Systèmes d'information encyclopédiques édités par les scientifiques
Partager le savoir pour l'excellence documentaire et scientifique

Encyclopedic information systems edited by scientists
Sharing knowledge for documentary and scientific excellence

Jacques Ducloy[1]

[1] Université de Lorraine, Université Paris 8 (Paragraphe), retraité du CNRS (LORIA, Inist), Jacques.Ducloy@univ-lorraine.frq

RÉSUMÉ. Nous présentons une bibliothèque numérique structurée par une infrastructure encyclopédique. Des chercheurs, peuvent y exercer de façon collaborative, un large spectre de pratiques numériques, comme des explorations de corpus d'articles en texte intégral. Les textes, les données et les terminologies peuvent être mutualisées pour constituer de grands services de partage de connaissances (bases bibliographiques, dictionnaires, encyclopédies). Elle est réalisée avec un réseau de wikis sémantiques complété par une ingénierie XML. La conception de ce démonstrateur s'appuie sur une analyse de situations rencontrées à l'Inist.

ABSTRACT. This paper introduces a digital library structured by an encyclopedic infrastructure. Scientists can simultaneously and collaboratively perform many digital practices. The article gives examples in musicology and in the environmental sciences. It can also federate editorial actions or terminology field to constitute large knowledge sharing services such as encyclopedias, or more technical as bibliographic bases. On a technical level it is realized with semantic wikis completed by an XML engineering. The design of this demonstrator is based in particular on an analysis of situations encountered at INIST.

MOTS-CLÉS. musicologie, bibliothèque numérique, exploration de corpus, changement de paradigme, édition diplomatique.

KEYWORDS. musicology. digital library, corpus discovering, paradigm shift, diplomatic edition.

1. Introduction

Dans les années 1970, avec Pascal, Francis ou le Trésor de la langue française, la recherche française a été pionnière à l'échelle internationale sur les grands systèmes d'information scientifique. Comment, en 2020, retrouver une telle ambition, en s'appuyant sur les outils et pratiques fédératives du XXI^{ème} siècle ?

Ces grands projets s'inscrivaient au sein d'une des missions fondamentales du CNRS : *Le partage du savoir pour la communauté scientifique, les médias, et le grand public.* Dans les années 2000, Wikipédia est venu bouleverser le paysage en remplissant une mission citoyenne de partage du savoir. Au moment où le monde académique, humaniste et politique, s'interroge sur l'explosion de la désinformation sur les réseaux sociaux, faut-t-il laisser le monopole de la connaissance mondiale à un système qui repose sur l'anonymat ?

Progressivement, la *Wikimedia Foundation* est devenue un acteur terminologique incontournable, notamment sur le *Web sémantique*, avec DbPedia et WikiData. Comment redonner aux chercheurs et praticiens la maîtrise de leurs ressources sémantiques avec des mécanismes de validation scientifique ?

Le projet ISTEX[1] vient précisément d'ouvrir un nouveau défi en offrant à tous les chercheurs la possibilité d'explorer et de traiter des dizaines de milliers de documents. Pour atteindre *l'excellence documentaire pour tous* (le slogan d'ISTEX) ils vont devoir se former massivement aux pratiques de fouilles de données (TDM[2]). Il y a 50 ans, les ingénieurs, les physiciens et les chimistes ont dû massivement troquer leurs règles à calcul contre des paquets de cartes perforées pour bénéficier de la puissance de l'informatique naissante. En 2020, ils vont devoir acquérir une expertise dans le traitement de corpus textuels, en relation avec leurs données numériques ou symboliques. Est-il possible de canaliser cette énergie pour construire de gigantesques systèmes d'information scientifique ?

Pour approfondir ces questions, en nous appuyant sur une expertise acquise au Loria, à l'Inist et à l'ATILF, nous avons lancé une action de « partage du savoir » avec le réseau de wikis sémantiques Wicri. Sur cette base, avec le projet LorExplor, soutenu par ISTEX, nous avons exploré des corpus de publications en texte intégral avec une bibliothèque de composants XML nommée Dilib. Maintenant, sur quelques domaines pilotes, comme la musicologie, nous étudions comment cette architecture peut être utilisée pour créer de la connaissance, d'abord éditoriale, mais également terminologique, tout en gérant les données nécessaires à une pratique de recherche. Ces mécanismes sont-ils généralisables pour construire, en 2020, une infrastructure de la connaissance, pilotée par des scientifiques, avec comme finalité le partage du savoir, et apportant des retombées multiples dans les pratiques scientifiques ?

Nous donnerons, dans un premier temps un retour d'expérience sur les grands projets d'information scientifique qui se sont développés à Nancy. Nous proposerons un éclairage sur les problèmes rencontrés avec les changements de paradigmes intervenus depuis l'irruption du numérique dans la connaissance. Pour y faire face, nous présenterons une expérimentation qui, partant d'une intention de partage de la connaissance, débouche sur un réseau cohérent de bibliothèques encyclopédiques. Nous conclurons par quelques pistes de réflexion pour l'avenir.

2. Retour d'expériences sur des grands projets d'information scientifique

Avant-propos concernant cette section

Il y a 50 ans, j'ai troqué ma règle à calcul contre des cartes perforées pour calculer des fonctions de transfert en électrotechnique. J'ai alors eu la chance de rejoindre l'équipe des pionniers de l'informatique à Nancy et de faire un bout de chemin avec les acteurs du Trésor de la Langue Française. Il y a 30 ans, j'ai rejoint un autre groupe de chevaliers partis à la quête d'un Graal : *un pôle mondial pour le partage du savoir grâce à la maîtrise de l'Information scientifique et technique*. Ce n'est plus tout à fait la vision officielle de l'Inist, mais, *quelque part*, c'était la nôtre…

Cette section est le témoignage d'un ingénieur qui a voulu s'appuyer sur les résultats de la recherche et s'est trouvé confronté à une réalité souvent déroutante. J'utiliserai le pronom « *je* » pour marquer des situations qui exigeaient une implication individuelle. Concernant l'état de l'art, j'ai bien entendu parcouru de multiples publications qui ont planté le décor. Mais les éléments fondamentaux qui ont guidé mes décisions viennent de rencontres et d'expérimentations. La bibliographie sera ici utilisée pour associer ceux qui ont participé à cette aventure, et pour montrer la légitimité scientifique de nos choix technologiques.

[1] Le projet ISTEX (Initiative d'excellence de l'Information Scientifique et Technique) s'inscrit dans le programme « Investissements d'Avenir » < http://www.cnrs.fr/dist/projet-istex.html >

[2] Text and Data Mining

2.1. *Les débuts de l'informatique pour les données de la recherche à Nancy*

L'informatique à Nancy démarre avec le mathématicien Jean Legras qui explore, dès 1954, les retombées de l'analyse numérique pour les ingénieurs. Il encourage la création des bibliothèques de fonctions pour les aider à s'approprier cette technologie naissante[3]. Il fonde ce qui deviendra l'Institut Universitaire de Calcul Automatique (IUCA) dans les années 1970, en s'appuyant sur une coopération avec le Centre de Recherche pour un Trésor de la Langue Française (CRTLF) du CNRS. Grâce à ce partenariat, l'IUCA acquiert, dès 1974, des compétences opérationnelles sur les moteurs de recherche, et par exemple Mistral, développé par la CII, une référence historique du savoir-faire français dans l'information numérique.

En 1963, un autre mathématicien, Claude Pair, bâtit les fondations d'une informatique plus théorique qui donnera naissance au Crin puis au Loria. Il lance des travaux autour des langages de programmation (Algol 68), des structures formelles ou des techniques de compilation. Cet ensemble s'avèrera particulièrement signifiant pour les débouchés autour des documents structurés et l'ingénierie XML[4].

En 1980, la création d'un Groupement Scientifique ayant pour sigle ANL a joué un rôle essentiel pour nos orientations. L'Agence de l'Informatique (ADI), le CNRS, le Ministère en charge de la recherche, l'Inria et le CNET[5] voulaient créer un *Atelier National du Logiciel* pour transférer les logiciels issus des laboratoires vers l'industrie. Une étude a été lancée pour identifier les candidats et rédiger un catalogue basé sur des visites de laboratoire et sur des démonstrations. À la fin de l'étude, l'ADI a organisé 2 jours de séminaires et démonstrations où une cinquantaine d'équipes ont rencontré une centaine d'industriels. Ceux-ci ont été très sceptiques sur les missions d'un tel « Atelier », en gros : « *Nous savons industrialiser des logiciels, vous ne savez pas dans quoi vous vous lancez !* ». En même temps, ils ont enchaîné : « ***Mais**, nous avons besoin de l'étude que vous venez de faire. Et ça, nous ne savons pas le faire !* ». L'ANL est donc devenue une « Association Nationale du Logiciel » qui partageait le savoir par des tests de logiciels, des catalogues, un serveur, et des expositions. Grâce au soutien logistique et politique (Jean-Claude Rault, Robert Mahl) de l'ADI, nous avons pu monter des expositions internationales et notamment aux USA[6].

L'ANL publiait des catalogues et générait des serveurs (Mistral puis Texto), à partir de métadonnées. Impliqués dans la filière française de stations Unix (SM90), nous avons expérimenté des outils d'intelligence artificielle (Lisp, Prolog) sur nos données. Nous avons aussi mené des études comparatives. En effet, le modèle SGBD[7] relationnel nous paraissait plus séduisant que des traitements de fichiers dans des hiérarchies Unix, que nous vivions un peu comme du bricolage. Mais nos essais comparatifs donnaient toujours un avantage aux traitements basés en fait sur une modularité par flux de données. Cet avantage sera déterminant avec XML.

La suppression de l'ADI a déséquilibré l'équilibre financier de l'ANL. Une bonne partie de l'équipe a alors saisi l'opportunité de rejoindre l'Inist.

---

[3] Signalons la bibliothèque Cartolab, de Jean-Laurent Mallet, qui sera la base du consortium GOCAD sur le traitement de données géologiques < http://www.ring-team.org/ >

[4] Voir la thèse de Jean-Claude Derniame, en 1966 : *Étude d'algorithmes pour les problèmes de cheminement dans les graphes finis.* Un document XML est précisément un graphe fini.

[5] Centre national d'études des télécommunications, devenu *Orange Labs* en 2007.

[6] Par exemple en 1984 à Orlando dans le cadre d'une conférence *software engineering* de l'IEEE, où nous avions 5 stands partagés chacun par un industriel et un laboratoire.

[7] Système de Gestion de Bases de Données.

2.2. *Des bulletins analytiques du CNRS à ISTEX en passant par Pascal et Francis*

Une mission du CNRS, nous l'avons évoquée, est le partage des connaissances[8] :

« Le CNRS donne accès aux travaux et aux données de la recherche car ils font partie d'un patrimoine commun. Ce partage du savoir vise différents publics : communauté scientifique, médias, grand public. »

Dès sa naissance, en 1939, le CNRS a donc créé un centre de documentation, afin de communiquer avec les partenaires de la recherche sur **l'essentiel**[9] des résultats obtenus au niveau international. Jean Wyart en a rapidement pris la direction en 1941. Il a été rejoint par Nathalie Dusoulier en 61. Elle dirigera les bases de données en 1967, avant de prendre la tête des bibliothèques de l'ONU (Genève puis New-York) en 1978, pour enfin de revenir en France et créer l'Inist en 88.

