Avant-propos destiné aux évaluateurs

Cet article propose une réflexion qui s'appuie en partie sur des témoignages issus d'un parcours professionnel relativement visible. L'auteur est un retraité qui n'a pas besoin d'être évalué pour des raisons de carrière.

De plus, le cadre d'application traité, un wiki du réseau Wicri sur la Chanson de Roland, est immédiatement identifiable.

Surtout, l'article propose un mode rédactionnel qui privilégie la lecture en ligne. Le texte sera recopié et annoté sur le site Wicri. En particulier, les notions qui ne sont pas universellement connues sont explicités par un lien vers la partie encyclopédique (ou sur d'autres wikis du réseau). Quelques figures sont actives sur le site (2 et 5).

Un regard sur la version en ligne est indispensable pour une évaluation correcte. Avant le résultat de l'évaluation ce lien est :
https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/HIS_(2023)_Ducloy_(proposition)

Pour cet ensemble de raisons, les contraintes du double aveugle sont impossibles à respecter et conduiraient à un résultat illisible. Nous avons donc décidé de ne pas appliquer les mécanismes d'anonymisation.

Si l'article est accepté le lien de la page sera le suivant :
https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/HIS_(2023)_Ducloy
Le texte de l'article sera le même que celui de la version acceptée (définitive). Il sera enrichi par des liens ou des images actives. Des notes de bas de pages pourront être supprimées pour être remplacées par des liens.

Le travail d'alignement de la version PDF avec le wiki est juste initialisé (quelques exemples de liens sont visibles) il sera mené pendant la première phase d'évaluation (avec peut-être de très légères adaptations du texte). Ce travail sera naturellement repris en réponse aux demandes de modification.

Humanités assistées par ordinateur, un exemple avec la *Chanson de Roland*.

Computer Assisted Humanities with the Chanson de Roland*.*

Jacques Ducloy,

Laboratoire Paragraphe, Université Paris 8.

Résumé. Cet article présente une bibliothèque numérique où la *Chanson de Roland* est la partie émergée d'un vaste ensemble de documents sur les poésies épiques du moyen-âge : des manuscrits, des éditions critiques, des traductions, des œuvres dérivées. Leur diversité et leur étroite complémentarité en font un vaste champ d'expérimentation pour les wikis sémantiques (Semantic MediaWiki) et l'ingénierie xml. Ces travaux s'appuient sur des expériences antérieures dans l'information scientifique et technique et France dans une réflexion stratégique sur la bibliodiversité.
Mots-clés. Chanson de Roland, Humanités numériques, Semantic MediaWiki.

Abstract. This article presents a digital library where the *Chanson de Roland* is the visible part of a vast set of documents on the epic poems of the Middle Ages: manuscripts, critical editions, translations, derived works. Their diversity and close complementarity make them a vast field of experimentation for semantic wikis (Semantic MediaWiki) and xml engineering. This work is based on previous experiences in scientific and technical information and France motivated by a quest for bibliodiversity..
Keywords. Chanson de Roland, Digital Humanities, Semantic MediaWiki.

Version en ligne. Une version annotée et avec des liens actifs est visible ici :
https://wicri-demo.istex.fr/Wicri/Europe/ChansonRoland/fr/index.php/HIS_(2023)_Ducloy

1 Introduction

A partir des premiers résultats d'un projet de bibliothèque numérique sur la *Chanson de Roland*, nous proposons des réflexions sur l'appropriation des technologies numériques pour la valorisation du patrimoine culturel, mais aussi sur la position française dans l'information numérique. En effet, il y a 50 ans, les bases de données Pascal, Francis et le dictionnaire du Trésor de la langue française ont été vécus comme très prestigieux au niveau international… avant d'être abandonnés. Face à la situation de monopole résultant de cet abandon, nous étudions des stratégies pour recréer une nouvelle bibliodiversité avec deux technologies complémentaires : l'ingénierie XML d'une part, les wikis sémantiques de l'autre.

Appliqués initialement aux bases bibliographiques, elles se sont avérées très performantes sur les données patrimoniales. Ainsi, un wiki dédié à la musique rassemble des références bibliographiques, des articles réédités en hypertexte, des œuvres musicales, des manuscrits et leurs transcriptions, avec un noyau encyclopédique pour la navigation. Nous travaillons maintenant sur un sujet plus spécialisé, la *Chanson de Roland* qui se révèle une fondation pour l'exploration d'un vaste ensemble de poésies épiques, complété par des écrits sur plus de 10 siècles de littérature, musique, linguistique, dans un contexte international (et multilingue).

La manipulation des manuscrits introduit une évolution fondamentale. En effet, une grande majorité de documents sont en dépendance étroite les uns avec autres. Nous avons donc décidé d'installer une bibliothèque numérique où l'on puisse expérimenter l'ensemble des actions liées à la recherche, depuis la transcription des données jusqu'à la diffusion de connaissances vers le grand public. Cette infrastructure est également utilisable pour des formations professionnelles destinées aux agents du soutien de la recherche et aussi pour les conservateurs ou les chercheurs impliqués dans les humanités numériques.

Dans cet article, nous présenterons nos motivations pour ces travaux et les solutions envisagées autour de l'Information Scientifique et Technique (IST). Nous montrerons ensuite comment elles s'appliquent aux humanités numériques et plus particulièrement dans l'exploration et la valorisation du patrimoine écrit.

2 Grandeur et décadence de l'IST en France

Notre témoignage comporte ici des assertions qui ne font pas forcément l'unanimité mais qui expliquent nos motivations et les options techniques retenues.

2.1 Se dégager du complexe inhibitif de rigueur pour aborder le numérique

Dans le contexte du Plan Calcul (1966), des initiations à l'informatique ont été créées dans les écoles d'ingénieur. Avec des collègues, nous sommes lancés dans le calcul numérique assisté par ordinateur avec les langages Algol ou Fortran.

