Wikipédia
Quinze ans de recherches

INTERNET

Les savoirs, outils et fonctionnements de l'encyclopédie en ligne sont devenus des objets d'études. Au cœur de celles-ci figure la création de bases de connaissances compréhensibles par les machines. Un enjeu pour les grands de l'industrie numérique

DAVID LAROUSSERIE

Quel succès! Quinze ans après son lancement, le 15 janvier 2001, par les Américains Jimmy Wales et Larry Sanger, l'encyclopédie en ligne Wikipédia reste le premier site non commercial du Web mondial, toujours dans le top 10 des sites les plus fréquentés avec près de 500 millions de visiteurs uniques par mois pour plus de 250 éditions linguistiques. 36,9 millions d'articles sont rédigés, corrigés, améliorés par quelque 2 millions de contributeurs. 800 nouvelles entrées en anglais sont ajoutées chaque jour, 300 en français. La version française tenant la troisième position, avec plus de 1,7 million d'articles, derrière l'anglophone (plus de 5 millions) et la germanique (1,8 million).

Mais Wikipédia, c'est moins connu, est bien plus qu'une encyclopédie qu'on consulte pour se documenter ou faire ses devoirs scolaires. Elle est devenue aussi un objet de recherche en tant que tel, à l'instar d'une tribu d'Amazonie, d'un programme informatique ou d'un patient. La base de données Scopus, l'une des trois plus importantes du monde, recense ainsi plus de 5400 articles ayant pour sujet ou pour objet Wikipédia publiés dans des revues, des actes de colloques ou des livres. Quatorze brevets mentionnent même le célèbre site, selon la même Scopus.

Autre preuve de l'intérêt académique pour le sujet, en juin 2013, à Paris, se tenait un colloque, coorganisé par le CNRS et le CNAM et intitulé «Wikipédia, objet scientifique non identifié», avec sociologues, spécialistes de sciences de la communication, informaticiens...

Depuis 2011, la Fondation Wikimédia, qui héberge Wikipédia, s'est même dotée d'un groupe de recherche. *«A l'origine, nous étions là pour fournir des outils d'analyse à la communauté. Maintenant, nous sommes un vrai département de recherche, testant de nouvelles technologies et collaborant avec les universités»*, résume Dario Taraborelli à la tête des dix personnes de ce département, en Californie. Fin novembre 2015, il annonçait ainsi un projet d'intelligence artificielle ecapable d'estimer la probabilité qu'une modification soit dommageable ou non à un article et donc susceptible d'être retirée. Auparavant, les systèmes automatiques de détection avaient tendance à trop souvent écarter les contributions pourtant bienveillantes, freinant l'entrée de nouveaux contributeurs. Le groupe essaie aussi de réduire les asymétries de contenus entre différentes langues en proposant automatiquement des articles à rédiger aux contributeurs des langues minoritaires.

L'ouverture et la transparence offrent aux chercheurs la vérifiabilité et la reproductibilité

Mais que font tous les autres chercheurs en tripatouillant Wikipédia? De récentes publications témoignent du large spectre couvert. Depuis novembre, une équipe japonaise s'est servie des articles de l'encyclopédie pour analyser les suicides de personnalités dans son pays. Des Britanniques ont construit automatiquement un glossaire technique. Des Turcs ont utilisé le site pour repérer à grande échelle des entités dans des corpus de leur langue. Des Français ont proposé un classement des universités reposant sur les citations des établissements au sein de plusieurs versions linguistiques de Wikipédia. Citons encore un article paru en mai, qui prévoit les pics d'apparition de la grippe grâce aux statistiques de visites des pages de l'encyclopédie.

Les raisons d'un tel engouement sont simples à comprendre. L'objet est vaste, une quinzaine de gigaoctets de textes (pour la version anglaise). D'utilisation gratuite, contrairement aux données de Facebook, Google ou Twitter, pourtant gigantesques et fournies gracieusement par leurs utilisateurs. Même les données de fréquentation sont disponibles pour chaque article! Les archives sur quinze ans permettent d'avoir du recul historique, tout en ayant un objet toujours rafraîchi. Des versions en plus de 200 langues ouvrent des perspectives pour des comparaisons ou des analyses culturelles. L'ouverture et la transparence offrent aussi ce que les chercheurs adorent: la vérifiabilité et la reproductibilité. Pour parfaire leur bonheur, l'encyclopédie, tel un iceberg, recèle plus de trésors que sa seule vitrine d'articles. Si la version française contient 1,7 million de pages d'articles, elle contient 4,5 fois plus de pages pour les historiques, les discussions et autres coulisses qui font le dynamisme et la réputation du site.

