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).

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.

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. Fournir un résumé en français. La taille du résumé peut éventuellement être telle que la première page puisse contenir le résumé et les mots-clés, aussi bien en français qu'en anglais.
Mots-clés. Chanson de Roland, Humanités numériques, Semantic MediaWiki.

Abstract. Fournir un résumé en anglais.
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*, un retraité propose ici des réflexions sur l'appropriation des technologies numériques pour la valorisation du patrimoine culturel. Cet article s'appuie sur 50 ans de rencontre avec le dictionnaire du Trésor de la langue française et les bases de données Pascal et Francis. Après une phase vécue comme prestigieuse, elles ont été abandonnées en raison de difficultés techniques. Pour étudier des voies de redressement nous avons travaillé sur deux technologies complémentaires : l'ingénierie XML d'une part, les wikis sémantiques de l'autre.

Appliquées initialement aux analyses de publications scientifiques, elles se sont avérées très performantes sur la gestion et la publication de données diversifiées dans les humanités numériques. Par exemple, nous avions installé un wiki dédié à la musique qui rassemble des références bibliographiques, des articles réédités en mode hypertexte sémantique, des œuvres musicales, des manuscrits et leurs transcriptions. Un noyau encyclopédique permet de naviguer dans cet ensemble.

Par des concours de circonstances, nous avons travaillé sur la *Chanson de Roland*. Cette thématique, apparemment très spécialisée, est en réalité un point d'entrée pour l'exploration d'un vaste ensemble de poésies épiques avec des développements sur plus de 10 siècles d'histoire, de littérature, de musique, de linguistique, dans un contexte international (et multilingue). Cette thématique introduit une différence fondamentale avec notre antériorité : une grande majorité de documents sont en dépendance étroite les uns avec autres.

Nous avons donc décidé d'expérimenter une vaste bibliothèque numérique où l'on puisse réaliser 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 professionnels 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 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 de plus en plus fondamentale en 2023 pour dominer les 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 notamment enseigné le langage Fortran. Puis j'ai intégré l'IUCA comme ingénieur système (en passant une thèse en compilation en parallèle). En 1973, j'ai été invité à assurer une initiation au langage COBOL pour les étudiants de l'IUT Carrières de l'Information à Nancy. Cette option avait été retenue comme une école de rigueur par mes prédécesseurs issus de l'informatique de gestion. L'écriture d'un programme COBOL était particulièrement rébarbative[2]. Par exemple, COBOL ne manipulait que des données de taille fixe. Une notice bibliographique devait donc être distribuée sur quelques dizaines de cartes perforées. Compte tenu de mon expérience en analyse numérique, il me paraissait impossible de motiver les étudiants dans ces conditions. Comme le compilateur Fortran de l'ordinateur ICL 1901 de cet IUT disposait d'une extension pour manipuler des chaines de caractères, j'ai décidé de me dégager du ***complexe inhibitif de rigueur*** porté par COBOL pour montrer aux étudiants, en utilisant Fortran, comment faire pour réaliser des filtrages dans des corpus bibliographiques.

2.2 Une première mondiale dans les humanités numériques avec le TLF

La langue française, dans sa dimension historique est un héritage fondamental. En 1960, Paul Imbs, également à Nancy, avait lancé 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 le plus puissant de cette époque un Gamma 60[3] de la compagnie Bull. Mais la programmation devait se faire dans un langage machine assez acrobatique, elle était donc inaccessible aux chercheurs en sciences humaines (ou même en calcul numérique). 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 stade initial. Dans les années 73, leurs programmes étaient devenus totalement obsolètes avec un ordinateur plus moderne, l'Iris 80, construit par la Cii[4]. Mais les experts étant partis et la transition a été très difficile.

