Database, magie et visualisations

Ou Les littératures de l'imaginaire pour développer des méthodes de datavisualisations littéraires

Ce texte est issu de l'intervention lors du colloque Sciences et Magie, organisé par le Laboratoire des imaginaires, à Rennes2 du 28 février au 1er mars 2023.

Introduction

Comme nombre de chercheur·se·s, je suis confrontée au quotidien à des corpus de textes et d'œuvres que je dois manipuler, analyser, comprendre dans un sens large.

Mes recherches sont aux frontières de la littérature, des narrations numériques et vidéoludiques, mais aussi plus largement des artéfacts culturels et narratifs, comme les jeux de tables ou de cartes, les vidéos ou pratiques narratives amateures… Cette diversité pose de nombreuses questions, sur lesquelles je reviendrai à travers cette communication.

Cette diversité est à l'origine du projet que je vais présenter aujourd'hui : le *Répertorium* pour les narrations de l'imaginaire. Ce chantier est né d'une recherche commune sur l'archivage des corpus littéraires contemporains, présenté au CRSH avec Marcello Vitali-Rosati et Raphaël Lauro, tous deux professeurs au département des littératures de langue française de l'Université de Montréal. Ce projet questionnait au départ deux aspects : comment adapter les méthodes bibliothéconomiques aux corpus numériques, et comment y garder traces des archives d'écrivain·e·s dans leur variabilité ?

Mon projet propose une troisième voie, plus hybride : comment documenter les narrations de l'imaginaire contemporaines ? La diversité de ces œuvres est un enjeu que nous connaissons bien au laboratoire, il n'y a qu'à voir le programme du présent colloque pour s'en assurer : nos objets sont divers !

Les littératures de l'imaginaire – et plus largement les narrations de l'imaginaire –, comme nombre de productions culturelles populaires multiplient les clins d'œil intertextuels, les

imbrications et la production nombreuse (Angenot, 2013 ; Besson, 2015). Des livres, imprimés ou numériques, aux films, jeux de société de cartes, de rôles, ou vidéo, des fictions sonores, des feuilletons radio, des fanfictions, *text based adventure*, romans vidéoludiques, ou séries (entre autres), le corpus est hétérogène et riche (Ensslin, 2014).

La diversité des supports des narrations de l'imaginaire confronte les chercheur·e·s à la nécessité de croiser les approches d'analyse et de documentation (Lescouet, 2022).
Or, à la suite d'Hélène Laurichesse en 2012, nous pouvons constater que cette intermédialité s'accompagne d'une production tentaculaire, compliquant encore une approche exhaustive. Pour cette vision d'ensemble, tant thématique qu'éditoriale, il est nécessaire de faire un pas de côté et de trouver des méthodes nouvelles, de combiner l'étude narrative ou littéraire à d'autres outils.
La littérature produit des archives et des lieux d'archive, et les bibliothèques, les dépôts légaux, etc., ont une place importante dans l'histoire de ces études (notamment à travers les travaux de Derrida). Cependant, être en mesure de rassembler des corpus trans-supports – multimodaux au sens plein de Nathalie Lacelle – en un seul lieu est encore un enjeu de recherche, et seule l'unification des œuvres dans des bases de données interopérables peut permettre d'étudier conjointement les œuvres.
Les archives littéraires numériques soulèvent également de nombreux enjeux de recherche, qu'il s'agisse des aspects d'indexation et d'organisation des connaissances jusqu'à ceux d'accessibilité et de valorisation de ces mêmes objets (cf. bibliographie sur les archives numériques).
Un autre exemple est la constitution d'archives universitaires qui mène à une forme de légitimisation du corpus, mettant en avant un corpus particulier et permettant sa sauvegarde.

