

Expérimentations sémantiques autour de la Chanson de Roland

Semantic experiments around the Chanson de Roland

Jacques Ducloy^{1,*}, Thierry Daunois² and Isabelle Turcan²

¹Université Paris 8, Laboratoire Paragraphe, F-93200 Saint-Denis, France

²Université de Lorraine, F-54000 Nancy, France

Abstract

This article introduces a hypertext digital library on the *Chanson de Roland*. It collects manuscripts, critical editions, translations, research articles and musical scores. It is both a workspace for specialists in humanities and a source of information for a curious but non-specialist readers. Articles and manuscripts are republished in hypertext mode with a common semantic structure. The current demonstrator is using 3 manuscripts (Oxford, Paris, Châteauroux) and critical editions (Francisque Michel, Léon Gautier, Edmund Stengel, Joseph Bédier). Two applications are presented. Specialists can work on part of the Paul Meyer collection. Curious amateurs, for example choristers, can explore the context of a secular oratorio by Gilles Mathieu. This diversity implies taking into account various digital approaches which are experimented here with Semantic MediaWiki, and XML engineering. The generalization of this approach is studied.

Keywords

Chanson de Roland, époque carolingienne, wiki sémantique, manuscrits, bibliothèque numérique, édition critique, semantic mediaWiki, musique

Avant-Propos

Une version numérique augmentée (page de discussion, liens hypertextes et sémantiques) est disponible sur le site Wicri/Chanson de Roland ¹.

1. Introduction

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. Ce fait d'armes a inspiré des cantilènes, des récits et une chanson de geste, la *Chanson de Roland*. Ce poème épique a été déclamé dans toute l'Europe par des jongleurs et des troubadours. Quelques

. *Workshop on Digital Humanities and Semantic Web*

*. Corresponding author.

. ✉ Jacques.Ducloy@univ-lorraine.fr (J. Ducloy); thierry.daunois@univ-lorraine.fr (T. Daunois); isabelle.turcan@univ-lorraine.fr (I. Turcan)

1. https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Article_Humanum_Nancy_2022.

manuscrits ont survécu et font l'objet d'une abondante production littéraire depuis le XIX^e siècle.

Mais ces écrits n'étaient pas toujours accessibles facilement. Les manuscrits étaient enfermés dans des bibliothèques dispersées (Oxford, Paris, Venise, Châteauroux...). Les ouvrages étaient souvent édités avec une diffusion modeste à destination d'un public d'érudits comme les élèves de l'École nationale des chartes, à côté d'éditions grand public. Le numérique permet aujourd'hui d'accéder à cette littérature. Mais cette dernière est toujours dispersée sur de multiples sites qui ont chacun leurs modes d'accès.

Le fonds Paul Meyer de l'Université de Lorraine contient un document particulièrement intéressant : une édition de 1869 de « *La Chanson de Roland, ou de Roncevaux, du XII^e siècle* » de Francisque Michel [1], annotée par Paul Meyer. Celui-ci a ainsi effectué un travail préparatoire à une de ses publications [2]. Pour confronter les points de vue des deux auteurs aux manuscrits originaux, des centaines de laisses², avec leurs transcriptions et leurs traductions, sont manipulées. Ce problème est apparu comme particulièrement pertinent pour le réseau Wicri, un projet sur les bibliothèques qui gèrent des collections de documents hypertextes.

Par un concours de circonstances, nous avons travaillé avec un musicien, Gilles Mathieu, qui a composé une suite musicale à partir des mêmes manuscrits, mais sur la base d'une autre traduction [3]. Cette composition amène un nouveau point de vue qui enrichit cet ensemble. Elle ouvre également le site à un nouveau public, les choristes amateurs, qui sont des lecteurs curieux mais pas forcément érudits. Cette contrainte implique notamment de rééditer d'autres documents plus explicatifs.

Nous avons donc décidé de constituer une bibliothèque numérique spécialisée autour de la *Chanson de Roland*. Ce projet a déjà été présenté, dans sa phase de démarrage, avec un éclairage de valorisation du patrimoine écrit [4]. Nous présentons ici les premières avancées et un éclairage sur les aspects sémantiques.

Après une description des relations sémantiques dans le réseau Wicri, nous détaillerons l'organisation retenue pour les manuscrits et leurs traductions. Puis nous montrerons les premières réalisations autour du fonds Paul Meyer et de la suite musicale.

2. Les relations sémantiques dans le réseau Wicri

Le projet Wicri (Wikis pour les communautés de la recherche et de l'innovation)³ a été créé en 2008. Pour les communautés de la recherche, il explore de nouvelles approches numériques en s'inspirant des mécanismes et pratiques mises en œuvre dans Wikipédia dont le moteur (MediaWiki) favorise un développement collectif et incrémental.

Un premier réseau d'une dizaine de wikis avait été expérimenté pour valoriser les résultats de la recherche en Lorraine autour des sciences et du génie de l'environnement. Une coopération avec le Loria a ouvert l'usage des extensions sémantiques (Semantic MediaWiki). Elle avait permis de modéliser les équipements financés par le Contrat de Projets État Région (CPER). Plus tard, un système d'information évolutif sur les projets européens en Lorraine a été développé.

2. Une laisse est un couplet composé de vers ayant la même assonance (voir plus loin).

3. <https://wicri-demo.istex.fr/Wicri/Wicri/fr/index.php?title=Accueil>.

Pour ces actions, un modèle initialisé sur l'ancien site Semantic Web⁴ a été adapté pour décrire les systèmes de recherche, notamment autour des colloques.

Ce modèle a été utilisé sur la plupart wikis, et notamment, pour ceux dédiés aux communautés de colloques (notamment CIDE⁵ ou H2PTM⁶). La figure 1 montre, en 2021, l'ensemble des wikis communs en français du démonstrateur Wicri. Ils sont généralement associés à un wiki en anglais (et parfois en allemand)⁷.



FIGURE 1: Le réseau Wicri en 2021.

2.1. Un réseau de bibliothèques sur base encyclopédique

Après cette première étape sur la valorisation des résultats de la recherche, deux séries d'études ont été menées.

Pour les sciences relevant de l'ingénierie, de l'environnement et de la santé des résultats très intéressants ont été obtenus avec l'analyse statistique de corpus bibliographiques.

Un financement ISTE⁸ a permis de créer plus d'une centaine de serveurs d'explorations. Un tel outil traite des milliers de références hétérogènes (ISTEX, Pascal, HAL, PubMed). Il est créé à l'aide d'une boîte à outil XML nommée Dilib [6] dont la conception initiale a été réalisée à l'INIST [7].

Dans sa version initiale, un serveur d'exploration était généré par des commandes Unix avec un paramétrage complexe et sans accès au texte intégral. Le wiki est maintenant utilisé pour le paramétrage, la visualisation de résultats significatifs, et la curation des données. Plus précisément, les relations sémantiques utilisées dans la valorisation des innovations demandent une grande précision dans l'identification des données. Celles-ci seront utilisées pour définir les

4. Ce modèle est maintenant soutenu sur le site OpenResearch.org [5].

5. Colloque international sur le document électronique.

6. Hypertexte et hypermédia Produits, Outils et Méthodes.

7. Malheureusement, avec des porteurs alors « proches de la retraite » les travaux ont été poursuivis, mais avec des moyens humains limités à un retraité (et le financement d'un demi-poste d'ingénieur d'études pendant la durée du programme ISTE). Paradoxalement, cet état de fait est significatif pour apprécier les performances de cette approche.

8. ISTE (Initiative d'excellence de l'Information Scientifique et Technique) projet retenu dans le cadre du programme « Investissements d'Avenir ».

règles de curation. Par exemple, l'Université de Groningue est localisée à Groningue dans une région éponyme des Pays-Bas. Le modèle sémantique contient alors des triplets tels que :

Groningue (ville) A pour région:: Groningue (province)

Les règles de curation vont utiliser cette nomenclature pour inférer des mentions géographiques à partir de la mention d'une université dans une affiliation. Voici un exemple qui peut être activé avec « *Rijksuniversiteit Groningen* ».

Université de Groningue	Rijksuniversiteit Groningen ; University of Groningen	country : Pays-Bas ; region : Groningue (province) ; settlement @type=city : Groningue (ville)
-------------------------	--	--

FIGURE 2: Un exemple de règles de curation exprimées avec des tables MediaWiki.

Pour les humanités, cette approche donne des résultats plus limités. En effet, des sources de données très structurantes comme Pascal ou PubMed ne sont plus utilisables. De plus, les corpus ISTEEX sont souvent constitués de « books review » qui traitent de sujets variés rassemblées dans un même document numérique. Les résultats statistiques donnent alors des corrélations aberrantes⁹. En revanche des résultats très pertinents ont été obtenus avec des rééditions hypertextes (et sémantisées) de documents anciens (libres de droit).

