Revisiter les textes anciens dans les bibliothèques numériques avec l'exemple de la Chanson de Roland

Revisiting old texts in digital libraries with the example of the Chanson de Roland

Jacques DUCLOY (1), Thierry DAUNOIS (2), Frédérique PÉGUIRON (2), Jean-Pierre THOMESSE (2), Isabelle TURCAN (2)

(1) Laboratoire Paragraphe, Université Paris 8
Jacques.Ducloy@univ-lorraine.fr

(2) Université de Lorraine

Résumé. Cet article décrit la création d'une bibliothèque numérique autour de la *Chanson de Roland*. Elle repose notamment sur une réédition hypertexte d'un ensemble de manuscrits et d'ouvrages proposant des transcriptions et des traductions de ces sources primaires. Elle contient également des ouvrages et articles de recherche contemporaine qui utilisent une même souche encyclopédique. Elle met également en interaction des pièces musicales avec notamment un oratorio profane composé par Gilles Mathieu.
Mots-clés. Bibliothèque numérique, Chanson de Roland, manuscrit.

Abstract. This article describes the creation of a digital library around the *Chanson de Roland*. It is based in particular on a hypertext re-edition of a set of manuscripts and works offering transcriptions and translations of these primary sources. It also contains works and articles of contemporary research which use the same encyclopedic strain. It also puts musical pieces in interaction with, in particular, a secular oratorio composed by Gilles Mathieu.
Keywords. Digital library, The song of Roland.

Avant-propos.
Cet article a été rédigé pour une meilleure lisibilité dans un environnement numérique. En particulier, ce qui relève des notes de bas de page a été volontairement omis pour être remplacé par des liens dans la version numérique.
Nous invitons le lecteur à privilégier une lecture numérique sur :
https://wicri-demo.istex.fr/wicri-chanson-roland.fr/index.php/ArticleHIS7

1 Introduction

Qu'est-ce qu'une bibliothèque numérique, au juste ?

Il y a 15 ans Carl Lagoze, un des pionniers des archives ouvertes aux États-Unis, posait cette question dans un article de référence (Lagoze 2005).

Alors qu'il les considérait comme entrées dans « *leur adolescence* » - nous étions en 2005 - Carl Lagoze soulignait que leur situation était déjà préoccupante, alors que l'idée infusait que « *Google a déjà tout résolu* ». Et il insistait notamment sur le fait que les bibliothèques ne sont pas seulement des endroits où l'on peut retrouver de l'information pour la consulter[1], mais aussi « *des lieux où des personnes se rencontrent pour accéder à un savoir qu'ils partagent et qu'ils échangent* ». Reprenant l'idée de David Levy (2000), il reliait déjà les bibliothèques aux communautés dont elles sont l'espace de référence. Il s'inscrivait dans les réflexions sur le quatrième paradigme de la recherche (Gray 2006).

Mais, quinze ans plus tard, loin d'être reconnue largement, cette idée de communauté, de la dimension « humaine » des bibliothèques, semble toujours s'effacer derrière la technique. Et pourtant !

Le terme même de bibliothèque devrait nous alerter, dans sa dimension éminemment polysémique. Car une bibliothèque, nous dit le TLF[2], est à la fois un lieu, pouvant désigner un espace (de différentes natures : « *bâtiment [...] où sont déposées, rangées, cataloguées diverses collections de livres* », « *cabinet de travail [...] qui renferme une collection de livres* », « *meuble à rayonnages destiné au rangement et au classement de livres* ». Elle peut également désigner la collection de livres elle-même (la bibliothèque de Paul Meyer a permis de constituer le fonds du même nom). Enfin, une « bibliothèque vivante » fait directement référence à l'érudit(e) dont la mémoire est particulièrement remarquable.

La transposition dans un espace numérique de ces différentes fonctions pointe à la fois vers des documents numériques, vers des entrepôts de données et de métadonnées et vers la version moderne des *scriptoria* que sont les *learning centers*.

Carl Lagoze concluait son article en signalant que les bibliothèques numériques ne devaient pas être seulement des endroits où trouver de l'information et y accéder, mais devaient également permettre « *d'ajouter de la valeur aux ressources internet* », un enrichissement lié à leur contextualisation, à leur mise en relation avec de nouvelles informations et par leur imbrication dans des réseaux de relations - modèles d'usage, savoir communautaire, réseaux sémantiques. Ainsi, disait-il, « *La bibliothèque numérique devient alors un espace pour l'information collaborative et l'enrichissement* ».

