Analyser et écrire la science multidisciplinaire dans un réseau d'hypertextes sémantiques avec Wicri

Jacques Ducloy[1]

1. Laboratoire Paragraphe, Université Paris 8
2 rue de la Liberté - 93526 Saint-Denis cedex, France
Jacques.Ducloy@univ-lorraine.fr

RESUME. Le projet WICRI travaille sur une alternative à la galaxie Wikipédia pour les communautés scientifiques. Il prend également en compte des besoins d'analyse de corpus avec une boîte à outil XML pour la création de systèmes d'informations, notamment dans la perspective de développer des bases de données bibliographiques. Enfin, dans ce réseau culturel, les rééditions hypertextes d'ouvrages ancien et de manuscrits sont particulièrement démonstratives des applications avancées dans les humanités numériques (avec notamment une bibliothèque numérique sur la Chanson de Roland).

ABSTRACT. The WICRI project aims at a potential alternative to the Wikipedia galaxy for scientific communities. It also takes into account corpus analysis needs with an XML toolbox for the creation of information systems. Finally, in this cultural network, hypertext reissues of old works and manuscripts are particularly suited to advanced applications in digital humanities (including a digital library on the Song of Roland).

Mots-clés : Semantic Mediawiki, Ingénierie XML, Réseau de Wikis, Exploration de corpus, Humanités numériques, Chanson de Roland

KEYWORDS: Semantic Mediawiki, XML engineering, Wiki network, Corpus exploration, Digital humanities, Chanson de Roland

1. Introduction

Le projet présenté est issu de travaux de R&D menés à l'INIST en 1990. Il s'agissait de construire un système d'IST (Information Scientifique et Technique) compétitif au niveau international. Il concernait notamment les bases Pascal et Francis (500.000 analyses par an par environ 400 ingénieurs) avec des mécanismes d'indexation assistée, intégrant les spécificités de chaque secteur scientifique, et des mécanismes de coopérations prenant en compte l'ensemble des besoins informationnels de la recherche. Mais en 1992, le CNRS a engagé un virage à 180° en visant un groupe commercial. Le département de R&D a été dissous. Les bases Pascal et Francis ont été arrêtées en 2015. Dans la même période, rien que sur la santé, aux USA, la *National Library of Medicine* (NLM) a doublé sa capacité de production (1.000.000 d'analyses par an par 750 ingénieurs, et un réseau de 8.000 collaborateurs). De même, dans les années 60, le CNRS avait lancé le dictionnaire du Trésor de la Langue Française qui a quasiment disparu face à la domination de la Wikimedia Foundation (750 personnes à San Francisco, plus de 200 millions de dollars de chiffre d'affaires).

Nous avons donc décidé d'engager une réflexion allant de l'ingénierie aux pratiques avancées des chercheurs et des partenaires de la recherche. Nous disposons d'un démonstrateur qui apporte un début de preuve de concept. Il permet déjà de réaliser des premiers services opérationnels et de mener de multiples expérimentations, soit technologiques, sur les systèmes de réseau hypertexte, soit éditoriales, comme l'écriture hypertexte collaborative ou multidisciplinaire.

2. Histoire des projets Dilib et Wicri

En 1991, le premier résultat obtenu à l'INIST a été une boîte à outil SGML (en préfiguration d'XML). Reprise par le Loria sous l'appellation Dilib, elle permettait de réaliser des serveurs d'exploration de corpus bibliographiques hétérogènes. On y associait des mécanismes de classification à des fonctions plus classiques, de type moteur de recherche. Il était ainsi possible de réaliser des applications de taille modeste avec un grand niveau d'interdisciplinarité (comme par exemple une base iconographie et bibliographique sur l'art nouveau) ou des services à volumétrie conséquente (l'intégralité des bases Pascal et Francis).

Un deuxième axe a été initialisé par une réflexion autour de Wikipédia qui apportait des éléments de réponse à la gestion des flux de contributions rencontrés dans la production des bases. Elle nous a amené à travailler sur une alternative à cette encyclopédie pour la production d'informations produites par la recherche (et donc souvent nouvelles et non sourcées). Nous avons envisagé une collection d'encyclopédies thématiques qui pourraient être pilotées et modérées par des comités scientifiques. En 2008, un démonstrateur (Wicri) a donc été construit sous la forme d'un réseau de wikis dopés par des mécanismes d'annotation sémantique

(avec Semantic MediaWiki). Grâce un financement CPER le réseau a acquis une dimension multidisciplinaire, notamment dans les sciences liées à la santé et l'environnement.