2.2.1. *Des bulletins analytiques aux bases du CDST et du CDSH*

En 1940, paraît le premier bulletin analytique. Il était réalisé par des ingénieurs qui résumaient des articles et les indexaient. Une anecdote illustre le caractère réellement éditorial de ces bulletins. Quand nous soumettions une note ou un rapport à Nathalie Dusoulier, nous avions régulièrement ce type de remarque : « *Votre deuxième paragraphe est le double du premier alors qu'il est dix fois moins important. Revenez donc avec un texte dans lequel la taille de vos paragraphes sera proportionnelle à l'importance de vos arguments !* ». Appliquée aux bulletins, cette pratique permettait au lecteur de repérer « **l'essentiel** » par un simple feuilletage. Autrement dit, comme le montre l'évolution de Gallica (Laborderie 2015) : un flux RSS, ou une base de données, ne remplace pas un travail éditorial !

La création des bases Pascal et Francis en 1971 est une préfiguration des applications de type *big data* au CNRS. Comme le confirment des témoignages (Burh 1977), les caractéristiques techniques étaient remarquables. La production atteignait déjà 500.000 références par an. Les notices numériques étaient générées dans un format normalisé ISO 2709 (voir plus bas) à partir des fichiers de photocomposition. En 1977 Pascal était déjà accessible sur 3 sites, via le réseau Cyclades, avec le progiciel Recon sur IBM 360 et avec Mistral sous Iris 80. Cette aventure se poursuivra avec la création de Télésystèmes qui deviendra Questel.

Ce succès initial fut suivi de difficultés qui ont joué sur les motivations du transfert à Nancy. Lors de ma nomination comme directeur informatique à l'Inist, et, en même temps au CDST et au CDSH[10], j'avais notamment constaté un blocage décisionnel très important. En effet, les fonctions qui me semblaient vitales pour la conception des services étaient complètement externalisées chez des sous-traitants.

Les deux centres avaient chacun un profil de fonctionnement assez différent. Le CDST, grâce à ses formats normalisés, pouvait nouer de multiples coopérations, basées sur des achats/ventes de notices, avec d'autres centres ayant la même technologie. Mais il était organisé autour d'une chaîne de production, qui reposait sur des sous-traitances externes ; ce qui paralysait l'unité. En revanche, le CDSH était organisé avec un système « plus rustique » mais qui offrait des possibilités de

[8] < http://www.cnrs.fr/fr/missions >

[9] Pierre Auger avait repris les ambitions des listes de grandeurs physiques : *Nous relevions l'essentiel de ce qui se faisait dans toutes les langues intéressantes à l'époque.* Cité par Jean Astruc dans : *le CNRS et l'information scientifique et technique en France* (Solaris 1997) < https://lorexplor.istex.fr/Wicri/Ticri/fr/index.php/Solaris_(1997)_Astruc >

[10] Le CDST et le CDSH étaient les 2 centres de documentations du CNRS en 1988, respectivement pour les sciences et techniques et pour les sciences humaines.

coopérations avec un réseau de laboratoires (principalement français). Ce modèle inspirera la conception en réseau du projet Wicri.

2.2.2. *ISO 2709, un socle normatif pour les bibliothèques de la recherche*

Nous avons cité la norme ISO 2709 (ou MARC, acronyme de *MAchine-Readable Cataloging*). Elle désigne une norme générique qui jouera un rôle très important dans nos choix technologiques. Elle décrit les données bibliothéconomiques sous la forme de zones, repérés par des codes, et de sous-zones. Chaque format d'application peut définir sa nomenclature propre. Par exemple la zone 210 dans le *Commons Communication Format* (CCF) de l'UNESCO correspond à un titre parallèle, exemple[11] :

210 0 1 **@a**Legislatives studies**@l**eng

Ce même code correspond au lieu de publication dans Unimarc (utilisé à la BnF).

210 ## **$a**Berlin**$a**Heidelberg**$a**New York**$c**Springer**$d**2004

Les formats MARC sont encore très largement utilisés dans le monde des bibliothèques (et de l'IST), grâce à une adaptation à la norme XML (XmlMarc et MARC 21). À la création de la base Pascal, le format UNISIST sous ISO 2709 a été choisi. Le CSDT utilisait également le format LCMARC, basé aussi sur ISO 2709, pour gérer sa bibliothèque.

Nathalie Dusoulier avait piloté la numérisation des bulletins signalétiques dans Pascal. Elle a ensuite informatisé le réseau des bibliothèques de l'ONU, en utilisant Unimarc. Elle dirigeait enfin un groupe de travail international de l'Unesco sur le CCF, dédié cette fois à la documentation. L'Inist disposait donc, à son démarrage, d'un socle solide pour des coopérations internationales, mais relativement complexe à maîtriser. En effet, l'installation exhaustive d'une norme MARC dans un SGBD s'avère très lourde. De plus, son implémentation définie dans les années 70 (à base de pointeurs) demandait un bon savoir-faire en codification et en algorithmique.

2.2.3. *De la création de l'Inist à l'aventure XML*

L'Inist a été créée sur Nancy en 1988 sous la direction de Goéry Delacôte, à la DIST du CNRS, et de Nathalie Dusoulier à la tête de l'institut.

Goéry Delacôte m'avait donné comme objectif de transférer la maîtrise de la chaîne de production, de la sous-traitance vers l'Inist. De même, il m'avait demandé d'intégrer une alimentation des bases par des coopérations avec les laboratoires (à la façon du CDSH). Enfin, il était tout à fait partisan de reprendre, au moins en partie, et d'améliorer les services en ligne, qui étaient également assurés en sous-traitance.

Nous disposions d'un schéma directeur qui planifiait les actions informatiques en 2 grandes étapes. Pendant 2 à 3 ans, il préconisait l'informatisation de la bibliothèque et la création d'une application fourniture de documents (FDP) intégrant un serveur d'archivage numérique. La refonte de la production des bases de données était envisagée dans un deuxième temps, en s'appuyant sur cette première infrastructure.

Sur un plan technique, le schéma directeur recommandait « un système totalement intégré par un SGBD, et *si possible sur un mainframe IBM*[12] ». À mon arrivée, l'ordinateur avait été commandé et

[11] Les conventions de présentation sont différentes suivant les formats. Dans le CCF @l désigne la sous-zone « l » pour langue, et dans Unimarc on utilise $d pour désigner la sous-zone date.

[12] Bien entendu, ce n'était pas formulé dans ces termes, mais presque ! En effet, mon premier travail en arrivant comme directeur a été de sauver le dossier d'acquisition d'une configuration IBM bloquée par une commission de contrôle des marchés (CSMI).

une première équipe avait été recrutée. Elle était composée d'ingénieurs très compétents, qui avaient fait leurs preuves dans des applications de gestion, mais pas encore dans la bibliothéconomie. La mise en place du schéma directeur a tout de suite révélé des dissensions au sein de l'institut. Le problème s'est posé dès le départ avec l'informatisation de la bibliothèque.

J'avais une expertise en documentation (et sur les gros systèmes informatiques pour scientifiques). Mais je ne connaissais rien à la bibliothéconomie ! J'ai bénéficié d'une formation accélérée en travaillant sur le dossier FDP avec Nathalie Dusoulier, et avec mes collègues de la bibliothèque sous la direction de Caroline Wiegandt.

S'appuyant sur son expérience à l'ONU, Nathalie Dusoulier n'était pas convaincue par l'intégration de la bibliothèque dans un ensemble intégré. Elle nous a demandé de travailler sur une alternative avec un système dédié, communiquant avec la FDP par un transfert de fichiers normalisés (Unimarc). De son côté, l'équipe de développement informatique souhaitait naturellement un système totalement intégré, intégrant donc la bibliothèque. Pour ma part, je n'avais pas d'avis a priori concernant la bibliothèque. En revanche, j'étais plutôt favorable à une constitution des bases de données par un réseau de machines départementales sous Unix. En fait, j'ai très rapidement rejoint la position de Nathalie Dusoulier, d'abord, en analysant la structure des formats Unimarc, et par des visites de quelques bibliothèques déjà automatisées.

Un appel d'offre a donc permis d'acquérir un système Geac d'origine canadienne[13]. L'informatisation de la bibliothèque de l'Inist a été vécue comme un succès total, en parallèle avec l'installation de la FDP et de son serveur d'archivage, sous la direction informatique finale de Francis André.

En effet, il me paraissait urgent, pour traiter les données bibliothéconomiques de la recherche, de maîtriser la manipulation des notices de métadonnées codées en ISO 2709. Goéry Delacôte avait prévu de doter l'Inist d'une structure de recherche et développement, le DRPN[14]. J'avais donc demandé de quitter la direction informatique pour prendre celle du DRPN. J'espérais ainsi, à court terme, résoudre la maîtrise de ces métadonnées complexes, puis passer ensuite à l'indexation assistée.

Au-delà de l'inadéquation d'un modèle relationnel pour gérer l'aspect générique des fichiers MARC, trois problèmes très concrets m'avaient alerté. Tous les services d'extractions sur les bases Pascal utilisaient un logiciel nommé VIRA, développé dans les années 70 sur IBM 360, et que personne ne maîtrisait. Des statistiques simples demandées par les ingénieurs documentalistes demandaient en moyenne 3 jours par demande (en effet, la technique consistait à trouver un programme correspondant à un cas voisin, le recopier et le modifier…). Enfin, personne, ni dans l'équipe SGDB interne, ni chez le sous-traitant n'avait su résoudre, dans un temps raisonnable, la connexion entre la bibliothèque et la FDP (j'ai dû intervenir directement, en une semaine, pour éviter 3 à 6 mois de retard).

Une rencontre avait précipité les évènements. J'avais commencé à étudier des formalismes de type LISP pour remplacer la souche ISO 2709. Mais je n'étais pas très satisfait de mes maquettes qui, cela dit, préfiguraient JSON ! Et puis, mon successeur à l'ANL, Jacques Guidon, m'a mis en contact avec François Chahuneau qui était responsable de l'innovation chez Berger-Levrault. En quelques dizaines de minutes, j'ai été convaincu qu'une ingénierie basée sur SGML était la solution prometteuse.

[13] Cette normalisation a permis son remplacement sans problème dix ans plus tard.

[14] Département Recherche et Produits Nouveaux

Pour les lecteurs non familiers avec le formalisme XML, la norme SGML permet de manipuler des arbres de profondeur quelconque, et donc des formats MARC. Par exemple la zone 210 d'une notice CCF citée plus haut peut être codée ainsi :

< f210 i1="1" i2="0"><sa>Legislatives studies</sa><sl>eng</sl></f210>

Au bout de quelques mois, nous disposions d'une plate-forme SGML (Ilib). Nous avons eu des retombées immédiates en termes de publications, au départ dans le monde du génie logiciel (Ducloy 1991). Nathalie Dusoulier a présenté notre approche à la communauté Unimarc/CCF[15] (Dusoulier 1991). Une équipe dirigée par Xavier Polanco s'appropriait Ilib pour des études infométriques. Avec Valérie Warth, nous avions réalisé un noyau de parser SGML avec une approche XML/DOM.