Cette démarche n'était pas anodine. En effet, en 1956 à Nancy, Jean Legras, le fondateur de l'IUCA[1] écrivait (Legras 1956) : « *L'ingénieur, le physicien se trouvent souvent devant les problèmes que les mathématiciens classiques n'ont pas pu résoudre. Il leur faut alors, ou renoncer à l'emploi de l'outil mathématique, ou* ***utiliser des méthodes moins strictes, que réprouvent les mathématiciens, mais qui sont seules capables de les dépanner.*** ».

[1] Institut Universitaire de Calcul Automatique, un service commun pour les facultés et laboratoires universitaires sur Nancy en 1965.

Pour illustrer un véritable changement de paradigme, il ajoutait : « *Il est alors indispensable que l'ingénieur, le physicien et tous ceux qui s'occupent de mathématiques appliquées, soient capables* ***de se dégager du complexe inhibitif de rigueur que leur a imposé leur éducation****, et qu'ils osent se lancer à l'aventure : la vérification expérimentale sera là pour leur crier casse-cou le cas échéant.* ».

Cette remarque sur le ***complexe inhibitif de rigueur*** nous paraît également fondamentale en 2023 pour les acteurs des humanités numériques.

Un premier exemple dans la documentation en 1973

En 1970, après un DEA en analyse numérique, j'ai démarré ma carrière comme assistant à Nancy (pendant un an) où j'ai enseigné le langage Fortran. Puis j'ai intégré l'IUCA comme ingénieur système (et thésard en compilation). En 1973, j'ai été invité à former au langage COBOL les étudiants de l'IUT Carrières de l'Information à Nancy. Cette option avait été choisie pour sa rigueur par mes prédécesseurs issus de la gestion. Or la programmation COBOL était très rébarbative (une notice bibliographique devait être distribuée sur quelques dizaines de cartes perforées). Il me paraissait impossible de motiver les étudiants dans ces conditions. Or, le compilateur Fortran de l'ordinateur ICL 1901 de cet IUT pouvait, par une extension, manipuler des chaines de caractères, j'ai décidé de me dégager du ***complexe inhibitif de rigueur*** pour montrer aux étudiants, en Fortran, comment il était possible de réaliser des filtrages dans des corpus bibliographiques.

2.2 Une première mondiale dans les humanités numériques avec le TLF

Entrant dans la valorisation de la langue française à l'ère post-numérique, Paul Imbs, à Nancy en 1960, lançait un projet sur 20 ans pour la réalisation informatique d'un dictionnaire de langue, le Trésor de la langue française. Le CNRS avait acquis l'ordinateur français alors le plus puissant, un Gamma 60[2] de la compagnie Bull. Mais la programmation, dans un langage machine assez acrobatique, était inaccessible aux chercheurs en sciences humaines. Le CNRS a donc appelé des informaticiens de haut niveau pour réaliser les développements. La compagnie Bull avait également affecté des ingénieurs pour cette vitrine technologique. Malheureusement, cette équipe a eu une durée de vie limitée au démarrage. En 1973, les programmes sont devenus obsolètes avec un nouvel ordinateur, l'Iris 80, construit par la Cii[3]. Mais, les experts étant partis, la transition a été très difficile.

Dans les années 80, Jacques Dendien, a rejoint le TLF pour y développer des services de haut niveau, Frantext et le TLFi (le TLF accessible par Internet). En dépit de ces succès, l'expérience du management de la production avait été mal vécue par le CNRS qui, en 1995, a renoncé à la mise à jour du TLF. Le TLFi qui avait un immense succès sur le Web dans les années 2000 est maintenant supplanté par Wiktionnaire, technique et juridiquement piloté à San Francisco.

2.3 Une référence mondiale en 1975 : Pascal sur Cyclades avec MISTRAL

En 1970, la Cii a développé MISTRAL, un système de recherche d'information qui plaçait la France en position mondiale dans l'IST. Compte tenu de la présence du TLF, la Cii nous a naturellement invité à acquérir ce progiciel.

Les étudiants de L'IUT ont été les pionniers à Nancy. En 1973, la première version ne fonctionnait qu'avec des bandes magnétiques (6 dérouleurs) et ne pouvait pas être utilisée en travaux pratiques. En revanche, en 1974, une version

[2] Cette puissance était en fait très modeste. En effet la mémoire centrale était de 130 K (octets) complétée par un tambour de 100 K. Le stockage de données utilisait exclusivement des bandes magnétiques (pas de disques).

[3] Compagnie Internationale pour l'Informatique

disque permettait déjà des extractions avec des équations booléennes. En parallèle, l'IUCA, grâce au TLF, étant devenu site pilote pour tester les nouvelles versions du système Siris 8 (et de MISTRAL), les étudiants ont bénéficié de conditions exceptionnelles pour l'époque. Par petits groupes ils pouvaient créer leur propre base (avec un thésaurus) et lancer des recherches en temps partagé.

Forts de cette première expérience, nous avons ensuite informatisé le BALF[4], associé au TLF. Avec un informaticien du TLF nous avons réalisé un transcodage des notices bibliographiques (de mémoire assez simple) et généré une base MISTRAL. En même temps, grâce à nos relations avec l'IRIA, nous avons été reliés au réseau Cyclades, la préfiguration française de l'Internet.

Mais la plus grande performance est venue du CDST[5] qui avait réussi, avec Nathalie Dusoulier, à créer la base Pascal à partir des bulletins signalétiques du CNRS. Elle avait choisi d'utiliser le format ISO 2709 qui venait d'être créé (en 1973) dont la manipulation était assez complexe mais qui garantissait une compatibilité internationale. Avec une production qui était déjà de 400.000 références par an, la base Pascal a pu être accessible sur le réseau Cyclades sous le logiciel MISTRAL.

Malheureusement cette position d'excellence a été de courte durée. Dans les années 80, le réseau Cyclades a été arrêté. Le logiciel Mistral n'a pas été repris par le groupe Bull. L'équipe MISTRAL a rejoint la société TéléSystèmes pour y créer les services Questel. De plus, forte de ce succès, Nathalie Dusoulier a été appelé à diriger les bibliothèques de l'ONU (Genève et New York). Elle a y assuré la fédération numérique de ses bibliothèques. De son côté, le CDST est devenu très dépendant, du savoir-faire de la société Jouve pour la constitution des bases de données et de la société Questel pour les services en ligne. Cette situation a causé de nombreux problèmes de management qui ont conduit à la création de l'INIST.

