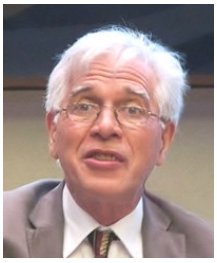

Wikipédia et la Science ouverte

Jacques DUCLOY
Ingénieur honoraire du CNRS ; Laboratoire Paragraphe (Univ. Paris 8)

(15 06 2020)

Par un beau matin de 2002, j'ai reçu un courriel de Jean-Michel SALAÜN, chercheur réputé en sciences de l'information. Il m'invitait à visiter un site encore inconnu. Ce site s'appelait **Wikipédia**.

Je me souviens encore de ma réaction : *"Ce truc semble assez génial, mais ça ne pourra jamais marcher ; ça ne peut que diverger dans tous les sens…"*.

Et puis, j'ai repris une activité apparemment plus sérieuse : la direction de la production des bases Pascal et Francis du CNRS (500.000 références documentaires par an). Une chaîne totalement intégrée, comportant de multiples étapes de validation avant mise en ligne, qui devait garantir l'intégrité d'une synthèse de l'information scientifique et technique mondiale…

WIKIPÉDIA
L'encyclopédie libre

Quelques années après, le 22 janvier 2007, à 10h22 très exactement (tout le monde peut vérifier), j'ai vaincu mes hésitations pour faire ma première contribution à Wikipédia. J'avais commenté l'article "Dublin Core", une norme pour l'Internet. Six mois plus tard, le 10 octobre 2007, j'ai osé créer mon premier article, le "Livre vermeil de Montserrat", un recueil de partitions musicales de la fin du Moyen Âge. Et puis, comme des centaines de milliers de personnes, je suis devenu dépendant de la "*wikicontributomanie*", une nouvelle drogue encore légère.

Le 27 août 2008, j'ai essayé une drogue encore plus dure en créant, au LORIA, la première pierre d'un réseau de wikis expérimental nommé WICRI, en utilisant MediaWiki, le moteur de Wikipédia.

En 2020, les bases Pascal et Francis se sont arrêtés. Wikipédia est devenu la principale source d'accès à la connaissance scientifique et culturelle. Pour ma part, j'étais gravement dépendant de WICRI.

A partir de ces expériences, je voudrais maintenant alerter sur les dangers, pour la science, la société et les citoyens, d'une encyclopédie fondée sur l'anonymat, et qui n'aurait plus aucune concurrence.

La galaxie Wikipédia

Tout utilisateur d'Internet un peu entraîné, sait reconnaître Wikipédia, qui est souvent le premier site proposé par Google pour donner une définition. En fait, l'internaute est conduit sur le site "Wikipédia en français", qui cache (au 17 mai 2020) 309 autres wikis, qui sont des versions de cette encyclopédie en différentes langues… L'ouvrage le plus volumineux est la version anglaise qui contenait, à cette date, un seul volume de 6.079.791 pages au contenu significatif.

En réalité le nombre réel de "pages" de cette version anglaise dépasse les 50 millions. En effet, une page peut contenir un ouvrage qui serait imprimé sur plus de 50 feuilles A4. Elle peut se limiter à un morceau de code pour dire, sur la page "Louis 14", que cette page est redirigée vers "Louis XIV". Comme pour les index matières d'une encyclopédie papier, Wikipédia gère des milliers de pages de terminologie. Nous verrons plus loin d'autres exemples.

La version française est plus modeste, mais déjà conséquente (2.217.129 pages de contenus sur un total de 10.000 pages). Parmi les centaines de langues disponibles, il existe par exemple une version en latin (132.000 pages, très intéressantes à feuilleter).

Toutes ces versions sont hébergées et maintenues par une organisation américaine (*non profit*), la ***WikiMedia Foundation***. Sa Présidente actuelle est Maria SEFIDARI, (38 ans, nationalité espagnole) et sa Directrice Générale est Katherine MAHER, (37 ans, nationalité américaine). Leurs photos figurent ci-dessus.

La Fondation supervise un autre ensemble d'applications qui utilisent la même architecture logicielle, le moteur **MediaWiki**.