Nous pouvions passer aux bases de données où, sur des créneaux différents, nous faisions jeu égal avec la National Library of Medicine aux Etats-Unis. En 1996, Olivier Bodenreider, un chercheur d'une équipe d'informatique médicale de Nancy avait d'ailleurs rejoint Bethesda où il est maintenant *Chief of the Cognitive Science Branch* de la NLM[16]. À l'Inist, en 1991, une équipe, menée par Laurent Schmitt, avait déjà réalisé STID, un prototype de station de travail pour l'indexation (Schmitt 1992). L'Inist avait donc de bons atouts dans cette compétition.

Nous avons aussi travaillé avec les bibliothécaires de l'Inist pour les aider à analyser des résultats de reformatages LCMARC vers Unimarc. Il ne restait plus qu'à former les informaticiens pour que la direction Informatique puisse déjà récupérer 2 postes sur le traitement des demandes des ingénieurs documentalistes.

☺ Et c'est là que les ennuis ont commencé !

Avant d'aborder une nouvelle étape, il faut signaler une situation tendue avec la direction de la production. Voici un exemple. Pour un membre de l'encadrement méritant, la récompense était « d'être chargé d'une mission de la prospective avec le CNRS, ou d'un marché avec un sous-traitant, ». Avec ses résultats potentiels, le DRPN, composé de jeunes recrues, venait perturber ce type de perspectives.

2.2.3. *La réforme Eisenmann*

En 1992, un changement important est intervenu dans la gouvernance de l'Inist avec le départ de Goéry Delacôte. La direction du CNRS a alors demandé à Etienne Eisenmann de faire des propositions « *pour l'adaptation des produits, services et structures de l'Inist* ***aux besoins du marché européen*** ». En effet, l'Inist s'était doté d'une filiale, Inist Diffusion, qui était chargée de la commercialisation des services. Mais, en dépit de la réussite technique, le marché n'était pas au rendez-vous. Etienne Eisenmann a donc mis en place une profonde réforme pour créer un « Groupe Inist ».

Les priorités des services ont été inversées. La fourniture de documents, qui était un service d'accompagnement, est devenue l'axe prioritaire de l'Inist. Concernant la base Pascal, Etienne Eisenmann venait du secteur pharmaceutique et biomédical où les problèmes d'antériorité sont primordiaux pour les prises de brevets. La fraicheur de l'information, est devenue le « critère de qualité » prioritaire pour satisfaire ce marché, au détriment du partage de connaissance[17].

[15] en présence de représentants de la Library of Congress (qui reprendra le concept quelques années plus tard avec XmlMarc).

[16] < https://lhncbc.nlm.nih.gov/personnel/olivier-bodenreider >

[17] Par exemple, les résumés d'analystes ont été remplacés par les résumés d'auteur.

Une partie de l'équipe de direction, dont Nathalie Dusoulier, a été remplacée par des cadres venus du secteur privé, essentiellement recrutés sur leurs compétences en marketing ou en gestion de production. La plupart n'avaient pas d'expérience en bibliothéconomie, ni dans les métiers de l'édition, ni dans la recherche. L'externalisation des fonctions complexes vers la sous-traitance est redevenue la solution pour résoudre les problèmes techniques. La nouvelle direction informatique préconisait un système intégré et réprouvait toute activité informatique hors de son périmètre. Le fait d'avoir obtenu des résultats exploitables a en fait précipité le démantèlement du Département Recherche et Produits Nouveaux.

J'ai été nommé « chargé d'études prospectives ». Je disposais de quelques minutes par mois, lors du repas de l'encadrement, pour communiquer avec mes collègues. Pour ma première intervention, au moment du dessert, j'ai fait une démonstration en montrant trois documents récupérés « gratuitement » sur un serveur FTP. La sanction a été immédiate : mon accès internet a été coupé dans l'heure qui a suivi !

2.2.4. *L'action autoroutes de l'information au Loria*

Nathalie Dusoulier et Jean-Pierre Finance m'ont tiré de ce mauvais pas. Muté au Crin, j'ai pu créer une nouvelle plateforme (Dilib) qui sera évoquée plus loin. Grâce à l'expérience infométrique acquise au DRPN, j'ai pu monter en 1992-1993 un premier scénario de génération d'un site Web (préfigurant les serveurs d'exploration) à partir d'un corpus bibliographique. J'ai alors été soutenu par le Crin et l'Inria Lorraine (Patrick Rambert) pour monter une « action autoroutes de l'information ».

Grâce aux contacts noués au temps de l'ANL (par exemple Georges Nissen à l'Inria), je suis entré dans les programmes européens de l'ERCIM, et, par ce biais, au DCMI (*Dublin Core Metadata Initiative*). Les retombées ont été intéressantes pour le Loria et ses partenaires avec une appropriation de la technologie XML et de mécanismes d'exploration de corpus, à travers des projets comme MedExplore[18] ou Biban[19]. Au niveau national, j'ai tenté de promouvoir XML auprès de mes collègues de l'Inria dont la stratégie reposait alors sur les bases de données objet avec O2. En fait XML n'y est devenu populaire qu'à partir de 1995, avec la prise de direction dans le W3C. Au niveau européen, nous avons soumis deux projets *Digital Libraries* (Samos[20] et Imesis[21]) qui étaient plutôt bien classés mais n'ont pas été retenus. Avec du recul, nous n'aurions pas eu les moyens de les assurer dans de bonnes conditions, faute d'un opérateur de R&D tel que ce qui avait été prévu avec le DRPN de l'Inist.

2.2.3. *L'arrêt des bases Pascal et Francis*

En 1999, le CNRS a mis fin à la réforme Eisenmann et chargé Alain Chanudet de réinsérer l'Inist au CNRS. En 2000, j'ai été rappelé à l'Inist pour diriger le département qui supervisait les services et la fabrication des bases Pascal et Francis.

La direction du CNRS avait préconisé une option « indexation automatique » en vue de réduire les effectifs. Elle avait déjà lancé une première vague de départs. Nous avons été confrontés à une situation où il fallait, en fait, plus de ressources humaines pour piloter des mécanismes d'indexation que pour assurer un traitement manuel. Nous avons fait passer l'idée d'une indexation assistée et

[18] Analyse des mécanismes psycho-cognitifs mis en œuvre pour explorer des bases médicales.

[19] Base Bibliographique et Iconographique Art Nouveau (navigation dans des bases d'images)

[20] SAMOS voulait réaliser une bibliothèque numérique distribuée à partir du protocole DIENST de Carl Lagoze (Cornell). Problème : le projet préfigurait le libre accès et… Elsevier faisait partie du consortium !

[21] IMESIS était un projet Euro-Méditerranéen en santé publique avec une dizaine de partenaires.

lancé un plan de formation appelé *mutation technologique*. En effet, il nous paraissait important que les ingénieurs documentalistes soient mieux armés pour gérer des spécificités thématiques (pour programmer des heuristiques par exemple). J'espérais (naïvement ?) inverser la tendance sur les effectifs en cherchant une forme d'excellence dans les secteurs rescapés.

Malheureusement, une vague de réductions budgétaires a rendu la situation encore plus difficile. Deux ans après ma nomination, il n'était plus possible d'embaucher des vacataires ou d'acheter des notices venant d'autres bases. Les ingénieurs documentalistes ont dû alors doubler les cadences et ont très mal vécu le sentiment de faire un travail dont la qualité se dégradait en permanence.

Dans le cadre d'une mission CNRS présidée par Bernard Pau, nous avons élaboré un plan de réforme de Pascal et Francis. L'idée était de passer d'une production de 500.000 notices de qualité médiocre par an, à 50.000 mais avec une excellente qualité, (avec 50% de couverture française, et l'essentiel de l'international). Pour les services en ligne de grande volumétrie, nous aurions utilisé des techniques d'apprentissage. Mais, avec du recul, la situation était ingérable. En effet, avec l'arrêt du DRPN, aucune solution interne n'était disponible. Dans un climat devenu tendu à la fois sur le terrain et avec la direction du CNRS, la réforme a été abandonnée. Les bases Pascal et Francis ont continué leur déclin pour être arrêtées 10 ans plus tard.

Avec Francis André, nous avons alors monté une cellule prospective au sein de l'Inist, où, en coopération avec Sylvie Lainé-Cruzel, nous avons créé une activité éditoriale autour de l'appropriation des technologies de l'IST par les laboratoires. Nous avons commencé avec un blog scientifique ARTIST[22]. Sur cette base nous avons lancé la revue AMETIST (avec un comité de rédaction international). Mais le climat devenait de plus en plus difficile à l'Inist[23] et j'ai rejoint la DRRT Lorraine. La revue AMETIST a été retirée du Web peu après dans des circonstances rocambolesques[24].

2.2.4. *Un rebond potentiel avec ISTEX*

La réussite actuelle d'ISTEX mérite d'être mise en avant. Rappelons qu'ISTEX met à la disposition des chercheurs plusieurs dizaines de millions de documents scientifiques en texte intégral. Pour les aspects techniques, nous donnerons quelques éléments plus loin à propos de l'expérience LorExplor. Par rapport aux épisodes précédents, la mise en œuvre d'ISTEX est intéressante. En effet, elle a été réalisée intégralement par des ingénieurs du CNRS ou encadrés par eux. Cette rupture avec la culture interne de l'externalisation a été un succès technique incontestable. La plateforme centrale est opérationnelle. Elle donne à l'Inist un ensemble de ressources qui constituent un trésor, au sens de ce mot à la Renaissance. Si l'Inist avait disposé de ce corpus dans les années 2000, le redressement de Pascal avec l'option apprentissage aurait été nettement plus facile à gérer ! Comment la communauté scientifique française peut-elle maintenant s'approprier ce Trésor ? L'histoire du Trésor de la Langue française peut donner quelques enseignements.

[22] Appropriation par la recherche des technologies de l'Information Scientifique et Technique

[23] Le fait de préconiser aux laboratoires de s'approprier les technologies entrait en contradiction avec l'équilibre financier de l'Inist qui reposait sur des prestations. Nous étions également en contradiction totale avec les principes de l'informatique d'administration.

[24] Le site ARTIST était géré avec le CMS SPIP qui confond les notions d'auteur et les autorisations à contribuer. La direction de l'Inist souhaitait m'interdire d'y intervenir, mais ce faisant, elle violait la loi car je n'apparaissais plus en tant qu'auteur sur les articles existants. La solution « juridique » a été expéditive : tout a été supprimé, y compris la revue !

2.3. *Du Trésor de la Langue Française à l'ATILF*

Le démarrage du Trésor de la Langue Française relève du roman d'anticipation[25]. Nous sommes en 1955. Les disques magnétiques ne sont pas encore inventés[26]. Des linguistes et des philologues (Bernard Quémada, Paul Ibms) utilisent des machines mécanographiques. En 1959, un projet de « *mise en chantier d'un Trésor général de la langue française ou Dictionnaire historique général de la langue française* » figure dans le rapport de conjoncture du CNRS. Fin 1960, le CRTLF est créé. En 1961, un des plus gros ordinateurs du monde, un Gamma 60, est commandé à la compagnie des machines Bull. Il arrivera à Nancy en 1963 avec 10 dérouleurs de bandes magnétiques, 2 lecteurs de ruban perforé, 3 imprimantes, mais sans disques magnétiques.

