et manipuler les données.

Nous disposons de plus du rapport de stage de Théo Gady dans lequel se trouve de nombreuses informations intéressantes. Elles serviront d'orientation pour notre travail. Théo Gady a davantage travaillé sur le côté technique du projet. Nous sommes conscients que dans le cadre de ce projet, les commanditaires souhaitent que l'on apporte une vision moins technique et davantage tournée vers les Sciences Humaines et Sociales tout en s'inscrivant bien sûr dans les Humanités Numériques.

## Forces et faiblesses du groupe projet

Comme nous l'avons mentionné, nous pensons que notre plus grande faiblesse, est pour le moment nos connaissances et compétences dans les différents outils informatiques qui pourraient nous être utiles pour le traitement des données. Cependant, grâce aux différents cours suivis cette année dans le master Humanités Numériques, nous pensons pouvoir palier à cette faiblesse.

Notre point fort est notre diversité de profil au sein du groupe. Chacun d'entre nous fait des études différentes et donc possède des connaissances différentes, ce qui est un avantage certain dans ce type de projet. Néanmoins, aucun d'entre nous n'est spécialiste en littérature (nativement numérique), il faudra également s'intéresser de plus près à ce qui existe déjà sur la littérature nativement numérique afin d'avoir de bonnes bases pour commencer nos analyses. Nous pourrons cependant apporter nos connaissances et nos regards divers à ce projet en proposant des orientations qui n'ont peut-être pas été envisagées jusqu'alors.