Du coup, presque tous les domaines sont couverts. La sociologie, bien sûr, fascinée par cette démocratie d'un nouveau genre, car auto-organisée et reposant sur quelques règles et le consensus. Les chercheurs, profitant de la transparence du site, y ont également étudié le rôle des «vandales» et autres «trolls» qui mettent leurs pattes malveillantes dans les articles. Les inégalités hommes-femmes particulièrement criantes, avec moins de 10 % de contributrices à l'encyclopédie, ont également donné lieu à beaucoup de littérature et de controverses.

Wikipédia est devenu une sorte de bac à sable dans lequel s'ébrouent les spécialistes du traitement automatique du langage qui disposent là d'un corpus immense pour tester leurs logiciels de reconnaissance de texte, de traduction, d'extraction de sens... C'est aussi le jouet de physiciens, statisticiens, informaticiens... prompts à dégainer leurs outils d'analyse pour en extraire de nouvelles informations ou aider à les visualiser.

«Après quinze ans, l'intérêt des chercheurs est toujours là. La première phase était très active car l'objet était nouveau. Cela a contribué à l'émergence de nouveaux domaines comme la sociologie quantitative ou l'informatique sociale, rappelle Dario Taraborelli. *Puis, à partir de 2007, l'apparition de nouveaux médias sociaux a détourné un peu les recherches, avant un renouveau depuis 2010. Notamment parce que nous sommes le seul site important à publier nos données quotidiennes de trafic.»*

Ce renouveau est aussi tiré par une révolution à venir. Wikipédia est devenu l'un des maillons indispensables à un projet particulièrement ambitieux: rassembler toute la connaissance mondiale et la rendre intelligible par des machines. *«Notre ambition est de rendre encore plus intelligents les ordinateurs afin qu'ils soient toujours plus utiles à l'humanité»*, s'enthousiasme Fabian Suchanek, enseignant-chercheur à Télécom ParisTech et artisan de cette évolution qui vise à transformer Wikipédia et d'autres riches corpus en une source accessible aux ordinateurs.

De tels changements sont en fait déjà à l'œuvre, discrètement. Dans les moteurs de recherche par exemple, lorsque l'utilisateur tape un nom de célébrité, apparaissent toujours une liste de liens mais aussi un encadré résumant la biographie de la personne cherchée. Et cela automatiquement: le programme a compris où, dans la page Wikipédia, se trouve l'information souhaitée. Mieux. On peut désormais poser des questions explicites, en langage naturel, à ces moteurs: quand Elvis Presley est-il mort? Où? Quel est l'âge de François Hollande?... et recevoir des réponses directes, sans avoir à lire la page contenant l'information.

Derrière ces prouesses qui n'ont l'air de rien se cachent de nouveaux objets : les bases de connaissance. Les plus célèbres sont Yago, DBpedia, Freebase ou Wikidata. Toutes se sont construites en triturant Wikipédia. Et, preuve des enjeux économiques, les plus grands du Web actuel investissent dans ces constructions. En 2010, Google a ainsi racheté Freebase, qui lui sert pour son Knowledge Graph, l'encadré qui fournit des réponses directes aux requêtes. L'entreprise soutient également financièrement Wikidata, une initiative de la fondation Wikimédia. Amazon a racheté EVI en 2012, anciennement connue sous le nom de True Knowledge, une base de connaissances.

En outre, derrière les assistants personnels vocaux des mobiles, Siri, Cortana ou Google Now, se cachent aussi ces fameuses bases de connaissances. Pour gagner au jeu Jeopardy en 2011, l'ordinateur Watson d'IBM a bien sûr assimilé bon nombre de données, en particulier de Wikipédia, mais dans une forme prédigérée fournie par la base de connaissances Yago.

Le sujet de ces bases ou graphes de connaissances est très actif. Le chercheur le plus prolixe sur Wikipédia, toutes activités confondues selon Scopus, est par exemple l'Allemand Gerhard Weikum de l'Institut Max-Planck de Sarrebruck, à l'origine de la première base de connaissances, Yago, en 2007. Le second est un Hollandais, Maarten de Rijke, professeur d'informatique à l'université d'Amsterdam, dont les récents travaux utilisent ces graphes. Il est capable de savoir de quoi parle un tweet en repérant les noms et les faits à l'intérieur et en les confrontant à Yago ou DBpedia. Il enrichit aussi les émissions de télévision automatiquement en fournissant des liens sur les tablettes ou téléphones, choisis en fonction du thème de l'émission, déterminé grâce aux bases de connaissances.