Au départ, il s'agit « simplement » de constituer « le Trésor ». En 1963, 22 opératrices-mécanographes commencent les opérations de saisie de textes à raison de 100.000 mots par jour. En 1970, une base initiale de 1000 textes dans lesquels chaque mot était étiqueté par sa catégorie grammaticale a été constituée.

En 1968, le projet de dictionnaire prend corps et des projets d'articles sont évalués. Puis la rédaction définitive démarre. Le premier tome, daté de 1971 est présenté au public en 1972. Le dernier tome sortira en 1994, à l'issue de longues pérégrinations. L'ensemble représente 16 volumes, 100.000 mots, 270.000 définitions, 430.000 exemples et 350 millions de caractères.

Les traitements informatiques ont été conçus dans les années 65, avec les bandes comme mémoire de masse[27] et les imprimantes pour l'interface homme machine. Pour chaque mot de faible fréquence, une liste de concordances suffisait au rédacteur. Pour les mots plus courants, des algorithmes basés sur des associations, les *groupes binaires*, ont été développés. Les chaînes de traitement étaient décomposées en étapes qui s'étalaient environ sur un mois, avec, en fin de phase des tris qui mobilisaient 6 à 8 dérouleurs pendant plusieurs heures. Ces contraintes historiques (pas de disque, peu de mémoire centrale au départ) ont favorisé un style de traitements basés sur une alternance de tris et de programmes relativement simples.

Curieusement, ces chaînes ont eu une forte influence sur Dilib. En effet, le CRTLF utilisait, à partir de 1972, le CII 10070 de l'IUCA où j'étais ingénieur. Nous avions été sensibilisés par la modularité imposée par la manipulation de corpus sur bande. Nous l'avons transposée sur Dilib avec d'excellentes performances en utilisant le mécanisme des « *pipes* » d'Unix et le tri standard (*sort*). Nous avons déjà cité Mistral qui avait été utilisé pour l'informatisation du BALF[28]. À titre anecdotique, nous avions aussi développé un « *jeu du mot le plus long* » avec environ 200.000 formes fléchies venant du TLF[29].

L'équipe informatique du CRTLF, rencontrait des problèmes assez voisins de ceux qui ont été cités à l'Inist à propos des formats MARC. Elle était constituée en majorité par des techniciens qui n'avaient pas reçu de formation approfondie en algorithmique. Ils étaient visiblement très compétents pour comprendre les problèmes lexicographiques et adapter les programmes qui constituaient les chaînes de traitement autour des tris. Mais les linguistes du CNRS n'avaient pas

[25] Voir notamment la thèse de Ruth Radermacher : http://www.atilf.fr/IMG/pdf/These_Radermarcher_Ruth_2004.pdf

[26] Les mémoires de masse alors testées sont les tambours magnétiques.

[27] Dans les années 75, la bandothèque contenait environ 2000 bandes magnétiques, chacune pouvait stoker 20 millions de caractères.

[28] Bulletin Analytique de la Langue Française

[29] Cette application est un parcours dans l'arborescence des anagrammes des formes fléchies. Elle posait des problèmes sur un Iris 80 où l'arbre ne tenait pas en mémoire et où il fallait éviter les appels aléatoires sur disque. Cette expérience a été très formatrice, pour aborder, 20 ans plus tard, des classifications sur une année de Pascal (500.000 références), ou maintenant sur ISTEX.

réalisé la complexité d'un univers de données où il fallait résoudre des parcours dans des graphes de taille considérable avec d'énormes contraintes techniques. Ce problème a été résolu dans les années 80 avec l'arrivée de Jacques Dendien. En 1986, le TLF disposait d'un moteur de recherche permettant de manipuler des éléments de grammaire sur la base FRANTEXT qui contient maintenant 5390 références soit 253 millions de mots. Dans la foulée, il a également développé une mise en ligne du dictionnaire (le TLFi), qui est maintenant en accès public depuis la direction de Jean-Marie Pierrel à l'ATILF.

Mais, comme Pascal, comme Francis, le dictionnaire TLF n'est plus maintenu.

3. Changements de paradigmes dans la connaissance numérique

Nous venons de décrire l'abandon de deux systèmes complexes d'exploitation de données de la recherche qui répondaient aux missions de transfert de savoir du CNRS et des universités. Ces réalisations ont mobilisé pendant des décennies des centaines d'ingénieurs avec un noyau conséquent de décideurs. Tous ces acteurs ont été recrutés en fonction de leurs compétences attestées dans d'autres circonstances. Avant de donner des pistes pour de nouvelles applications, nous proposons une réflexion sur les changements de paradigmes pour une analyse systémique de ces évènements.

3.1. *Les quatre paradigmes de Jim Gray*

Pour alerter les décideurs américains sur la révolution numérique, Jim Gray (Gray 2005) avait défini quatre paradigmes dans les pratiques de la recherche.

1 : Pendant des millénaires, les premiers érudits avaient une méthodologie empirique basée sur l'observation.

2 : Puis, depuis quelques siècles, avec Maxwell ou Newton, les scientifiques utilisent des modèles théoriques faisant appel aux abstractions et aux généralisations, afin d'établir des « lois universelles ».

3 : Depuis environ 1950, quelques décennies, ils utilisent des ordinateurs pour modéliser des phénomènes complexes. La programmation devient un outil de travail et d'expression du chercheur.

4 : Nous entrons maintenant dans une nouvelle étape où les chercheurs doivent maîtriser le déluge de données.

3.2. *Les ingénieurs et physiciens face au troisième paradigme de Jim Gray*

La façon avec laquelle les chercheurs ont géré ces mutations il y a cinquante ans donne des pistes pour analyser les problèmes rencontrés maintenant sur les données numériques. Par exemple, l'histoire de l'informatique à Nancy, révèle des conflits entre mathématiciens et pionniers de l'analyse numérique. En effet, en 1956, dans son livre sur la résolution des équations aux dérivés partielles, Jean Legras, écrivait :

« 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. »*

Assumant pleinement cette réprobation, 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. »*

3.3. *Le document structuré face au paradigme relationnel*

Par rapport aux grandes étapes tracées par Jim Gray, les pratiques documentaires ajoutent de nouveaux paradigmes « secondaires ». Concernant l'Inist, les conflits informatiques des années 1990 sont révélateurs d'un changement de paradigme mal identifié. Pour l'immense majorité des formateurs, des décideurs et des sociétés de service, le ***complexe inhibitif de rigueur*** était « un système de gestion de données intégré géré par un SGBD relationnel ». En effet, en 1990, ceux-ci offraient une approche globale avec des outils méthodologiques comme MERISE. Issues du monde de la compilation, les technologies du document structuré ont dû attendre 1996 et la généralisation d'XML pour obtenir le même niveau de complétude.

L'émergence d'une technologie de rupture implique alors des prises de décisions qui ne peuvent pas encore être aidées par un soutien méthodologique. Par exemple, dans l'informatisation de la bibliothèque de l'Inist, Nathalie Dusoulier, nous a amené à changer notre vision sans pouvoir faire une démonstration formelle. Ceux qui avaient vécu une expérience comme celle de l'ANL sont facilement arrivés à une « conviction commune » sur la séparation des applications. La suite nous a donné raison, et, nous aurions perdu au moins 3 ans avec un système intégré. Mais, nous avons pris cette décision sur une forte **conviction** ! Nous aurions été incapables de produire un argumentaire recevable par une commission de validation composée d'informaticiens des systèmes d'information du CNRS. De même une grande partie des personnels de l'Institut, notamment chez les informaticiens, ne partageaient pas notre point de vue. En revanche, Goéry Delacôte nous a fait confiance.

3.4. *Autres ruptures liées à la nature du document et de la connaissance en 2020*

En 2019, le succès rencontré par ISTEX avec une infrastructure « *file system* » sur le moteur de recherche *Elasticsearch* montre une évolution considérable dans la conception des services de recherche d'information. Mais de nouvelles technologies de rupture apparaissent.

Nous allons présenter une expérimentation basée sur une technologie du XXI^{ème} siècle, illustrée par MediaWiki. Elle introduit trois ruptures conséquentes. Le simple usage du wiki rompt déjà le principe de validation *a priori,* pour des mécanismes de modération *a posteriori*. Ensuite, le paramétrage de MediaWiki met l'algorithmique à la disposition de l'utilisateur. Il permet à chaque discipline scientifique de définir ses propres applications, mais il brise la séparation des rôles entre les informaticiens et les utilisateurs. Enfin, la généricité de MediaWiki permet aux acteurs de travailler ensemble mais en brisant les périmètres traditionnellement gérés par différents chefs de projets bien identifiés. De son côté, l'exploitation des corpus ISTEX à des fins de recherche (et pas seulement d'évaluation) montre le besoin d'appropriation de compétences TDM par les chercheurs eux-mêmes - ce qui n'est pas toujours bien perçu. Nous avons donc rencontré en 2015 le même type de difficultés que celles de 1991. Le projet LorExplor voulait analyser en profondeur les besoins des chercheurs dans les explorations de corpus. Initialement calibré pour une dizaine de permanents, et soutenu par ISTEX, son effectif en permanents s'est finalement réduit à un retraité.

4. Wicri, une architecture de bibliothèque numérique basée sur une encyclopédie

Le projet Wicri est né en 2008 avec la mission Ticri, lancé par la DRRT Lorraine dirigée par Jean-Pierre Thomesse[30]. Nous voulions sensibiliser les acteurs régionaux sur les enjeux du

[30] Ticri signifie : Technologies de l'Information et de la Communication pour les Communautés de la Recherche et de l'Innovation. La DRRT est le sigle de la Délégation Régionale à la Recherche et à la Technologie (Ministère de l'Enseignement supérieur et de la Recherche)

numérique pour valoriser les résultats de la recherche. En même temps, nous voulions aussi nous approprier une technologie qui aurait pu être utilisée à grande échelle, par exemple, pour la production des bases Pascal et Francis.

Pendant quatre ans environ, nous avons bénéficié d'un montage CPER (contrat de projets état région) et de la logistique de l'INPL. Sur les conseils du Loria, Thierry Daunois et Alice Hermann, ont utilisé les extensions sémantiques *Semantic Mediawiki* pour réaliser un inventaire chiffré et commenté des équipements financés par le CPER. Puis, nous avons monté des expérimentations dans les sciences du génie de l'environnement. Enfin, dans la perspective ISTEX, nous avons élaboré le projet LorExplor pour analyser en profondeur les besoins des chercheurs face aux corpus numériques. Nous avons été labélisés par ISTEX pour développer une bibliothèque de composants XML. Nous bénéficions ainsi d'un hébergement informatique à l'Inist.

Pour présenter nos travaux, nous utilisons une métaphore : Wicri devient une bibliothèque où murs et étagères sont remplacés par une encyclopédie. On y introduit des documents numériques, comme on dépose des livres sur les rayonnages. Mais ici, les textes ne sont pas seulement juxtaposés mais ils sont en interrelation par des liens sémantiques. Avec ISTEX, la bibliothèque devient, ou plutôt redevient[31], un espace de travail où chercheurs et praticiens peuvent mener ensemble, dans le même espace, des activités autrefois dispersées. Ils peuvent aussi mutualiser des documents, des ontologies ou des données, pour constituer une vaste bibliothèque numérique.