Dans le cadre d'ISTEX, le projet LorExplor en 2013 a permis de rapprocher les deux approches. Plus précisément, des serveurs d'exploration ont été intégrés à la base Wiki sémantique. En amont, des mécanismes de curation sont basés sur des formalisations gérées dans les wikis. En aval, la bibliothèque XML offre des procédures de génération de modèles en utilisant par exemple les outils de visualisation géographique de Wikipédia. La bibliothèque XML permet également de développer des robots pour assister les actions éditoriales.

La fin des financements ISTEX a réduit les capacités de coopération. Nous avons donc recherché des thématiques que nous pouvions explorer sans l'obligation technique de recourir à une expertise extérieure (comme la santé). Une première série d'expériences en musique a amené à enrichir notre panoplie de services avec des rééditions hypertexte d'ouvrages avec des éléments musicaux (comme le Dictionnaire de Jean-Jacques Rousseau). En 2020, une nouvelle étape a été franchie avec une bibliothèque numérique sur la Chanson de Roland. Ici, pratiquement chaque document (strophe d'un manuscrit, chapitre d'une édition critique, article de recherche, partition) demande un traitement numérique spécifique. Chaque mot d'un manuscrit (ou d'une note de philologue) peut devenir un élément hypertexte dont les explications peuvent de développer dans plusieurs wikis.

3. Le démonstrateur Wicri

Le réseau actuel est un ensemble encyclopédique étendu (avec par exemple des rééditions d'ouvrages et des extraits de bases de données) développé sur 150 wikis. Il offre également des applications stabilisées comme la revue les mots de l'agronomie de l'INRAE.

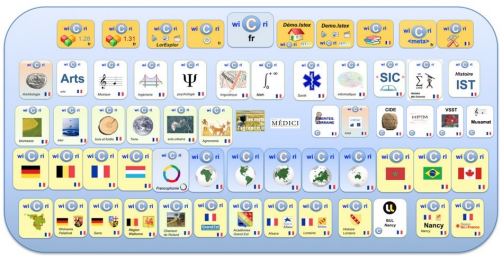

Figure 1. Le réseau Wicri

En termes de volumétrie, ce réseau expérimental contient 200.000 pages wiki (avec 40.000 articles conséquents et 13.000 fichiers multimédia). En complément, 150 serveurs d'exploration donnent accès à plus d'un demi-million de documents.

4. Démonstrations proposées

Ce réseau propose un très vaste champ de démonstrations. De façon générale il est possible de voir la multiplicité des relations sémantiques et les mécanismes de cohérence dans le réseau de wikis. Au-delà de cette base, quatre applications significatives sont proposées. Sur le COVID une vingtaine de serveurs d'explorations ont été réalisés avec une procédure rapide de mise en place (quelques minutes). Sur la Chanson de Roland nous avons construit un ensemble qui contient déjà près de 4.000 pages significatives (chapitres d'ouvrage, versets de manuscrit, analyses critiques, etc.). L'Histoire de l'Information scientifique et technique est une thématique en cours de démarrage et qui doit se décliner dans l'ensemble des wikis du réseau. Enfin, un travail sur l'intégration de la très vaste Histoire naturelle de Buffon dans une bibliothèque hypertexte vient d'être initialisé.

5. Remarques et perspectives

Le démonstrateur Wicri montre la faisabilité d'un déploiement de l'IST française pour retrouver une dimension internationale. Il montre aussi l'intérêt de cette technologie dans les applications de la recherche où les approches classiques des systèmes d'information échouent ou donnent lieu à des développements particulièrement laborieux.

Nous démarrons une nouvelle étape dans laquelle nous allons étudier une répartition du réseau Wicri sur plusieurs sites physiques. Avec des moyens limités (un seul permanent, le retraité auteur de cet article), nous allons démarrer avec 2 sites dans un premier temps et avec une augmentation du nombre des wikis (pour prendre en compte, par exemple, l'ensemble des régions françaises).

Pour aborder une vraie dimension internationale, compétitive avec la volumétrie de la Wikimedia Fundation, il faudrait passer à quelques milliers de wikis sur une centaine de sites. Cet objectif nous semble difficile et ambitieux mais techniquement abordable pour la communauté universitaire. Mais la base actuelle offre déjà une infrastructurel de formation, expérimentation et même de services…

6. Liens et bibliographie

Liens vers le réseau Wicri :

– le wiki d'accueil : https://wicri-demo.istex.fr/Wicri/Wicri/fr ;

– Une reproduction hypertexte de cet article avec accès aux démonstrations.
https://wicri-demo.istex.fr/Wicri/Sic/fr/index.php/INFORSID_Nancy_(2004)_Ducloy

Bibliographie du projet Wicri :

https://wicri-demo.istex.fr/Wicri/Wicri/fr/index.php/Bibliographie_du_projet_Wicri