Comment ces éléments peuvent-ils constituer un cadre profitable à une œuvre comme la *Chanson de Roland* ? Quelles sont les conditions de mise en œuvre d'une telle expérimentation, quel retour d'expérience peut-on en tirer ?

2 Un projet autour de la *Chanson de Roland*

Le 15 aout 778, de retour d'Espagne, Charlemagne perd son arrière-garde, tombée, à titre de représailles, sous le feu des troupes des seigneurs basques dont il a attaqué les possessions. Lors de la bataille de Roncevaux, l'arrière-garde est écrasée, provoquant la mort de nombreux braves de l'entourage de Charlemagne, dont celle de Roland, préfet de la Marche de Bretagne. On peut imaginer, mais la

[1] Ce qui fait référence à deux problématiques, celle de la recherche d'information et celle de son accessibilité, qui constituent encore aujourd'hui des enjeux forts.

[2] Trésor de la langue française, dictionnaire du CNRS.

tradition orale en a perdu la trace, que ce fait d'armes ait été l'objet de chansons de geste et d'épopées, qui ont circulé, au gré de l'errance des jongleurs ou des troubadours, de seigneurie en seigneurie. Quoi qu'il en soit, la légende de Roland (avec par exemple la traitrise de Ganelon, le son du cor, ou l'épée Durandal qui brise le rocher) refait surface et s'inscrit matériellement sur parchemin au XII^e^ siècle, les basques ayant, pour des raisons probablement opportunistes, laissé la place dans le récit aux sarrasins.

2.1 Un corpus riche et varié

De la *Chanson de Roland* et de ses transcriptions médiévales, on connait aujourd'hui sept versions, et trois fragments. La version considérée comme la plus ancienne et la plus proche d'un hypothétique « texte initial » est le manuscrit conservé à la Bibliothèque Bodléienne d'Oxford (Digby, 23, f. 1r-72r). Communément daté du deuxième quart du XII^e^ siècle, ce manuscrit a suscité plusieurs dizaines d'éditions modernes, depuis le début du XIX^e^ siècle, a été traduit dans de nombreuses langues, et été l'objet de plusieurs centaines d'études[3].

Une analyse même sommaire des versions manuscrites de la chanson de geste permet immédiatement de comprendre la situation. Là où le manuscrit d'Oxford compte 4002 vers répartis en 291 laisses (ou couplets), la version Venise 4 - datée du XIII^e^ siècle - en compte 6011, pour 419 laisses, la version de Châteauroux, 8201 vers et 449 laisses, le manuscrit Venise 7 rassemble 8395 vers organisés en 445 laisses. Les manuscrits de Paris, Cambridge et Lyon, pour leur part, comptent respectivement 6828, 5695 et 2932 vers, distribués en 375, 354 et 216 laisses.

Cette sérieuse diversité pose donc d'entrée de jeu la question de l'alignement des textes. Sans même parler des études, dont nous avons déjà signalé le grand nombre, on comprend aisément qu'avec la *Chanson de Roland*, ses éditions modernes, et les analyses sur ces éditions, on dispose d'un corpus riche à la fois en volume et en complexité.

Mais la création artistique autour des exploits et de la mort du « neveu » de Charlemagne ne s'arrête pas aux traductions : elle a également pris la forme de diverses mises en vers, en prose, en musique, avec là aussi de nombreuses productions au fil des siècles.

2.2 Une expérience pilote en 2014

La bibliothèque universitaire du Campus Lettres et sciences humaines de l'université de Lorraine à Nancy dispose d'un fonds Paul Meyer. Celui-ci, diplômé de l'École des Chartes, philologue et romaniste, spécialiste de littérature romane, a notamment travaillé à la Bibliothèque nationale. Élu au Collège de France en 1876, il prend la direction de l'École des Chartes en 1882. À sa mort, en 1917, il choisit de léguer sa bibliothèque à l'université de Strasbourg, mais, celle-ci étant soumise aux mouvements de frontières que l'Alsace et la Moselle connaissent depuis 1870, c'est la bibliothèque de l'université de Nancy qui est chargée de l'accueillir, par mesure de précaution. C'est ainsi qu'elle abrite le fonds Paul Meyer, composé de 4222 titres de monographies et d'environ 7700 brochures, tirés-à-art et petites publications.

[3] La consultation de la bibliographie proposée sur le site arlima.net est éclairante sur la richesse et la diversité des écrits sur et autour de la Chanson de Roland.

Dans ce fonds figurent plusieurs éditions de la *Chanson de Roland*, dont certaines sont annotées de la main de Paul Meyer. Nous pouvons montrer qu'il s'agissait là d'un travail préparatoire à la publication d'ouvrages[4].