Le premier résultat significatif a été obtenu avec un ouvrage sur le Palais ducal de Nancy¹⁰. À partir d'un facsimilé en mode « image + OCR » sur Gallica, nous avons notamment montré comment transformer en hypertexte une gravure de fin de volume (Fig. 3). Elle contenait des liens, matérialisés par des lettres, qui pointaient vers un hypertexte de paragraphes descriptifs qui eux-mêmes renvoyaient à des pages du livre.

Cette approche a été généralisée dans les articles scientifiques pour quelques colloques. Puis, dans un wiki sur la musique, des entrées du dictionnaire de Jean-Jacques Rousseau ont été réédités avec la possibilité d'écouter les partitions. Ainsi, un dictionnaire devient alors un document totalement hypertextuel (là où par exemple Gallica conserve une vision linéaire).

En appliquant cette approche au dictionnaire TLF¹¹, les auteurs cités deviennent alors des points d'entrée potentiels. Ainsi, sur un wiki dédié à la santé nous avons pu associer à une réédition d'un ouvrage de Claude Bernard¹² de nombreux articles du TLF.

Dans le réseau Wicri, un site wiki devient donc une bibliothèque spécialisée qui utilise une base encyclopédique pour mettre en relation des ouvrages réédités. Il devient également un

9. Par exemple, un corpus ISTEEX de 1 500 documents sur le compositeur William Byrd donne 360 mentions de l'Islam (dont aucune n'est significative).

10. [https://wicri-demo.istex.fr/Wicri/Europe/France/GrandEst/Lorraine/Nancy/fr/index.php/Le_Palais_ducal_de_Nancy_\(1852\)_Lepage](https://wicri-demo.istex.fr/Wicri/Europe/France/GrandEst/Lorraine/Nancy/fr/index.php/Le_Palais_ducal_de_Nancy_(1852)_Lepage).

11. Trésor de la langue française, dictionnaire du CNRS, Nancy, 1971-1994, publication papier; 2004, édition numérisée sur cédérom; consultation en accès libre sur : <https://www.atilf.fr/ressources/tlfi>.

12. [https://wicri-demo.istex.fr/Wicri/Sante/fr/index.php/Introduction_médecine_expérimentale_\(1865\)_Bernard](https://wicri-demo.istex.fr/Wicri/Sante/fr/index.php/Introduction_médecine_expérimentale_(1865)_Bernard).

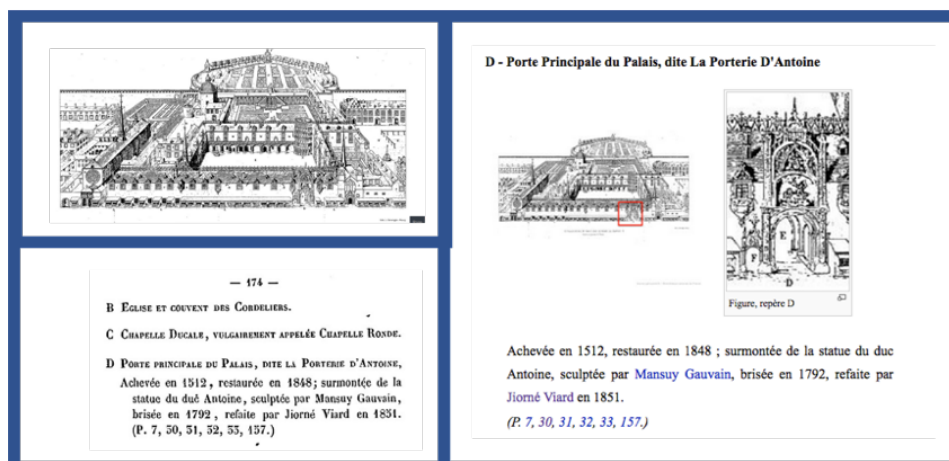


FIGURE 3: Le Palais ducal : à gauche la gravure et une rubrique (D) avec des renvois ; à droite, le développé de la rubrique D en hypertexte.

espace de travail, où il est, par exemple, possible de piloter collectivement des explorations de corpus.

2.2. Les relations sémantiques en réseau

Dès le lancement du réseau Wicri la cohérence terminologique et sémantique du réseau a fait l'objet d'investigations [8]. A titre d'exemple simple, la figure 4 montre l'alignement des éléments géographiques entre les wikis du réseau.

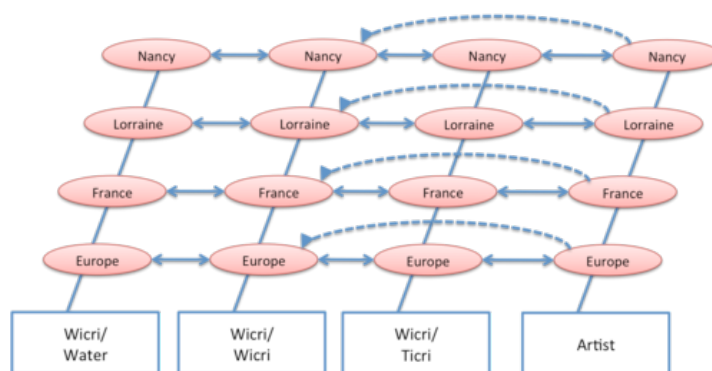


FIGURE 4: Alignement des relations géographiques entre les wikis.

Cette cohérence est basée sur un alignement sur le Web sémantique. Plus précisément les noms de page sur les wikis sont, si possible, les mêmes, que ceux de Wikipédia. Pour favoriser cet alignement, de nombreux modèles (par exemple la « palette des régions administratives de

France ») sont importés de Wikipédia et éventuellement adaptés. Ces modèles communs sont gérés sur un des wikis du réseau (Wicri/Base). Ils sont regroupés en collections pour favoriser des opérations d'exportation (depuis Wicri/Base) vers les wikis cibles. Actuellement, tous les wikis sont sur le même site physique, et ces actions sont réalisées par des traitements par lots.

Un autre mécanisme, nommé wiki de référence, est également utilisé. Par exemple, l'Université McGill a naturellement Wicri/Canada pour wiki de référence. Lorsqu'une activité significative de cette université est détectée sur un autre wiki, par exemple sur Wicri/Musique, une page spécialisée y est alors créée. Sur celle-ci, un lien interwiki pointe vers la page de référence (sur Wicri/Canada). Enfin, sur ce dernier, un lien est établi vers Wicri/Musique. Ces opérations sont en fait très rapides pour des entités déjà signalées. Cela dit, la création d'un nouveau wiki demande une adaptation du réseau. Par exemple, avant la création de Wiki/Canada, les entités canadiennes étaient sur Wicri/Amérique. Il a donc fallu passer quelques heures pour mettre à jour le réseau de liens¹³. Le maintien de la cohérence du signalement des universités françaises en mutation permanente s'avère nettement plus complexe et montre la nécessité d'une administration terminologique, et surtout éditoriale.

3. Les manuscrits et leurs éditions critiques

Nous venons de présenter la structure d'accueil de l'expérimentation sur la *Chanson de Roland*. Nous allons maintenant introduire les ressources bibliographiques fondamentales de ce sujet : les manuscrits originaux et les éditions critiques associées. Dans une bibliothèque universitaire classique, ce sujet occupe quelques décimètres de rayonnage sous la forme de quelques livres de références (Francisque Michel, Léon Gautier, Joseph Bédier, Joseph Duggan, etc).

Ici, pour permettre des études comparatives, ces quelques livres vont alimenter, à moyen terme, un réseau hypertexte de plusieurs dizaines de milliers d'articles.

3.1. Un corpus riche et varié

3.1.1. Les manuscrits

De la *Chanson de Roland* et de ses transcriptions médiévales, on connaît 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 fait l'objet de plusieurs centaines d'études¹⁴.

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,

13. Une telle opération pourrait assez facilement être partiellement automatisée par un robot.

14. La consultation de la bibliographie proposée sur le site arlima.net est éclairante sur la richesse et des écrits sur la *Chanson de Roland* : https://www.arlima.net/qt/roland_chanson_de.html.

comptent respectivement 6828, 5695 et 2932 vers, distribués en 375, 354 et 216 laisses. Chaque manuscrit possède sa propre variante linguistique — par exemple, Venise 4 est en italien francisé. Les mécanismes de versification sont variables, de l'assonance à la véritable rime.

Ces manuscrits sont organisés en laisses. Une laisse est une suite de vers avec une unité de versification (assonance sur le manuscrit d'Oxford), et généralement matérialisée par une lettrine (Fig. 5). Dans le manuscrit d'Oxford, elles se terminent par une mention mystérieuse [Aoi], sur laquelle aucune explication ne semble unanimement acceptée [9].