2.4 Stations de travail Unix XML pour l'exploration de corpus

Un bouquet d'outils logiciels Unix sur la SM90

Dans les années 80, les ordinateurs Multics ont remplacé les Iris 80. Multics étant géré à Phoenix, nous n'avions plus les relations privilégiées avec les experts de la Cii (Siris 8 ou MISTRAL) ou de l'Iria (Cyclades). J'ai alors quitté l'IUCA pour rejoindre un projet nommé ANL pour Association Nationale du Logiciel. L'ANL était pilotée par l'Agence de l'Informatique (ADI) et le CNRS avec comme partenaires le CNET, l'INRIA et le Ministère de la Recherche. Suite à la réalisation d'une enquête, l'ANL est devenue un Groupement Scientifique[6] pour la valorisation informationnelle des logiciels issus de la recherche. Nous gérions un inventaire de logiciels et nous organisions des expositions en France et à l'international.

Nous étions ainsi en première ligne pour repérer des logiciels innovants pour le traitement de l'information technique. Ainsi, en 82-83 nous pouvions générer des catalogues et alimenter un serveur (sous le logiciel Texto). Un virage très important a été pris avec le pilotage des actions SM 90 par l'ADI. La SM 90 était une station de travail sous Unix, issue des études du CNET. L'ANL a alors été sollicitée pour faire un inventaire des logiciels français disponibles sous Unix avec le montage de démonstrations. Notre inventaire numérique est devenu une matière première pour de nombreux tests de logiciels. En particulier, les équipes travaillant sur les compilateurs de compilateurs commençaient à appliquer aux documents leurs outils initialement conçus pour des programmes structurés. L'équipe technique ANL a

[4] Bulletin Analytique de la Langue Française.
[5] Centre de Documentation Scientifique et Technique du CNRS, alors basé à Paris.
[6] Dont j'ai pris la direction en 1981.

donc fait une utilisation intensive d'analyseurs lexicaux (lex) pour adapter nos données à des logiciels d'intelligence artificielle (Lisp ou Prolog).

La fouille de données bibliographiques avec une ingénierie XML à l'INIST

Coup de tonnerre, en 1987, Alain Madelin décide la dissolution de l'ADI qui assurait 50% du soutien de l'ANL. Je me suis alors rapproché de l'INIST. Débauché par Goéry Delacôte et sous la direction de Nathalie Dusoulier[7], j'ai assuré au départ la direction Informatique. L'INIST avait hérité d'un schéma directeur basé sur un système intégré avec un SGBD relationnel sur un mainframe IBM. Nous étions très loin d'unix pour l'indexation des bases bibliographiques mais cela me semblait raisonnable pour les services de fournitures de documents. Par chance, Nathalie Dusoulier tenait à un système dédié pour la bibliothèque. Elle m'a invité à plonger dans les normes de catalogage, et plus précisément dans l'étude du format Unimarc sous la norme ISO 2709[8]. J'ai ainsi découvert que, malgré mon expérience documentaire antérieure j'avais tout à découvrir en bibliothéconomie ! Renonçant au complexe inhibitif de rigueur, l'INIST a donc fait l'acquisition, pour la bibliothèque, d'un système Geac qui a donné entièrement satisfaction.

Grâce aux relations issues de l'ANL, j'ai découvert (début 89) la norme SGML qui me paraissait bien adaptée à la norme ISO 2709. Nous avons alors développé une boîte à outil (iLib) pour le développement rapide d'applications. Avec un mécanisme préfigurant xPath nous avons démarré par des filtrages de corpus ISO 2709. Nous avions environ 5 ans d'avance sur MarcXml de la Library Of Congress.

Une équipe animée par Xavier Polanco utilisait un ensemble de programmes pour la détection des fronts de recherche. Ils avaient été élaborés dans le cadre de thèses, souvent programmés avec les données en mémoire, ce qui limitait la taille des corpus. En m'inspirant des chaines du TLF (qui utilisait le tri standard) et de l'architecture MISTRAL nous avons spécifié des modèles SGML pour les données internes (fichiers inverses par exemple). Nous avons développé des briques de base (comme Lego, mais en langage C) pour générer des systèmes de recherche avec des mécanismes de classification, dénommés serveur d'exploration. Suffisamment stabilisée, après mon départ (voir plus bas), cette équipe l'a utilisée pour réaliser le système Stanalyst. En 1996, le programme Miriad (toujours basé sur iLib) a permis le retour à l'INIST d'un système de recherche documentaire pour la totalité de Pascal (comme en 1976 avec Mistral).

Ilib était cependant limitée par un format SGML dédié à la norme ISO 2709. A partir de 1993, au Loria, j'ai développé une nouvelle version (Dilib). La première version préfigurait le modèle DOM du format XML. Avec une stratégie de compatibilité avec le W3C, elle permettait de mener des classifications sur des sources de plus en plus diversifiées (Medline, Dublin Core). Une action patrimoniale a été menée avec la base Biban (Base iconographique et bibliographique sur l'Art Nouveau). En 2000, lors de mon retour temporaire à l'INIST, Dilib y a été utilisée pour un programme de formation (mutation technologique) et pour la création d'un service de veille et d'édition numérique.

2.5 Le démantèlement des missions stratégiques en IST du CNRS

Goéry Delacôte m'avait donné comme mission de redonner à moyen terme l'indépendance technologique de l'INIST. L'action SGML entrait dans cette stratégie, mais dans un climat souvent très conflictuel. En effet, les cadres impliqués

[7] Qui avait été rappelée par le CNRS pour la création de l'INIST à Nancy.

[8] Plus connue sous l'appellation MARC. D'un point de vue informatique, une notice MARC est un ensemble de petits arbres où toutes les données structurelles sont variables.

dans les relations avec les sous-traitants, voyaient une menace directe dans cette stratégie d'indépendance.