Dans les années 80, Jacques Dendien, a rejoint le TLF pour y développer des services de haut niveau comme Frantext ou le TLFi (le TLF accessible par Internet). En dépit de ces succès en 95, l'expérience difficile du management de la production avait été tellement mal vécue par le CNRS qu'il a renoncé à engager une mise à jour du TLF. Le TLFi qui était le dictionnaire français de référence avec un immense

[2] Les professionnels avaient introduit la catégorie analyste pour sous-traiter la programmation à des programmeur-codeurs.

[3] 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).

[4] Compagnie Internationale pour l'Informatique

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 pour placer 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 elle ne pouvait pas être utilisée en travaux pratiques. En revanche, en 1974, une nouvelle version, disque cette fois 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[5], associé au TLF. Ce bulletin existait sous la forme de notices bibliographiques. Avec un informaticien de l'INIST nous avons réalisé un transcodage (de mémoire assez simple) et généré une base Mistral. Nous avons organisé des séances de formation où je me souviens d'avoir insisté sur l'intérêt de la navigation dans le thésaurus.

En même temps, grâce aux bonnes relations que Claude Pair avait avec l'IRIA, nous avons pu disposer d'un strapontin dans les réunions techniques du réseau Cyclades, la préfiguration française de l'Internet, pour y être finalement connecté en 78, Nancy.

Mais la plus grande performance se situait du côté du CDST qui avait réussi, grâce notamment à l'impulsion de 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 mené une carrière dans les bibliothèques de l'ONU, Genève et New York. Elle a notamment assuré la fédération numérique de l'ensemble des bibliothèques de l'ONU. Le CDST est devenu très dépendant, en amont 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 gestion qui ont conduit les pouvoirs publics à la création de l'INIST.

2.4 Stations de travail Unix pour l'ingénierie XML

Dans les années 80, les ordinateurs Multics ont remplacés les Iris 80. Avec la Cii basé à Louveciennes, nous avions des relations privilégiées avec les experts (Siris 8 ou MISTRAL) ou avec les équipes Iria (Cyclades). Multics étant géré à Phoenix, le travail à l'IUCA n'avait plus le même intérêt. J'ai alors rejoint une équipe sur un projet nommé ANL pour Association Nationale du Logiciel.

[5] Bulletin Analytique de la Langue Française.

Ce projet était piloté 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'un inventaire de logiciels issus de laboratoires public, l'ANL est devenu un Groupement Scientifique pour la valorisation informationnelle des logiciels issus de la recherche. J'en ai pris la direction en 1981. Nous avions constitué un inventaire (informatisé) d'un millier de dossiers logiciels et nous organisions des expositions de logiciel en France et à l'international.

Nous étions 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 issu des études du CNET pour concevoir une station de travail sous le système Unix qui commençait son expansion. L'ANL a alors été sollicitée pour faire un inventaire des logiciels français disponibles sous unix, avec le montage de démonstration. 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 leurs outils initialement conçus pour des programmes structurés aux documents. Du côté de l'équipe technique ANL nous avons donc fait une utilisation intensive d'analyseurs lexicaux (lex) pour adapter nos données à des logiciels d'intelligence artificielle (Lisp ou Prolog).

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[6], 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. Cela ne me paraissait pas adapté à l'indexation des bases bibliographiques mais raisonnable pour les services de fournitures de documents. Or, 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[7]. J'ai ainsi découvert que, malgré mon expérience documentaire antérieure j'avais tout à découvrir en bibliothéconomie ! L'INIST a donc fait l'acquisistion, pour la bibliothèque, 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. Bien avant MarcXml de la Library Of Congress, 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. Puis en s'inspirant des chaines du TLF et de l'architecture MISTRAL nous avons spécifié des modèles SGML pour les données internes (fichiers inverses par exemple). Nous avons développé des modules (en langage C) qui permettaient de générer des systèmes de recherche avec des mécanismes de classification, dénommés serveur d'exploration.