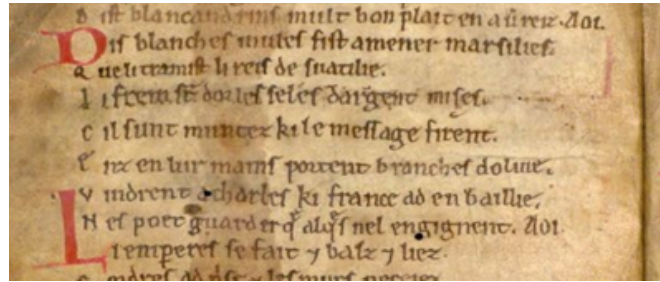


FIGURE 5: Un enchaînement de 3 laisses, 2 lettrines (D et L) et 2 mentions Aoi en fin de ligne.

Au-delà de la forme poétique chaque laisse contient une partie du récit. Une grande majorité de laisses traitent des mêmes faits (avec cependant des variantes locales) sur les différents manuscrits. Voici par exemple le début de la première laisse dans le manuscrit d'Oxford :

Carles li Reis, nostre emperere magnés,
Set anz tuz pleins ad estet en Espagne
Cunquist la tere tresqu'en la mer altaigne.

Sur le manuscrit de Châteauroux, ce passage devient :

Challes li rois à la barbe grifaigne
Sis anz toz plens a esté en Espagne,
Conquist la terre jusque la mer alteigne

Pour les lecteurs non familiers avec la langue romane voici la traduction donnée par Léon Gautier pour le manuscrit d'Oxford :

Charles le roi, notre grand empereur,
Sept ans entiers est resté en Espagne :
Jusqu'à la haute mer, il a conquis la terre.

Les chiffres donnés plus haut sur le nombre de vers et de laisses montrent une très grande variété de situations (ajout ou retrait de vers, éclatement de laisses, etc.).

Ces laisses sont distribuées sur des feuillets avec un découpage basé généralement sur un nombre de lignes par page ou par colonne. Une laisse peut ainsi être à cheval sur plusieurs feuillets.

D'un point de vue informatique, la colonne vertébrale de ce rayonnage numérique est donc une juxtaposition de 2 arborescences (avec ou sans le niveau feuillet) et des relations pas toujours

binaires entre les laisses. Sur cette base, l'interprétation donnée par les philologues introduit un nouveau niveau de complexité.

3.1.2. Divergences entre les transcriptions et éditions critiques

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, alors qu'ils sont censés avoir travaillé sur le même manuscrit de départ (en l'occurrence, le manuscrit d'Oxford).

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 sont plus attentifs à la versification. Certains ont eux-mêmes commis une erreur de numérotation. D'autres enfin considèrent que le copiste a fait des erreurs qu'il faut rectifier. Le feuillet 43 verso est exemplaire de ce point de vue car il 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 cieus.

Le manuscrit contient curieusement un point (en guise de lettrine ?), avant ce vers (Fig. 6).

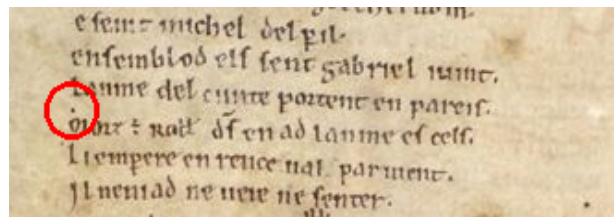


FIGURE 6: Le verset 43 verso.

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).

3.2. Gestion numérique des manuscrits et des éditions critiques

À partir d'investigations menées dans le cadre d'un stage, nous avons confronté le manuscrit d'Oxford avec les versions de Francisque Michel, Léon Gautier et Joseph Bédier.

Dans un premier temps, nous avons demandé à l'étudiant de réaliser un alignement entre le manuscrit d'Oxford et la version de Francisque Michel. Plus précisément, les laisses étaient identifiées (au sens numérique) en utilisant la numérotation de Michel. Malheureusement cette approche était insuffisante pour prendre en compte de façon précise les analyses de Gautier et de Bédier. Nous avons donc décidé de gérer les manuscrits en nous appuyant sur les laisses visibles par un public non forcément érudit, et avec notre propre numérotation.

En même temps, l'exploration des sources a mis en évidence un ouvrage d'Edmund Stengel [10] dans lequel la pagination suit le découpage en laisses du manuscrit d'Oxford.

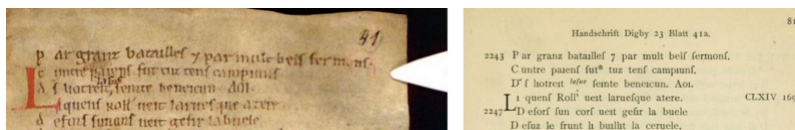


FIGURE 7: Le haut du feuillet 41 aligné entre le manuscrit et l'ouvrage de Stengel

En utilisant les mécanismes de modification propres aux wikis, nous avons pu transformer l'organisation numérique de l'application¹⁵. La gestion du manuscrit d'Oxford s'articule maintenant autour d'une première structure hypertexte basée sur les feuillets. A chaque feuillet est associée une page wiki qui est généralement organisée en 3 parties¹⁶ :

- 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. La suite de l'article montrera qu'elle contient également un ensemble d'informations permettant de confronter les points de vue. Sur cette base, nous allons maintenant aborder deux expérimentations.

- La restitution des annotations de Paul Meyer sur l'édition critique de Francisque Michel. Plutôt destinée à un public de spécialistes (érudits), elle permet de tester l'organisation décrite ici.
- Le traitement de l'oratorio de Gilles Mathieu, et plus particulièrement de son livret afin de permettre au choriste de comprendre le contexte de ce qu'il interprète.

4. Le fonds Paul Meyer pour les spécialistes

En 2014, un étudiant de la filière "Métiers du livre" avait eu pour mission de stage l'exploration et l'analyse de l'édition critique de Francisque Michel de 1869 annotée par Paul Meyer. Suite au travail sur le Palais ducal cité plus haut, le projet Wicri a été sollicité pour aider à produire une version numérique de cette annotation. Ce travail initial a été réalisé, au sein d'un wiki dédié aux collections de la bibliothèque de l'Université de Lorraine, et donc dans un contexte très général.

Cette réalisation est maintenant intégrée dans une bibliothèque spécialisée sur la *Chanson de Roland*, où elle bénéficie d'interactions hypertextes et sémantiques avec les manuscrits et les autres ouvrages sur le sujet.

15. Ceci entraîne naturellement quelques incohérences temporelles mais évite un arrêt de l'application.

16. Exemple le feuillet 41 : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Chanson_de_Roland/Manuscrit_d'Oxford/Feuillet_41.

4.1. Le fonds Paul Meyer

La bibliothèque universitaire du Campus Lettres et sciences humaines de l'université de Lorraine à Nancy dispose d'une archive nommée *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 a choisi de léguer sa bibliothèque à l'université de Strasbourg; mais celle-ci était soumise aux mouvements de frontières que l'Alsace et la Moselle connaissent depuis 1870. C'est donc la bibliothèque de l'université de Nancy qui a été 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-part et petites publications, dont une cinquantaine d'éditions de la *Chanson de Roland*.

4.2. Francisque Michel annoté par Paul Meyer

Dans ce fonds figurent donc plusieurs éditions de la *Chanson de Roland*, dont certaines sont annotées de la main de 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 corrections et des indications d'édition.

4.2.1. Une première expérimentation sur une partie d'ouvrage

Dans cette première expérimentation (en 2014), l'objectif principal était de produire une version OCR correcte d'un texte imprimé avec des annotations.

La mission de stage consistait donc à traiter les annotations manuscrites pour les restituer sur le web. Les pages ainsi traitées présentaient quatre versions :

- un fac-similé de l'original annoté,
- le texte numérisé de Francisque Michel avec ses commentaires,
- le texte avec les annotations de Paul Meyer,
- la version obtenue en intégrant les annotations.

Le travail effectué par l'étudiant ne portant que sur 7 pages, nous avons effectué en parallèle la réédition des 115 autres pages du livre, afin de disposer d'un espace d'expérimentation plus complet.

Dans le contexte ISTEEX, deux serveurs d'exploration ont été développés : un sur la *Chanson de Roland*, l'autre sur la philologie.

Enfin, en annotant sémantiquement les variantes des noms de Charlemagne et de Roland, un système d'information a été construit (liste, nombre de pages sur lesquelles chacune est utilisée...) en utilisant des relations sémantiques. Ici 2 types de relations ont été utilisées :

- « A pour variante de Charlemagne:: » entre une page de F. Michel et la page wiki d'une variable donnée, par exemple « Carles ».
- « Est une variante orthographique de :: » entre les différentes variantes et la page Charlemagne.