De plus, en 1991, nouveau coup de tonnerre, Goéry Delacôte, en conflit avec la Direction Générale, quitte le CNRS pour aller diriger l'Exploratorium de San Francisco. Parmi les raisons du conflit, l'INIST avait créé une filiale INIST Diffusion pour commercialiser la fourniture de documents. En dépit de la bonne qualité du nouveau service, le marché n'a pas suivi. Le CNRS a donc voulu créer un Groupe INIST, piloté par la filiale (et donc par son chiffre d'affaires). Le CNRS a recruté des cadres issus du secteur de la vente en ligne et favorable à un retour à un modèle informatique centralisé avec maintien des aspects techniques à la sous-traitance. Suite à l'échec de cette stratégie, Le CNRS a fait machine arrière en 2000 (j'ai alors été rappelé comme directeur des produits et services). Mais un nouveau changement de direction de l'INIST est intervenu en 2004. Une nouvelle stratégie encore trop dépendante de la filiale s'est révélée catastrophique sur le long terme pour les bases Pascal et Francis qui ont finalement été démontées. Pendant ce temps, les américains, et notamment la NLM[9] avec qui nous faisions jeu égal dans les années 90, a plus que doublé sa production et possède un monopôle stratégique.

3 Une stratégie de mobilisation générale avec le projet Wicri

Un an après le démarrage de Wikipédia, en 2002, le moteur MediaWiki apporte un ensemble d'innovations fondamentales qui vont devenir réellement disponibles à partir 2006. En désaccord avec la politique de l'INIST, j'ai alors rejoint la DRRT Lorraine. Nous y avons lancé une expérimentation sur l'usage des wikis sémantiques pour la promotion des résultats de la recherche. Nous avons monté un premier démonstrateur avec un wiki pour la région Lorraine et d'autres sur les priorités du CPER[10] (eau, bois et forêts, sols urbains…). Puis, souvent pour des raisons démonstratives le réseau s'est enrichi pour atteindre maintenant 150 wikis.

Figure 1. *Le réseau Wicri en 2023*

La nature des wikis s'est diversifiée autour des publications scientifiques avec la revue les mots de l'agronomie (avec l'INRAE) ou les wikis des colloques CIDE ou H²PTM. Bien entendu, j'ai aussi exploré la faisabilité d'une réponse pilotée par des coopérations entre les unités de la recherche face aux monopoles américains dans les bases bibliographiques. En même temps, à l'occasion d'une action sur la Renaissance en Lorraine, nous avons commencé à rééditer des ouvrages anciens, ouvrant ainsi une dimension héritage culturel. Nous avons fait un premier essai en 2010 avec un ouvrage d'Henri Lepage sur le Palais Ducal de Nancy. Il a été réédité avec une organisation hypertexte, en reprenant notamment une gravure avec des

[9] National Library of Medicine qui diffuse le service PubMed.
[10] Contrat de Projet Etat Région.

renvois dans le texte de l'ouvrage où des annotations orientent le lecteur vers des explications complémentaires dans la souche encyclopédique (Ducloy 2019). Pour des expérimentations multimédia, nous avons monté le wiki Wicri/Musique où nous avons réédités des partitions (avec le langage LilyPond), des articles encyclopédiques (Jean-Jacques Rousseau par exemple) et des extraits du TLF.

La figure 2 montre en partie gauche l'ensemble des documents numériques qui cohabitent sur un wiki. Les catégories et liens sémantiques font interagir divers modèles ontologiques (dont celui de Mistral). Enfin les modèles et modules permettent de programmer tout ce qui est spécifique à un domaine donné[11]. Une équipe de recherche qui investit en formation devient alors totalement autonome.

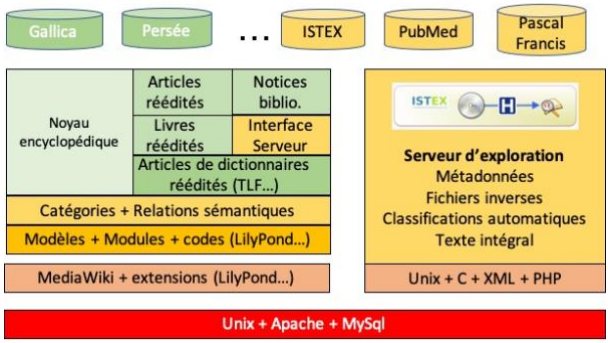


Figure 2. *Éléments d'un ensemble wiki + serveur d'exploration*

Aidés par un financement ISTEX, nous avons réalisé le couplage d'un wiki avec des serveurs d'exploration. Avec le programme LorExplor, de nouvelles versions ont vu le jour avec une innovation assez fondamentale. En 2000, le paramétrage des applications et le nettoyage des données (curation) relevait du bricolage informatique. Avec la version LorExplor les wikis ont été utilisés pour la navigation (cartes dynamiques), le paramétrage et la curation des données. Près de 150 serveurs manipulant de 1000 à plus de 20.000 documents, dans tous les domaines couverts par Wicri ont été développés.

4 Roland au combat pour le patrimoine numérique

L'innovation n'est pas un long fleuve tranquille, et sans entrer dans les détails, l'usage des wikis ne fait pas l'unanimité dans les services de soutien à la recherche. L'INIST, avait été mandatée et financée par ISTEX pour héberger le réseau Wicri. Mais à la fin du pilotage académique d'ISTEX, elle a refusé d'examiner une collaboration. Je me suis donc rapproché du laboratoire Paragraphe, pour tester le potentiel Wicri au sein d'une équipe de recherche. Avec comme moyens ma seule productivité, je me suis immergé dans la position d'un étudiant de master en musicologie, philologie ou médiévistique, tout en restant bibliothécaire, éditeur, et informaticien[12]. La Bibliothèque Universitaire de Lettres de Nancy, dépositaire du fonds Paul Meyer, a fait émerger une thématique : *La Chanson de Roland.*

[11] Par exemple, sur Wikipédia, les outils géographiques sont réalisés par les contributeurs.
[12] J'agis donc comme un « chercheur praticien » en SHS, capable de programmer des modules récursifs, comme un chimiste résout des équations aux dérivées partielles.