2.5 Le démantèlement des missions stratégiques du CNRS en IST

Goéry Delacôte m'avait donné comme mission à l'informatique de redonner à moyen terme l'indépendance technologique (numérique donc) de l'INIST. L'action SGML entrait dans cette stratégie, mais dans un climat souvent très conflictuel. En effet, de nombreux cadres à qui le CNRS avait demandé de rester pour assurer une

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

[7] 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.

continuité de services étaient très impliqués dans les relations avec les sous-traitants et voyaient comme une menace la 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 les bases Pascal et Francis et les prestations de fourniture de documents. En dépit de la bonne qualité du nouveau service de FDP, le marché n'a pas suivi. Le CNRS a créé un Groupe INIST, piloté de fait par la filiale (et plus précisément par son chiffre d'affaires). Pour cela le CNRS a recruté des cadres plutôt 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. 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 dans un contexte difficile à la direction générale. Une nouvelle stratégie 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-là, les américains, et notamment la NLM avec qui nous faisions jeu égal en dans les années 70 jusqu'au premières années de l'INIST a plus que doublé sa production et possède de fait 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 où nous avons lancé une expérimentation sur l'usage des wikis sémantiques pour la promotion des résultats de la recherche. Nous avons monté un démonstrateur avec un wiki pour la région Lorraine et d'autres sur les priorités du CPER (eau, bois et forêts, sols urbains…). Puis, souvent pour des raisons démonstratives le réseau s'est enrichi pour atteindre maintenant 150 wikis.

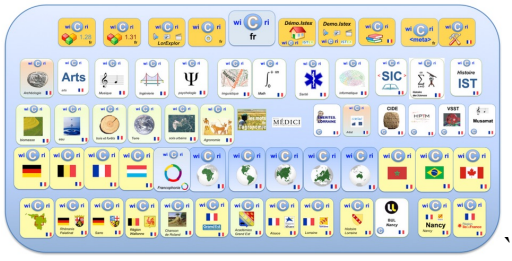

Figure 1. *Le réseau Wicri en 2023*

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. Pour être transparent, j'ai aussi travaillé la faisabilité d'une réponse pilotée par des coopération entre les laboratoires de 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 de Lepage sur le Palais Ducal de Nancy. Il a été réédité avec une organisation hypertexte, en reprenant notamment une gravure avec des renvois dans le texte de l'ouvrage où des annotations orientent le lecteur

vers des explications complémentaires dans la souche encyclopédique. 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[8]. Une équipe de recherche qui investit en formation devient donc totalement autonome.

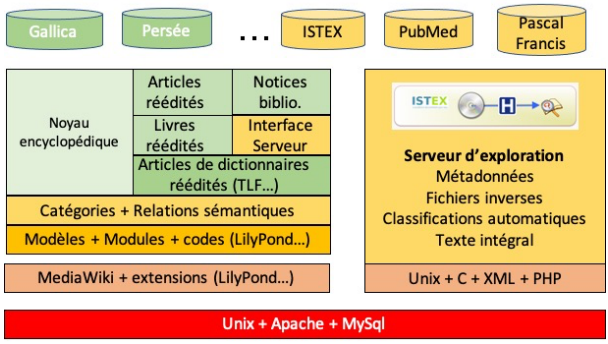

Figure 2. *Éléments d'un ensemble wiki + serveur d'exploration*

Aidés par un financement ISTEX, nous avons réalisé le couplage d'un wiki avec des serveurs d'exploration. Citée plus haut, cette approche a démarré à l'INIST en 1991. L'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. La boîte à outil iLib, à la façon du TLF, utilisait le tri standard pour traiter de gros corpus. Suffisamment stabilisée, après mon départ, 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. Une dizaine d'année plus tard, avec le programme LorExplor financé par ISTEX, 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.

[8] Par exemple, sur Wikipédia, les outils géographiques sont réalisés par les contributeurs.