4.1. *Un réseau de wikis pour les communautés de la recherche et de l'innovation*

Notre architecture repose donc sur un réseau de wikis opérés par MediaWiki, le moteur de Wikipédia. Nous offrons ainsi la même interface et nous avons la garantie d'une bonne volumétrie. En revanche, les usages sont fortement différenciés. Par exemple, pour faire de la recherche, les chercheurs doivent exposer des connaissances nouvelles, et donc non *sourcées*. De même, toute équipe de recherche, tout chercheur simplement confirmé, doit aussi bénéficier d'une page – ce qui est contraire aux critères de notoriété de Wikipédia. Bien entendu, les contributions sont transparentes et l'anonymat est proscrit. Dans l'avenir, une ligne éditoriale (et une modération) pilotée par des comités scientifiques s'impose naturellement. Sur un plan éditorial, Wikipédia regroupe toute la connaissance dans une seule encyclopédie ; de son côté, Wicri offre une collection de wikis thématiques ou régionaux.

– une quinzaine de wikis thématiques, allant d'un large domaine comme la santé, à un thème plus spécialisé comme les sols urbains, en passant par la musicologie ;

– une vingtaine de wikis régionaux allant du niveau continent, comme l'Afrique, à l'agglomération, comme la métropole du Grand Nancy.

Inspiré par la philosophie du CDSH, nous voulions favoriser l'appropriation d'un site par une communauté. Par exemple, le wiki Wicri/Lorraine[32] serait modéré par un comité scientifique parrainé par l'Université de Lorraine et pourrait bénéficier de soutiens de la Région Grand-Est. En fait, l'expérience a montré que cette organisation jouait aussi un rôle fondamental au niveau terminologique et rédactionnel[33]. La figure ci-dessous donne la « carte des wikis communs ». La ligne du haut contient des wikis techniques, comme le réservoir d'image de l'ensemble du réseau.

[31] Un film comme *Le nom de la rose* montre qu'avant l'imprimerie, la bibliothèque était bien un lieu de lecture, d'écriture et de travail.

[32] La notation Wicri/Lorraine désigne un wiki commun du réseau Wicri. Dans le cas d'une communauté comme par exemple H^2PTM, nous utiliserons la formule : le « wiki H^2PTM ».

[33] Une notion, comme « la danse » ne sera pas décrite de la même façon sur Wicri/Musique, Wicri/Santé ou Wicri/Psychologie.

Figure 1. *Les familles de wikis communs sur le réseau Wicri en 2019*

Le réseau contient également des wikis gérés par des communautés qui y définissent leurs propres règles. On trouvera également des wikis très spécialisés pour gérer des inventaires de données de la recherche. Par exemple, le wiki « bobines de l'Est » repère des scènes de films ayant un rapport avec la Lorraine. Avec les composantes multilingues, la taille du réseau voisine les 150 wikis.

Par rapport à l'approche « monolithique » de Wikipédia, nous avons travaillé sur les spécificités du passage en réseau. Pour l'utilisateur, le premier problème rencontré a été le repérage dans un réseau de wikis. Il a fallu deux ans de tâtonnements pour arriver au système de navigation iconographique actuel.

Pour maintenir la cohérence terminologique, nous avons défini la notion de « wiki de référence ». Par exemple, l'Université McGill a naturellement Wicri/Canada pour wiki de référence. Lorsqu'une activité significative de cette université est détectée sur un wiki comme Wicri/Musique, une page spécialisée est alors rédigée. Sur celle-ci, un lien interwiki est créé vers la page de référence (sur Wicri/Canada). Enfin sur ce dernier, un lien est établi vers Wicri/Musique. Ces opérations sont en fait très rapides pour des entités déjà signalées[34]. Bien entendu, la création d'un nouveau wiki demande une adaptation du réseau. Par exemple, avant la création de Wiki/Canada, les entités canadiennes étaient sur Wicri/Amérique. Il a donc fallu passer quelques heures pour mettre à jour le réseau de liens. Une telle opération peut assez facilement être automatisée par un robot. Le maintien de la cohérence du signalement des universités françaises en mutation permanente s'avère nettement plus complexe et montre la nécessité d'une administration terminologique, et surtout éditoriale.

Les mécanismes d'indexation de MediaWiki (les catégories) offrent un dispositif puissant pour maintenir la cohérence. Nous utilisons ainsi le thésaurus EuroVoc de la Communauté européenne comme souche commune pour le réseau. Des ressources plus ciblées sont utilisables sur quelques wikis, comme le MeSH sur Wicri/Santé.

Le réseau actuel est « calibré » pour des formations et des expérimentations sur le réseau des coopérations de la Lorraine (et également pour une première étape d'un déploiement ISTEX). Les mécanismes de navigation et de maintien de la cohérence sont compatibles pour une exploitation ayant la surface qu'avait le CDSH (une cinquantaine de wikis communs, avec la présence d'administrateurs). Pour aller au-delà, la maîtrise des robots devient indispensable. Cette technologie fonctionne parfaitement sur Wikipédia et ne constitue pas un verrou technologique. Elle devrait notamment permettre une exploitation sur plusieurs sites physiques. Nous avons vécu deux expériences de portage du réseau Wicri impliquant une phase multi sites. Elles ont montré une difficulté liée au partage des fichiers multimédia[35]. Il y a 8 ans, une première expérience (sur une quinzaine de wikis) avait été franchement laborieuse. En 2017, le portage de 150 wikis nous a

[34] Par exemple pour signaler l'Université McGill sur Wicri/Mathématiques à partir d'un copier-coller de la page sur Wicri/Musique.

[35] Le « wiki référentiel d'images » doit être physiquement sur le même site que les wikis clients.

amené à maitriser ce problème. Avec un support logistique relativement léger, on peut envisager quelques centaines de familles de wikis communs[36] distribués sur une dizaine de sites physiques. Cela dit, le passage à un niveau mondial avec des milliers de wikis constitue un défi technologique particulièrement intéressant…

Nous allons maintenant montrer que l'intérêt cette technologie dépasse très largement le niveau de la simple encyclopédie pour devenir une structure d'accueil pour exploiter des collections de données et documents de la recherche.

4.2. *Wicri, une bibliothèque pour les publications scientifiques*

Dès le démarrage, pour repérer les points forts de l'innovation en Lorraine, nous avons « wikifié » des documents structurants comme des chapitres du CPER. Nous avons ainsi édité des contenus encyclopédiques en nous appuyant, via les documents de référence, sur les priorités négociées entre l'Université et ses partenaires.

L'arrêt de la revue AMETIST a joué un rôle de détonateur quand nous avons décidé de la rééditer[37]. Nous avons récupéré les articles à partir de HAL, dans des archives personnelles et… grâce aux exemplaires papier. Nous avons pu comparer les interfaces de saisie de MediaWiki et de Lodel. Passé le barrage du *wikitexte,* les difficultés techniques[38] semblent plus faciles à résoudre sur MediaWiki. L'apport de la base encyclopédique a été immédiat par son « glossaire partagé ». Les pages de discussions permettent, le cas échéant, de remettre des informations en perspective. Et enfin, les liens mettent les articles en relation avec d'autres documents ; par exemple, les auteurs se retrouvent dans les comités scientifiques (et réciproquement).

Par la suite, nous avons ouvert des wikis pour des colloques scientifiques, comme CIDE, VSST ou H^2PTM. Le réseau Wicri offre ainsi des espaces où une communauté peut « hypertextualiser » sa production scientifique. Il permet de valoriser les meilleurs articles (ou les plus structurants) par des recopies sur des wikis communs (avec quelques adaptations au niveau des liens). Ainsi, le wiki Wicri/Musique contient des articles sur « le document numérique en musique » issus de CIDE ou de H^2PTM.

Dans cette dynamique, nous avons repris des expérimentations que nous avions menées sur ARTIST avec la rédaction en public[39] d'articles à soumettre pour des congrès internationaux (Ducloy 2006). Un article sur la gestion des métadonnées dans notre réseau a été ainsi publié pour la conférence DC 2010 (Ducloy 2010). Très récemment, pour le colloque CIDE 2019, nous avons franchi une nouvelle étape avec un exercice éditorial portant sur la rédaction d'un article avec la construction simultanée de son environnement encyclopédique. Par exemple, nous voulions qu'il soit lisible par des musicologues. Dans ce but, une expression technique dans l'article comme « fichier texte » donne lieu à la rédaction d'une page de vulgarisation qui montre comment la musique peut aussi être codée dans un tel fichier.

Sur un wiki spécialisé, des chercheurs de l'INRA ont lancé une revue en mode encyclopédique : *Les Mots de l'Agronomie*[40]. Pour la suite, on peut imaginer des mécanismes de publication dans une

[36] Par exemple, un wiki par communauté d'universités françaises, un wiki par pays et une cinquantaine de thématiques.

[37] Sur un wiki dédié aux applications des TIC pour la recherche (Wicri/Ticri)

[38] Par exemple, des formules avec des caractères ensemblistes.

[39] La rédaction en public modifie en profondeur le mode éditorial. Ainsi l'équipe des auteurs peut changer en cours de rédaction, et la version finale est fort différente de l'intention initiale.

[40] < https://lorexplor.istex.fr/Wicri/Europe/France/InraMotsAgro/fr >

bibliothèque encyclopédique, avec une évaluation par les pairs, mais avec des règles un peu modifiées (pas d'anonymat) ou avec quelques développements[41]. En fait, il faut repenser la publication scientifique dans un paradigme réellement différent du support papier.

Par rapport à notre métaphore, nous disposons bien d'une bibliothèque encyclopédique où le chercheur peut écrire et déposer des résultats de recherche. Il peut également manipuler des données de la recherche comme des livres anciens.

4.3. *Un espace pour les rééditions diplomatiques*

Une action régionale autour de la Renaissance en Lorraine nous a permis de rééditer un ouvrage de 1850 sur le Palais Ducal de Nancy[42]. À partir d'un original en mode « image + OCR » sur Gallica, nous avons montré comment transformer en hypertexte une gravure de fin de volume (fig. 2). Elle contenait des liens, matérialisés par des lettres, qui pointaient vers une vingtaine de paragraphes qui renvoyaient vers des pages du livre. Cet exemple a inspiré d'autres actions comme le traitement d'un ouvrage, annotée par le philologue Paul Meyer, sur la chanson de Roland[43].

