Wikimedia, les bibliothèques, et l'entraînement des IA

Matthieu Tarpin

Wikimedia, les bibliothèques, et l'entraînement des IA

WikiConvention francophone, du 31 octobre au 3 novembre 2024 à Québec (Canada)

Matthieu Tarpin

Élève conservateur des bibliothèques, DCB 33

Entre le 31 octobre et le 3 novembre 2024 s'est tenue à Québec la WikiConvention francophone, organisée par Wikimedia Canada. Pendant quatre jours, contributeur·ices aux différents projets wiki, bibliothécaires, archivistes, technicien·nes de la donnée, salarié·es et bénévoles de la Wikimedia Foundation ou encore représentant·es des pouvoirs publics de la francophonie ont pu échanger au cours de soixante et une conférences et une journée thématique dédiée aux Communs du numérique et à l'intelligence artificielle (IA). Presque tous les projets wiki ont fait l'objet d'interventions, des inévitables Wikipedia et Wikidata aux plus confidentiels Wikispecies ou Wikivoyage.

Si cette convention n'a pas manqué de discussions fécondes et de questionnements partagés, c'est bien l'IA qui s'est retrouvée au cœur du plus grand nombre de débats. La journée du 1er novembre lui a été entièrement dédiée, avec comme fil conducteur une question : comment la francophonie peut-elle tirer son épingle du jeu dans le contexte du développement des IA dans le monde, et en particulier dans le monde anglophone ?

Entraîner des IA avec Wikipedia

Les projets wiki représentent une masse de données considérables. Le Wikipedia anglophone compte à lui seul plus de 6,5 millions d'articles, pour 2,4 pour le Wikipédia francophone. Wikidata, de son côté, compte 114 millions d'items dûment répertoriés. Si Wikipedia était réputée peu fiable au début des années 2000, elle est devenue, de très loin, l'encyclopédie la plus lue au monde. Sa visibilité, ainsi que les efforts constants des bénévoles et salariés pour traquer les erreurs, les articles non sourcés, les tentatives de désinformation et les manipulations, en font aujourd'hui un outil plutôt fiable, et le point de départ de nombreuses recherches, y compris académiques.

En tant que telles, les bases de connaissances et de données que sont Wikipedia et Wikidata sont le terrain de jeu rêvé pour quiconque souhaite entraîner un algorithme. Les données qui y sont contenues sont rédigées en langage naturel, de bonne qualité, en évolution constante, et richement interconnectées. Dès lors, comme l'exprime Rémy Gerbet, directeur exécutif de Wikimédia France dans sa conférence, il est regrettable que Wikimédia en France ne soit pas plus proactive dans l'entraînement des IA. On sait que Wikipedia souffre de nombreux biais, qui reproduisent ceux de nos sociétés : biais de représentation linguistique, effacement des communautés marginalisées – particulièrement les communautés autochtones au Canada, problème dont sont très conscientes les institutions québécoises –, biais genré très important…

Entraîner des IA sur les données de Wikipedia, c'est prendre le risque de voir ces IA amplifier et perpétuer les biais de son modèle. Des projets sont en cours pour résorber le fossé de genre dans Wikipedia, où les articles ayant des femmes pour sujet sont moins de 15 % des pages biographiques. C'est le cas par exemple du projet « Les sans pagEs », qui vise à « rédiger et améliorer des pages biographiques sur des femmes, mais aussi sur les féminismes, le biais de genre ou d'autres sujets sous-représentés » selon la page Wikipedia du projet. Ce travail, outre son intérêt intrinsèque pour rendre l'encyclopédie plus représentative des réalités de nos sociétés, a donc toute sa place dans une stratégie wiki pour l'IA.

Liam Wyatt, senior manager des partenariats techniques à la Wikimedia Foundation, précise que la fondation n'a pas le choix de s'impliquer dans le développement des IA. Wikipedia et Wikidata ne peuvent pas se couper de leur existence : depuis leur conception, ces projets ont été pensés comme libres et ouverts à tous les usages, y compris commerciaux. Par conséquent, explique-t-il, la question n'est pas de savoir si Wikipedia doit participer à l'entraînement des IA, mais comment faire pour que, dans l'utilisation que celles-ci font des données issues de la communauté wiki, l'humain reste au centre. La Wikimedia Foundation entend se poser comme un maillon central de ce sujet,

interlocuteur privilégié à la fois des grandes institutions culturelles et des acteurs technologiques et commerciaux de l'IA. Le poids de la fondation lui permet de négocier un à un avec les GAFAM (acronyme de Google, Apple, Facebook, Amazon et Microsoft), mais aussi avec Open Ai et les autres acteurs majeurs de l'IA, et sa qualité de fondation à but non lucratif en fait un interlocuteur crédible pour les bibliothèques et les archives qui souhaitent s'impliquer dans l'ouverture des données.