Or, ces modalités de conservation doivent correspondre aux différentes caractéristiques et spécificités des œuvres littéraires : évolution des formes et réinterprétation des modèles ; bouleversement des figures d'auctorialité, des supports de création et de circulation des œuvres ; évolution et consolidation des métatextes culturels et génériques contemporains ; sans oublier l'évolution des formes de productions du métatexte, telle que les univers étendus, les œuvres collectives et fandom… (comme nous pouvons le voir à travers les travaux de Henry Jenkins).
Les humanités numériques mettent à disposition des outils permettant la documentation et l'étude de corpus tentaculaires de la sorte : le but de ce projet est donc de croiser les approches et méthodologies des deux traditions académiques pour parvenir à établir une méthodologie stable permettant une analyse quantitative de la production/publication.
En construisant cette base de données, j'espère parvenir à documenter largement des corpus extensifs, à rendre accessibles ces œuvres facilement – grâce à un travail de métadonnées – pour de futures recherches.

En effet, la création d'archives demande l'établissement d'un système tout autant que de processus (Lauro, Vitali-Rosati, 2022). Un tel établissement doit également s'accompagner d'une réflexion sur les responsabilités qu'implique la mise en place de ces

règles, à l'image de la création de tout standard. En effet, une règle qui ne ferait pas consensus ne pourrait être pérenne. S'il faut dans un premier temps savoir quoi référencer, il est également important, une fois le corpus délimité, d'établir les éléments à documenter : une liste de champs ou de critères qui vont permettre de faire le tour d'un objet. Ces derniers vont permettre de cerner les éléments et les différents aspects considérés comme pertinents pour des études postérieures ou pour des analyses analytiques. Pour chacun de ces critères, il faut ensuite établir une liste de termes possibles : une taxonomie permettant de lister les différentes valeurs ou caractéristiques de chacun de ces champs. C'est par l'établissement de ces prérequis qu'adviendra la possibilité de trouver des systèmes correspondants aux œuvres littéraires adaptés aux particularités du champ disciplinaire ainsi que d'établir des taxonomies souples pouvant accompagner les évolutions techniques et formelles qui vont advenir. Cette matérialité de documentation du savoir concret de la lecture apparait comme essentielle, non seulement dans une visée de transmission et d'utilisation interuniversitaire, mais également par l'établissement d'une littératie des supports et des gestes dédiés, ainsi que par la recherche de l'évolution des formes et des genres littéraires. Ce sont ces particularités de la littérature à s'inscrire dans une histoire au travers de ces nouveaux outils qu'il est nécessaire de capter et de documenter.

L'indexation revêt également une grande importance dans ce projet. En effet, celle-ci offre à un environnement numérique une plasticité sémantique forte, mais demande néanmoins l'établissement rigoureux de standards – nous pouvons ici rejoindre les travaux du laboratoire NT2, notamment de Gina Cortopassi et Bertrand Gervais — pour être interopérable et compréhensible. Elle implique également de penser des modèles permettant de décrire au plus près les objets étudiés. Ces standards nous permettront aussi de conserver des traces et d'archiver, au sens propre de documentation et de conservation des œuvres, et de rendre les archives interopérables – à la suite, par exemple, des réflexions de Gilles Bonnet ou de mon équipe autour de Marcello Vitali-Rosati. Il est particulièrement important de pouvoir croiser les documents et les fiches documentaires entre les bases de données – Pierre Marc de Biasi.

Je vais commencer cet exposé par présenter l'intérêt des humanités numériques pour cette recherche, et les définir très brièvement.
Le cœur de mon propos va être sur la méthodologie qui est testée ici. Ainsi, après une présentation des corpus mobilisés, je vais plonger dans une partie plus technique, en prenant soin de vulgariser au maximum mon propos, où je vais détailler comment le projet Répertorium se déploie.
Je présenterai ensuite des résultats intermédiaires dans cette collecte de donnée, à l'aide du corpus de Star Wars, qui me permettra d'incarner concrètement les notions évoquées précédemment avec un ensemble populaire. N'oublions pas que le projet est en cours d'élaboration, tous les résultats sont donc imparfaits pour l'instant.
Enfin, ma conclusion sera un appel à collaborations, mêlée d'un peu d'espoir académique.