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'engager une discussion sur une collaboration éventuelle. Je me suis donc rapproché du laboratoire Paragraphe, pour tester le potentiel Wicri au sein d'une équipe de recherche. Ne disposant d'aucune équipe à 75 ans, je me suis mis dans la peau d'étudiants de master en musicologie, philologie ou médiévistique, tout en restant aussi bibliothécaire, éditeur, et informaticien. Globalement j'agis un « chercheur praticien » en SHS, capable de programmer des modules récursifs, comme un chimiste résout des équations aux dérivées partielles. Des relations avec la BUL de Nancy ont fait émerger une thématique : *La Chanson de Roland.*

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.

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.

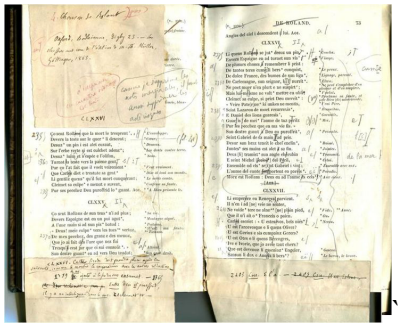

Figure 3. *Exemples d'annotation*

Le démarrage a été très rapide. Un expert en numérique (votre serviteur) a développé quelques modèles MediaWiki (mise en page…). Il a formé et encadré l'étudiant qui a produit des premiers 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 a conduit à mettre en place un projet plus conséquent avec un nouveau public : les choristes. En effet dans le cadre de travaux

sur Wicri/Musique, nous avons localisé une suite pour chœur et orchestre, composée par Gillles Mathieu et 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, en introduisant, à titre d'illustration, des facsimilés de feuillets du manuscrit.

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. Le manuscrit devient alors le composant fondamental au cœur de l'organisation numérique.

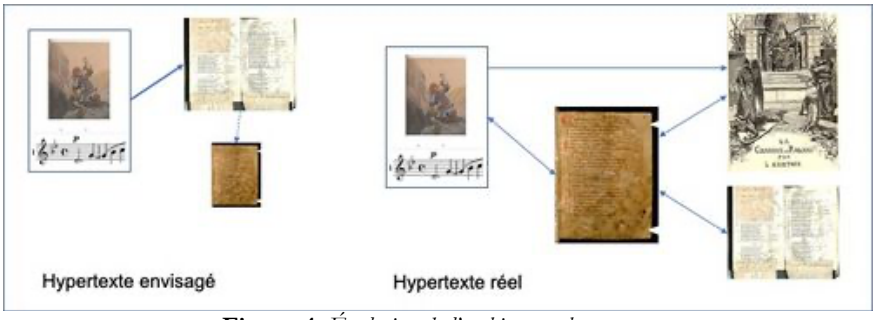

Figure 4. *Évolution de l'architecture hypertexte*

Nous avons donc modifié en profondeur le modèle initial. En quelques mois, nous disposions d'un ensemble déjà démonstratif. En préparant un séminaire de travail avec des philologues, nous avons localisé l'ouvrage cible des annotations de Paul Meyer, avec une nouvelle piste d'enrichissement du modèle.

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 avec des informaticiens, nous aurions été bloqués pour plusieurs mois en quelques jours. Nous avons donc décidé d'analyser le potentiel de cette thématique pour un projet conséquent de bibliothèque numérique. Nous avons donc décidé d'étudier une infrastructure numérique utilisable par des chercheurs pour leurs investigations et pas seulement pour la diffusion des résultats.

4.2 Des documents hétérogènes, très diversifiés et très interconnectés

Dans un premier temps, nous allons donner quelques exemples sur la diversité des documents de cette bibliothèque, et sur 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 types de 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 de description. Elle est plutôt destinée à la gestion. Par exemple, les images extraites (comme une lettrine pour alimenter un article) mentionnent leur appartenance à ce facsimilé. Réciproquement MediaWiki gère les liens inverses. Il est ainsi possible de connaître tous les extraits et de savoir où ils 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, un facsimilé de ces pages dans les pages feuillets pour permettre au chercheur de confronter une page du manuscrit avec une première interprétation. Nous introduisons ici un nouvel ouvrage avec un mode de traitement numérique très spécifique.

