# Élaborer, numériser, mettre en ligne et exploiter un corpus d'auteur : exemples de deux cas pratiques en littérature

Marianne Froye*†[1] and France Marchal Ninosque*‡[1]

[1]université de Bourgogne Franche-Comté – Edition, Littératures, Langages, Informatique, Arts, Didactique, Discours [Besançon] (ELLIADD) – France

**Résumé**

À travers un dialogue entre deux expériences de chercheurs sur des corpus d'auteurs (un romancier, Louis-Combet et un poète, Frénaud), France Marchal-Ninosque et Marianne Froye vont tenter de conceptualiser l'approche scientifique de la constitution et de l'exploitation d'un corpus littéraire, selon la méthode inductive mise en place par ces deux enseignantes-chercheuses à l'Université de Bourgogne Franche-Comté (EA Elliadd).
Un retour expérimental sera porté sur les trois paradigmes du travail du chercheur : l'expertise (pour l'établissement d'un corpus), le conseil (pour la plateforme numérique conçue par un ingénieur informaticien), l'interprétation (pour l'exploitation du corpus). Un regard sera porté sur la plateforme numérique Fanum qui accueille des corpus littéraires et artistiques.

Puis la réflexion portera sur la nature du corpus littéraire devenu numérique, à la fois signifiant et matière. Une fois numérisé, un corpus implique qu'il soit pensé comme tel : sous son format numérique, ce corpus implique des choix qui lient l'iconicité et la plasticité d'un nouveau document, devenu numérique. Les choix graphiques, les liens hypertextes peuvent-ils transformer le discours qu'on peut faire à partir de l'objet originel, c'est-à-dire le manuscrit ? L'interface homme-machine qu'est une plateforme numérique de corpus littéraires (l'exemple de la plateforme Fanum de l'UFC sera développé) peut-elle influencer l'exploitation traditionnelle des archives littéraires (génétique, éditions de texte...) ?

## I. Présentation des projets : de la diversité des corpus vers une plateforme commune

À partir d'un bref historique de deux expériences, les enseignantes-chercheuses tenteront de montrer les principaux enjeux des projets en humanités numériques sur des corpus d'auteurs. Les deux expériences montrent une variété certaine : un romancier et un poète, des pratiques diverses dans le processus créatif : des brouillons relativement propres pour Claude Louis-Combet et des manuscrits très raturés et fournis pour Frénaud, mais un but commun : exploiter et interpréter les états intermédiaires de la création littéraire et concevoir une plateforme générique de visualisation et d'exploitation.

I. 1. Claude Louis-Combet (CLC) : expérience de France Marchal-Ninosque

*Intervenant
†Auteur correspondant: marianne.froye@univ-fcomte.fr
‡Auteur correspondant: france.marchal-ninosque@univ-fcomte.fr

Le travail effectué s'est composé de six principales étapes : il a fallu dans un premier temps collecter les manuscrits et les brouillons de l'auteur, photocopiés. Le projet de conservation a été initié dès les années 1990. S'en est suivie l'entreprise de numérisation d'un fonds *in vivo,* que l'auteur vivant alimentait régulièrement. Différents personnels ont numérisé le fonds et ont informé les fiches, tels des archivistes. L'ingénieur informaticien a ensuite conçu l'interface pour mettre à disposition de la communauté scientifique l'ensemble des données, tout en réfléchissant avec l'institution aux enjeux juridiques. Depuis lors, les chercheurs novices et confirmés exploitent le fonds dans le cadre de thèses, de colloques monographiques ou thématiques.

I.2. Frénaud numérique : expérience de Marianne Froye

L'expérience sur l'œuvre de Frénaud est différente dans son processus ; il s'est agi dans un premier de reconstituer le fonds qui avait été désorganisé après le legs opéré par la veuve du poète. À cette étape, il fallait essentiellement inventorier le fonds pour connaître le contenu de toutes les pochettes et boites léguées. La reconnaissance du projet par le consortium Cahier a permis des avancées très importantes dans le traitement du fonds. La numérisation exigeait des subventions peu conséquentes que les institutions ou les appels à projet peinaient à financer, car la rentabilité scientifique à court terme ne leur semblait pas patente. La confiance de Cahier en ce projet a permis d'avancer rapidement une fois la numérisation effectuée. Le processus créatif de Frénaud mêle écriture de poèmes et exégèse. Le corpus qui semblait alors le plus approprié pour affiner la connaissance de son geste d'écriture était composé de deux œuvres : *La Sorcière de Rome* et *Gloses à la Sorcière.* Le choix a mêlé arguments scientifiques, pratiques, pragmatiques et génétiques. Notre volonté était d'étudier le processus de création de Frénaud. Nous souhaitions donc avoir en regard l'œuvre poétique et son exégèse rédigée par le poète lui-même, pour mettre en évidence sa pratique d'écriture. L'ensemble le plus achevé et le plus accessible physiquement correspond à ces deux ensembles de brouillons.