« Avec ces bases de connaissances, on peut faire des choses qui étaient impossibles auparavant », estime Fabian Suchanek, cofondateur de Yago. Par exemple ? *« Extraire de l'information du quotidien* Le Monde *: combien de femmes en politique au cours du temps ? Quel est l'âge moyen des politiciens ou des chanteurs cités ? Quelles compagnies étrangères sont mentionnées ? »*, énumère ce chercheur en citant un travail publié en 2013 avec la collaboration du journal. Le *New York Times* construit sa propre base de connaissances tirées des informations de ses articles. Autre exemple, il devient possible de poser des questions aussi complexes que : qui sont les politiciens également scientifiques nés près de Paris depuis 1900 ? Ou, plus simplement, quelle est la part des femmes scientifiques dans Wikipédia ?

L'évolution actuelle vise à transformer Wikipédia et d'autres riches corpus en une source accessible aux ordinateurs

Mais quelle différence entre ces objets et une base de données ou même une page Wikipédia ? Si un humain comprend que dans la phrase « Elvis Presley est un chanteur né le 8 janvier 1935 à Tupelo, Mississippi », il y a plusieurs informations sur son métier, sa date et son lieu de naissance, une machine ne le comprend pas, et ne peut donc répondre à la question simple, pour un humain, « Quand Elvis est-il né ? ». *« C'est un peu paradoxal, mais pour un informaticien, notre langage n'est pas structuré et donc un ordinateur ne peut le comprendre ! »*, souligne ironiquement Fabian Suchanek. Il faut donc transformer les pages en les structurant différemment, en commençant par repérer les entités, les faits et les relations entre eux. Presley est une entité. Sa date de naissance ou son métier sont des faits. « Né le » et « a pour métier » sont les relations. Tout cela peut être codifié en langage informatique.

Une autre particularité de ces objets est qu'ils ne répertorient pas ces faits et entités dans des tableaux, comme la plupart des bases de données, mais en les organisant en arborescences ou en graphes. Les branches correspondent aux liens entre les entités et les faits. Les informaticiens et mathématiciens ont bien sûr développé les techniques pour interroger ces graphes et y faire des calculs comme dans un vulgaire tableur. Aujourd'hui, Yago « sait » plus de 120 millions de choses sur 10 millions d'entités (personnalités, organisations, villes...).

L'avantage-clé est que le rapprochement devient plus simple entre plusieurs bases de connaissances, celles construites sur Wikipédia mais aussi d'autres concernant les musiciens, les coordonnées GPS, les gènes, les auteurs... Le site Linkeddata.org recense ces nouvelles bases et leurs liens entre elles. Petit à petit se tisse un réseau reliant des faits et des entités, alors que, jusqu'à présent, la Toile connecte des pages ou des documents entre eux. Cela contribue au rêve de ce que Tim Berners-Lee, le physicien à l'origine du Web, a baptisé « Web sémantique » en 2001. *« Les défis ne manquent pas. La troisième version de Yago est sortie en mars 2015. Nous avons déjà traité la question du temps. Nous traitons aussi plusieurs langues. Il faut maintenant s'attaquer aux "faits mous", c'est-à-dire moins évidents que les dates et lieux de naissance, les métiers, le genre...*, estime Fabian Suchanek. *En outre, tout ne peut pas se mettre dans un graphe ! »*

Bien entendu, faire reposer la connaissance future de l'humanité sur Wikipédia n'a de sens que si ce premier maillon est solide. La crédibilité de l'encyclopédie a donc été parmi les premiers sujets d'études. Dès 2005, *Nature* publiait un comparatif entre l'encyclopédie en ligne et sa « concurrente » *Britannica*, qui ne montrait pas d'énormes défauts pour la première. D'autres études ont été conduites depuis pour estimer l'exactitude, en médecine par exemple, Wikipédia étant l'un des premiers sites consultés sur ces questions. Les résultats sont bien souvent satisfaisants.