Dans cette optique, la Wikimedia Foundation travaille actuellement à rendre ses services capables d'absorber la hausse de trafic de données que le siphonage massif par l'entraînement des IA implique, en proposant un forfait payant qui distinguerait l'utilisation humaine et l'utilisation robotisée de Wikipedia et Wikidata. La fondation espère ainsi encourager le développement de plus petits programmes indépendants et open source, tout en mettant au service du partage des connaissances leur gigantesque masse de données.

Le second objectif visé par la fondation est d'œuvrer à la transparence des IA : une utilisation des données des projets wiki devrait entraîner un travail de responsabilisation important, notamment pour les Large Language Model (LLM, en français grand modèle de langage), dont on attend qu'ils soient en mesure de sourcer sans hallucination l'origine des données citées. Cela n'est pour l'instant pas encore le cas, et un certain nombre de contributeurs à Wikipedia et Wikidata ont pu faire part de leur méfiance. En effet, le travail qui a été fourni depuis plusieurs décennies pour enrichir l'encyclopédie participative, tout bénévole qu'il soit, n'en est pas moins un travail, qui mérite d'être reconnu et valorisé en tant que tel. Les wikimédiens sont conscients de cette problématique, et la Wikimedia Foundation a entendu ces remarques.

Et les bibliothèques, dans tout ça ?

La WikiConvention francophone est un lieu où tous les acteurs des projets wiki peuvent échanger, mais elle est aussi un moment où les institutions culturelles, au premier rang desquelles les bibliothèques et les archives, sont invitées à débattre. Celles-ci n'ont pas fait défaut, puisque des représentant·es de la Bibliothèque nationale de France (BnF), de Bibliothèque et Archives nationales du Québec (BAnQ), ainsi que d'autres bibliothèques de la francophonie, étaient présent·es. La journée dédiée à l'IA et aux Communs du numérique leur a donné la parole pour discuter des stratégies de leurs établissements en matière d'innovation technologique, et de leurs éventuelles interactions avec les projets wiki.

Viriya Thach, responsable du secteur stratégie numérique et intelligence d'affaires à BAnQ, a souligné l'importance de travailler sur les LLM pour assurer que les futurs systèmes d'IA employés par les Québécois répondent à leurs besoins en matière de données francophones de qualité. BAnQ entend s'impliquer dans l'entraînement des IA, pour garantir que les données employées correspondent aux normes éthiques et environnementales des établissements publics québécois, et que les IA soient nourries par des données qui reflètent la diversité des cultures représentées au Québec. Il s'agit donc pour BAnQ de se positionner comme un partenaire important pour les GAFAM et les autres acteurs de l'IA, en leur proposant des données de bonne qualité et en quantité très importante, de manière à influencer positivement le développement des prochaines IA.

Madame Thach a par ailleurs insisté sur l'importance qu'une stratégie commune avec la France revêt pour BAnQ. Les IA sont un atout majeur pour la découvrabilité des contenus dont disposent les bibliothèques, mais ces dernières ne peuvent espérer peser dans les évolutions technologiques qu'en collaborant, y compris à l'international, pour faire valoir les intérêts des bibliothèques francophones. Dans un contexte où le Québec souhaite défendre sa seule langue nationale, le français, face à l'anglais, des données francophones de qualité sont un enjeu majeur. Hors de la seule francophonie, si le Québec veut représenter équitablement l'ensemble de ses communautés, il doit faciliter la sauvegarde des langues autochtones du Canada, et par conséquent faire un important travail de documentation et de sauvegarde des données dans ces langues, pour qu'elles soient découvrables pour les IA.