Figure 1. *Pages annotées par Paul Meyer.*

En 2014, saisissant l'opportunité d'un stage, Isabelle Turcan confiait à l'un de ses étudiants de la filière "Métiers du livre" la tâche d'explorer et d'analyser l'édition de Francisque Michel de 1869 annotée par Paul Meyer. En effet, sur sept pages du recueil, on retrouve des notes, des indications d'édition, des paperolles... Le travail de l'étudiant a consisté à rééditer sur un site web les sept pages concernées, avec trois principales présentations du texte : le texte initial de Francisque Michel ; le texte avec les annotations de Paul Meyer ; le texte tel qu'il apparaîtrait une fois appliquées les modifications indiquées par Paul Meyer.

À cette occasion, nous avons décidé d'assurer en parallèle la réédition de l'ensemble de l'ouvrage. L'étudiant ayant travaillé sur sept pages, nous nous sommes chargés de cent quinze autres pages, non annotées. Et nous avons profité de l'occasion pour effectuer une expérimentation : en annotant sémantiquement les variantes des noms de Charlemagne et de Roland, nous avons pu construire un système d'information sur les variantes (liste, nombre de pages sur lesquelles chacune est utilisée...). L'ensemble de cette expérimentation a été menée avec un seul document de référence : l'édition de 1869 de la *Chanson de Roland* par Francisque Michel.

2.3 Une bibliothèque numérique sur la *Chanson de Roland*

En mai 2021 un nouveau stage a conduit à mettre en place un projet de plus grande envergure. L'idée est de voir comment explorer et exploiter le corpus décrit précédemment, dans toutes ses dimensions et dans toute sa complexité. En effet, grâce à la numérisation d'un nombre croissant de documents, à la mise en ligne de ressources, il est désormais possible d'accéder à plusieurs de ces sources.

Tels des copistes de l'époque médiévale dans un *scriptorium*, ou comme les membres d'une société savante occupant la salle de travail d'une bibliothèque, l'ambition est de pouvoir travailler sur ces textes, d'observer leurs différences et leurs rapprochements.

De plus, grâce à un travail musical effectué précédemment sur une messe irlandaise *(Irish Mass)* du compositeur Gilles Mathieu, nous avons découvert qu'il avait également composé un oratorio profane sur la base du manuscrit d'Oxford et

[4] *Recueil d'anciens textes bas-latins, provençaux et français, accompagnés de deux glossaires*, publié en 1874.

de sa transcription par Léon Gautier. Nous avons donc eu l'idée d'effectuer le rapprochement numérique de la partition et de la transcription du manuscrit.

La structure apparente du manuscrit d'Oxford - et que l'on peut aisément imaginer en consultant l'une ou l'autre des transcriptions - s'organise autour de "laisses" - des couplets - rassemblant un nombre variable de vers. Chacune contient des vers en assonance, et commence habituellement par une lettrine. Dans le manuscrit d'Oxford, elles se terminent généralement par une mention mystérieuse, sur laquelle aucune explication n'est acceptée largement : [Aoi].

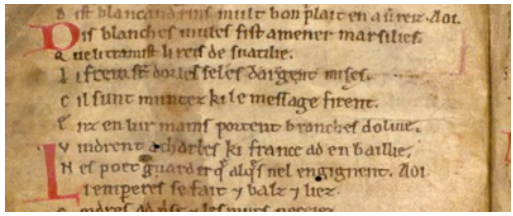

Figure 2. *Un enchaînement de 3 laisses, 2 lettrines (D et L) et 2 mentions Aoi en fin de ligne*

Lorsque l'on commence à vouloir aligner les textes des manuscrits et leurs transcriptions, on constate rapidement des divergences dans la numérotation des laisses. Ainsi, la dernière laisse du texte est numérotée CCXCI chez Joseph Bédier, CCXCIII chez Edmund Stengel, CCXCVI chez Francisque Michel et CCXCVII chez Léon Gautier.

En effet, certains philologues se réfèrent à la différenciation des laisses à l'aide des lettrines et des marques [Aoi] telle qu'elle est dans le manuscrit d'Oxford. D'autres considèrent que le copiste a fait des erreurs qu'ils cherchent à rectifier. Le feuillet 43 verso est exemplaire de ce point de vue car ne contient ni lettrine, ni mention [Aoi]. En revanche, il contient un vers qui marque une charnière essentielle entre deux parties de l'épopée : la mort de Roland.