« C'est finalement un peu une question vaine scientifiquement, car les comparaisons sont souvent impossibles. On confronte les articles tantôt à des encyclopédies, tantôt à des articles de revues scientifiques... », estime Gilles Sahut, professeur à l'Ecole supérieure du professorat et de l'éducation, de l'université Toulouse-Jean-Jaurès. *« La question a un peu changé de nature. Il faut passer d'une appréciation globale à une appréciation au cas par cas, et donc éduquer afin d'être capable de dire si un article semble biaisé ou complet »*, précise ce chercheur, qui a soutenu une thèse en novembre 2015 sur la crédibilité de Wikipédia. Il adosse ce constat à une étude menée sur plus de 800 jeunes entre 11 et 25 ans, pour tester la confiance accordée à l'encyclopédie. Celle-ci s'érode avec l'âge et le niveau de scolarité, mais elle remonte dès lors que les élèves participent. *« Ils découvrent d'ailleurs, comme leur enseignant, qu'il n'est pas si facile d'écrire dans Wikipédia ! »*, sourit le chercheur en faisant allusion aux difficultés à entrer dans la communauté. *« Certes les wikipédiens sont des maîtres ignorants sur les savoirs, comme le dit le sociologue Dominique Cardon, mais ils sont très savants sur les règles et les procédures ! »* ■

« Un objet scientifique non identifié »

Lionel Barbe est enseignant-chercheur en sciences de l'information et de la communication à l'université Paris-Ouest-Nanterre. En juin 2013, il a coorganisé un colloque sur « Wikipédia, objet scientifique non identifié ». En 2015, un ouvrage des Presses universitaires de Paris-Ouest a rassemblé, sous le même titre, les contributions d'une douzaine de chercheurs participants.

Que représente Wikipédia pour vous ?

C'est un objet en constante transformation, un objet non identifié, ainsi que nous l'avions désigné pour notre colloque de 2013, car il est singulier. Selon le bout par lequel on l'aborde, on peut parvenir à des conclusions différentes. Si vous vous intéressez à la qualité des pages dans certaines disciplines, comme l'informatique ou la géographie, vous trouverez Wikipédia très bien. Si vous vous intéressez à l'actualité ou à la sociologie, vous la trouverez peut-être plus faible, bien que cela soit de plus en plus relatif car le niveau qualitatif est désormais élevé dans de nombreuses disciplines.

Par ailleurs, il faut comprendre qu'il n'y a pas de version finale d'un article, les pages sont en constante évolution. Elles sont le résultat d'un processus issu d'une dynamique de consensus. Cette caractéristique fascine les chercheurs depuis l'origine : comment comprendre ce miracle de l'auto-organisation qui est une démocratie reposant sur le consensus au-delà du vote ? Dans Wikipédia, on ne peut rien imposer, tout se discute. C'est un objet technique, le wiki, qui permet ce dialogue permanent. Et puis, constater que des ignorants peuvent construire de la connaissance en a étonné plus d'un !

Peut-on alors parler d'encyclopédie ?

Il y a eu des polémiques un peu stériles refusant cette dénomination. Evidemment, ce n'est pas comme un livre fermé et définitif, mais l'information y est beaucoup plus structurée qu'on ne le pense, finalement assez proche d'une encyclopédie classique. Pour l'anecdote, dans la version anglophone, en suivant à partir de n'importe quelle page le premier lien, on finit par tomber sur la page consacrée à la philosophie, c'est-à-dire la base de la connaissance. On aboutit toujours à des macroconcepts. Néanmoins, à cause du caractère changeant et ouvert des pages, certains préfèrent parler d'un encyclopédisme d'usage.

Comment Wikipédia a-t-elle évolué ?

Il y a eu plusieurs phases dans son histoire. D'abord une période d'incrémentation quantitative, entre 2001 et 2006, où le nombre d'articles a énormément augmenté. On parlait alors d'« effets piranha » pour montrer qu'une foule de contributeurs pouvait rapidement créer de nouveaux articles à partir d'une base d'une ou deux phrases sur n'importe quel sujet. Puis à partir de 2007, le site évolue vers plus de qualitatif en demandant aux contributeurs de multiplier les sources et références. C'est l'apparition d'une balise bien connue, « références nécessaires ».

A l'heure actuelle, nous sommes dans une phase où le magma se solidifie. Le nombre d'articles nouveaux a tendance à diminuer dans les Wikipédia les plus avancées. Quant au nombre de contributeurs réguliers, il se maintient mais ne progresse plus. C'est le temps de l'expertise. De plus en plus d'articles sont protégés en écriture, tandis que des débats ont lieu sur la pertinence de laisser des utilisateurs n'ayant pas créé de compte modifier les articles. Les contributions sans inscription deviennent de plus en plus marginales. Un nouveau palier qualitatif est recherché. Par ailleurs, il y a désormais tellement de règles que si on repartait sur ces bases, Wikipédia ne pourrait pas se développer comme elle a pu le faire dans ses premières années.

Faut-il recommander Wikipédia pour l'éducation ?