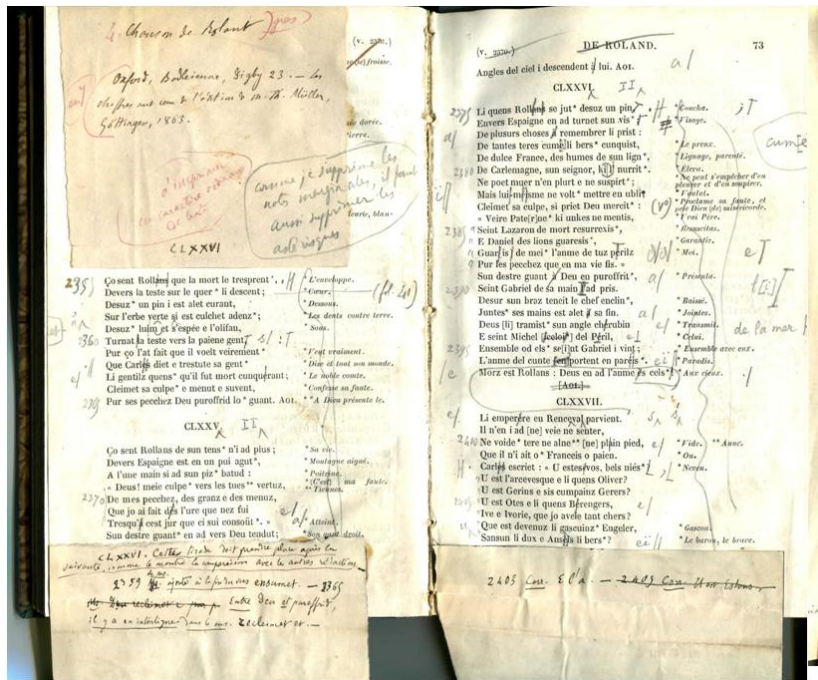


FIGURE 8: Exemples d'annotations.

Cette première expérience a été montée sur un wiki (collections de la BU Lettres de Lorraine) avec une simple juxtaposition avec d'autres travaux relativement indépendants. Elle est maintenant intégrée à une bibliothèque spécialisée.

4.2.2. Où la bibliothèque ouvre le paysage

Les travaux sur Paul Meyer ont été menés en parallèle avec l'expérience musicale décrite plus loin. Celle-ci repose sur une autre transcription : celle de Léon Gautier.

Nous avons décrit dans la section précédente la gestion des éléments des manuscrits, au départ celui d'Oxford. Les laisses du manuscrit sont devenues un lieu d'interconnexion entre un manuscrit, deux versions critiques et des commentaires.

Sur cette base, nous avons entrepris de compléter le traitement des annotations de Paul Meyer sur Francisque Michel. En effet, l'ouvrage de Francisque Michel contient deux parties. La première est dédiée au manuscrit d'Oxford. La deuxième, nommée *Roman de Roncevaux*, est principalement basée sur le manuscrit de Paris. Elle est également annotée. Nous avons donc décidé de traiter le manuscrit de Paris. Le début de celui-ci est malheureusement tronqué, et Francisque Michel a utilisé le manuscrit de Châteauroux (qui est donc également traité) pour le début de son *Roman de Roncevaux*.

Un chantier est donc en cours pour généraliser l'approche testée avec le manuscrit d'Oxford. Le modèle numérique s'est avéré stable. En revanche la maîtrise de l'hétérogénéité des sources est plus complexe. Par exemple, pour le manuscrit de Châteauroux, seule la première page est

disponible avec un fac-similé de bonne qualité à l'IRHT, mais les autres pages, accessibles via le site des bibliothèques de Châteauroux, sont encombrées par une inscription de propriété. En fait chaque manuscrit (Venise, Cambridge) dépend de son propre service de visualisation.

Nous commençons donc à bénéficier d'un dispositif qui permet de confronter deux expertises sur trois manuscrits. L'étape suivante est l'ouverture vers d'autres éditions critiques, et notamment celles de Léon Gautier ou de Joseph Bédier. Les unités numériques de « confrontation » sont naturellement les laisses, mais également les vers et les notes.

Trois principales sources sont actuellement utilisées : Gallica, Internet Archive et Wikisource. Les deux premières offrent un OCR linéaire brut, avec, là encore, des protocoles différents. Wikisource est une source particulièrement intéressante car elle fournit du document « prêt à l'emploi ». Ainsi, avec l'édition critique de Léon Gautier à *l'usage des classes de seconde* publié en 1881, on peut générer un hypertexte de plusieurs milliers de nœuds potentiels (laisses transcrites, traduites, vers, notes sur les vers).

Avec plusieurs documents de ce type, le problème est de concilier une bonne lisibilité par un lecteur humain et la possibilité de réaliser des traitements informatiques. L'approche actuellement testée est basée sur une duplication partielle de ces documents. D'une part, une version arborescente du document est générée en s'appuyant sur les chapitres avec une mise en paragraphe des laisses. Les notes, initialement repoussées dans les annexes (ou dans un autre tome) sont intégrées dans les chapitres numériques. D'autre part, les éléments intéressants sont intégrés dans le graphe des laisses des manuscrits. Ainsi, une laisse dans cet espace expose la diversité des points de vue sans chercher à l'exhaustivité de points de vue communs.

Pour les traitements informatiques, MediaWiki permet d'insérer des annotations en XML. Elles sont notamment utilisées pour réaliser des programmes d'extractions sur des ensembles de pages (sélectionnées par exemple sur un critère sémantique).

4.2.3. Un premier résultat sur les annotations

La valorisation du fonds Paul Meyer a conduit à rechercher ses travaux sur la *Chanson de Roland* pour les intégrer à la bibliothèque numérique. Or Paul Meyer a édité un recueil d'anciens textes bas-latins [2], provençaux et français. On y trouve des extraits relatifs à la *Chanson de Roland*. Nous avons pu constater que des annotations portées sur la version de Francisque Michel se retrouvaient dans le recueil.

4.3. Autour de la revue *Romania*

Paul Meyer est le fondateur de la revue *Romania* qui contient de très nombreux articles sur la *Chanson de Roland*. Ces articles portent naturellement sur l'ensemble des manuscrits et sur les analyses critiques. Ils font de très nombreuses références aux laisses et aux vers. Toutes ces références seront implantées sous forme de liens qui vont compléter cet hypertexte.

Pour le réseau Wicri, cet ensemble devient une description d'un système de recherche dont les relations sémantiques sont relativement classiques. La figure 9 donne un exemple autour d'un article qui traite de « l'accident du vers 2242 ». Ce papier montre comment le copiste a mis par erreur un vers en fin d'un autre feuillet que celui où il devait être copié. La réédition de l'article de *Romania* va donc contenir également des liens vers les manuscrits.

Autour de Paul Meyer

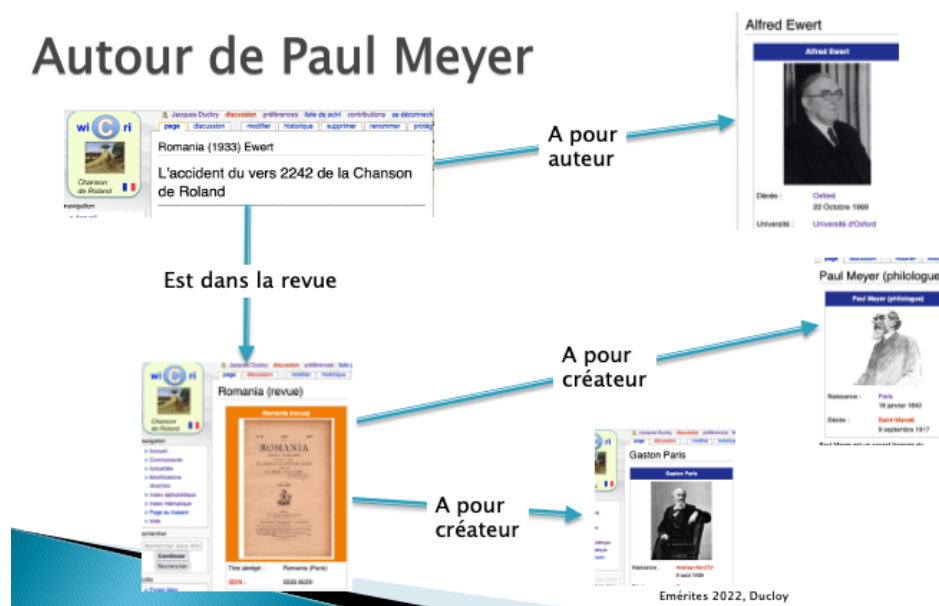


FIGURE 9: Autour d'un article de la revue *Romania*.

5. Un oratorio pour un public amateur

En complément de ce travail pour philologues, nous souhaitons ouvrir notre réflexion vers un plus large public. Dans un autre contexte, nous avons réédité en hypertexte une messe irlandaise (*Irish Mass*) du compositeur Gilles Mathieu¹⁷. Or celui-ci a composé un oratorio profane sur la base du manuscrit d'Oxford (dans la version de Léon Gautier). Nous avons donc entrepris d'étudier le rapprochement numérique de la partition et de la transcription du manuscrit.