4.1 La défaite de Roncevaux et les premières étapes du projet

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. Tels sont les faits racontés par Éginhard au chapitre neuvième de sa *Vita Karoli Magni* (Vie de Charlemagne), et rappelés par Léon Gautier dans son édition populaire de 1895 (Gautier 1895).

Un stage d'une filière* Métier du livre *pour un ouvrage annoté

En 2014, suite à nos travaux sur la réédition de livres, nous avons été sollicités par Isabelle Turcan pour accompagner un étudiant d'une filière "Métiers du livre" dans la numérisation d'une édition critique du manuscrit dit d'Oxford, publiée en 1869 par Francique Michel (qui l'avait découvert), et annoté par Paul Meyer.

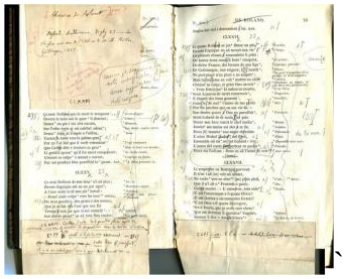
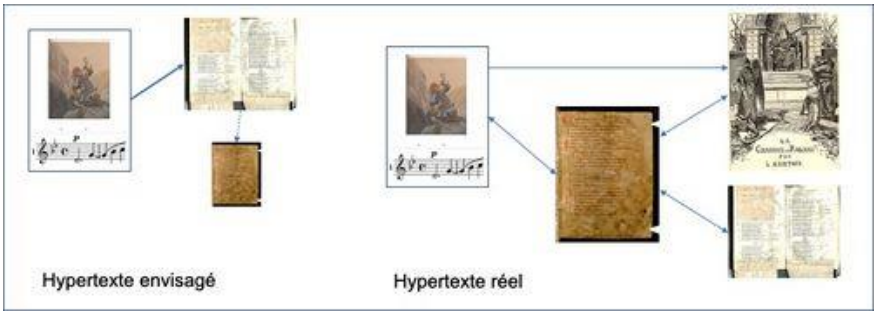


Figure 3. *Exemples d'annotation*

Le démarrage a été très rapide. J'ai développé quelques modèles MediaWiki (mise en page…) et encadré l'étudiant qui a produit des résultats en quelques jours. A la fin du stage, toutes les pages annotées avaient été traitées et une partie conséquente de l'ouvrage avait été transcrit en code wiki. Nous disposions d'un démonstrateur à destination des philologues sur l'utilisation des wikis sémantiques.

Un stage apparemment anodin, mais décisif

En mai 2021 un nouveau stage[13] a conduit à un projet plus conséquent avec un nouveau public : les choristes. En effet sur Wicri/Musique, nous avions travaillé sur une messe irlandaise avec Gilles Mathieu qui avait aussi composé une suite pour chœur et orchestre, basée sur le manuscrit d'Oxford. J'ai demandé aux stagiaires de mettre en relation les vers de l'oratorio avec le texte de Francisque Michel et des facsimilés de feuillets du manuscrit.



[13] Pour dépanner des collègues en période Covid

Figure 4. *Évolution de l'architecture hypertexte*

Après un démarrage très satisfaisant sur les premières strophes, des incohérences de numérotation de vers sont rapidement apparues. En effet, Gilles Mathieu avait travaillé à partir d'une autre édition critique (Léon Gautier). Le modèle hypertexte s'est donc enrichi, avec 2 éditions critiques. Nous avons donc modifié en profondeur le modèle initial. En quelques mois, nous disposions d'un ensemble déjà démonstratif. De plus, en préparant un séminaire de travail avec des philologues, nous avons localisé l'ouvrage cible des annotations de Paul Meyer.

Une bibliothèque numérique aux objectifs multiples

Dans notre réflexion sur l'appropriation du numérique, ce premier problème, découvert au bout de quelques jours de développement, nous a semblé particulièrement démonstratif. En effet, dans un protocole de sous-traitance, basé sur un cahier des charges, nous aurions été bloqués, au bout de quelques jours, pour plusieurs mois. Nous avons donc décidé d'analyser le potentiel de cette thématique pour un projet conséquent de bibliothèque numérique. Plus précisément, cette infrastructure numérique doit être utilisable par des chercheurs pour leurs investigations et pas seulement pour la diffusion de leurs résultats.

4.2 Des documents hétérogènes, très diversifiés et très interconnectés

Dans un premier temps, voici un aperçu de la diversité des documents traités, et de leurs multiples imbrications.

Combien de mètres de rayonnage pour la Chanson de Roland ?

Dans la bibliothèque des lettres de l'Université de Lorraine, la *Chanson de Roland* occupe trente centimètres de rayonnage dont une dizaine pour les trois tomes (2 940 pages) de Joseph Duggan (Duggan 2005). Sur Google Scolar, la requête « "Roland" "Chanson" "Charlemagne" », sans les citations, donne environ 14.000 références, soit des centaines de mètres... La création d'une bibliothèque significative n'est donc pas une entreprise anodine. Le modèle général du réseau Wicri donne déjà une base d'organisation pour les documents courants. Nous allons détailler les documents originaux, en commençant par les manuscrits.

Le manuscrit d'Oxford

Le manuscrit d'Oxford, fondamental pour tous les auteurs, occupe une place particulière au cœur du dispositif, avec 3 types de « pages wikis » :

Chaque **facsimilé de page** (144 au total) donne lieu à une page wiki de description, plutôt destinée à la gestion. Par exemple, les images extraites (comme une lettrine pour alimenter un article spécialisé) mentionnent leur appartenance à ce facsimilé. Réciproquement, MediaWiki gérant les liens inverses, il est possible de connaître tous les extraits et de savoir où les facsimilés ont été utilisés.