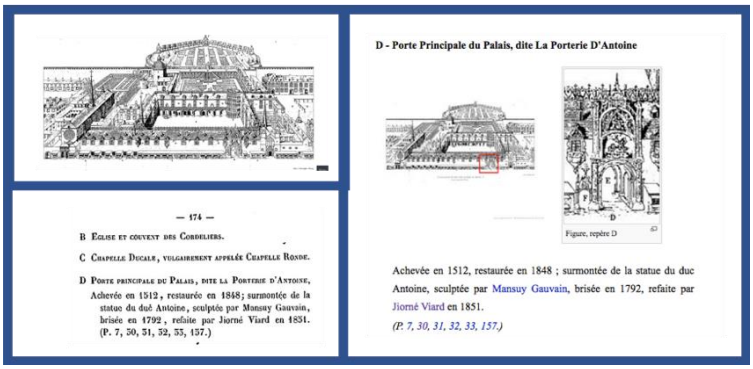

Figure 2. *Le Palais ducal : à gauche la gravure et une rubrique (D) avec des renvois ; à droite, le développé de la rubrique D en hypertexte*

En musicologie, nous travaillons à rééditer un ensemble d'ouvrages biographiques sur un même compositeur (actuellement Roland de Lassus[44]). Des livres de diverses périodes sont ainsi croisés avec des entrées de dictionnaires de musiciens ou d'artistes. Dans les annexes, les bibliographies et les listes de compositions peuvent être comparées (et liées) avec des inventaires contemporains déjà numérisés[45]. Parmi les documents qui structurent cet espace éditorial, nous avons fait des expériences très intéressantes en rééditant des entrées du Trésor de la Langue Française. En particulier, les termes musicaux sont très riches en exemples et en syntagmes qui fournissent autant de pistes d'investigations. Réciproquement, l'analyse d'anciens écrits de musicologie fait émerger de nouveaux exemples[46].

---

[41] Par exemple avec un couple de deux wikis en réplication (public et privé). Le manuscrit peut alors être rédigé dans un espace privé qui clone le wiki public.

[42] < https://lorexplor.istex.fr/Wicri/Europe/France/Lorraine/fr/index.php/Le_Palais_ducal_de_Nancy_(1852)_Lepage >

[43] On y montre notamment le texte avec ses annotations numérisées en mode wiki avec le résultat : <https://lorexplor.istex.fr/Wicri/France/Lorraine/CollectionsBul/fr/index.php/FPM,_Chanson_de_Roland_(1869)_F._Michel,_page_72>

[44] Roland de Lassus (1532-1594), compositeur de l'école franco-flamande de la Renaissance.

[45] Une amélioration considérable porte sur la lisibilité des listes (bibliographies, répertoires) qui sont aérées en numérique ; là où les contraintes de coût les rendent illisibles sur l'original.

[46] Par exemple nous avons enrichi l'entrée *chœur* dans le syntagme *enfant de chœur* par un exemple montrant que cette expression désignait les jeunes chantres sous la Renaissance.

Le substrat encyclopédique de « notre bibliothèque » permet ainsi de relier des documents anciens avec des publications actuelles qui les exploitent. Ceci constitue une différence importante avec la plupart des services numériques de la recherche qui sont organisés par types de publication. Par exemple, les documents anciens sont sur Persée, les revues sont éditées sur OpenEdition, avec un modèle éditorial classique, et l'archivage se fait sur Hal. De même, la fondation WikiMedia offre une gigantesque encyclopédie, un gigantesque dictionnaire, et une gigantesque bibliothèque de textes numérisés. Wicri propose un réseau de sites par domaine scientifique où sont regroupées et mises en relation les données et productions éditoriales de la recherche.

4.4. *Une boîte à outils éditoriale avec les modèles MediaWiki*

Notre approche amène donc à traiter des types de documents diversifiés dans un spectre de thématiques variées. De nombreux CMS obligent à figer *a priori* un modèle de document. Ici, MediaWiki offre la possibilité de travailler en mode incrémental. En effet, le socle SQL gère uniquement les relations entre les contributeurs et les pages du wiki. Tout ce qui relève du contenu et de la présentation est assurée par un ensemble de catégories et de modèles, modifiable à tout moment par les contributeurs.

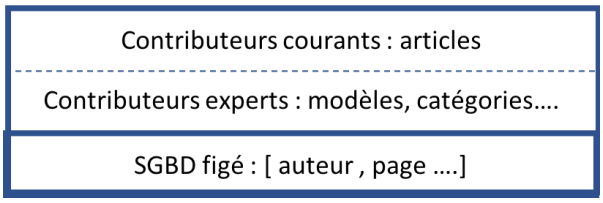

Figure 3. *Architecture interne MediaWiki*

Wikipédia, avec les Infobox ou les applications géographiques, montre tous les jours la puissance des modèles[47] et la possibilité de leur appropriation par le plus grand nombre. De même, sur Wicri, les contributeurs disposent d'un environnement extrêmement riche pour personnaliser leurs applications. À la création d'un wiki, l'utilisateur dispose déjà d'un millier de modèles[48]. Chaque thématique peut développer ses propres outils, comme les tables de Mendeleïev en chimie. Bien entendu, une utilisation optimale suppose une formation conséquente des utilisateurs.

Les modèles jouent aussi un rôle fondamental pour la cohésion terminologique. Par exemple, les palettes géographiques de Wikipédia, (comme *les régions françaises*) facilitent le travail des contributeurs et les amènent naturellement à respecter les règles terminologiques. Ils vont ainsi utiliser une terminologie compatible avec celle de Wikipédia, et donc du *Web Sémantique*. Sur Wicri, un wiki est ainsi dédié à la gestion des modèles qui portent la terminologie commune. Ce mode de fonctionnement est totalement compatible avec une distribution sur un réseau physique porté par plusieurs communautés géographiquement dispersées.

Cette architecture est caractéristique du changement de paradigme des systèmes d'information. Dans une application traditionnelle, le modèle des données est décrit au niveau de la couche SGBD. Ici, tout ce qui relève de l'application est géré par les modèles. Des programmes (php) construisent les relations dans la couche SGBD, à partir d'éléments de syntaxe dans les pages. Les supports méthodologiques sont alors bouleversés. Les extensions sémantiques permettent d'aller encore plus loin !

[47] D'un point de vue informatique, les modèles sont des sous-programmes ouverts (à la façon des macros en Word) qui peuvent appeler des modules en langage Lua également modifiables par les contributeurs.

[48] Une grande partie, notamment en cartographie, provient de Wikipédia (avec adaptations).

4.5. *Les extensions sémantiques pour fédérer données et publications*

L'extension *Semantic MediaWiki*, réalisée par l'Université de Karlsruhe, permet de créer des liens sémantiques (dans une philosophie RDF[49]) entre les pages.

La figure ci-dessous en donne un exemple. Dans la page sur les *Noces de Figaro*, Lorenzo da Ponte est cité comme auteur de livret. Un lien sémantique, ayant pour attribut « *a pour auteur de livret* », est alors créé entre l'opéra et la page du librettiste. Dans la page « Lorenzo da Ponte », une liste de ses prestations est générée automatiquement par une requête sémantique – ce qui permet de découvrir que Mozart et Salieri avaient le même librettiste…

Figure 3. *extrait d'une diapositive montrant la génération d'une liste d'opéras*

Nous avons mentionné une première utilisation avec les équipements financés par le CPER. Chaque équipement et chaque bénéficiaire disposait d'une page dans laquelle toutes les données nécessaires ont été ainsi codifiées. De multiples listes et tableaux ont pu être générés. Cette approche a été reprise pour des éléments d'observatoire de la recherche comme un inventaire des projets européens en Lorraine. Elle a été testée sur les données de la recherche, par exemple sur le réseau hydrologique de la Moselle. On retrouve un intérêt majeur de ce dispositif : la possibilité d'une définition incrémentale du modèle de données, avec en plus, la cohabitation possible entre les données et les parties textuelles[50].

Pour les publications, le croisement des bibliographies d'article avec les membres des comités de programme donne des éléments très intéressants pour modéliser les réseaux de grands acteurs d'une communauté. Nous avons pu aussi vérifier qu'il était tout à fait possible de créer des notices Pascal ou Francis[51]. Par rapport à l'approche chaîne de production, les avantages sont considérables. La prise en compte de contraintes spécifiques aux coopérations ne pose plus de problème. De même, il est possible de gérer et modifier en permanence le vocabulaire d'indexation[52], avec une grande souplesse pour créer des relations. Là encore, ce dispositif est d'autant plus performant qu'il est construit par les chercheurs eux-mêmes – ce qui implique une formation des acteurs, pas forcément bien perçue par les décideurs.

---

[49] Resource Description Framework

[50] Quelques limites méritent d'être signalées. MediaWiki gère mal les contraintes de visibilité partielle comme des données expérimentales partagées entre des industriels (donc confidentielles) et des laboratoires. De même, pour les contraintes transactionnelles fortes.

[51] Avec cependant quelques adaptations car le modèle sous-jacent est RDF et non RDFa. Autrement dit, un attribut est affecté à une page et pas à un de ses élément. Il n'est pas possible par exemple de formaliser les relations entre les auteurs, les revues et les affiliations au sein d'une bibliographie dans un article sur une page. Pour cela, il faudrait éclater la bibliographie sur plusieurs pages – ce qui n'est plus compatible avec la lisibilité de l'article.

[52] Le fait d'admettre de légères incohérences temporaires évite une situation où à l'Inist, le vocabulaire d'indexation pouvait se trouver figé, avec ses erreurs, pendant plusieurs années.

4.6. *Explorer des corpus avec une bibliothèque XML et des wikis sémantiques*

Nous avons été soutenus par ISTEX pour étudier des pratiques de TDM en liaison avec une activité éditoriale. Pour cela, nous générons des « serveurs d'explorations » à haut niveau de paramétrage grâce à Dilib, une boîte à outils XML, dont les premiers développements remontent à 1993.

Le socle de Dilib est un *parser* XML de type DOM[53]. Il est complété par des fonctions qui permettent de réaliser des systèmes de recherche d'information. Voici un exemple introductif. Nous avons défini une organisation de documents basée sur l'arborescence Unix. Par exemple, une notice de métadonnées ayant comme clé interne « 012345 » est rangée dans le fichier d'adresse : *MaBiblio.hfd/01.fd/23.dd* à la 46ème place. Sur cette base nous représentons des listes inverses dans un formalisme XML. Ainsi, celle de Mozart devient :

<*index*><*key*>Mozart**</*key*><*list*><*li*>**012345**
…</list></index>**

Sur ce type de spécifications, nous avons développé une bibliothèque de composants. Dans les années 1995, nous pouvions construire un moteur de recherche, type Mistral, qui était plus performant pour les types de données traitées, avec la prise en compte des métadonnées en MARC codées en XML. Avec ISTEX, nous avons amélioré les performances pour traiter des flots de documents volumineux, en texte intégral et issus de diverses sources[54]. Nous avions également développé des outils infométriques (clustérisassions) sous la forme de fonctions qui manipulent ces listes inverses en XML. Pour ISTEX, avec cet ensemble d'outils de base, nous avons développé un générateur de plateformes de curation et d'exploration.

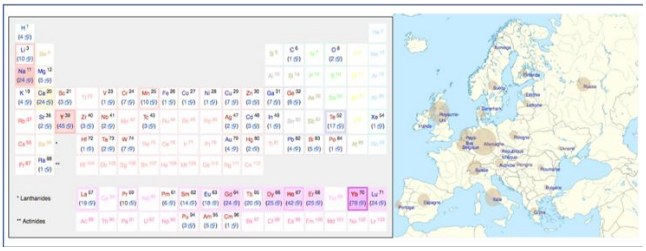

Figure 4. *visualisations sur un wiki*

Les interfaces d'exploration ont été améliorées. Nous avons d'abord implanté des composants infométriques (ou de navigation) sous la forme de fonctions sur la classe DOMDocument de php 5[55]. Puis nous avons fait un saut déterminant en utilisant les wikis. En effet, à partir des documents XML et des listes inverses, des fonctions Dilib génèrent des modèles (comme, par exemple, la liste des principaux auteurs, ou une carte de projection) qu'un rédacteur peut insérer dans les pages. L'utilisateur peut alors démarrer une exploration dans le wiki, puis aller dans le serveur. La figure 4 montre, à droite, une projection d'activités sur une carte d'Europe et, à gauche, une projection de composés chimiques sur une table de Mendeleïev. Toutes les données ainsi générées peuvent être liées aux pages du wiki.