I.3 FANUM[1] / FANA[2] : Multiarch

L'ingénieur informaticien de l'EA ELLIADD a conçu deux interfaces de visualisation l'une FANA[3] pour les arts du spectacle vivant, l'autre FANUM pour les manuscrits de corpus littéraires. L'évolution souhaitée est la réunion des deux plateformes en une seule, pour visualiser et exploiter l'ensemble des fonds sur MultiArch. Ces plateformes ont permis de rendre accessible la totalité de ces données à l'ensemble de la communauté scientifique. La création d'un outil générique serait un apport décisif. L'intégration récente du projet Frénaud à cette structure de recherche permet de faire évoluer l'offre de visualisation offerte par FANUM et vise à appliquer le TAL à ces œuvres littéraires pour en renouveler l'approche et en approfondir l'interprétation génétique.

II. En amont et en aval : le *corpus* comme colonne vertébrale

Le premier bilan tiré de ces différentes expériences en humanités numériques consacre l'importance du *corpus.* Elle se justifie pour différentes raisons : sans la constitution d'un corpus raisonné et réfléchi, aucune exploitation n'est possible. Des compétences nécessaires sont multiples : elles sont celles des chercheurs, des ingénieurs informaticiens, des juristes et des archivistes.

II.1 La constitution du corpus. Le chercheur expert : l'alpha

Le chercheur endosse plusieurs casquettes lorsqu'il entreprend de traiter numériquement un corpus littéraire. Il est, en amont du projet, celui qui possède l'expertise par sa connaissance du fonds à traiter, par sa capacité à déchiffrer l'écriture parfois illisible de l'auteur étudié. Son expertise sur l'histoire de sa discipline, de la critique, les études génétiques et sur l'histoire littéraire est également essentielle à la constitution optimale d'un corpus d'étude. Enquêteur à la recherche d'indices dans les manuscrits, il reconstruit le cheminement de la création, il devient de ce fait l'épicentre de multiples ressources qui sont éparses. Il est donc celui qui, en construisant un corpus fiable et solide, légitime la recherche. Étape et rôle

essentiels finalement peu visibles dans l'expertise scientifique

II.2 L'exploitation et la valorisation du corpus. Le chercheur interprète : l'oméga

L'autre rôle essentiel du chercheur se situe en aval du projet, au moment de l'interprétation des données recueillies. En conjuguant ses pratiques de recherche à celles d'autres disciplines, comme la linguistique et le TAL par exemple, le chercheur interprète des occurrences lexicales, des modifications opérées par l'auteur, ou encore des correspondances textométriques.

L'autre domaine d'expertise du chercheur est l'interprétation historique qu'il mène sur le corpus ainsi constitué. La numérisation facilite l'étude des influences d'un auteur, met en évidence les phénomènes d'intra- et d'intertextualité et permet de constituer plus facilement le réseau intellectuel explicite ou implicite de l'auteur étudié.

Ces données, dont l'accès est grandement facilité, permettent une interprétation stylistique de plus grande ampleur, grâce au traitement numérique du fonds. L'outil informatique permet une visualisation optimisée de différents manuscrits, de différentes versions d'un même extrait.

Finalement, le chercheur peut interpréter la genèse de l'œuvre. Or, la numérisation et son exploitation linguistique opèrent au moins un déplacement, voire une évolution de cette critique. Elle est possible *in vivo*, elle impose de réfléchir désormais aux statuts des différents états physiques du manuscrits : le papier, la version numérisée, le fichier sur la plateforme...

II.3 L'étape médiane. Le traitement du fonds : le chercheur consultant

Mais le chercheur et le projet scientifique ne seraient rien sans l'aide d'autres ressources humaines primordiales au traitement du fonds. Le chercheur se met en retrait pour être davantage un consultant qu'un acteur lors de cette étape médiane. L'ingénieur informaticien est la cheville ouvrière de la conception de la plateforme de visualisation. L'apport novateur du numérique est sans conteste l'accessibilité d'un fonds à l'ensemble de la communauté au-delà des limites de temps et d'espace. Or, sans sa visualisation optimale et optimisée, le fonds numérique resterait tout aussi inaccessible que ne l'est le fonds papier. L'informaticien permet donc un véritable accès au fonds constitué. Les projets en Humanités numériques seraient donc des projets Janus, à la fois tournés vers les sciences humaines et vers les sciences informatiques. L'expertise de l'ingénieur comme des chercheurs en informatique pour développer de nouvelles fonctionnalités est essentielle. Malgré les formations que les littéraires pourraient suivre, leur connaissance du code restera toujours en-deçà de celle d'un informaticien.