Chaque **feuille**t (72) possède sa page de description (avec insertion des images recto et verso). Pour ce manuscrit, le philologue allemand Edmund Stengel a édité une version critique en 1878 où chaque page imprimée recouvre exactement le contenu d'une page du manuscrit. Nous avons donc inséré, au niveau feuillet, une copie de cette interprétation à l'usage du chercheur. Notons ici la spécificité du traitement éditorial de l'ouvrage de Stengel.

Chaque **laisse** (396, avec notre numérotation) donne lieu à la création d'une page wiki où sont reproduits les facsimilés des pages du manuscrit. Nous complétons systématiquement avec une version de référence (Gautier 1872) afin

d'associer à chaque couplet sa transcription et sa traduction. Nous verrons que chacune de ces pages wiki héberge d'autres informations de provenances diverses.

Les vers sont généralement identifiés par des numéros (de 1 à 4401). Nous avons été amenés à créer des pages vers qui sont des redirections au sens wiki. Dans d'autres articles, nous avons signalé les problèmes rencontrés par la diversité de numérotations des vers et des laisses par différents auteurs. Ces faits sont mentionnés dans chaque page laisse (et font l'objet de traitements spécifiques).

Les autres manuscrits

Une étude même rapide de la littérature montre qu'il faut considérer une dizaine de manuscrits sur la *Chanson de Roland*, et presque autant de dizaines sur des dizaines de poèmes épiques. Or, chaque manuscrit implique une étude particulière[14].

Quelques-uns sont en cours de traitement dans leur intégralité. Par exemple, le premier stage ne portait que sur la partie de l'ouvrage de Francisque Michel dédiée au manuscrit d'Oxford. Or, il en traite deux autres. Celui, dit de Paris, est édité dans son intégralité (6828 vers décasyllabiques sur 375 laisses monorimes), mais il est incomplet, tout le début ayant été égaré. Francisque Michel a donc comblé cette lacune avec le début de celui de Châteauroux (85 laisses, 1332 vers[15]). Ces deux manuscrits seront donc traités intégralement sur le wiki, mais avec une organisation différente. De plus, un travail d'alignement avec celui d'Oxford doit être réalisé pour comprendre l'histoire de ce poème. Cette multitude de petits problèmes va donner lieu à une multitude de petites initiatives de structuration.

D'autres manuscrits comme ceux dits de Cambridge et de Venise 4 donneront lieu au même type de traitement. Citons également le manuscrit de Conrad qui est en allemand avec de très intéressantes illustrations (qui sont absentes sur les autres).

De nombreux textes établissent des comparaisons avec d'autres manuscrits du moyen âge ou même de la Renaissance qui sont traités de façon partielle.

Citons également des manuscrits pour des partitions (Charpentier).

L'édition critique de Léon Gautier (1872)

La plupart des manuscrits donnent lieu à des éditions critiques. Là encore, nous rencontrons une grande variété de modèles éditoriaux. Celle de Léon Gautier joue un rôle particulier par son utilisation par Gilles Mathieu et pour sa notoriété. Il s'agit d'un ouvrage conséquent (1000 pages sur 2 tomes). Nous l'avons vu, les 300 pages dédiées à l'édition critique proprement dite (et sa traduction) sont retranscrites, laisse par laisse (au lieu d'une répartition « page paire, page impaire »).

Les autres 700 pages sont réparties entre un glossaire de quelques milliers d'entrées avec des liens vers les vers du manuscrit ; une table des matières qui pointe vers de numéros de page du manuscrit ; plus d'un millier de notes qu'il faut associer aux vers ; et une introduction d'une vingtaine de chapitres avec des contenus qui ouvrent vers des dizaines d'autres documents. De plus, des dizaines d'entrées de l'index sont en fait de véritables articles qui deviennent des pages wikis.

Ce document demande donc un traitement très spécifique pour chaque partie. Le même type de problème se pose pour la plupart des autres éditions critiques. Dans son article *traduire la chanson de Roland* Christopher Lucken (Lucken 2018) donne un chiffre de 50 ouvrages significatifs de traduction.

---

[14] Rappelons que le porteur de ce projet était un peu béotien sur les Chansons de Geste au début de projet. Il a donc fallu repenser en permanence (et sans problème majeur) l'architecture informationnelle – ce qui démontre la souplesse de l'approche wiki.

[15] Le manuscrit complet fait 8201 vers sur 452 laisses.

Du côté de la musique

La technologie utilisée repose sur le logiciel de gravure musicale LilyPond, avec un langage formel dont la syntaxe rappelle celle de TeX pour les mathématiques. Voici par exemple le début du thème « Au clair de la lune » en si bémol majeur.

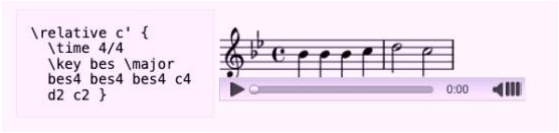


Figure 5. *Exemple de codification en LilyPond*

La suite musicale de Gilles Mathieu est rééditée dans la continuité de ce que nous avions fait pour Irish Mass, du même compositeur, sur Wicri/Musique. La partition « conducteur » est composée de 10 fichiers PDF dans l'original, de même pour les partitions SATB + piano. Sa restitution hypertexte est composée de 10 arbres hypertextes dont les éléments sont des phrases musicales qui correspondent à des vers d'une même laisse. Ils sont ainsi être reliés au glossaire de Léon Gautier. Ainsi, le choriste comprend ce qu'il chante et comment le prononcer.

Encyclopédies et le dictionnaire Trésor de la langue française

Les encyclopédies ou les dictionnaires bénéficient d'un traitement particulier. Nous avions une expérience avec le Dictionnaire de Musique de Jean-Jacques Rousseau sur Wicri/Musique. Ici, le Grand Dictionnaire universel du XIX^{e} siècle de Larousse est une source d'articles particulièrement intéressante (et facile à traiter) pour donner des explications aux lecteurs non érudits.