> Morz est Rollant, Deus en ad l'anme es cels.
> *Roland est mort ; Dieu a son âme dans les cieux.*

Le manuscrit contient curieusement un point en guise de lettrine.

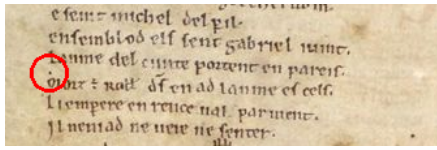

Figure 3. *Un simple point avant le vers charnière*

Bédier et Gautier considèrent ce vers comme le début d'une nouvelle laisse. Michel en fait la fin de la précédente et Stengel propose une version sans changement de laisse (et donc avec un décalage dans la numérotation).

Ainsi, et c'est ce qui motive cet article, ce travail s'est avéré à la fois plus complexe et plus riche que nous l'imaginions.

3 Expérimentation, difficultés rencontrées et solutions retenues

3.1 Une organisation numérique à définir

Nous avions déjà eu l'occasion de procéder à des rééditions numériques avec le moteur MediaWiki. Nous n'avons donc pas été surpris par la première question qui se pose lorsque l'on s'attaque à ce type d'exercice : celle du choix d'une structure éditoriale et informationnelle. En effet, depuis pratiquement 2200 ans, l'organisation classique des codex, puis des livres, longtemps reprise pour les fichiers numériques simples est celle d'un assemblage de feuillets, dans lequel on tourne des pages, avec la possibilité de feuilleter.

Cette organisation avait succédé à près de 300 ans durant lesquels le format classique était celui de la page, qu'il s'agisse de tablettes d'argile ou de papyrus.

L'année 1985 marque un premier tournant, avec l'apparition du format SGML, suivi, 10 ans plus tard, par XML, qui se caractérise par un modèle arborescent. Mais le changement n'est en général pas perceptible pour les utilisateurs.

En parallèle, et au rythme des innovations technologiques, une nouvelle approche se popularise, celle de l'hypertexte où « *l'usager navigue d'information en information par un jeu de liens d'associations entre les îlots d'informations* » (Vignaux 2001). Cette structuration s'articule autour de blocs de textes liés entre eux de manière non séquentielle. Elle modifie fondamentalement le "parcours" de lecture : de linéaire il devient non-linéaire.

Dans un hypertexte, donc, l'unité n'est plus la classique page : elle est choisie - négociée ! - au cas par cas. Là où une page donnée peut contenir très peu d'information, une autre peut contenir l'équivalent d'un livre entier. De façon très prosaïque, le premier choix est celui du découpage du texte : quelle est l'unité élémentaire la plus pratique à manipuler ? Ici, la laisse nous a semblé être la bonne unité.

3.2 Modéliser l'arborescence des sources

Le deuxième questionnement qui intervient très rapidement, c'est celui de la façon de traiter l'arborescence des sources. Et le phénomène décrit par Carl Lagoze trouve ici une illustration remarquable : certes, depuis 20 ans, nombre de ces sources ont été numérisées, et sont désormais trouvables et accessibles sur le web. Mais cela ne forme pas une bibliothèque pour autant, du fait de l'hétérogénéité des formats et des protocoles sous lesquels elles sont disponibles.

Ainsi, le manuscrit d'Oxford est accessible - en format photo - dans son intégralité sur Wikimedia Commons, avec une organisation séquentielle. Pour le manuscrit de Châteauroux, seule la première page est accessible avec un fac-similé de bonne qualité à l'IRHT[5], mais les autres pages, accessibles via le site des bibliothèques de Châteauroux, sont encombrées par une inscription de propriété. En fait presque chaque manuscrit (Venise, Cambridge) dépend de son propre service de visualisation. Il n'est pas trivial d'atteindre à un traitement identique des différents manuscrits.

De façon parallèle, pour les livres du XIX^{e} et du XX^{e} siècle, trois principales sources permettent d'accéder aux textes : Gallica, Internet Archive et Wikisource. La qualité de la numérisation et la performance des logiciels d'océrisation employés varient sensiblement entre Gallica et Internet Archive. Sur Wikisource, les

[5] Institut de recherche et d'histoire des textes

documents sont faciles à récupérer en texte intégral. En effet, ils ont déjà fait l'objet d'une curation par des contributeurs (humains !) et ils sont structurés avec des modèles MediaWiki.

3.3 Gestion des manuscrits

De façon évidente, la priorité a été donnée à la gestion des sources primaires (manuscrits), leurs transcriptions et leurs traductions. En effet, la plupart des articles plus récents contiennent des références à ces documents, souvent sous la forme de numéro de vers ou de numéro de laisse (ni les uns ni les autres n'existant dans les manuscrits).