La plus intéressante est probablement "***WikiMedia Commons***", une gigantesque base d'objets multimédias. Ce réservoir contient 60 millions d'images ou de fichiers vidéo. Pour chacun d'eux, un document le décrit et précise ses conditions juridiques. La plupart sont sous des licences *Creative Commons* et donc réutilisables par n'importe qui. Attention cependant, c'est la loi américaine qui est y est appliquée (y compris, d'ailleurs, pour Wikipédia en français).

Enfin la *Wikimedia Foundation* offre d'autres services. Par exemple ***Wiktionaire*** est un dictionnaire, composé comme pour Wikipédia, d'un site wiki par langue. La version française contient près de 4 millions d'entrées, pour des mots, dans toutes les langues, avec leurs définitions en français. On trouve également une bibliothèque d'ouvrages également multilingue, etc.

Ci-dessous la galaxie des projets de la *Foundation* (plus de 500 wikis).

Les raisons d'un succès, un logiciel qui bouleverse les pratiques.

Dès 1993, des projets d'encyclopédies alimentées par des internautes ont été tentés. Le plus ancien a pour nom Interpedia. En 2000, Jimmy WALES lance un projet nommé Nupedia. Il était basé sur un processus rigoureux avec notamment un comité scientifique qui contrôlait l'introduction de nouveaux articles. Ce projet a été abandonné en 2003.

En 2001, à côté de ce projet "très sérieux", Jimmy WALES, en association avec Larry SANGER, a créé un projet plus populaire : Wikipédia…

Au départ, Wikipédia utilisait *UseModWiki*, un moteur efficace, mais rudimentaire (écrit en Perl). En 2002, Magnus MANSKE, un étudiant allemand en bio-informatique à l'Université de Cologne, a développé MediaWiki, un moteur de wiki bâti sur les technologies maintenant classiques : php et MySQL. Il y a implanté un mécanisme "de modèles" qui permet à n'importe quel internaute (bien formé) de développer des fonctions propres à son domaine de spécialité. Par exemple, toute la cartographie de Wikipédia est réalisée par des "utilisateurs" et non par "le service informatique de Wikipédia". Ces utilisateurs, concepteurs des modèles de cartographie, doivent maîtriser des bases de programmation (et peuvent être, *"dans le civil"*, des informaticiens).

Grâce aux ordinateurs de terrain, le GPS et le laser télémètre, il est possible d'effectuer directement la cartographie sur le terrain. À gauche le sentier du sommet Pinnacle dans le parc national du mont Rainier (État de Washington).

Le logiciel MediaWiki a donc été particulièrement bien conçu pour faire résoudre par les utilisateurs les problèmes techniques posés par l'immense diversité des domaines scientifiques, techniques ou culturels abordés dans cette gigantesque encyclopédie.

Il dispose d'outils également sophistiqués pour gérer les malversations et les erreurs. Mais l'anonymat et l'absence de réelle ligne éditoriale marque des limites.

Les vandales sont repoussés mais les manipulateurs peuvent circuler…

Est-il possible d'écrire n'importe quoi sur Wikipédia en s'y prenant n'importe comment ? *La réponse est non !*

Est-il possible de manipuler le contenu en s'y prenant bien ? *La réponse est plutôt oui…*

Avec près de 20 ans d'expérience, les développeurs de MediaWiki ont programmé des robots qui savent bloquer les vandalismes brutaux. Par exemple, il est possible d'intervenir sur Wikipédia sans être identifié, autrement que par son "adresse IP" (adresse numérique interne au réseau Internet). Les robots sont alors particulièrement attentifs.

Voici une expérience vécue avec mon petit-fils, prénommé Arthur. Nous étions sur la page correspondant à son prénom et je voulais lui montrer, en regardant l'historique, que cette page avait été

rédigée par un grand nombre de personnes. En effet, cet historique permet de connaître en détail la suite de toutes les contributions depuis la création d'un article.

Nous avons eu la chance de tomber sur une tentative de vandalisme. A partir d'une adresse IP, un apprenti vandale avait complètement effacé le contenu pour le réduire à "*Arthur est un imbécile*" (un autre mot avait été utilisé). Dans le même espace de temps, moins d'une seconde, un robot avait restauré le contenu initial. Le vandale n'avait même pas pu voir le résultat de sa tentative…

D'autres mécanismes sont basés sur les "listes de suivi". Tout contributeur enregistré peut se construire une liste d'articles pour lesquels il sera alerté en cas de modification.