Le dictionnaire du TLF, compte tenu de son histoire, bénéficie d'un traitement spécifique. D'une part, nous voudrions démontrer qu'il peut être mis à jour dans une approche type Wicri. D'autre part, de nombreux articles du TLF font référence, dans la partie étymologie, à la Chanson de Roland, via les notes de Joseph Bédier. Sur Wicri/Chanson de Roland un article y est structuré en 2 parties. La première contient un extrait (ou l'intégralité) d'un article du TLF. Une deuxième (optionnelle) contient des propositions de mise à jour. Nous avons aussi créé un wiki spécialisé (Wicri/Francophonie) pour tester l'ensemble du TLF et des textes associés.

Dans notre démarche informationnelle, et sur le thème de la Chanson de Roland, le TLF est parfois un moteur de sérendipité particulièrement intéressant. En effet, un article du TLF contient des exemples d'auteurs des deux derniers siècles qui ont écrit des textes en relation avec la Geste de Charlemagne. Nous avons notamment pu mettre en évidence Victor Hugo, Alfred de Vigny et Anatole France. Cet exemple montre que le wiki devient, même dans une phase intermédiaire, une vraie source d'information pertinente pour le chercheur.

5 Bilan et perspectives

En 1991, Goéry Delacôte nous avait invité à travailler sur la future « Station de travail du chercheur ». L'INIST devait alors mettre à sa disposition de vastes ressources avec une excellence dans les traitements numériques exploratoires. Nous devions alors implanter des mécanismes pour permette au chercheur d'enrichir naturellement, cette architecture informationnelle. Cet objectif est-il atteint ?

5.1 Un bilan techniquement très satisfaisant (vue du porteur du projet)

La puissance de MediaWiki, le moteur de la galaxie Wikipédia est maintenant incontestable. Des centaines de milliers de volontaires enrichissent une

infrastructure informationnelle commune. De son côté, l'expérience Wicri montre qu'un individu souvent isolé peut mettre en place un réseau de plusieurs dizaines de familles de wikis multilingues, complété par 150 serveurs d'explorations (500 000 documents). Sur la Chanson de Roland, les chiffres de productions sont éloquents :

Tableau 1 *– Indices de production sur les wikis (janvier 2022).*

	Pages wiki	Avec contenu	Modif.	Sémantique
Chanson de Roland (01/2022)	5 056	1 731	15 738	18 560
Chanson de Roland (08/2023)	11 213	3 240	50 219	52 371
Wicri/Musique	4 332	1 415	11 028	52 472

Les chiffres sur *la Chanson* montrent la productivité d'une seule personne pendant 18 mois. De son côté, Wicri/Musique est un wiki en concurrence avec une cinquantaine de familles. La première colonne comptabilise toutes les pages au sens wiki (par exemple un lien de redirection pour atteindre un vers). La colonne sémantique montre la différence de profil entre un wiki contenant de nombreuses fiches (compositeur, œuvre, villes…) et la Chanson qui contient beaucoup de textes.

D'un point de vue plus éditorial, des résultats concrets ont été atteints. Par exemple, la table de concordance du manuscrit d'Oxford (et donc toute son architecture interne) est terminée (avec un complément pour les originalités de Léon Gautier). Toutes les séquences musicales SATB pour l'oratorio sont disponibles. Nous avons organisé en aout 2023 à Aussois une manifestation musicale où le contexte culturel est explicité dans le wiki qui devient un outil de médiation culturelle.

5.2 Les raisons des batailles perdues

Les résultats techniques et scientifiques sont incontestables, mais notre bilan institutionnel est celui d'une bataille perdue. Par exemple, pendant la crise du COVID, nous avons amélioré nos protocoles d'analyse des publications de santé, avec des résultats que nous jugions comme très intéressants. Dans la continuité du programme ISTEX, nous avons demandé au CNRS de nous aider à trouver de nouveaux champs de coopération en santé. Nous avons été confrontés à une fin très ferme de non-recevoir. Sans expertise pointue dans les sciences du vivant (génomique…), nous avons donc choisi de changer de sujet d'application, avec bonheur sur unplan scientifique, mais pas au niveau des retombées sociétales.

Cela dit, il convient de relativiser notre projet dans ce qu'il faut bien désigner par désastre au niveau national avec la perte de Pascal (face à MEDLINE notamment), de Francis (face à Oxford) et du TLF (face à Wiktionnaire)[16]. A ce niveau, la réponse est politique. Au temps du Plan Calcul, de l'ADI et du démarrage de l'INIST les moyens humains dédiés au TLF et à l'IST voisinaient les 700 personnes (500 titulaires, essentiellement CNRS et 200 occasionnels). En 2023, la *WikiMedia Fundation* affiche un effectif comparable de 700 personnes. Les effectifs français de l'éducation, de l'enseignement et de la recherche sont de 1.500.700[17] agents. La France a largement les moyens de dégager un millième de ses moyens

[16] Au niveau logiciel, pour Mistral, la concurrence est moins concentrée USA avec Geac au Canada ou Elasticsearch aux Pays-Bas.

[17] https://www.insee.fr/fr/statistiques/2493501#tableau-figure1

pour retrouver une place de leader au niveau international. La France est à la confluence de l'Europe et de la francophonie, où des moyens peuvent être fédérés.

Au niveau des services de terrain, il conviendrait d'étudier les freins psycho-sociologiques qui amènent les agents et leur encadrement à rejeter l'approche wiki. Citons deux hypothèses. Les informaticiens intervenant en gestion ou en communication institutionnelle ont été formé à « la validation a priori ». Le ***complexe inhibitif de rigueur*** leur interdit d'envisager une modalité de modération a posteriori. Un autre obstacle vient de l'expertise requise. La manipulation d'un wiki avec une interface WISIWIG parait simple. L'expérience de la Chanson montre la permanence de besoins relativement imprévisibles de tout type d'expertise. Cette analyse est partagée par le projet PolimaWiki (Chastan 2020), mais avec un objectif moins ambitieux. Les enjeux liés à un positionnement international méritent de travailler à un programme de formation des chercheurs aux techniques avancées du numérique.