À l'occasion d'un stage, un alignement a été tenté entre la version de Francisque Michel (1869) et le manuscrit d'Oxford. Des premiers travaux d'alignement entre l'oratorio de Gilles Mathieu et le manuscrit ont été réalisés. Dans les deux cas, des divergences ont été observées, qui n'ont pas été solutionnées, bien au contraire, en faisant appel à la version de Léon Bédier.

En même temps, l'exploration des sources a mis en évidence un ouvrage d'Edmund Stengel dans lequel la pagination suit le découpage en laisses du manuscrit d'Oxford.

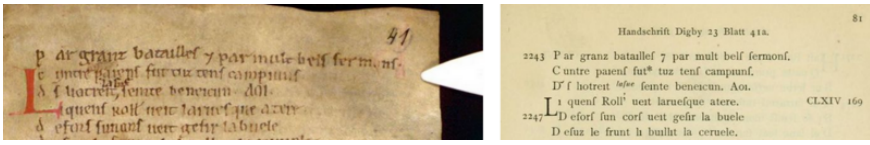

Figure 5. *Le haut du feuillet 41aligné entre le manuscrit et l'ouvrage de Stengel*

Nous avons donc décidé de gérer les manuscrits en nous appuyant sur les laisses avec notre propre numérotation.

3.4 Un hypertexte collaboratif pour supporter des traitements complexes

Plus la complexité et la richesse du corpus apparaissait, plus il devenait évident que le gestionnaire des données devait permettre de naviguer dans un hypertexte de façon collaborative. Le choix s'est porté sur la technologie des wikis sémantiques (Semantic MediaWiki), que nous avions déjà expérimentée avec le projet Wicri. Un wiki « Wicri/Chanson de Roland » a donc été créé au sein du réseau.

La gestion du manuscrit d'Oxford s'articule autour d'une première structure hypertexte basée sur les feuillets. A chaque feuillet est associé une page wiki qui est généralement organisée en 3 parties :

- pour le recto, l'association entre le fac-similé de la page du manuscrit et la transcription de Stengel (les liens sur les images sont actifs et permettent des navigations parallèles) ;
- même chose pour le verso ;
- la liste des laisses (avec des liens) avec notre numérotation.

Pour chaque laisse, une page wiki permet de retrouver le ou les feuillets dans lesquels elle est contenue.

Elle contient également un ensemble d'informations permettant de confronter les points de vue (notamment diverses transcriptions ou traductions).

3.5 Les interactions entre un oratorio profane, les manuscrits et les traductions

Pratiquement tous les documents offrant des analyses du manuscrit d'Oxford sont organisés autour des laisses. Les pages wikis correspondantes jouent donc un rôle essentiel : « la colonne vertébrale de l'hypertexte associé à un manuscrit ».

Pour constituer son oratorio, Gilles Mathieu s'est appuyé sur la transcription de Léon Gautier qui structure sa composition. Il a opéré un découpage du texte et de l'histoire en dix mouvements. Ceux-ci sont souvent proches de la mise en chapitre de l'ouvrage (exemple : *La cité sur la colline* qui correspond au *conseil tenu par Marsile à Saragoss*e). Il a ensuite sélectionné quelques vers significatifs pour les mettre en musique. Cette musique donne un éclairage particulier aux couplets concernés. La réédition de l'oratorio va donc contenir des liens vers les laisses concernées (avec parfois un décalage de numérotation).

Réciproquement, dans chaque laisse concernée, le thème musical est explicité par une ligne mélodique.

La même approche est utilisée pour les traductions et transcriptions. Les textes sont souvent dupliqués pour être plus facilement accessibles aux lecteurs. Mais, de fait, un ouvrage donné est découpé en environ 300 fragments.

Ainsi, les pages associées aux laisses gèrent les interactions entre les contenus suivants :

- des liens vers le ou les feuillets concernés,
- la liste des numérotations dans les divers ouvrages,
- quelques versions significatives, éventuellement avec les notes associées,
- s'il y a lieu, un extrait musical.