Utilité des humanités numériques

Les Humanités numériques, comme leur nom l'indique partiellement, visent à utiliser les technologies numériques pour aider la recherche en humanités – à la suite des travaux du CRIHN, de Michael Sinatra et Dominic Forest, notamment. Elles établissent un ensemble de méthodologies et d'outils pour mener des recherches difficiles autrement ; par exemple, en littérature, les usages d'intelligences artificielles pour des analyses textuelles, comme le travail mené sur les variations au sein de l'Anthologie grecque par Mathilde Verstraete et Marcello Vitali-Rosati.
Ces recherches nous permettent généralement d'appliquer des méthodes d'analyse rapprochées à des ensembles trop grands pour être humainement étudiés d'aussi près.

L'avantage pour des corpus larges et divers est évident : si je ne peux, seule, expérimenter et documenter sur des fiches dans la base de données toutes les œuvres narratives des narrations de l'imaginaire, ou ne serait-ce qu'un segment d'entre-elles, les outils numériques peuvent m'aider à gagner un temps précieux.
Les données collectées peuvent l'être plus efficacement ; en manipulant des listes professionnelles, comme, par exemple, les listes de publications des libraires, ou les métadonnées de sites de publication ou d'autopublications.
Elles peuvent aussi être manipulées plus facilement, notamment à l'aide de l'arsenal d'outils statistiques disponibles : il est alors possible de poser des questions aux données et avoir des réponses chiffrées rapidement. Si cet output est limité, ne donnant qu'une réponse factuelle, sans contexte ou significations, elle peut permettre de rebondir sur une assomption, confirmer une intuition ou simplement pointer des faits invisibles ou difficilement visibles autrement.

Établissement des corpus

Avant de manipuler des données, il est nécessaire de choisir quelles données nous voulons.
Pour ce projet, j'ai fait le choix de documenter des ensembles cohérents par blocs. J'entends par là un ensemble d'œuvres ayant un point commun thématique ou auctorial fort. Aujourd'hui, nous verrons ainsi les corpus liés à l'univers de Star Wars, mais un autre choix aurait pu être fait.
Une fois l'ensemble délimité, choisi, il faut également délimiter les médias potentiellement utiles : quelles formes narratives allons-nous mobiliser ? Ici, les films, séries et livres, bandes dessinées ou encore les jeux vidéo officiels étaient les plus évidents. Les fanfiction et fangames étaient bien sûr incontournables. Il semblait impossible de laisser de côté les jeux de plateau, de cartes ou de rôles qui entourent l'univers.

Pour une méthodologie de documentation des corpus contemporains

Une fois le corpus choisi et une idée large de ce que je vais vouloir intégrer à la base de données, il est temps de mettre en place les templates qui vont permettre la documentation des œuvres, et de travailler concrètement avec toutes ces informations si durement accumulées.

Pour ce projet, j'utilise une base de données construite avec OmekaS. Ce système de gestion de contenus est pensé et développé pour la gestion d'archives et de bibliothèques, ce qui correspond plutôt bien aux enjeux du projet. De plus, il s'agit d'une technologie libre et ouverte, permettant d'avoir accès aux techniques mobilisées et de développer nos propres outils à partir de cette base. Les vocabulaires contrôlés sont construits sur Opentheso, un outil de thésaurus (vocabulaires) collaboratif permettant de discuter les termes employés et de les faire évoluer au besoin.

La possibilité de discussion ouverte se limite au·à la contributeur·rice du projet, mais permet tout de même d'ouvrir à d'autres personnes, tout en évitant le trolling. De plus, cette collaboration s'applique également aux traductions et définitions des divers termes, permettant d'ouvrir un espace d'une grande richesse pour parvenir à un consensus scientifique.

Quelles informations sont à conserver ?