Chaque **laisse** (396, avec notre repérage) donne lieu à la création d'une page wiki où sont reproduits les facsimilés des pages correspondantes dans le manuscrit. Nous avons choisi 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 provenance diverse.

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

Rappelons que le porteur de ce projet était un peu béotien sur les Chansons de Geste pendant le stage initial. Une étude assez rapide de la littérature (sur Persée par exemple) montre qu'il faut considérer une dizaine de manuscrits sur la Chanson de Roland, et presque autour de dizaines sur des dizaines de poèmes épiques. Or chaque manuscrit implique de fait une étude particulière.

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 qui était dédiée à la version critique du manuscrit d'Oxford. Il traite également deux autres manuscrits. Le manuscrit dit de Paris dans son intégralité (6828 vers décasyllabiques répartis en 375 laisses monorimes). Mais ce manuscrit est incomplet, tout le début a été égaré. Francisque Michel a donc comblé de manque avec le début du manuscrit dit de Châteauroux (85 laisses, 1332 vers). Le manuscrit complet fait 8201 vers sur 452 laisses. Ces deux manuscrits seront donc traités intégralement sur le wiki. Mais l'organisation sera différente, par exemple une page du manuscrit d'Oxford est limitée à une vingtaine de vers pour une centaine sur celui de Paris. De plus un travail d'alignement avec celui d'Oxford est fondamental pour comprendre l'histoire de ce poème. Ceci pose en fait une multitude de petits problèmes qui vont 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. Celui de Léon Gautier joue un rôle particulier en raison de son utilisation par Gilles Mathieu (et de sa notoriété propre. Il s'agit d'un ouvrage conséquent (1000 pages sur 2 tomes). Nous avons vu que les 300 pages dédiées à l'édition critique proprement dite (et sa traduction) sont distribuées, laisse par laisse (au lieu d'une répartition « page paire, page impaire » dans l'original.

Les autres 700 pages sont réparties entre un glossaire de quelques milliers d'entrées avec des liens vers les vers du manuscrit d'Oxford ; 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 et manuscrits. Quelques dizaines d'entrées de l'index sont en fait de véritables articles. Ils sont alors extraits de l'index.

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.

Du côté de la musique

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.

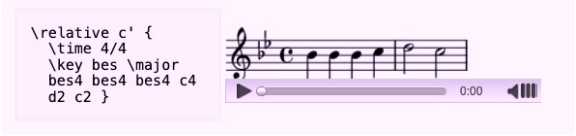

Figure 5. *Exemple de codification en LilyPond*

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. Du coup ils peuvent être reliés au glossaire de Léon Gautier (et 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 sont des documents qui bénéficient d'un traitement particulier. Nous avons évoqué le Dictionnaire de Musique de Jean-Jacques Rousseau sur Wicri/Musique. Ici, nous devons donner des explications pour des lecteurs non érudits. Le Grand Dictionnaire universel du XIXe siècle de Larousse est une source d'articles particulièrement intéressante (et facile à traiter).

Le dictionnaire du TLF, compte tenu de son histoire, bénéficie d'un traitement particulier. D'une part, nous voudrions démontrer qu'il peut être mis à jour dans une approche type Wicri. D'autre part, une spécificité pour ce wiki, 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 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 également créé un wiki spécialisé (Wicri/Francophonie) qui pourrait contenir l'ensemble du TLF et les 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 m'avait invité à travailler sur la Station de travail du chercheur à long terme. Dans son idée, l'INIST commençait à mettre à sa disposition de vastes ressources avec une excellence dans les traitements numériques exploratoires. Il m'avait invité à aller plus loin en travaillant sur des mécanismes qui auraient permis au chercheur d'enrichir, sans efforts, additionnels, cette architecture informationnelle. Cet objectif est-il atteint ?