Pour la musique, la technologie utilisée repose sur le logiciel de gravure musicale LilyPond. La musique y est codée dans un langage formel dont la syntaxe rappelle celle de TeX pour les mathématiques. Voici par exemple les premières notes du thème « *Au clair de la lune* » en si bémol majeur

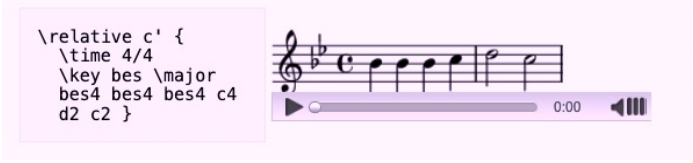

Figure 6. *Au clair de la lune en Lilypond.*

Ce mode d'interaction permet un travail collaboratif sur une ligne musicale et de faire des assemblages en fonction du contexte (présentation d'un thème relatif à un vers ou outil d'apprentissage pour choriste).

Enfin, un blog, installé sur le wiki, et intitulé « dialogue avec un compositeur », permet d'échanger avec Gilles Mathieu sur ses choix musicaux ou sa perception de l'épopée.

3.6 La base encyclopédique

Pour reprendre une analogie développée précédemment (Ducloy 2019), cet ensemble de textes est « déposé » dans une bibliothèque où une fondation encyclopédique remplace les traditionnels rayonnages et fichiers.

Lors de son lancement en 2009, le réseau Wicri était une fédération d'observatoires sur les recherches en cours. Un site était alors une encyclopédie spécialisée alimentée de façon pragmatique (en fonction de l'intérêt des sujets) et non par ordre alphabétique. Avec l'introduction des rééditions d'articles (exemple, les wikis des conférences CIDE ou H2PTM) ou d'ouvrages cette couche encyclopédique a également joué un rôle de glossaire collectif amélioré.

Ce mode de fonctionnement s'applique sur les travaux actuels sur la *Chanson de Roland*, à partir par exemple de l'identification de grandes revues (exemple *Romania*) ou des grands colloques (exemple : *Rencesvals*). Cet ensemble est structuré par des relations sémantiques devenues classiques (exemple A pour auteur pour une relation entre une page article et une page de type fiche chercheur).

La *Chanson de Roland* induit une dimension historique forte allant du Moyen Âge au XIX^{e} siècle, avec une forte dimension linguistique. Les philologues du XIX^{e} siècle sont à la fois des chercheurs qui analysent des manuscrits et des personnalités analysés par les auteurs contemporains. Tout ceci peut être modélisé par des relations sémantiques. Les aspects légendaires impliquent également de traiter des faits hypothétiques (le lieu de la bataille de Roncevaux) ou purement imaginaires. Enfin, la *Chanson de Roland* a inspiré une multitude d'œuvres romanesques, théâtrales, musicales ou cinématographiques dans différentes régions pour différents publics. La dimension imaginaire varie en fonction des œuvres dérivées (le Roland du Manuscrit n'est pas tout à fait le même que celui de *l'Orlando furioso*).

La base encyclopédique est donc potentiellement très riche et pose des problèmes de modélisation sémantique très intéressants. Pour aider à la construire, nous nous appuyons sur la réédition d'articles de dictionnaires comme celui du Trésor de la langue française avec une utilisation hypertexte des exemples et de l'étymologie. Enfin, le wiki Wicri/Chanson de Roland est inséré dans un réseau de sites utilisant la même approche. Il bénéficie par exemple d'un ensemble initial de 1.000 modèles et relations sémantiques, qu'il contribue réciproquement à enrichir.

4 Résultats, analyses et perspectives

Que montrent ces premiers mois de travail ? D'abord, ils soulignent le fait que, lorsque l'on s'aventure sur un terrain dont les frontières ne sont pas encore précisément définies, il est indispensable de disposer d'un outil adaptable et qui peut, par itérations, être reconfiguré en fonction de l'actualité du terrain. Si nous avions voulu fixer à l'avance un cadre rigide, plusieurs des interrogations signalées précédemment n'auraient pas trouvé de réponse. Le traitement des différents manuscrits nécessite une adaptabilité permanente ; les publications du XIX^{e} siècle, si elles sont un peu plus homogènes, soulèvent d'autre question ; quant à l'alignement entre texte et partition, il s'agit d'un exercice sinon inédit, du moins peu courant, et pour lequel il n'existe pas réellement de cadre réflexif stabilisé.

L'intérêt du wiki, dans ce contexte, est triple. Avant tout, il répond au cahier des charges initial, en proposant un espace hypertextuel collaboratif. D'autre part, il offre effectivement cette capacité d'adaptation, grâce à la possibilité de faire évoluer en fonction des besoins à la fois les modalités de gestion du contenu et d'affichage

des informations traitées. Enfin, la possibilité d'exploiter sémantiquement les données donne accès à des capacités de calcul particulièrement importantes.