Jean-Philippe Moreux, chef de mission IA à la BnF, a rappelé que la question se posait de savoir dans quelle mesure ces nouveaux usages de l'IA sont – et seront – compatibles avec l'environnement réglementaire de l'Union européenne. L'augmentation significative du trafic de données qu'implique l'entraînement des IA sur les données des bibliothèques va représenter un coût qui n'est pas négligeable, et qui retombe pour l'instant sur les acteurs publics, tout en générant d'importants bénéfices pour les acteurs privés du secteur. L'ouverture massive des données des bibliothèques a des implications très complexes, puisque nos collections publiques se retrouvent valorisées par des acteurs privés, sans garantie suffisante sur la nature des réutilisations des données publiques à l'heure actuelle.

Monsieur Moreux a ainsi rappelé que, si la BnF s'implique activement dans le secteur des IA en suivant sa feuille de route 2021-2026, et si elle souhaite participer à la réutilisation et à l'accessibilité de ses jeux de données, en particulier dans le contexte de la francophonie, elle demeure consciente de l'impact financier et environnemental que l'IA risque d'avoir, et de ses implications pour le monde des bibliothèques.

Wikipedia, un intermédiaire pour l'ouverture des données des bibliothèques ?

Au terme de cette journée Communs du numérique, la question qui est posée aux bibliothèques est donc plutôt claire. Si Wikipedia et les projets de la Wikimedia Foundation se positionnent comme de fervents partisans de l'ouverture sans limite de leurs données pour l'entraînement des IA, ils souhaitent aussi être un interlocuteur privilégié des bibliothèques dans ce processus d'ouverture. Que ce soit par l'intervention de formateurs issus de la fondation, de wikimédiens en résidence, ou bien tout simplement par l'utilisation des jeux de données et métadonnées des bibliothèques pour enrichir les notices de Wikidata ou inversement, les bibliothèques et les projets wiki sont amenés à travailler en commun.

Il importe donc particulièrement que notre profession s'interroge sur ce qui nous lie au mouvement wiki, mais aussi sur ce qui nous en distingue. Wikipedia possède une visibilité et une force de frappe dont peu de bibliothèques peuvent se targuer ; elle reste le premier réflexe de nombre d'internautes, bien avant d'interroger leur bibliothèque favorite pour n'importe quelle recherche, professionnelle ou personnelle. Le mouvement wiki a toujours promu une culture participative, où chaque citoyen peut être un acteur du monde de la culture en ligne, et où toutes les informations sont accessibles à chacun·e, sans restriction. Sur tous ces points, les bibliothèques sont donc tout à fait alignées avec le mouvement wiki – et nombre d'entre elles ont depuis longtemps des liens avec les wikimédiens, liens que le nouveau Label Culture Libre proposé par Wikimedia France devrait renforcer. Ces partenariats sont précieux et féconds, et le mouvement wiki peut dans de nombreux cas être notre allié.

Cependant, l'ouverture proactive que la Wikimedia Foundation annonce avec les acteurs de l'IA doit nous faire réfléchir à ce qui nous distingue du mouvement Wiki. En effet, si le Label Culture Libre permet de valoriser les efforts des bibliothèques en faveur de l'ouverture des données, il est attribué entre autres aux établissements qui contribuent à Wikipedia, Wikidata ou Wikimedia Commons, et qui téléversent leurs contenus sur les plateformes wiki. Comme l'a fait remarquer Jean-Philippe Moreux, la question des retombées économiques de l'utilisation de nos données par les IA dans un contexte de réduction des budgets publics et de souveraineté numérique doit nous interroger. Le mouvement wiki, tout ouvert qu'il soit, n'est pas un service public et ne répond pas aux mêmes impératifs. Il peut très certainement être un allié fiable et un interlocuteur technique pertinent dans l'ouverture des données des bibliothèques, dans l'amélioration de la qualité des données et de la découvrabilité de nos collections, mais on ne saurait tracer un signe égal entre les intérêts de la Wikimedia Foundation et ceux des bibliothèques. Au terme de cette journée de discussions, la question reste donc ouverte, et il reviendra à l'appréciation de chacun de contribuer, à l'échelle de notre profession, à ce débat. ◉

Photo : Habib Mhenni. Disponible sur Wikimedia Commons en licence CC :
https://commons.wikimedia.org/wiki/File:Photo_de_groupe_WikiConvention_francophone_2024_par_Dyolf77,_ZVE05241.jpg