5.1. Réédition hypertexte d'un oratorio

Pour constituer son oratorio, Gilles Mathieu s'est donc appuyé sur la transcription de Léon Gautier. Il a organisé son livret en dix mouvements. Ceux-ci sont souvent proches de la mise en chapitre de l'ouvrage (exemple : *La cité sur la colline* correspond au *conseil tenu par Marsile à Saragosse*). Il a ensuite sélectionné quelques vers significatifs pour les mettre en musique. Cette musique donne alors un éclairage particulier aux couplets ainsi concernés.

La réédition de l'oratorio va donc contenir des liens vers les laisses correspondantes (avec souvent un décalage de numérotation entre celle qui est citée dans le livret et celle donnée par Wicri). Ainsi, pour chaque mouvement, un paragraphe regroupe, par laisse dans un tableau, l'ensemble des vers utilisés¹⁸.

De plus, l'analyse de la partition montre que, dans un mouvement donné, les phrases musicales

17. [https://wicri-demo.istex.fr/Wicri/Musique/fr/index.php/Irish_Mass_\(Gilles_Mathieu\)](https://wicri-demo.istex.fr/Wicri/Musique/fr/index.php/Irish_Mass_(Gilles_Mathieu)).

18. Voici un exemple avec le deuxième mouvement : [https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Chanson_de_Roland_\(Gilles_Mathieu\)/2_-_La_cité_sur_la_colline](https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/Chanson_de_Roland_(Gilles_Mathieu)/2_-_La_cité_sur_la_colline).

sont généralement associées à une laisse du manuscrit. Pour chaque mouvement, nous avons donc introduit un ensemble de pages de détail, identifiées par un intervalle de mesures. Dans une telle page, les vers sont rappelés avec leur traduction et un pointeur donne accès à laisse correspondante. Les partitions sont données par voix et par instrument, avec une version *tutti*.

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

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 10: Au clair de la lune en Lilypond.

Ce mode d'interaction permet un travail collaboratif sur une ligne musicale et la réalisation des assemblages en fonction du contexte (présentation d'un thème relatif à un vers du manuscrit 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.

5.2. Apports encyclopédiques et sémantiques liés à la vulgarisation

La réédition de cet oratorio veut offrir au choriste, ou au mélomane, une meilleure compréhension du contexte de l'œuvre interprétée ou écoutée. Mais les éditions critiques sont avant tout destinées à un lectorat érudit. Notre bibliothèque doit donc offrir des ouvrages accessibles à un large public.

Le site étant en accès ouvert, les contraintes juridiques limitent très fortement l'utilisation d'éditions modernes¹⁹. Nous avons réédité une version dite populaire et rédigée par Léon Gautier en 1895²⁰. Elle est effectivement abordable par un public amateur. Mais elle fait appel à de très nombreuses connaissances, parfois décalées (comme les connaissances religieuses entre les XIX^e et XX^e siècles). Son contenu va donc servir de base pour identifier la base d'un glossaire au niveau du wiki (et pas seulement de l'ouvrage), et ainsi enrichir la base encyclopédique.

Pour améliorer un espace explicatif, un conservateur procède à des acquisitions. Sur Wicri, le documentariste réalise de nouvelles rééditions pour que l'amateur qui découvre le monde des manuscrits puisse en savoir plus. Par exemple, nous envisageons de rééditer le texte d'Eginhard

19. Paradoxalement, les articles de recherche, donc destinés aux érudits, sont plus facilement exploitables avec les nouvelles pratiques de la Science Ouverte.

20. https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/La_Chanson_de_Roland/Léon_Gautier/Édition_populaire/1895.

(*Vita Karoli Magni*) qui cite la bataille de Roncevaux en 830. La même remarque s'applique à Rutebeuf qui cite Roland dans la *Complainte d'Outremer* au XIII^e siècle.

De même, l'exploration du paysage correspondant à l'œuvre de Gilles Mathieu conduit à situer cette pièce dans l'histoire poétique et musicale de Roland, comme par exemple l'*Orlando Furioso* de Ludivico Arioso qui a inspiré Vivaldi, Lulli ou Charpentier.

Par rapport à la valorisation du fonds Paul Meyer qui relève d'un contexte professionnel, les besoins de la vulgarisation demandent en fait un approfondissement bien plus important. C'est également vrai sur le plan de la structuration sémantique de la bibliothèque numérique.

Ici, le contexte très spécialisé de *la Chanson* dans le réseau multidisciplinaire Wicri offre un champ d'expérimentation très intéressant. Plus précisément, pour de nombreuses notions, il faut faire cohabiter plusieurs contextes historiques. Par exemple, avec une vision nationaliste, l'Europe au temps de Charlemagne n'est pas celle de notre temps, ni celle de Léon Gautier en 1881 après la Guerre de 1870. La page Europe sur le wiki *Chanson de Roland* sera donc très différente de celle du wiki Wicri/Santé.

Le problème se complique avec les relations sémantiques. Dans pratiquement tous les wikis la capitale de la France est Paris. Ici Paris est bien la capitale de la France pour les auteurs d'articles français. En revanche, l'Empire de Charlemagne, qui n'est pas exactement la France, a pour capitale Aix-la-Chapelle.

Un autre niveau de complexité est introduit par la nature des faits. « Charlemagne, fils de Pépin le bref, est mort à Aix-la-Chapelle » est historiquement vrai. « Roland est mort à Roncevaux » est probablement vrai. « Turpin a été archevêque de Reims » est vrai. Enfin, « Turpin est mort à Roncevaux » est historiquement faux mais légendaire.

6. Bilan et perspectives

Le projet Wicri étudie (au sens preuve de concept) la diversification de l'offre de connaissance scientifique ou culturelle face au monopôle Wikipédia. La figure 11 donne la croissance financière de la Wikimedia Foundation. Elle montre un profond changement de profil financier depuis sa création. Avec un chiffre d'affaires de 120 millions de dollars, basée sur des contributions anonymes, cette compagnie peut-elle garantir le maintien de sa politique citoyenne initiale ?

Sur l'information stratégique, le projet Wicri s'appuie sur les mécanismes d'interopérabilité qui avaient fait le succès des réseaux de coopérations autour des bases Pascal et Francis en France. Ils sont mis en œuvre actuellement pour la santé par le NIH aux États-Unis autour de PubMed. Nous avons un axe de réflexion pour étudier comment un ensemble d'opérateurs européens comme l'INIST pourrait organiser un réseau de valorisation des résultats de la recherche en relation étroite avec les communautés scientifiques.

De façon complémentaire, le projet Wicri/*Chanson de Roland* veut plutôt étudier l'usage des technologies wikis sémantiques (et ingénierie XML) dans les humanités numériques, par exemple pour fédérer les travaux de chercheurs travaillant sur un même sujet.

6.1. Performances techniques

En 2021, les moyens affectés à ces 2 projets ont été limités à un retraité à temps plein et à 2 stagiaires pendant 2 mois (soit 3 semaines de formation pour 15 jours effectifs). Signalons

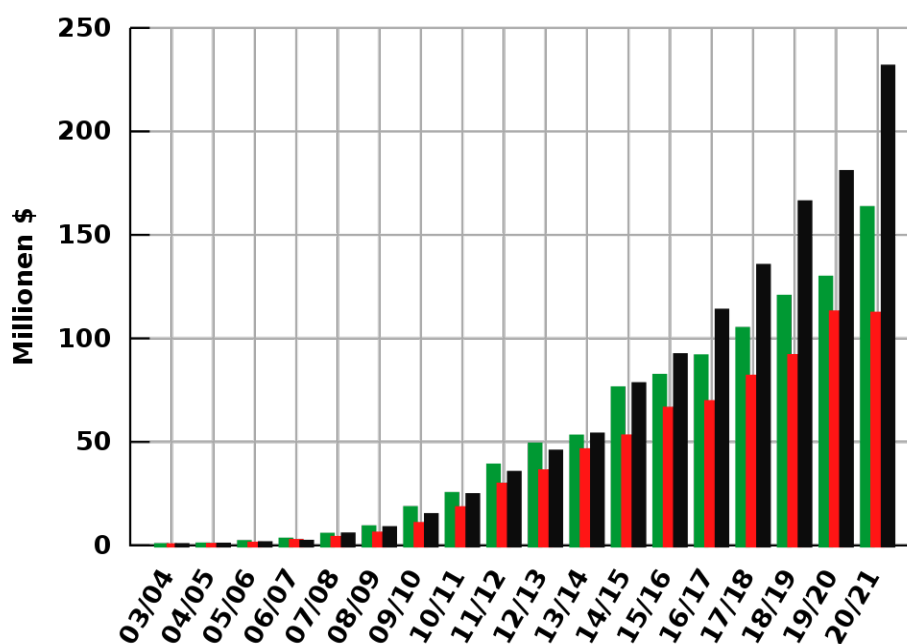


FIGURE 11: La croissance du chiffre d'affaires de la Wikimedia Foundation.