Pour chaque œuvre, il faut décider quelles informations seront nécessaires pour les recherches à mener. S'il est possible par la suite de modifier des informations ou de les compléter, il est bien plus simple de créer des templates, des ensembles de champs d'informations qui vont servir à décrire les œuvres, les plus pertinents possibles dès le début.

Une première partie des fiches est consacrée aux informations de base de l'œuvre : le titre, le·a ou les auteur·rice·s, l'instance de publication, la date de publication ainsi que l'aire géographique d'origine ainsi que la nature de l'œuvre. Ce dernier point permet de qualifier largement l'œuvre : est-ce un roman ? Un film ? Un jeu ? À quelle grande catégorie cette chose appartient.

Ensuite, une partie se concentre sur les particularités médiatiques de l'œuvre : cette partie est propre à chaque forme – énoncées ci-dessus. Pour les œuvres vidéo, il s'agira de la durée du contenu, de l'appartenance ou non à une série, le cas échéant du nombre d'épisodes, par exemple. Ou de la forme d'un jeu : y a-t-il des cartes ? Un plateau ? Des pions ou meeples ? Quel genre de gameplay est mobilisé ? Etc.

Enfin, une troisième partie s'intéresse aux caractéristiques de l'imaginaire de l'œuvre : à quels genre et sous-genre appartient-elle ?

Et à ses particularités propres : certains corpus nécessitent des informations particulières pour être abordés : pour les réécritures mythologiques, cela peut -être les divinités ou

héros présents ; pour Star Wars, notre exemple d'aujourd'hui, il s'agira des ères historiques de cet univers.

Importer les données

Une fois les templates construits, il faut les remplir.

Généralement, en documentant les branches de corpus nécessaires, il a été possible de trouver des avenues de moissonnages de données. C'est-à-dire des endroits où il sera possible de connecter la base de données pour récupérer les informations, ou des points de sorties où il sera possible d'exporter des données pour permettre de les documenter dans notre projet sans avoir à tout remplir à la main.
Une grande part de ce travail passe par les sites de distribution à destination des libraires, qui permettent d'obtenir des listes – des documents au format CSV notamment – où toutes les publications accessibles sur le sol canadien vont être disponibles.
Cela pose plusieurs limites :

- Les œuvres non disponibles dans cette aire géographique devront être importées séparément.
- Les œuvres qui ne sont plus disponibles vont ainsi demander des recherches plus approfondies. Et parfois, l'appel à des bases de données privées, construites par des fans, qui sont souvent, pour nos corpus, tant la mémoire vivante des œuvres que les spécialistes les plus pointus.

Curer les données

Une fois toutes ces données importées dans la base de données, elles ont généralement besoin d'être nettoyées – curées, disons-nous en Humanités numériques – pour être pertinentes. Les champs sont souvent différents et demandent d'être remis en ordre.
Il faut généralement aussi ajouter des informations qui ne sont pas utiles aux bases de données d'origines : ainsi les ères historiques des narrations de Star Wars sont souvent disponibles pour les fanfictions, afin de permettre aux lecteur·rice·s spécialistes de naviguer dans l'univers étendu et tentaculaire… mais rarement pour les romans, notamment pour les plus anciens. Il faut alors les remplir à la main et vérifier globalement les données amassées.
La moindre erreur se répercutant dans les étapes suivantes, il est important de minimiser ces aléas.
Cependant, une faible quantité d'imprécision fait partie de la méthode : il faut être conscient de cette part pour pouvoir la prendre en compte lors de l'analyse finale des données.

Questionner les données

Finalement, commence la partie véritablement passionnante de tout cela.

Questionner les données est l'aboutissement de tout ce travail. Les données une fois bien documentées, les œuvres clairement labellées sont mobilisables pour répondre aux questions de recherche que chacun·e d'entre nous peut leur poser.
Chaque question doit être formulée comme un facteur documenté par rapport à un ou plusieurs autres facteurs documentés. Les méthodes de documentations actuelles mobilisent majoritairement des méthodes de comparaisons – à la manière de tableaux à doubles entrées, mais en plus complexes, en somme.
Possibilités de recherches
Comme évoqué précédemment, pour que ce projet exprime sa complète utilité, il faut lui poser des questions dont les réponses peuvent être des ensembles de nombres ou de graphiques simples. Des approches plus complexes sont en cours d'établissement, mais prenons pour l'instant des exemples simples.