Il faut également signaler que tous les contributeurs n'ont pas besoin d'avoir le même niveau de maîtrise de l'outil. S'il faut naturellement un noyau d'experts pour piloter l'ensemble, chaque participant peut contribuer, quel que soit son niveau de connaissance de la syntaxe spécifique de l'outil. En revanche, il faut associer également des spécialistes du domaine (ici des médiévistes ou des philologues).

4.1 Premiers résultats : quelques chiffres

Lancé en mai dernier, le site a été alimenté essentiellement par trois personnes jusqu'ici : deux stagiaires et un utilisateur expert. En seulement cinq mois, donc, ce sont près de 850 pages à contenu significatif (fiches ou articles) et 2.500 pages techniques (modèles, métadonnées) qui ont été créées.

Le résultat de la première expérimentation de 2014 a été rapatrié ; le traitement de fond d'environ 50% du manuscrit d'Oxford a été effectué, et, dans une optique de démonstration, de premiers liens ont été créés vers d'autres éléments plus disparates. Une partie significative de la partition de l'oratorio composé par Gilles Mathieu a également été traité, de façon à permettre le travail d'alignement évoqué précédemment.

On le voit, avec des ressources humaines limitées, il est possible de constituer rapidement un premier cadre démonstratif.

4.2 Gérer l'incomplétude sans égarer le visiteur

L'approche wiki permet donc de diffuser très rapidement des premiers résultats, même inachevés. L'intérêt très clair : des non-spécialistes de la technologie bénéficient ainsi d'un substrat concret sur lequel ils immédiatement travailler. Ainsi, un expert de la musicologie, un linguiste, ou un médiéviste, peut rapidement s'emparer du projet sans avoir à passer par la technique. Le revers de cette médaille est la gestion de l'incomplétude qui devient un problème omniprésent.

Dans le wiki sur la *Chanson de Roland*, la volumétrie est déjà consistante. Une amélioration minime sur le contenu des laisses (qui demanderait par exemple 2 minutes par action) peut se traduire par une dizaine d'heures de travail. Wikipédia rencontre des problèmes analogues et sait les traiter en organisant des chantiers (ou en programmant des robots). Le même type d'approche doit pouvoir se dégager ici.

Deux approches, ou deux types de chantiers, sont amenées à coexister. Pour certaines opérations, par exemple, pour numéroter les laisses d'un nouveau manuscrit, il est indispensable de travailler dans une continuité totale. À l'inverse, certaines expérimentations ont, par nature, une nature transversale, et nécessitent de parcourir quelques pages sur lesquelles sont effectuées des opérations ponctuelles. L'enjeu majeur devient alors d'assurer la cohérence du traitement malgré son émiettement.

Dans les deux cas, les visiteurs peuvent être confrontés, lors d'une exploration inopinée, à des erreurs, à des liens brisés, à une navigation rendue complexe par des situations de "rupture de phase". Il faut donc travailler sur la constitution de complétude partielle autour d'un thème donné (par exemple les quatre premiers vers du manuscrit).

Cela souligne que le wiki a une dimension éditoriale, qui ne peut et ne doit pas être négligée.

4.3 Traitement automatique des langues et fouille de données

La *Chanson de Roland* est un sujet d'intérêt notable pour des applications de traitement automatique des langues. Déjà en 1969, Joseph Duggan a publié, chez *Ohio State University Press* un ouvrage sur les concordances dans le manuscrit d'Oxford. Ce travail avait été réalisé sur un ordinateur IBM 7090 de l'Université Stanford. Plus près de nous, Jean-Baptiste Camps propose aux élèves de l'École Nationale des Chartes des travaux pratiques de codification du manuscrit d'Oxford en TEI (Camps 2017). A l'Université de Padoue, une thèse sur la lemmatisation du manuscrit de Venise (Bellotto 2020) utilise le logiciel Pyrrha de l'École des Chartes.

Lors de l'expérimentation de 2014, nous avions ainsi travaillé sur les variantes des noms de Charlemagne et de Roland (en utilisant les relations sémantiques et les requêtes associées). Le travail effectué sur le manuscrit d'Oxford montre que, sur celui-ci, le nom de Roland est en général abrégé Roll. ; pourtant, cela n'a pas empêché Francisque Michel, ou Léon Gautier, sur la base de ce manuscrit, d'employer différentes versions, Rollans ou Rollant. Et, sur une même laisse, transcrite par l'un et par l'autre, d'opter pour des versions différentes. Ainsi, si l'on prend la laisse CCVIII chez Francisque Michel, qui correspond à la laisse CCX chez Léon Gautier, là où la laisse commence par Ami Roll, le premier a opté pour Ami Rollans, le second pour Ami Rollant.