TABLE 1

Indices de production sur les wikis (janvier 2022).

	Pages wiki	Avec contenu	Modifications	Sémantique
<i>Chanson de Roland</i>	5 056	1 731	15 738	18 560
Histoire de l'IST	1 839	455	3 798	36 301
Association des émérites	1 562	216	2 250	17 479

également un soutien logistique, limité à quelques demi-journées, mais de haut niveau technique, pour l'hébergement sur le réseau de l'INIST.

Pendant cette période une procédure de changement de version a été entreprise sur une centaine de wikis et trois nouveaux sites significatifs ont été créés (celui sur la *Chanson de Roland*, un sur l'Histoire de l'Information Scientifique et Technique et un autre pour l'association des émérites de Lorraine). Sur ces 3 wikis, le tableau 1 donne les chiffres de production.

Le nombre de pages fait l'objet d'un double comptage. La première colonne donne un nombre total, avec par exemple les pages qui contiennent les modèles ou les déclarations de catégories. Lorsqu'un wiki est initialisé, environ 900 pages de ce type sont chargées (depuis Wicri/Base cité plus haut). La mention « avec contenu » repère les pages de l'espace principal. Cette production a été atteinte en fait avec 6 environ hommes-mois.

Dans le projet Wicri, les mêmes personnes sont donc intervenues sur un ensemble de wikis avec des indices de production significatifs. Les 2 stagiaires (L3 MIASHS ²¹) n'avaient jamais travaillé sur des wikis ou dans un environnement Unix. Ils n'avaient aucune connaissance des sujets traités par les wikis. Au bout d'une quinzaine de jours, ils ont pu installer et cataloguer, sur le site des émérites, des dizaines de publications. Au bout d'un mois, ils commençaient à faire des travaux simples mais significatifs sur leurs sujets respectifs (histoire de l'IST en francophonie d'une part, musique et Roland de l'autre), avec un bon début d'autonomie en fin de stage.

Pour la *Chanson de Roland*, les chiffres doivent être rapprochés de la volumétrie des manuscrits. En particulier, l'ensemble des laisses constitue un « plan de travail » qui permet par exemple de rééditer des articles en résolvant les références par des liens hypertextes.

- Le manuscrit d'Oxford contient 170 feuillets, 300 laisses et 4000 vers.
- Un traitement complet de l'ouvrage de Francisque Michel ajoute le manuscrit de Paris (370 laisses) puis le début du manuscrit de Châteauroux (85 laisses).
- L'ensemble des manuscrits de la *Chanson de Roland* se chiffre en centaines de feuillets, en milliers de laisses et en dizaines de milliers de vers.

Le chiffre de 1 731 pages (2 162 en avril) contient notamment la totalité des laisses (et donc des feuillets) du manuscrit d'Oxford. Les annotations de Paul Meyer sont localisées sur les laisses 140 à 160 du manuscrit de Paris. Mais il a été nécessaire de traiter « au sens plan de travail pérenne » les 250 laisses antérieures (plus celles de Châteauroux). Le squelette informationnel concerné par Francisque Michel vient d'être terminé, soit près de 1000 laisses au total. Pour constituer une base pour couvrir un ouvrage hypertexte comparable à celui de Joseph Duggan [11] il faudrait traiter environ 1 000 laisses complémentaires.

La colonne sémantique identifie la production des catégories ou des relations sémantiques. Ceci va être discuté dans la section suivante.

6.2. Relations et web sémantiques

Concernant les aspects sémantiques, les chiffres du tableau montrent une disparité révélatrice. Deux des wikis (IST, émérites) sont relatifs à des systèmes relativement contemporains avec une forte activité éditoriale. L'approche héritée du Semantic Web peut y être déployée. Nous avons cherché à la compléter à partir des approches CRIS ²² et plus précisément du modèle CERIF [12]. Dans les deux cas nous avons rencontré quelques difficultés liées à la francisation de ces systèmes. De même, la généralisation de ce modèle à toutes les disciplines scientifiques demande des adaptations (ou des simplifications) notamment dans les humanités qui ont une grande variété de fonctionnement de comités scientifiques ou éditoriaux.

Dans le cas de la *Chanson de Roland*, le modèle sémantique de ce sujet médiéval, analysé pendant plusieurs siècles, doit être construit quasi intégralement. Pour cela, l'utilisateur, en situation de concepteur, dispose de nombreux outils qu'il peut combiner :

- MediaWiki offre un mécanisme d'indexations hiérarchisées à base de catégories.
- Celles-ci peuvent être créées ou manipulées avec des modèles ou des modules (en langage informatique Lua).

21. Licence en mathématiques et informatique appliquées aux sciences humaines et sociales.

22. *Current Research Information Systems* (systèmes d'informations sur les recherches en cours).

- Semantic MediaWiki permet de créer des triplets sémantiques. Ils peuvent également être manipulés avec les modèles et combinés avec les catégories.
- Dilib, la boîte à outil XML initialement conçue pour des analyses statistiques de corpus qui contient maintenant des modules d'interface qui peuvent utiliser l'API de MediaWiki. Cet ensemble est naturellement utilisable de façon incrémentale.

Le modèle sémantique de la *Chanson de Roland* repose déjà, pour chaque manuscrit, sur un graphe fiabilisé au niveau des feuillets et des laisses. Ce travail implique une numérotation fiabilisée, et donc un traitement séquentiel de l'ensemble manuscrit (évoqué plus haut). Le résultat est un ensemble de pages identifiées par une nomenclature arborescente, par exemple :

Chanson de Roland/Manuscrit d'Oxford/Laisse CCX

Cette nomenclature permet déjà de piloter un robot. Un chantier est maintenant ouvert sur l'élaboration d'un ensemble sémantique permettant à un contributeur de poser des requêtes au sein du wiki. Les modèles permettent déjà de créer de façon implicite des relations normalisées. Par exemple, la figure 12 montre un bandeau qui est généré par l'appel suivant :

```
{{Manuscrit de Paris/Header laisse | sort=004 | id=IV | précédent=III |
suivant=V}}
```

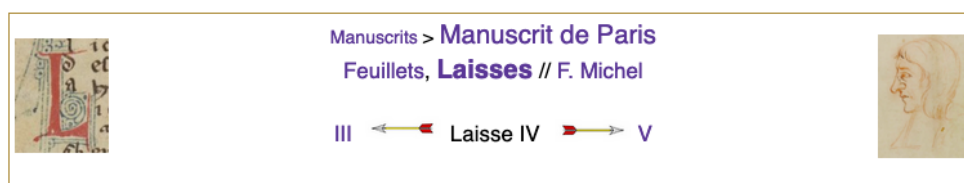


FIGURE 12: En tête généré par l'appel d'un modèle.

Cet appel va provoquer l'ajout d'une catégorie « Laisses » et une relation sémantique « A pour manuscrit:: » avec la page « Chanson de Roland/Manuscrit de Paris ».

Cette organisation sémantique doit être maintenant confrontée avec les applications comme par exemple l'étude des annotations de Paul Meyer. Puis il faudra identifier les relations génériques ou celles qui doivent être diversifiées. L'apport fondamental de l'approche wiki est la possibilité d'une démarche incrémentale où le modèle peut être élaboré dans un processus « essai – erreur ».

Par rapport au Web sémantique, nous avons déjà mentionné des stratégies d'alignement sur des terminologies existantes, notamment Wikipédia en français qui est aligné avec son équivalent anglais souvent présenté comme base terminologique du Web sémantique.

Cela dit, nous avons pu constater la difficulté d'alignement sur des sites de référence comme Worldcat de l'OCLC. Nous avons notamment essayé d'identifier les versions de Léon Gautier « de son vivant » soit avant 1897. Une requête, pas forcément exhaustive, donne 82 entrées sur 129 au total²³. La version 1881 contient dans son titre la mention « 23^e édition ». Il y a donc

23. <https://www.worldcat.org/search?q=chanson+de+roland+léon+gautier>.

environ 60 doublons sur un ensemble de 80 notices. Plus inquiétant, 30 entrées ont 1872 comme date d'édition. Or la revue *Romania* donne en 1873 une analyse relative à la troisième édition datée de 1872 (Paris 1873) et qui, de plus, n'a pas été imprimée... Donc, 30 notices portent en fait sur 2 exemplaires réels. Ces exemples montrent les limites d'une utilisation non contrôlée des triplestores produits par accumulation de sources de données.

En revanche, pour le réseau des communautés du réseau Wicri, les stratégies collaboratives basées sur des noyaux de métadonnées « modérables » semblent plus prometteuses. Signalons par exemple les travaux des Smithsonian Institutions [13] aux États-Unis, du fichier d'autorité commun (GHD) de la Bibliothèque Nationale d'Allemagne [14], et de la Bibliothèque nationale de France [15]. Les trois approches sont basées sur des solutions MediaWiki avec l'extension WikiBase. Dans ces trois cas, il s'agit de mutualiser des catalogues de collections. Le projet Wicri s'attaque à un autre problème : la mutualisation de connaissances, portées par des articles présélectionnés. Le modèle développé pour des collections, grandes mais finies, doit être revisité pour des connaissances potentiellement infinies.