5.1 Un bilan techniquement très satisfaisant pour… le porteur du projet

La puissance de MediaWiki, le moteur de la galaxie Wikipédia est maintenant incontestable. 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 personne pratiquement à temps plein pendant 18 mois ; là où Wicri/Musique est un wiki en concurrence avec la cinquantaine de familles. La première colonne comptabilise toutes les pages au sens hypertexte (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 culturel, 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 organisons fin aout 2023 à Aussois une manifestation musicale où le contexte culturel est explicité dans le wiki.

5.2 Les raisons des batailles perdues

Si les résultats techniques et scientifiques sont incontestables, le bilan institutionnel est celui d'une toute petite bataille perdue. En effet, 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), Francis (face à Oxford) et du TLF (face à Wiktionnaire). Pour Mistral, la concurrence est moins concentrée USA avec Geac au Canada ou Elasticsearch aux Pays-Bas. Ici, la réponse est politique. Dans une démarche purement française, au temps du Plan Calcul, de l'ADI et du démarrage de l'INIST les moyens humains dédiés au TLF et à l'IST voisinait les 700 personnes (500 titulaires, essentiellement CNRS et 200 occasionnels). La WikiMedia Fundation affiche également un effectif de 700 personnes. Les effectifs français de l'éducation, de l'enseignement et de la recherche est de 1.500.700[9]. La France est à la confluence de l'Europe et de la francophonie.

Cela dit, il conviendrait d'étudier les freins psycho-sociologiques qui amènent les agents du support de la recherche à rejeter l'approche wiki. Citons deux exemples. Les informaticiens intervenant en gestion ou sur des sites de communications institutionnelles sont des « adeptes de 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. Une expérience telle celle de la Chanson montre la permanence de besoins relativement imprévisibles de tout type d'expertise.

5.3 Des perspectives avec Wicri/Chanson de Roland

L'ensemble « Wicri + Dilib » est un démonstrateur et non un service opérationnel. Concernant Dilib par exemple, à l'INIST j'avais affecté 2 ingénieurs à mi-temps pour son évolution pour une utilisation au sein de l'INIST. Pour une opération au niveau européen, il faudrait environ 5 personnes. Le même type d'ordre de grandeur s'appliquerait à un moteur type MediaWiki indépendant des alea de la stratégie américaine. Dans son état actuel, Wicri/Chanson de Roland permet de réaliser des services expérimentaux et des expérimentations. Voici quelques exemples.

La réédition fine d'articles de recherche

La valorisation du fonds Paul Meyer offrait un premier axe de développement. Paul Meyer n'est pas seulement l'annotateur de Francique Michel, il a été directeur de l'École des Chartes en 1882 et un des fondateurs des revues *Romania* et de la *Revue critique d'histoire et de littérature*. La réédition de nombreux articles de ces revues est une voie de développement de la bibliothèque. Par rapport à nos expériences précédentes, une particularité est l'abondance des liens qui mériteraient d'être résolus (accès par un clic).

L'écriture d'articles de recherche

La rédaction de cet article (celui que vous lisez) est un exemple d'une publication savante qui devrait être intégrée à un ensemble numérique. En un seul clic sur un mouvement de la suite de Gilles Mathieu, le lecteur découvre en quelque secondes la structure d'un document. Cette connaissance au sein d'un article demande un page d'explication et plusieurs minutes laborieuse pour sa compréhension. D'un autre côté les contraintes héritées de l'édition papier oblige à

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

produire une étape cohérente, propre à être citée dans son cheminement scientifique.

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[10]. En 2022, la grande majorité de nos visiteurs de moins de 40 ans ignoraient tout de la bataille de Roncevaux.

Apports et limites du libre accès à l'information culturelle

La gestion des droits d'accès à l'information avec un haut niveau de confidentialité introduit un très haut niveau de contraintes qui entrainent des développements coûteux et complexes. Nous ne mettons donc en ligne que des documents libres de droits. Tout le monde peut lire les textes et seuls des contributeurs dûment identifiés peuvent intervenir sur les contenus.

6 Conclusion

La Chanson de Roland.

7 Bibliographie

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

[10] 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).