Lorsque j'ai réalisé mes premières interventions, j'ai eu la bonne surprise de voir qu'elles étaient relues, et que mes fautes d'orthographes étaient corrigées. Mais, cela n'a duré qu'une dizaine de jours. J'ai compris pourquoi quand je me suis constitué une liste regroupant les articles que j'avais créés ou sur lesquels j'avais fait des interventions significatives. En effet, au bout de quelques temps, j'allais vérifier les interventions faites sur mes textes, par des anonymes se cachant sous une adresse IP. En revanche, faute de temps, je ne vérifiais plus celles des pseudonymes qui me paraissaient fiables.

Cette procédure est donc relativement efficace mais présentent des dangers. Pour un groupe de manipulateurs, il est relativement facile de commencer par des contributions apparemment banales (ou ouvertes) pour s'approprier un espace informationnel. L'industrie du film pornographique a su notamment utiliser Wikipédia avec une stratégie impressionnante. L'exploration des catégories est significative. Par exemple, sur Wikipédia en français, la catégorie "Psychothérapeute français" indexe seulement 47 personnes (une seule sur la version anglaise). En revanche, la catégorie "Actrice pornographique américaine" indexe 620 pages sur la version française. Les pages sont parfois rédigées par des professionnels qui donnent une filmographie complète et des liens vers tous les sites payants. Paradoxalement, pour cette catégorie, la version anglaise semble mieux régulée avec simplement 294 pages (la version espagnole dépasse les 1.000 pages).

Apprendre à coconstruire de la connaissance avec Wikipédia

L'histoire de Wikipédia montre comment des contributeurs, au départ amateurs, arrivent à s'autogérer pour mener des projets éditoriaux conséquents et, parfois, de grande qualité.

Les sujets qui ne portent pas à polémique, comme les communes de France, constituent un bon exemple.

En consultant les onglets "historique", il est possible de suivre les évolutions du style. Au début, les pages étaient limitées à une courte définition, et quelques paragraphes. Puis, les auteurs ont unifié les déclarations factuelles (population, région) avec des tableaux. En s'appuyant sur des pages de discussion, la communauté a su créer des éléments plus sophistiqués comme les Infobox (voir ci-dessous pour Chaligny).

Ce type de présentation s'est généralisé. Les communautés spécialisées ont su s'organiser pour définir des outils communs (comme des modèles dits génériques pour ces "Infobox").

Les contributeurs expérimentés ont créé des palettes, telles que "les régions d'un pays" qui permettent aux débutants de respecter la terminologie commune. Ils ont défini des outils pour gérer la qualité des articles, (par exemple en apposant des bandeaux "ébauches").

Tout le contenu de Wikipédia a donc été intégralement défini et réalisé dans un mode communautaire. Cette démarche est à 180° par rapport à l'état de l'art de l'édition numérique qui enferme les auteurs dans un modèle défini *a priori*. Ici tout est construit *a posteriori* de façon incrémentale.

Mais là encore, l'anonymat peut poser problème. Voici une illustration avec le "sourçage" (un terme du jargon wikipédien). Le sourçage définit l'obligation de fournir des sources. En effet, toute affirmation doit être étayé par des sources d'informations fiables, comme des articles scientifiques.

Voici un exemple, montrant l'intérêt et la limite de ce mécanisme. Un groupe de mathématiciens pilotés par l'École Normale Supérieure de Lyon avait rédigé un article sur l'infini. De retour d'un séjour en Égypte, avec un brin de provocation par rapport à une communauté quelque peu portée sur l'athéisme, j'ai

ajouté une rubrique "L'infini en théologie". Je citais Akhenaton, promoteur du monothéisme en Égypte. En effet la notion de Dieu unique induit, c'était mon hypothèse, la notion d'éternité.

Ma contribution a été immédiatement (en quelques heures) agrémentée, probablement par un ami normalien, d'une bannière d'avertissement "citation nécessaire".

Coup de chance, en quelques heures, j'ai trouvé un ouvrage de Sigmund Freud, qui contenait plusieurs chapitres sur ce thème. J'ai pu effacer le bandeau en ajoutant une citation.