Un article de Luca Mauri [16] étudie (favorablement) la cohabitation des extensions Semantic MediaWiki et WikiBase. Il met également en avant, pour Semantic MediaWiki, la possibilité offerte au contributeur de formuler des requêtes.

A l'heure actuelle, pour le projet Wicri, l'ensemble des wikis est implanté sur la même machine virtuelle (et donc sur la même machine physique). Il paraît donc souvent plus efficace d'utiliser des procédures batch. Dans une évolution vers un réseau physiquement distribué, la situation sera naturellement différente, et l'ouverture apportée par WikiBase semble effectivement séduisante.

6.2.1. Modèles sémantiques propres à l'histoire, à la culture et à la légende

Nous venons d'évoquer principalement des relations relevant plutôt de la modélisation de systèmes de recherche (ou éditoriaux). Nous y avons appliqué des outils et pratiques élaborées dans d'autres disciplines scientifiques notamment dans la santé ou de l'environnement. Or la *Chanson de Roland* est un sujet majeur dans les langues romanes et autour de l'histoire ou des légendes de Charlemagne. Elle est donc un sujet privilégié pour l'étude des approches sémantiques sur les données textuelles des humanités numériques.

Dans un premier temps, nous avons traité des articles scientifiques de référence (de la revue *Romania* par exemple). Nous avons ainsi consolidé les connaissances autour du réseau d'acteurs spécialistes du sujet et montré l'intérêt des liens vers les laisses des manuscrits. Nous avons identifié d'autres documents ou manuscrits à ajouter à notre bibliothèque numérique. Nous avons par exemple commencé à traiter un manuscrit en allemand : le *Rolandslied* du curé Konrad, écrit vers 1170. Malheureusement, nos compétences étaient encore trop approximatives pour réaliser un travail d'indexation pertinent, compte tenu du niveau d'érudition du contenu des articles.

En revanche, comme évoqué plus haut, nous avons démarré un travail qui apparaît très prometteur avec la réédition numérique de l'édition populaire publiée par Léon Gautier en 1895. En effet, cet ouvrage contient un très grand nombre de notes explicatives qui identifient un ensemble de concepts essentiels. Les sujets sont multiples : personnages, lieux géographiques, particularités linguistiques, visions historiques.

Nous démarrons donc des expérimentations avec les personnages. Faciles à identifier, ils font émerger immédiatement les difficultés (et l'intérêt) de la multiplicité des situations, des types de discours et des besoins des contributeurs.

Voici quelques exemples sur Roland et Charlemagne. Ces deux héros sont omniprésents sur ce wiki et seules les relations spécifiques sont *a priori* intéressantes. Il faut donc distinguer par exemple les textes narratifs et les articles historiques. Ensuite, les textes narratifs sont de style et de granularité différentes. Les personnages peuvent y apparaître sous leur référence historique ou avec un éclairage légendaire. Nous avons introduit deux relations sémantique différentes pour marquer cette différence, notamment dans les pages affectées aux laisses des manuscrits.

Ainsi la première laisse du manuscrit d'Oxford contient « Charles le roi ... sept ans est resté en Espagne ». Sa présence en Espagne est historiquement juste, même si la durée de sa campagne est plus modeste. Nous avons donc choisi d'affecter à cette laisse un attribut préexistant :

Laisse 1 > A pour personnalité citée : : Charlemagne

En revanche, dans la laisse CIX, le vers 1404 nous dit « Charles le Grand en pleure et se lamente ». Nous sommes ici dans la légende et nous avons introduit un nouvel attribut indiquant l'aspect « personnage de légende ». L'attribut devient :

Laisse CIX > A pour personnage cité : : Charlemagne

Dans certaines parties du récit, les personnages parlent. Ils peuvent même chanter dans l'oratorio. Nous trouverons alors des attributs tels que :

Mouvement 3 /Mesure 29 à 35 > A pour personnage chantant:: Roland

La présence de cet attribut est limitée aux pages dédiées aux courtes séquences musicales (et non, par exemple, aux descriptions des mouvements).

Comme indiqué plus haut, ce travail est en phase de démarrage. En effet, pour être pertinent il doit s'appuyer sur l'ensemble des laisses d'un manuscrit donné. Nos premières observations montrent que cette indexation va être très dépendante des préoccupations scientifiques des utilisateurs de cette bibliothèque. Plusieurs modèles sémantiques devront probablement cohabiter.

Par rapport aux travaux sur le web sémantique, une première réflexion semble s'imposer. Il est relativement facile d'appuyer sur un bouton de paramétrage pour offrir dès maintenant 25 000 triplets RDF. Mais, de notre point de vue, ceci n'aurait aucun sens. Par rapport aux offres de contenus relativement homogènes (exemple un site dédié aux partitions musicales) notre bibliothèque propose des ensembles de ressources différentes sur un même sujet. Il faudra donc plutôt imaginer un ensemble de triplestores.

6.3. Aspects multimédias

Le traitement des illustrations actuellement utilisé dans le réseau Wicri est relativement classique : des insertions d'images, avec parfois, notamment pour les serveurs d'exploration, des cartes interactives. Pour naviguer dans les bases d'images, il y a une vingtaine d'années,

nous avons utilisé Dilib pour générer des graphes de navigation [17]. Sur la plupart des wikis, ces images sont des objets relativement indépendants les uns des autres. Nous n'avons donc pas été incités à faire de la navigation dans les images. En revanche, sur Wicri/*Chanson de Roland*, les éditions populaires offrent une variété d'images sur un même thème avec, par exemple des graveurs de référence comme Merson, Ferat ou Zier. Un projet dans cette direction pourrait rapidement être mis en place.

Concernant la musique, sur Wicri/Musique, nous avons repris les extensions basées sur LilyPond avec un double point de vue : permettre à un lecteur de se faire une idée d'une pièce musicale et offrir à un choriste des outils de répétition. Avec la *Chanson de Roland*, les extensions musicales sont aussi utilisées pour accompagner le texte des laisses ; ce qui est bien perçu lors des démonstrations. Dans l'avenir, un point fondamental doit être abordé : un éclairage sur la prononciation des textes en ancien français. Or de nombreux articles traitent de ce sujet autour de la *Chanson de Roland*. La liaison entre les aspects phonétiques de la langue des manuscrits avec la musique chantée ouvre un champ d'application très intéressant ²⁴.

La notation LilyPond offre également une possibilité que nous n'avons pas encore utilisée : la recherche de séquences ou de particularités musicales.

Dans ces différents exemples, la précision de l'indexation et des modèles sémantiques joue un rôle fondamental.

6.4. Gérer l'incomplétude et la multiplicité des besoins

L'approche wiki permet 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 peuvent immédiatement travailler. Ainsi, un expert de la musicologie, un linguiste, ou un médiéviste, peuvent 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 minimale sur le contenu des laisses (qui demanderait par exemple 2 minutes par action) peut se traduire par des dizaines d'heures de travail. Cela dit, 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 types de chantiers, sont amenés à coexister. Pour certaines opérations, comme cité plus haut « 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 de thèmes démonstratifs. Cela souligne la forte dimension éditoriale, qui ne peut et ne doit pas être négligée.

24. Voir un premier exemple : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/La_Chanson_de_Roland/Léon_Gautier/Édition_populaire/1895/Introduction/La_verseification.

Enfin, les moyens affectés au projet Wicri sont relativement insignifiants par rapport aux enjeux. Nous donnons donc la priorité aux aspects « preuve de concepts » dans la variété des disciplines scientifiques. Nous assumons une situation où nous laissons un sujet en sommeil quand il ne pose plus de difficultés, pour aborder un autre problème qui nous paraît important dans une stratégie de déploiement institutionnel.

6.5. Aspects institutionnels, du contrôle à l'accompagnement

Le premier auteur de cet article a exercé des responsabilités à l'INIST (informatique, R&D, produit et services) et dans un centre de calcul partenaire du Trésor de la langue Française dans les années 75 [18]. Ces deux unités du CNRS ont depuis renoncé à leurs missions citoyennes (notamment pour le TLF) ou stratégiques (Pascal) après avoir été confrontées à des difficultés de production. Une motivation fondamentale du projet Wicri est donc la recherche de solutions qui permettraient de reprendre ce type d'activités avec une vision européenne.

Dans les deux cas (INIST ou TLF), la simple gestion des chaînes de production demandait des forces de développement et d'exploitation considérables (par exemple, à l'INIST en 2000, la taille du département informatique a dépassé 50 personnes). Notre expérience, comme celle de Wikipédia, montre que ces investissements peuvent être réduits de façon considérable. En revanche, cette approche demande un haut niveau d'expertise informatique (par exemple la capacité de réaliser un robot dans une structure arborescente de documents structurés).