Pour des raisons liées au moyens humains, nous avons mis la priorité sur la cohérence globale de l'architecture, et pas sur l'excellence des algorithmes. Nous avons donc repris notre premier

[53] *Document Object Model*. Cette spécification du W3C décrit comment représenter un document XML par une arborescence interne à la mémoire.

[54] ISTEX avec ses variantes internes, Pascal, Francis, PubMed, PubMed Central, HAL. Le parser traite donc des flots de documents ayant des DTD multiples.

[55] La même base de documents XML est alors utilisée en c/shell et en PHP (ou en Python).

ensemble de fonctions dans une architecture de type Mistral avec des méthodes de classification de type simple lien. Il serait bien entendu pertinent d'aller plus loin pour synthétiser par exemple un moteur de type Elastic Search, ou spécialisé sur des formules chimiques ou des partitions musicales[56].

Le couplage avec les wikis nous a amené à réaliser des travaux consistants autour de la curation des données. En effet, nous devons homogénéiser la navigation dans les serveurs d'exploration avec la terminologie utilisée dans les wikis (avec les contraintes de conformité avec Wikipédia et donc avec celles du *Web sémantique*). Par exemple la ville où Mozart est décédé doit être manipulée avec la graphie « Vienne (Autriche) ». Nous avons donc développé des outils qui s'appuient sur des informations gérées et validées dans les wikis. Ainsi la table associant les noms de pays et leurs codes ISO, recopiée depuis Wikipédia, est utilisée pour traduire les codes de pays venant de Pascal (2 lettres) ou de HAL (3 lettres). De même, toujours avec des pages issues de Wikipédia, les codes postaux sont utilisés pour repérer les régions – ce qui implique des algorithmes différents suivant les pays, avec des cas intéressants, par exemple l'ex Yougoslavie...

Le curateur peut également définir des règles, éventuellement contextuelles, qu'il exprime, avec un formalisme proche d'XML/TEI[57], dans des tables du wiki. Par exemple, en musicologie, la zone *auteur reconnu* « David Stevens » est associée à la graphie « David Stevens (musicologue) » et à une affiliation implicite :

affiliation @from=1964 @to=1976 : Université Columbia

Dans une autre table, sur Wicri/Amérique, l'entrée « Université Columbia » va enrichir l'affiliation de David Stevens avec ce type de structure :

country : États-Unis ; region @type=state : État de New York ; settlement @type=city : New York

Les noms d'entité sont des liens (et ici donc en bleu) – ce qui garantit une bonne correspondance terminologique. Nous avons environ une centaine de pages qui contiennent ce type d'informations. Elles sont de nature très variées (noms de pays, de région, de villes par pays, de codes postaux, etc.) et en évolution permanente. Nous avions naturellement rencontré ce type de problèmes au service de veille de l'Inist pour des prestations dans un contexte théoriquement plus simple (métadonnées Pascal). Mais nous n'avions pas réussi à mettre en place une interface collaborative de saisie satisfaisante, ni avec des fichiers, ni avec des programmes, ni avec des applications SQL. Le réseau de wikis permet de gérer plus facilement cette dispersion évolutive, avec la possibilité de mettre des commentaires, et surtout grâce à des mécanismes immédiats de validation par les liens wiki vers les pages de référence.

La génération d'un serveur d'exploration devient un processus itératif où le veilleur analyse des listes d'éléments non reconnus, introduit des règles et relance la génération. Toutes les pages éditoriales qui contiennent des appels de modèle sont alors automatiquement mises à jour avec des données actualisées. Nous avons pu vérifier avec des étudiants, ou des stagiaires, que ces techniques de spécifications étaient assez faciles à assimiler (avec une supervision indispensable par des experts).

4.7. *Une infrastructure pour la formation et l'expérimentation*

Concernant ISTEX, nous voulions voir s'il était possible de faire du TDM avec des contraintes de délais (par exemple quelques jours dans le cadre d'un appel à propositions). Nous avons donc mené

[56] Par exemple, en notation abc ou avec le schéma MusicXml.

[57] Plus précisément les tables contiennent des éléments avec une syntaxe moins verbeuse que celle d'XML, mais qui seront converties par Dilib en éléments conformes.

environ 200 explorations de corpus dont le volume variait de quelques centaines à 35.000 documents (4.000 en moyenne). Une partie a été réalisée avec des étudiants de Master de l'Université de Lorraine ou de Paris 8. Les étudiants, par petits groupes, ont choisi des sujets avec une grande liberté dans les choix des thématiques[58]. Pour chaque groupe, nous avons généré, en quelques heures par corpus, un serveur dont le coût moyen aurait été de 120.000€[59] par groupe avant ISTEX ! Après, une série de TP (de type curation) les étudiants ont présenté oralement leurs observations avec une approche « rapport d'étonnement ». Nous avons fait également de nombreux essais « opportunistes » en profitant par exemple d'une conférence ou à partir de demandes d'utilisateurs. Ces multiples expériences nous ont alerté sur l'immense variété des situations rencontrées et sur l'importance des actions de curation en préalable aux analyses statistiques.

En musique nous avons rencontré des « *avenues Mozart* », des « *Mozart de l'informatique* », ou des projets dont l'acronyme est *Mozart*. Toujours sur Mozart, les articles médicaux le concernant sont rédigés par des équipes bien rodées sur la déclaration des affiliations. Mais les musicologues réputés se contentent de donner leur nom voire même des initiales... Les résultats statistiques sur la musicologie sont alors insignifiants ou masqués par des problématiques de santé. Pour améliorer les résultats nous avons par exemple introduit des règles contextualisées par des ISSN.

0027-4631:P. H. L. ***désigne*** Paul Henry Lang ***avec*** affiliation : Université Columbia

Dans la recherche de coopérations, par exemple entre la Lorraine et l'Australie, nous avons rencontré des dizaines d'articles avec plus de 1000 auteurs et affiliations. Ils faussaient toute analyse statistique des coopérations. Nous avons développé un petit outil pour les détecter de façon à générer des règles de curation par élimination.

Pour étudier les travaux sur le Cobalt au Maghreb, nous avons développé un petit outil de pondération conjoncturelle avec des termes géographiques de l'Atlas pour ne pas être débordés par les coopérations franco-marocaine sur le Cobalt. Ce type d'outil a été utilisé pour extraire une petite quinzaine de pages concernant réellement la région d'Aussois parmi 1500 articles issus de colloques ayant eu lieu à Aussois.

Parmi les records, un corpus de 10.000 articles sur la méthode *Scrum* était pollué à plus de 80% par des problèmes d'OCR. En effet, la simple apparition du terme *sérum* en bibliographie d'un article en anglais générait une occurrence de « *scrum* ».

Il a donc fallu introduire des mécanismes de curation, pour éliminer des revues ou des articles, dans un corpus. Ils sont basés sur des tables implantées dans les wikis, afin, comme précédemment, de permettre un travail itératif sur un corpus donné.

Nous avons donc mené des explorations sur environ 200 corpus, dont la volumétrie allait de quelques centaines à quelques dizaines de milliers de documents. Nous avons rencontré des problèmes de curation dans pratiquement tous les cas. La majorité des résultats résultant d'une génération initiale était inexploitables. En revanche, nous avons pu obtenir des informations très intéressantes après quelques étapes de curation, mais, presque toujours, avec de grandes incertitudes sur la fiabilité numérique des faits observés. En fait les données statistiques du serveur et les algorithmes de classification se comportent comme un moteur de serendipité qui émet des hypothèses qu'il faut vérifier ensuite…

[58] Le générateur étant un prototype non encore stabilisé, nous avons généré les plateformes de curation et d'exploration à partir des « consignes » des étudiants.

[59] Cette somme est évaluée à partir d'un prix moyen de 30€ par document.

Les résultats les plus intéressants ont souvent été obtenus par des techniques de filtrages sur des corpus fiabilisés, mais sous Unix, en programmant des procédures souvent très spécifiques. Par exemple, pour identifier les œuvres de Mozart les plus citées, nous avons procédé à des extractions sur des expressions régulières liées au catalogue Köchel. Nous avons ainsi identifié la sonate pour deux pianos (citée sous KV 448 ou K.448 ou Kv.448 etc.) qui a fait l'objet de nombreux articles japonais en raison… de ses vertus curatives pour l'épilepsie ! Cet exemple montre l'importance de l'appropriation des outils par le musicologue. Un informaticien qui ne connaitrait rien en musique ne peut imaginer une telle heuristique. Réciproquement, un chercheur qui ne connait pas la notion d'expression régulière ne peut envisager un tel procédé.

Nous avons été interpelés par le niveau très important de variabilité des situations. Pratiquement chaque corpus pose des problèmes de curation spécifiques, et chaque procédure de fouille est également spécifique. Autrement dit, les outils de TDM doivent prioritairement être appliqués sur des corpus d'application et non sur le serveur central. Les outils doivent aussi être suffisamment robustes et paramétrables (boîtes à outils) pour être manipulés par des utilisateurs nécessairement formés.

4.8. *Multilinguisme et expérimentations internationales*

Enfin, MediaWiki offre des dispositifs pour gérer le multilinguisme. Pour chaque wiki commun, un wiki est créé en anglais, en vue d'être réactif pour toute démonstration internationale. Voici deux exemples où nous sommes allés plus loin.

Dans une coopération avec l'IHEST, nous avons monté un wiki bilingue (français portugais) pour valoriser des relations entre la France et le Brésil. En parallèle, nous avons ouvert 3 wikis communs sur le Brésil, en français en anglais et en portugais.

Une autre expérience a été menée sur la Grande Région, un espace de coopération européenne entre la Lorraine, le Luxembourg, la Région Wallonne, la Sarre et la Rhénanie-Palatinat. Lors d'une réunion sur l'Université de la Grande-Région, nous avons été sollicités pour créer un wiki montrant les complémentarités des systèmes académiques des participants. Nous l'avons donc créé. Mais les premières expérimentations nous ont montré qu'il était nécessaire de créer un wiki par système et par région, afin de pouvoir décrire séparément chaque système en profondeur, pour être capable, ensuite, de travailler sur les complémentarités. Pour cette expérimentation nous avons ouvert un ensemble de wikis en allemand. Nous avons commencé à élaborer des stratégies de coopération entre des étudiants de langues différentes[60]. Compte tenu de la migration des efforts sur ISTEX nous ne sommes pas allés très loin, mais la faisabilité du multilinguisme, et donc de larges coopérations internationales est démontrée.