Autrement dit, il suffit de trouver une "publication de référence" pour avancer une opinion… Ce mécanisme peut avoir des conséquences un peu imprévisibles dans des domaines sensibles comme la santé.

Actuellement, la crise sanitaire liée au Covid-19 crée une véritable explosion pandémique sur le web. Sur Google, le terme Covid était déjà présent dans 4.320.000.000 pages le 21 mai 2020 à 9h45. Vous avez bien lu **quatre milliards**, là où "VIH" n'indexe "que" 40 millions de pages". Sur Google Scholar on trouve déjà 300.000 articles (ou citations). Sur les bases de la NLM (National Libray of Medicine), 11.000 articles ont déjà été sélectionnés. Autrement dit, un internaute qui voudrait valider n'importe quelle affirmation sur Wikipédia dispose de centaines de milliers de sources potentielles de citation. Un groupe de manipulateurs agissant sous couvert de pseudonymes apparemment crédibles peut se constituer quelque chose qui pourrait ressembler à un "comité scientifique".

De nombreuses universités américaines, par exemple le MIT, ont une politique de publication sur Wikipedia qui dépasse l'aspect institutionnel. L'influence relativement modérée de la pornographie aux États-Unis en est un indice.

Les universités francophones ont intérêt à être plus présentes sur Wikipédia.

Mais on peut aller plus loin. Les universités et les établissements de recherche pourraient créer une alternative en s'appuyant sur les acquis de Wikipédia, mais avec des contributions transparentes modérées par des comités scientifiques.

MediaWiki dans la science et la culture

Nous avons évoqué les qualités techniques du moteur MediaWiki pour Wikipédia.

En pratique, ce logiciel libre est très facilement accessible. Plus de 30.000 applications sur Internet utilisent cette base. Par exemple, en musique, le site ChoralWiki fondé en 1998 utilise MediaWiki depuis 2005 et offre près de 35.000 partitions libres, souvent de très grande qualité, composées par près de 3.500 auteurs. En biologie, le site Proteopedia donne accès à 200.000 pages de contenu, avec des molécules sur lesquelles le chercheur peut naviguer en 3D.

De multiples extensions ont été développées. Une des plus intéressantes s'appelle Semantic MediaWiki, construite par l'Université de Karlsruhe. Voici en deux mots, l'apport de cette extension. Avec MediaWiki La gestion des contenus est structurée par un langage de programmation qui permet de définir des modèles.

Semantic MediaWiki ajoute une couche de type "base de données applicative". Pour cela, il utilise un mécanisme de liens sémantiques. Par exemple, le contributeur écrit formellement *"la ville de Nancy est explicitement dans la région Grand Est"*, au lieu de dire simplement : *" il y a un lien entre la page Nancy et la page Grand Est"*. Voici un exemple de lien simple, dans la page Nancy :

```
Nancy est une ville du [[Grand Est]]
```

Ce code génère un lien entre la page Nancy et la page Grand Est.

Avec Semantic MediaWiki, le lien peut maintenant être qualifié pour préciser sa nature.

```
Nancy est une ville du [[A pour région::Grand Est]]
```

Ce mécanisme permet d'insérer des éléments dynamiques dans le contenu des pages. Par exemple, il est possible d'éditer la liste des villes qui ont plus de 20.000 habitants dans une région donnée. Si la population d'une ville augmente, toutes les pages concernées seront modifiées.

Au-delà de l'aspect éditorial, MediaWiki et ses extensions permettent par exemple de gérer les données de la recherche.

Nous avons utilisé Semantic MediaWiki dans un projet encyclopédique, Wicri, qui a été créé au LORIA, puis successivement porté par l'INPL, l'INIST et l'Université Paris 8. Wicri est un réseau de 150 wikis bibliothèques thématiques, régionales ou spécialisées.

Chaque bibliothèque est gérée par un wiki sémantique, où les documents, réédités en hypertexte, sont déposés sur une couche encyclopédique.

Cette bibliothèque est enrichie par une application de fouille de données. Sur un sujet donné, il est possible de compléter la base éditoriale par des "serveurs

d'explorations" qui permettent de naviguer dans des corpus de documents scientifiques. Par exemple, sur un wiki dédié à la santé, nous avons monté une démonstration autour du Covid-19. Elle permet d'explorer une collection de près de 60.000 articles ou documents scientifiques.