Compte tenu de la jeunesse de notre initiative, nous n'avons pas encore réalisé de traitements très significatifs. Nous avons réalisé un serveur d'exploration portant sur une extraction sur ISTEX d'environ 3.000 articles en texte intégral. Nous avons utilisé une procédure disponible sur notre boîte à outils XML (Dilib). Celle-ci a été enrichie avec des modules pour réaliser des traitements sur des sites MediaWiki (par exemple en vue de développer des robots). Elle pourra donc être utilisée quand le travail de gestion des manuscrits sera terminé.

5 Conclusion

Le projet en cours autour de la *Chanson de Roland* montre d'abord l'intérêt de nouvelles approches, sémantiques, hypertextuelles, pour les bibliothèques - et les bibliothèques numériques - dans le contexte des humanités numériques.

Il met également en évidence l'explosion de nouvelles barrières dans le contexte du quatrième paradigme de la recherche. Dans les grandes bibliothèques numériques (Gallica, HAL, Persée, OpenEdition), la bibliothèque reste un entrepôt avec ses outils de classement, et les livres sont toujours des codex. Avec MediaWiki, le même logiciel peut gérer un immense ouvrage encyclopédique (Wikipédia), un dictionnaire (Wiktionaire) ou une bibliothèque (Wikisource) mais les frontières sont respectées. Avec le réseau Wicri, et plus spécialement sur la *Chanson de Roland*, le même site héberge une bibliothèque hypertexte qui met en relation des ouvrages réédités en hypertexte.

Comme le montre cet article, les textes peuvent être construits *ex nihilo* sans qu'il soit nécessaire de les « sourcer » (comme sur Wikipédia).

Au fil de cet article, et des expérimentations que nous avons menées, une dimension a été régulièrement évoquée. Cette dimension, c'est celle de l'humain. En effet, nous avons souligné à la fois comment le projet autour de la Chanson de Roland s'apparente pour nous au travail des copistes médiévaux ou des érudits du XIX^{e} siècle ; comment la présence de contributeurs humains sur Wikisource influe sur la simplicité de réemploi des textes traités ; comment notre projet repose sur l'intervention de spécialistes de diverses disciplines.

Or on a parfois l'impression que les humanités numériques sont nées de l'étonnement de ce que la technique peut faire, accentuant la part du numérique dans les humanités. Mais une fois dissipée l'illusion d'un "tout numérique" qui remplacerait le "tout automatique" qui animait les informaticiens du XX^{e} siècle, c'est bien d'acteurs des humanités numériques au sens complet de l'expression que nous avons aujourd'hui besoin.

Enfin, les derniers articles de Carl Lagoze et de sa communauté scientifique vont de plus en plus loin dans l'interopérabilité (Plantin 2018) mais sans remettre en cause les barrières entre les concepteurs des systèmes de bibliothèque et les utilisateurs. Depuis Wikipédia, il n'y a plus de séparation formelle entre les contributeurs de la connaissance et les programmeurs spécialistes des domaines scientifiques. Cet état de fait explique la réactivité des développements sur les wikis.

Il est donc légitime de revisiter une question vieille de 15, ou 50 ans :

Qu'est-ce qu'une bibliothèque numérique, au juste ?

et d'en poser une nouvelle :

Qu'est-ce qu'un acteur des humanités numériques, au juste ?

Remerciements

Nous remercions vivement Gilles Mathieu pour sa coopération constante sur le projet. Merci aux valeureux stagiaires Dalila Ladli et Léonard Braux qui ont défriché le terrain. Merci aux équipes techniques de l'INIST et aux anciens décideurs d'ISTEX pour l'hébergement du réseau Wicri.

6 Bibliographie

Ducloy J. et al. (2019), Systèmes d'information encyclopédiques édités par les scientifiques, Revue ouverte d'ingénierie des systèmes d'information, 1, 2019

Gray J., et al. (2006). Scientific Data Management in the Coming Decade, ACM SIGMOD, New York, NY, USA

Lagoze, C. et al. (2005). What Is a Digital Library Anymore, Anyway? In: *D-Lib Magazine*, 11 2005

Levy, D. (2000), Digital Libraries and the Problem of Purpose, *Bulletin of the American Society for Information Science*, 26 (6), 2000.

Vignaux, G. (2001), *L'hypertexte. Qu'est-ce que l'hypertexte. Origines et histoire*. http://www.msh-paris.fr, 2001. edutice-00000004