4.9. *Un fil conducteur pour la suite : la musicologie*

Avec le soutien ISTEX nous avions, comme priorité, les services généraux pour les chercheurs dans leurs diversités des thématiques, (avec des contraintes de temps). Nous voulons maintenant nous immerger dans un domaine scientifique, pour aller plus loin dans l'analyse des pratiques numériques sur les données dans une action de recherche. L'orientation initiale donnée au réseau Wicri (CPER) était centrée sur les disciplines stratégiques des sciences de l'environnement. Mais nos essais ont montré qu'une expertise solide dans les fondements théoriques était indispensable pour interpréter les résultats d'une expérimentation et émettre de nouvelles hypothèses. Nous avions naturellement les compétences nécessaires dans les approches numériques des sciences de l'information. Mais ce domaine présente une spécificité paradoxale. En effet, il n'est pas nécessaire

[60] Par exemple des étudiants lorrains pourraient, sur Wicri/Sarre(fr) traduire des pages sur la Sarre rédigées en allemand par des étudiants sarrois sur Wicri/Saar(de) et réciproquement.

de lire un article pour comprendre la démarche algorithmique d'un collègue. Une discussion suffit ! Le fait de soumettre un article à un colloque est fondamental pour clarifier des concepts. La présentation publique est importante. Mais, la lecture des travaux n'est pas toujours indispensable.

Nous avons donc cherché un domaine, impliquant des traitements de corpus, dans lequel nous n'avions pas forcément de fortes compétences, mais une forte motivation pour acquérir expertise et érudition. Nous avons fait des premières investigations sur la musique de la Renaissance. Elles mettent en évidence des besoins d'exploration de corpus qui dépassent nos espérances. La manipulation des données musicales, leur codification et leur exploitation s'avère être un champ d'investigation multidisciplinaire très intéressant. L'association entre une codification XML des partitions (MusicXML) et MediaWiki, sur lequel une telle partition peut être interprétée, est prometteuse.

Nous avons cité les publications scientifiques, les rééditions diplomatiques et les explorations de corpus sur la même base encyclopédique. Nous expérimentons aussi les articles de vulgarisation pour publics motivés[61]. Dans un article PDF, l'auteur peut « faire des impasses ». Sur Wicri/Musique, il doit aussi travailler sur l'ensemble hypertexte qui accueille l'article. Cette incitation à une forme de perfection soulève parfois des questions plus fondamentales. Par exemple, nous avons voulu vulgariser les mécanismes de transcriptions de partition originales de musique renaissance. Dans pratiquement tous les cas, nous avons rencontré des problèmes (erreur d'imprimeur sur une altération par exemple) qui ont amené à pousser plus loin les investigations (et par exemple explorer des corpus ISTEX). Tout se fait sur une même infrastructure qui évolue au fur et à mesure (à condition que le chercheur puisse la maîtriser).

5. Bilan et perspectives

Nous venons de présenter les résultats d'un démonstrateur réalisé par une équipe de permanents réduite à un ingénieur retraité (qui ne peut intervenir que depuis son domicile compte tenu des dispositions administratives du CNRS !). Que pourrait faire une équipe « de taille normale » dans des « conditions normales » pour mettre en synergie des activités éditoriales collaborative et un déploiement de pratiques d'exploration de corpus répondant aux missions du projet ISTEX ?

5.1. *Vers un réseau potentiellement vaste de sites encyclopédiques*

D'un point de vue purement technique, pour toutes les fonctions à caractère éditorial (ou de construction de connaissance), la mise en œuvre d'un réseau à l'échelle nationale, incluant un ensemble limité de coopérations internationales ne pose pas de problème particulier. De même, au niveau des équipements, la solution repose uniquement sur des logiciels ouverts, implantés sur les réseaux universitaires. Enfin, l'expérience de Wikipédia montre qu'un parcours de formation progressif est possible. Des acteurs comme les bibliothèques universitaires sont des potentiellement relais locaux dans une philosophie de type *learning center*.

Le défi à relever est plutôt institutionnel pour inciter les chercheurs à contribuer, les convaincre de s'approprier des technologies « assez simples », créer des comités scientifiques, lancer des appels à communication sur des publications hypertextuelles etc. La dynamique d'un déploiement ISTEX peut-elle contribuer à relever ce défi ?

5.2. *Rôle potentiellement moteur d'un déploiement ISTEX*

Le couplage des wikis avec des architectures d'exploration et de curation modifiables utilisables par des chercheurs de toutes disciplines, demande une bibliothèque XML très robuste. Cette « TDM

[61] Comme par exemple les choristes qui veulent comprendre le contexte des pièces interprétées.

de proximité » demande également un travail consistant pour améliorer des outils de laboratoire pour les rendre utilisables par une large communauté. Au lancement du projet LorExplor nous avions envisagé une équipe de 5 personnes en soutien logistique transversal. Cette hypothèse est plutôt confirmée par nos travaux.

Comme nous l'avons évoqué précédemment, nous avons privilégié de vérifier la complétude des pratiques, et pas forcément la recherche d'excellence dans les solutions. Par exemple, du côté des moteurs de recherche, nous nous sommes limités à une synthèse de moteurs plutôt traditionnels. De même, du côté des traitements linguistiques, nous nous sommes limités à des filtrages et à des heuristiques. Enfin, dans les mécanismes de curation, nous avons surtout travaillé sur des mécanismes utilisant des données et règles internes au réseau de wiki et assez peu sur l'utilisation d'ontologies externes. Ce sont donc les compétences qu'il faudrait réunir pour une équipe chargée de constituer une infrastructure de haute performance à partir de solutions disponibles dans la recherche. Le point qui nous paraît le plus sensible est une forte expertise en génie logiciel pour résoudre les multiples problèmes d'interopérabilité.

Nous avons aussi montré l'importance d'une forte acquisition de compétences au niveau des chercheurs eux-mêmes, pour atteindre « l'excellence documentaire pour tous ». Ceci implique un plan de déploiement conséquent, avec impérativement des relais régionaux. Nous avons montré l'intérêt d'un substrat encyclopédique multidisciplinaire comme un plan de travail pour des explorations de corpus. Nous pensons que la synergie potentielle entre une grande opération de transmission de savoir associée à un déploiement de techniques de TDM mérite d'être envisagée.

5.3. *Quelques besoins complémentaires de recherches multidisciplinaires*

Nous avons évoqué les difficultés, pour les décideurs, d'appréhender la complexité des grandes applications liées aux données de la recherche. Entre 1845 et 1906, les thermodynamiciens ont inventé les trois principes de la thermodynamique pour permettre aux ingénieurs de mieux appréhender la complexité des problèmes auxquels ils étaient confrontés. Les chercheurs en informatique théorique peuvent-ils élaborer de tels principes autour des données et documents de la recherche ?

L'usage des wikis impose aux organisations de faire exactement l'inverse de leurs pratiques antérieures : passer de la validation *a priori* à la modération *a posteriori*. Comment les sociologues peuvent approfondir cette rupture pour permettre aux organisations d'évoluer de façon apaisée vers ces nouvelles pratiques ?

Au niveau individuel, les chercheurs sont confrontés à la publication, même temporaire, de leurs erreurs dans des écrits. La relation à l'erreur (ou à la faute), ou même au livre ou à l'écrit est différente suivant les cultures religieuses ou philosophiques. Comment les psychologues, les cogniticiens et les anthropologues peuvent aider les chercheurs à partager leurs doutes dans un paysage numérique ? De même, la relation individuelle avec l'infini nous semble également essentielle à prendre en compte. En effet, elle est omniprésente dans les algorithmes récursifs. Elle l'est également dans l'appréhension du web qui donne accès à un hypertexte quasi infini avec un milliard de sites pouvant donner accès à des dizaines de millions de pages… Là encore, le contexte philosophique, voire théologique est important.

Enfin, nos expériences, dans toutes les disciplines font émerger des problèmes différents. La multidisciplinarité des thématiques est aussi un facteur fondamental.

6. Conclusion

Dans une première partie nous avons évoqué l'impasse dans laquelle nous nous sommes trouvés en 2002, à l'Inist, pour faire face à une réduction considérable des moyens affectés à cette unité.

Nous aurions « peut-être » pu réussir si nous avions pu nous appuyer sur des résultats obtenus par sa structure de Recherche et Développement, malheureusement dix ans auparavant. Mais nous n'avions pas de solution miracle pour sortir d'une chaîne de production qui interdisait des coopérations permettant à nos partenaires de répondre aussi à leurs propres besoins.

Avec les projets Wicri et LorExplor nous avons montré qu'une infrastructure basée sur des wikis sémantiques complétée par une ingénierie XML résout, dans un mode collaboratif, tous les verrous technologiques auxquels nous avons été confrontés. En cherchant à montrer la faisabilité de systèmes d'information encyclopédiques pilotés par les scientifiques, nous avons mis en avant la puissance de l'intention citoyenne du partage des savoirs. Elle amène à la recherche d'excellence dans la façon de communiquer les résultats de la science. Enfin comme Wikipédia donne du sens aux loisirs de milliers de contributeurs, le partage des savoirs peut donner un sens au travail des ingénieurs des données numériques de la recherche, et peut-être aussi, au-delà des facteurs d'impact, à celui des chercheurs.

Remerciements

Merci à celles et ceux avec qui nous sommes parti à la quête du Graal Inist, sans oublier l'ANL, Dilib, Wicri, LorExplor. Merci à Francis André, Thierry Daunois et Jean-Pierre Thomesse pour leurs conseils à propos de cet article.

Bibliographie

Buhr J., Degen. C. (1977) PASCAL: Une base de données multidisciplinaire son utilisation en physique atomique et moléculaire et physique des fluides et des plasmas. *Journal de Physique Colloques*, 1977, 38 (C3), pp.C3-249-C3-251. https://hal.archives-ouvertes.fr/jpa-00217115

Ducloy J., Charpentier P., François C., Grivel L. (1991) - "Une boîte à outils pour le traitement de l'information scientifique et technique", *Génie logiciel et systèmes experts*, nº 25, pp 80-90, Paris.

Ducloy, J., Nicolas Y., Le Hénaff D., Foulonneau M., Grivel L., Ducasse J.-P. (2006) Metadata towards an e-research cyberinfrastructure: the case of francophone PhD theses. Proceedings of *DC 2006*, Manzanillo, Mexico, 2006.

Ducloy J., Daunois T., Foulonneau, M., Hermann, Lamirel, J.-C., Sire, Thomesse, J.-P., Vanoirbeek C. (2010). Metadata for WICRI, a Network of Semantic Wikis for Communities in Research and Innovation, *DC 2010*, Pittsburgh.

Ducloy J., Daunois T., (2019) Pratiquer la musicologie dans une bibliothèque numérique encyclopédique, CIDE 2019, Djerba

Dusoulier, N., Ducloy, J. (1991): «Processing of data and exchange of records in a scientific and technical information center. Formats: what for?» UNIMARC/CCF Workshop - Florence (IT) (IFLA/UNESCO), 05-07 June 1991

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

Laborderie, A (2015). Éditorialisation des bibliothèques numériques : le cas des Essentiels de Gallica, In: *CIDE 2015*, Montpellier

Schmitt L., Olivetant B., Landi B., Royauté, J., Ducloy J. (1992) STID: Une station de travail pour une indexation assistée. Conference Internationale d'Avignon, 1992