Si Wicri est un prototype, de grands projets viennent d'être démarrés dans le monde de la recherche en utilisant une autre ressource issue de Wikipédia : le réservoir terminologique WikiData. Il s'agit d'une base de connaissance de 86.000.000 d'éléments, extraite des serveurs de *la WikiMedia Foundation*. Ce réservoir est le pilier actuel du Web sémantique qui contrôle de fait la terminologie sur le Web.

Deux projets, lancés il y a un an, illustrent ce nouveau phénomène :
- aux États-unis, les grands acteurs des bibliothèques numériques de la recherche (OCLC, NLM) s'allient avec de grandes universités (Harvard, Cornell) pour une interconnexion de leurs collections en utilisant MediaWiki et WikiData,
- en France, l'Abes et la BnF démarrent une opération analogue.

Première conclusion : Wikipédia est un ensemble incontournable qui aurait besoin de concurrence

Dans cet article, j'ai essayé d'attirer l'attention sur les dangers du monopole actuel de Wikipédia.

Je ne voudrais pas donner l'impression de noyer le bébé avec l'eau du bain. Pour celui qui sait l'utiliser avec prudence (en consultant par exemple les citations), Wikipédia est une fantastique source d'information qui contient un grand nombre de très bons articles.

En revanche, le CNRS et les Universités affichent des slogans vantant les mérites de la Science ouverte, et surtout de la bibliodiversité. Il est peut-être urgent de les mettre en pratique pour offrir à la société et aux citoyens une alternative plus fiable à Wikipédia.

Deuxième conclusion : il faut restaurer un dispositif ouvert de traitement de l'information scientifique

Au moment de terminer cet article, j'ai eu connaissance d'une initiative de la Maison Blanche pour utiliser à grande échelle l'intelligence artificielle dans la lutte contre le Covid 19. Voici un extrait d'un communiqué de l'Ambassade de France à ce sujet.

Via l'OSTP (Office for Science and Technology Policy), la maison blanche coordonne une initiative public-privé de grande envergure pour l'analyse de la littérature scientifique. Ce consortium regroupait initialement : a) la National Library of Medicine (NLM), la plus grande bibliothèque médicale du monde, b) le Allen Institute for AI, un institut de recherche indépendant à but non lucratif, c) la Chan Zuckerberg Initiative (CZI), qui se définit comme une "entreprise philanthropique" visant à apporter des solutions technologiques à la société, d) le Center for Security and Emerging Technology (CSET) de Georgetown University et Microsoft Research. De nouveaux partenaires ont rejoint cette initiative : bioRxiv, medRxiv, Amazon Web Services (AWS) et tout récemment IBM Research.

Ce dispositif met au premier plan la NLM. Il montre l'importance cruciale des bibliothèques numériques scientifiques pour une stratégie d'intelligence artificielle en Médecine.

Dans le cadre de l'expérimentation autour du Covid 19 citée plus haut, nous avons constaté notre totale dépendance des États-Unis, via la NLM. Cette dépendance résulte du démantèlement par le CNRS des activités sociétales de l'INIST (représentées par les bases Pascal et Francis).

Un grand projet comme un réseau d'encyclopédies, reposant sur la sélection et l'analyse de l'Information scientifique mondiale permettrait à la France (et à l'Europe) de retrouver une indépendance stratégique.

Un seul exemple : dans la course aux vaccins contre le Covid-19, l'intelligence artificielle va jouer un rôle déterminant. Celui qui arrivera le premier pourra faire de ce vaccin un bien public ou un produit marchand protégé par un brevet.

Références

- https://www.mediawiki.org/wiki/MediaWiki
- https://commons.wikimedia.org/wiki/Accueil
- https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki
- https://france-science.com/covid-19-lintelligence-artificielle-pour-accelerer-la-recherche-scientifique/
- https://lorexplor.istex.fr/Wicri/Sante
- J. Ducloy et al. "Systèmes d'information encyclopédiques édités par les scientifiques : partager le savoir pour l'excellence documentaire et scientifique". In : Ingénierie des systèmes d'information, numéro 1, 2019
- https://www.openscience.fr/Numero-1-526