5.3 Les perspectives actuelles de Wicri/Chanson de Roland

Dans son état actuel, le wiki Wicri/Chanson de Roland dispose d'un noyau relativement stabilisé autour de 3 manuscrits et des quelques éditions critiques. Voici quelques exemples de travaux qui sont engagés dans une nouvelle étape.

La réédition hypertexte d'articles de recherche

Paul Meyer n'est pas seulement l'annotateur de Francique Michel, il a été directeur de l'École des Chartes en 1882 et l'un des fondateurs de la revue *Romania* et de la *Revue critique d'histoire et de littérature*. La réédition d'articles de ces revues est une priorité d'extension de la bibliothèque. Par rapport à nos expériences précédentes (Wicri/Santé…), une spécificité est l'abondance de liens profonds qui doivent être résolus (accès par un clic). Chaque type de lien (vers, couplet, page…) est dépendant du type de cible. De nombreux articles comparent des sources et amènent à enrichir la bibliothèque de nouveaux extraits d'ouvrages ou de manuscrits. Cette activité demande une très grande réactivité et donc une forte maitrise technologique (architecture du wiki, écritures de modèles…).

L'écriture d'articles de recherche

La rédaction de cet article (celui que vous lisez) est un exemple d'une publication savante qui est intégrée à un ensemble numérique. Nous avions testé ce modèle avec la revue Ametist. D'un côté, les contraintes temporelles liées à la publication papier (ou PDF) obligent à produire une étape cohérente qui peut être citée. Sur la version numérique, il est possible de multiplier les annotations explicatives et les démonstrations. Au niveau du noyau encyclopédique, cette pratique amène à développer un ensemble de nouvelles pages.

L'écriture d'articles pour le grand public

A l'occasion de la « Fête de la Science » nous avons été confrontés à un phénomène de pertes de racines culturelles en quelques génération. En 1881, notre poème était officiellement un texte à travailler par des élèves de seconde, Nous montrons sur le wiki un exemple courant d'une revue de grande diffusion pour la jeunesse de 1906, avec une bande dessinée sur Roland. Dans les années 50 à 60, la Chanson était au programme des lycées. Elle était également présentée dans les cours d'histoire pour les cours élémentaires[18]. En 2022, la grande majorité de nos

[18] Le manuel d'Histoire de France diffusé par Nathan en 1955 consacre 2 pages (sur 80) à Roland (autant que pour Charlemagne, Louis XIV fait mieux avec 4 pages).

visiteurs de moins de 40 ans ignoraient tout de la bataille de Roncevaux. Ce constat demande à enrichir le wiki par un ensemble de textes destinés au grand public.

Du côté des ontologies

Le contexte des poèmes épiques est une immense mine de problèmes. En effet, chaque manuscrit, chaque ouvrage de référence possède de fait sa propre ontologie. Par exemple, l'archevêque Turpin a probablement existé. Dans le manuscrit d'Oxford, il meurt avant Roland. Mais, dans un autre contexte il est l'auteur de la Chronique du Pseudo-Turpin, où il raconte une bataille à laquelle il a survécu… Les catégories de MediaWiki, les extensions sémantiques offrent une large batterie d'outils pour exprimer des ontologies dans la diversité de leurs contextes. Notons que le contenu du wiki devient suffisant pour mener des instigations conséquentes.

Aspects informatiques, robots et fouille de données

Nous avons très peu abordé ce sujet dans cet article. En effet, nous avons mis la priorité sur la constitution d'un noyau significatif pour lequel il était fondamental de travailler « manuellement » pour bien comprendre la variabilité des sources. De même nous n'avons pas commenté le montage d'un serveur d'exploration réalisé il y a quelques années. Comme précédemment le contenu du wiki devient suffisamment significatif pour lancer des expériences consistantes (programmation de robots…).

En guise d'épilogue

Au moment de soumettre ce papier, nous avons introduit le langage Fortran dans le wiki en relation avec le début de notre article. Nous avons donc cherché des travaux numériques sur notre Chanson avec une référence à Fortran. Après une recherche infructueuse sur Internet, une exploration, très informaticienne avec Dilib sous unix, d'un corpus ISTEX crée sur la Chanson de Roland, nous a permis de découvrir des travaux datés de 1973 par Gian Piero Zarri (alors en Italie) sur une interpolation sur plusieurs manuscrits de la Chanson avec… le langage Fortran. Nos étudiants de l'IUT étaient des précurseurs en France sur l'appropriation des langages de programmation pour les humanités.

6 Conclusion

La plupart des services dédiés à la communication scientifique, comme HAL ou OpenEdition, diffusent des documents dont l'écriture est terminée. Ils permettent aux chercheurs de publier sans rien réellement changer dans des pratiques vieilles de plusieurs siècles. Avec la *Chanson de Roland*, comme dans tous les travaux d'étude et de valorisation du patrimoine écrit, la bibliothèque numérique devient un espace de travail, et pas seulement de diffusion ou de lecture.

Face à ce changement de paradigme, nous avons montré la puissance de deux approches technologiques, les wikis sémantiques et l'ingénierie XML. Cette expérience vécue comme passionnante est cependant exigeante. Au niveau d'une équipe de recherche elle demande une maitrise avancée dans tous les aspects du numérique. L'appropriation des technologies, autrement dit, les pratiques des humanités assistées par ordinateur offrent alors au chercheur une immense liberté de conception, et donc d'investigation dans leurs stratégies de recherche.

7 Bibliographie

Chastan, P., (2020) PolimaWiki : un Wiki sémantique pour l'analyse de listes au Moyen Âge, limites et apports. In *Memini. Travaux et documents,* 26, 2020

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

Duggan, J. (2005). *La Chanson de Roland. The Song of Roland. The French Corpu*s, Brepol 2005.

Gautier, L. (1895). *La chanson de Roland, Traduction, précédée d'une introduction et accompagné et d'un commentaire*

Legras, J. (1956). *Résolution des équations aux dérivés partielles*, Dunod 1956.

Lucken, C. (2018), Traduire la Chanson de Roland, In *Mediévales*, t. 75, 2018.