Bien sûr, même s'il y a des réticences. Etre prof est déjà difficile, et cet outil, en quelque sorte, dépossède l'enseignant de la connaissance. De plus, comme c'est un objet dynamique, il introduit l'idée que les savoirs évoluent, sont mouvants. Cela aussi est problématique. Pourtant, je défends la place de Wikipédia dans l'innovation pédagogique. D'ailleurs, la fondation Wikimédia développe Wikiuniversity, qui est une manière de jeter des ponts entre les deux mondes. Avec mes étudiants en master, depuis quatre ans, je propose de participer à l'élaboration de Wikipédia. Je m'efface le plus possible, passant du diseur de vérité au catalyseur d'apprentissage.

Mon rôle est de favoriser la collaboration dans le groupe, de faire en sorte qu'ils trouvent les réponses entre eux. Je leur montre aussi qu'on peut s'amuser en apprenant. Pour les évaluer, à la fin, j'étudie leur contribution à Wikipédia : création de pages, interaction avec la communauté, corrections, traductions... Au fond, pour moi, c'est le processus autour du fait qui compte, plus que le fait lui-même. La compétence acquise doit pouvoir être réutilisée dans un autre contexte, c'est avant tout un état d'esprit facilitant l'apprentissage. ■

PROPOS RECUEILLIS PAR D. L.

Boîtes à outils

Plusieurs sites permettent de se faire une idée de la richesse et du dynamisme de Wikipédia.

L'encyclopédie en chiffres Une seule adresse pour tout savoir du nombre de pages, de contributeurs actifs ou non, de la fréquentation dans chaque langue mais aussi de la taille des articles, du nombre de mots, d'images... Un autre site, Stats.grok.se, montre en outre les visites sur la page demandée. L'option est accessible également depuis la page d'un article dans l'onglet « historique », puis « statistique de consultation ».
Stats.wikimedia.org

Prendre le pouls de Wikipédia Aussi fascinant qu'inutile, Wikistream propose un défilement en temps réel de l'activité des éditeurs sur les pages du monde entier. L'utilisateur et le nombre de mots corrigés ou ajoutés sont indiqués. Sur les dernières vingt-quatre heures, le site Fr.wikiscan.org, plus statique, résume les activités des pages françaises.
Wikistream.wmflabs.org

L'exploration des concepts Grâce à l'université d'Osaka, il est possible (sur les pages en anglais) d'explorer différemment l'encyclopédie en visualisant les quelque 1,7 millions de concepts répertoriés et les 78 millions de liens entre eux.
Sigwp. org/en/index.php/Wikipedia_Thesaurus_Visualizer

Controverses et polémiques Contropedia, projet de recherche piloté par le Medialab de Sciences Po et financé par l'Union européenne, visualise les activités éditoriales sur des pages controversées, afin d'aider à comprendre la nature et le déroulement d'une polémique. Elle se fonde sur les multiples révisions et annulations permises par Wikipédia et archivées. Les mots les plus « attaqués » apparaissent en couleur. Une chronologie montre le rythme des changements. Et il est évidemment possible de retrouver chacune des modifications effectuées. Une centaine de pages ont été passées au crible (changement climatique, révolution tunisienne, les Beatles, la catastrophe de Fukushima...), mais d'autres peuvent être ajoutées en contactant les chercheurs.
www.contropedia.net

Bataille d'édition Comment se forme un consensus sur un article ? Notabilia offre quelques réponses esthétiques. L'activité d'une page s'y déploie comme la branche d'un lierre. Les « pour » la suppression d'une page font pencher la croissance d'un côté ; les « contre » de l'autre. En fonction de leur poids et de leur rythme d'intervention, la branche adopte plusieurs formes. Elle peut filer droit, traduisant la quasi parfaite alternance des opinions contraires et la lenteur du consensus. Ou elle se tord en spirale, sous l'effet de positions unanimes qui finissent par converger sur une position. Ou encore elle fait des « S », symbolisant une progression par vagues. Le projet a été développé notamment par Dario Taraborelli, qui est devenu ensuite directeur du département de recherche de la fondation Wikimédia.
Notabilia.net

Popularité sur courts Deux chercheurs des universités de Boston et de Harvard ont plongé dans le monde du tennis et dans Wikipédia pour en extraire les données de popularité et de performance de 500 joueurs. Leur visualisation permet de retrouver les rangs atteints durant les quinze dernières années et de les comparer à la popularité de leur page Wikipédia. La célébrité semble bien corrélée à la performance, sauf pour certains, plus « cliqués » que talentueux sur le terrain, comme Lleyton-Hewitt par exemple. Les chercheurs ont en outre développé un modèle pour prédire le nombre de visites d'une page en fonction des résultats sportifs.
Untangling-tennis.net