Dans ces deux cas, les procédures de contrôle alourdissaient, de façon parfois ahurissante, la productivité (et la motivation) des ingénieurs. Par exemple, à l'INIST, les protocoles définis par une cellule de qualité étaient conçus pour des applications financières. La simple correction d'une faute d'orthographe dans un résumé monopolisait deux techniciens avec contrôle de leurs hiérarchies (autrement dit, des dizaines de minutes d'intervention et des mois de délai pour une opération qui ne demande que quelques secondes sur un wiki).

Concernant les activités scientifiques proprement dites, les tentatives de remplacement des ingénieurs par des algorithmes se sont traduites par des échecs qui ont abouti à l'arrêt de Pascal (là où le NIH²⁵ aux USA soutient la production par la *National Library of Medicine* de 900 000 analyses documentaires par an où chaque article est lu par 3 experts avec une indexation assistée – mais pas automatique). La technologie wiki permet de faire travailler un réseau d'experts sur un ensemble d'applications (élaboration d'ontologies, édition encyclopédique, indexation, secrétariat de rédaction) dans le même environnement et en mode réseau. Elle demande une expertise technique et multidisciplinaire plus élevée que celle des chaînes de production. Notre expérience (par exemple avec les stagiaires) montre que les protocoles d'assistance et de modération fonctionnent, à condition qu'ils soient continus. Un dispositif de taille comparable à celui de Pascal/Francis en 1990 (400 personnes) est donc nécessaire, mais avec un mode de fonctionnement totalement différent.

Un tel dispositif doit être le plus proche possible des centres de recherche. Pour cela, la production de revues permet de créer des comités scientifiques qui peuvent être mobilisables dans la modération du wiki (nous avons testé cette approche avec la revue *AMETIST*). Dans les activités encyclopédiques, il est également envisageable d'impliquer des thésards dans leur phase d'étude de l'existant (avec la participation de leurs directeurs de thèse).

25. NIH : National Institutes of Health, institutions qui dépendent du Département de la Santé aux USA.

De son côté, l'expérimentation autour de la *Chanson de Roland* montre une exigence d'expertise multidisciplinaire. Là encore, un dispositif d'accompagnement conséquent, notamment en termes de permanence, s'avère indispensable. Il n'est pas à la portée d'une équipe de recherche ou d'un petit laboratoire. En revanche, il peut être obtenu par mutualisation d'équipes au sein d'une université, autour d'une Maison des sciences de l'homme, ou avec une bibliothèque universitaire en mode *learning center*.

Bien entendu, une organisation européenne (et/ou francophone), avec une solution telle que celle qui est étudiée avec Wicri, peut jouer un rôle d'accompagnement pour des projets, même de petite taille dans un contexte international. Par exemple, si la reprise en ordre de marche d'initiatives porteuses des ambitions de Pascal (informer sur l'essentiel de la science) peut être imaginée par discipline scientifique, avec si possible, une mutualisation des expertises.

7. Conclusion

Nous avons présenté un projet numérique autour de la *Chanson de Roland* dans une infrastructure multidisciplinaire.

Notre projet 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 un changement de paradigmes dans les mondes de la recherche ou de la connaissance.

La *Chanson de Roland* nous fait voyager dans le temps entre le Moyen-Âge et le troisième millénaire en passant par le XIX^e siècle.

Au temps de Roland, les lettrés, copistes ou bibliothécaires, disposaient d'une très grande marge d'initiative qu'ils ont perdu parfois avec l'imprimerie. Le numérique leur permet aux humanistes de retrouver cette autonomie. En particulier, les rééditions hypertextuelles revisitent les pratiques des copistes qui avaient une part d'interprétation d'un texte à la façon d'un musicien sur une partition.

Au temps de Roland, les sciences de la matière étaient dominées par les alchimistes qui sont devenus chimistes en s'appropriant des outils mathématiques de plus en plus sophistiqués, au XIX^e siècle avec les équations différentielles, et maintenant avec le numérique des big data. Avec le numérique les chercheurs et praticiens des humanités doivent se dégager de la domination des informaticiens en s'appropriant à la fois, l'algorithmique, les techniques sémantiques et les manipulations avancées de corpus...

Tout un programme qui peut s'avérer passionnant !

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, à sa direction et aux instances d'ISTEX pour l'hébergement du réseau Wicri. Merci aux groupes de travail Wicri pour leur soutien amical.

Références

- [1] F. Michel, *La chanson de Roland ou de Roncevaux du XIIe siècle*, Silvestre, Paris, 1837.
- [2] P. Meyer, *Recueil d'anciens textes bas-latins, provençaux et français*, F. Vieweg, Paris, 1874.
- [3] L. Gautier, *La chanson de Roland*, Traduction, précédée d'une introduction et accompagnée d'un commentaire, par Léon Gautier, Édition populaire, Tours, 1895.
- [4] J. Ducloy, T. Daunois, J.-P. Thomesse, F. Peguiron, I. Turcan, *Revisiter les textes anciens dans les bibliothèques numériques avec l'exemple de la chanson de roland*, 2021. URL : https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/HIS_2021_Casablanca/Atelier_Wicri/Article_soumis.
- [5] S. Vahdati, N. Arndt, S. Auer, C. Lange, *Openresearch : Collaborative management of scholarly communication metadata*, in : E. Blomqvist, P. Ciancarini, F. Poggi, F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management*, Springer International Publishing, Cham, 2016, pp. 778–793.
- [6] J. Ducloy, *Lorexplor : une bibliothèque open source de composants xml d'exploitation du corpus. séminaire de bilan du projet istex*, 2018. URL : https://wicri-demo.istex.fr/Wicri/Wicri/fr/index.php/Utilisateur:Jacques_Ducloy/Blog/Séminaire_ISTEX_2018.
- [7] J. Ducloy, P. Charpentier, L. Grivel, *Une boîte à outils pour le traitement de l'information scientifique et technique*, *Génie logiciel et systèmes experts* 25 (1991) 80–90.
- [8] J. Ducloy, T. Daunois, M. Foulonneau, A. Hermann, J.-C. Lamirel, S. Sire, J.-P. Thomesse, C. Vanoirbeek, *Metadata for wicri, a network of semantic wikis for communities in research and innovation*, in : *DC-2010–Pittsburgh Proceedings*, 2010, pp. 94–102.
- [9] J. Horrent, *La chanson de Roland dans les littératures française et espagnole au Moyen Âge*, Presses universitaires de Liège, Liège, 1951, pp. 323–333. URL : <https://books.openedition.org/pulg/1327>.
- [10] E. Stengel, *Das altfranzösische Rolandslied. Genauer Abdruck der Oxfordorder Hs. Digby 23*, Gebr. Henninger, Heilbron, 1878.
- [11] J. Duggan (Ed.), *La Chanson de Roland. The Song of Roland. The French Corpus*, Brepols, Turnhout, 2006.
- [12] O. Azeroual, J. Schöpfel, *Quality issues of cris data : An exploratory investigation with universities from twelve countries*, *Publications* 7 (2019). URL : <https://www.mdpi.com/2304-6775/7/1/14>. doi :10.3390/publications7010014.
- [13] J. Shieh, *Smithsonian libraries and archives & wikidata : Using linked open data to connect smithsonian information*, 2022. URL : <https://blog.library.si.edu/blog/2022/01/19/smithsonian-libraries,Archives&Wikidata:UsingLinkedOpenData/archives-wikidata-using-linked-open-data-to-connect-smithsonian-information/#.YfKzg2DjKlM>.
- [14] B. Fischer, *Towards an open and collaborative authority control*, *JLIS.it* 13 (2022) 283–290. URL : <https://jlis.it/index.php/jlis/article/view/438>. doi :10.4403/jlis.it-12767.
- [15] V. Boulet, *How to build an «identifiers' policy» : the bnf use case*, *JLIS.it* 13 (2022) 177–184. URL : <https://www.jlis.it/index.php/jlis/article/view/429>. doi :10.4403/jlis.it-12768.
- [16] L. Mauri, *The best of both worlds : Wikibase and semantic mediawiki*, 2021. URL : <https://mediawikiexperts.blog/the-best-of-both-worlds-wikibase-and-semantic-mediawiki/>.
- [17] J.-C. Lamirel, J. Ducloy, G. Oster, *Adaptive browsing for information discovery in an*

iconographic context, in : Content-Based Multimedia Information Access - Volume 2, RIAO '00, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, FRA, 2000, p. 1657–1675.

- [18] J. Ducloy, Systèmes d'information encyclopédiques édités par les scientifiques, Revue ouverte d'ingénierie des systèmes d'information 1 (2020). URL : <http://www.openscience.fr/Systemes-d-information-encyclopediques-edites-par-les-scientifiques>. doi :10.21494/ISTE.OP.2020.0488.