Les besoins humains dépassent ces deux disciplines et appellent d'autres ressources de personnel universitaire, notamment les juristes. L'accessibilité rendue possible par le support numérique pose la question des droits. Consulter un manuscrit en bibliothèque nécessitait jusqu'à présent une autorisation des ayant-droit, qui est à repenser lorsqu'une vue numérisée du même document est publiée sur un site internet. Les juristes permettent donc d'encadrer l'accessibilité à la plateforme et traitent également avec les maisons d'édition pour les auteurs qui ne sont pas tombés dans le domaine public.

Enfin, les besoins humains sont également au sein de la communauté scientifique : les projets en humanités numériques mobilisent des compétences pluridisciplinaires : les littéraires ont besoin des linguistes, des informaticiens, des historiens... Cette pluridisciplinarité favorisée par ce renouvellement épistémologique tend à une transdisciplinarité qui reste pour le moment utopique.

III. Une terminologie pour une méthodologie

L'apport du numérique pour traiter les fonds d'archives d'auteurs est indéniable, mais il implique aussi des changements de paradigme importants, que nous tenterons de concep-

tualiser dans un dernier temps. Les conséquences épistémologiques sont profondes et demandent à être éclaircies. Le numérique a notamment comme enjeu une redéfinition de la critique génétique. Il impose donc de réfléchir à une terminologie qui permet d'ordonner la pensée et d'organiser la méthodologie façonnée de manière intuitive, inductive, empirique, au gré des expériences en génétique. Elle permet de poser un regard rétrospectif pour mettre du sens sur le travail du chercheur généticien et essayer d'avancer plus rapidement et plus efficacement sur les projets suivants.

III.1 Une terminologie pour donner à voir les archives

La constitution d'un corpus, comme nous l'avons évoqué à partir de deux exemples précis, impose des choix. Le chercheur qui constitue son corpus en vue de le numériser donne une orientation scientifique à son projet. Le fruit de la numérisation est une lecture du chercheur de l'ensemble du fonds. En agissant ainsi, le chercheur donne à voir à l'ensemble de la communauté scientifique des vues numérisées qui ne correspondent plus finalement à un fonds vierge de toute entreprise scientifique comme les chercheurs pourraient y avoir accès lorsqu'ils se déplacent en bibliothèques pour découvrir des manuscrits. Le chercheur à l'origine d'un projet influe la suite des recherches génétiques sur un auteur ou sur un ensemble de manuscrits. La mise à disposition n'est pas totalement neutre. Il convient donc de réfléchir aux statuts de ces différentes archives physiques et / ou dématérialisées pour comprendre la perspective scientifique de la démarche du chercheur. En ce sens, la mise à disposition sur Nakala de données fairisées modifie en profondeur les rapports aux archives. Leur accessibilité et leur réutilisabilité sont deux apports indéniables qui invitent à ce que nous réfléchissions aux statuts des différents états ou statuts des archives.

III.2 Une terminologie pour découper / pour décrire le texte et pour l'encoder

Travailler sur des fonds littéraires de genres différents et sur des archives audiovisuelles montre toute la difficulté à trouver le vocabulaire adéquat pour décrire la réalité du document. Or, ce n'est pas uniquement une question lexicale. Le vocabulaire employé est consubstantiel à la réalité décrite. Espérer concevoir une plateforme générique impose une grammaire commune. Plusieurs raisons en sont la cause, la première est sans aucun doute l'étape de l'encodage. La seconde, celle de la visualisation. Les choix des balises de l'encodage dépendent de cette grammaire ; l'arborescence du site et sa granularité également. Que définir comme " texte " ? Que signifie " document " ?

III.3 Une terminologie pour exploiter le texte

Finalement, le changement de paradigme apporté par le numérique entraine la conception de nouveaux outils d'analyse, notamment ceux développés avec les chercheurs en informatique ou les ingénieurs de recherche. Le numérique automatise et fiabilise certains résultats, comment le chercheur en SHS s'approprie-t-il ces données ? Le décompte d'occurrences est certainement plus fiable lorsqu'une machine l'effectue que lorsque c'est un humain. Les outils conceptuels à disposition gardent-ils toute leur efficacité ? Comment peut-on les faire évoluer ?

Fonds d'Archives Numériques

Fonds d'Archives Numériques Audiovisuelles

https://fanum.univ-fcomte.fr//fana/?f=1

**Mots-Clés:** génétique, Louis Combet, Frénaud