De nombreuses questions traversent une grande partie des corpus de l'imaginaire : les proportions d'autrices et les évolutions de cette représentation ?
La proportion d'œuvres appartenant à des séries par rapport aux œuvres fonctionnant par elles-mêmes ?
Les évolutions du nombre de publications des différents sous-genres par an ?
Ou encore, la représentation de ces sous-genres en fonction des aires géographiques de publication…

Ces questions sont simples, mais peuvent nourrir la réflexion et la recherche sur ces corpus, permettant de montrer concrètement des évolutions globales de notre champ d'études. Gardons en tête que ce ne sont que des exemples, je n'ai clairement pas le temps de développer toutes les pistes possibles ici.

D'autres questions sont plus particulières à un corpus particulier :
Pour noter l'exemple de Star Wars, les évolutions des ères historiques représentées sont-elles un reflet des ères présentées dans les films ? Les séries ?
Le rachat de la franchise par Disney, a-t-il eu des impacts sur les aires historiques présentées ? Sur le nombre de publications ? Sur les supports mobilisés ?

Toutes les données sont libres de droits, en licence CC-BY – je tiens à la citation du projet dans l'usage de ces données pour valoriser la source et le travail investit dans la documentation menée, mais aussi pour permettre de retracer les données. Ces données étant disponibles via un *endpoint* d'API ou encore sur demande pour des exports particuliers ou des représentations qui peuvent être ajoutées au projet, j'espère pouvoir dans le futur appliquer d'autres questions au corpus constitué, et sans doute des questions qui ne se dessinent pas encore dans mon imaginaire, mais pourront aider d'autres recherches.

Conclusion et collaborations

Le but de cette présentation était de donner un aperçu d'un projet de documentation de notre corpus de recherche.
Ce projet est en cours de construction – nous attendons la réponse de financement du CRSH notamment – et n'est donc pas totalement bullet proof : des questionnements sont encore ouverts à travers lui, et demandent des recherches plus approfondies pour y apporter des réponses probantes.

Cependant, prenons un moment pour adresser un point qui me tient particulièrement à cœur à travers tout cela :
Je suis une chercheuse orientée vers les techniques et méthodologies numériques pour permettre la recherche. J'aime penser mon travail de recherche littéraire comme généraliste, puisque je m'intéresse aux mécaniques de lecture et non à un corpus précis – lié à un sous-genre, j'entends.
De plus, n'étant qu'un humain, mes capacités temporelles et conceptuelles sont limitées. Ce genre de projet de grande ampleur appelle donc à la collaboration.
Ainsi, si mon projet de travail usuel pour la Chaire de recherche du Canada sur les écritures numériques – le Répertoire des Écritures Numériques – un projet de documentation cousin du Répertorium, est un travail commun entre de nombreux laboratoires et chercheur·se·s, nous n'avons pas cette ampleur pour le Répertorium…

Je pense donc que nous sommes à un moment important d'ouverture : si je veux que ce travail soit utile plus largement que pour mes propres obsessions, il est fondamental d'impliquer d'autres chercheur·ses. Des spécialistes d'autres ensembles ou sous-ensembles d'œuvres, d'autres médias, ou encore simplement pour contribuer et parvenir à une base de données pouvant dans le futur être utile à d'autres chercheur·se·s…
Si votre recherche peut bénéficier de réponses statistiques apportées à un corpus large, ou si vous avez déjà un large corpus documenté et souhaitez contribuer à la base de données globale : écrivez-moi et discutons-en ensemble !

Bibliographies

Archiver la littérature numérique
Penser Star Wars