Pensé comme un outil de pilotage et de suivi, le Baromètre de la science ouverte utilise l'intelligence artificielle pour optimiser ses missions.

L'utilisation de l'apprentissage automatique dans le Baromètre de la science ouverte : *une façon de réconcilier bibliométrie et science ouverte ?*

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté Égalité Fraternité

Dès le lancement du Plan national pour la science ouverte (PNSO) en 2018, le Baromètre de la science ouverte (BSO) a été pensé comme un outil de suivi et de pilotage de politiques publiques. D'abord centré sur l'accès ouvert aux publications, le BSO a permis en quelques mois d'objectiver un « point de départ » du taux d'ouverture des publications françaises. Le BSO a vocation à élargir son périmètre, en s'intéressant à d'autres productions que les seules publications, et à approfondir ses analyses pour fournir des éléments d'aide à la compréhension et à la décision pour ses différents utilisateurs (décideurs au niveau national ou établissement, négociateurs, financeurs, chercheurs).

UNE ALLIANCE OBJECTIVE AVEC LES OUTILS D'APPRENTISSAGE AUTOMATIQUE

D'autres outils (internationaux notamment) de suivi de la science ouverte préexistaient au BSO, alors pourquoi en créer un nouveau ? D'abord, parce que les outils existants utilisent des données issues de bases propriétaires, ni partageables, ni réutilisables et introduisent des biais de couverture. De plus, le Ministère voulait disposer d'un outil souverain, adapté à ses propres besoins de suivi. Néanmoins, le pragmatisme conduit la plupart des institutions à se tourner vers les bases de données propriétaires. Notamment, les champs disciplinaires et les affiliations ne sont pas ouvertes en général (les affiliations sont absentes de Crossref dans plus de 75 % des cas en 2021). Impossible de suivre l'ouverture des publications sans analyser les tendances d'une discipline à l'autre qui sont le reflet de pratiques différentes entre les communautés. De même, il est nécessaire de connaître les pays d'affiliation : seules les publications avec une affiliation française sont analysées dans le BSO, encore faut-il savoir si une publication a une affiliation française !

Le manque de métadonnées ouvertes est parfois vécu comme un obstacle infranchissable, plaçant certaines institutions dans la situation de devoir avoir recours à des données propriétaires, dans l'attente de la mise en place, sur un temps plus long et incertain, d'infrastructures ouvertes et centralisées de métadonnées riches. Une troisième voie mérite d'être pensée, et l'apprentissage automatique s'avère être un outil déterminant dans sa mise en œuvre.

TRANSFORMER DES BASES EXISTANTES EN DONNÉES D'APPRENTISSAGE

L'apprentissage automatique recouvre de nombreuses méthodes. Mais un invariant demeure : l'apprentissage automatique utilise des données d'apprentissage (d'entraînement) pour construire un modèle, permettant ensuite d'enrichir de nouvelles données non rencontrées dans les données d'apprentissage. Il faut donc d'une part des données d'apprentissage suffisamment riches pour construire un modèle pertinent et d'autre part des données à enrichir grâce au modèle. Ces données à traiter doivent porter un minimum d'informations, sans quoi le modèle sera incapable de calculer quoi que ce soit d'utile.

Dans le cadre du BSO, la détection d'accès ouvert repose sur l'outil (ouvert) *Unpaywall*[1]. Les problématiques principales restantes portent sur l'inférence des champs disciplinaires et des pays d'affiliation. Nous faisons l'hypothèse qu'il est possible de déterminer une discipline à partir du titre du document et de la revue. Le module *scientific tagger*[2] utilise les bases PASCAL et FRANCIS[3] comme bases d'apprentissage. Le modèle est construit avec l'algorithme *fastText*[4] qui a le mérite d'être très léger et rapide. Une approche similaire est mise en place pour inférer la langue ainsi qu'une classification spécifique au domaine biomédical[5].

Pour les pays d'affiliation, l'obstacle à franchir est plus haut : les métadonnées ouvertes ne contiennent en général pas d'information sur les affiliations. Le problème ne porte pas sur les données d'entraînement mais bien sur les données à traiter. Sans données à traiter, l'algorithme se retrouve dans une impasse. Impasse qui semble pourtant paradoxale : les affiliations sont à la fois sous nos yeux à la première page des publications et invisibles dans les métadonnées ouvertes. Un outil de collecte et

[1] https://unpaywall.org
[2] https://github.com/dataesr/scientific_tagger
[3] https://pascal-francis.inist.fr
[4] https://fasttext.cc
[5] E. Jeangirard; Content-based subject classification at article level in biomedical context; 2021; hal-03212544
[6] https://github.com/dataesr/affiliation-matcher

Crédit Adobe stock

d'analyse des pages Web publiques des publications a été développé. Il en extrait les affiliations plein texte. Le module *affiliation-matcher*[6] permet à partir d'une affiliation plein texte (Université de Paris Dauphine, France) de deviner le pays associé (France). Cela peut être plus subtil qu'il n'y paraît dans cet exemple. Ainsi, la présence du mot « France » dans l'affiliation n'est ni nécessaire (CERMICS Université Paris Est), ni suffisante (Hôtel Dieu de France, Beirut, Lebanon).
Le module *affiliation-matcher* s'appuie sur des données référentielles (notamment le RNSR et ROR[7]) qui jouent le rôle de données d'apprentissage. L'algorithme en place ne relève néanmoins pas entièrement de l'apprentissage automatique car les règles d'appariement ne sont pas décidées par la machine, mais contrôlées par l'utilisateur du module.

L'EXTENSION DU BSO AUX DONNÉES DE LA RECHERCHE ET CODES LOGICIELS AMÈNE À L'UTILISATION DE NOUVEAUX OUTILS

Le deuxième Plan national pour la science ouverte (PNSO2)[8] fixe comme objectif au BSO de proposer de nouveaux indicateurs de suivi au-delà des publications. Dans sa déclinaison santé, le BSO analyse notamment les essais cliniques. À présent, nous travaillons à l'analyse des données de la recherche et des codes logiciels. Une piste suivie consiste à tenter de repérer dans le texte des publications, les références aux logiciels et aux données de la recherche. Il faut donc avoir accès au *full-text* des publications, et disposer d'un moyen pour y repérer une mention de logiciel ou de jeu de données. C'est un niveau supplémentaire de complexité. L'accès aux *full-text* est encore très difficile (hors accès ouvert) malgré l'existence d'accords TDM (*Text and Data mining*) dans les contrats avec certains éditeurs et les dispositions liées à la fouille de texte dans un décret[9] récent. De plus, ce type d'outil de détection fait appel à des techniques dites « d'apprentissage profond » (*deep learning*). Nous travaillons avec Patrice Lopez (*science-miner*[10]), un des experts internationaux de l'utilisation des techniques d'apprentissage profond sur les textes scientifiques.

NE PAS LÂCHER LA « PROIE » DU RÉEL POUR « L'OMBRE DES VÉRITÉS ALGORITHMIQUES »[11]

Quelles que soient les données et les techniques utilisées, les algorithmes d'apprentissage automatique produisent des erreurs, que nous tentons de contrôler et de mesurer. Une étude récente de Lauranne Chaignon et Daniel Egret[12] a validé l'efficacité de la méthode de détection des affiliations françaises du BSO en menant une comparaison avec les principales bases bibliométriques.
L'utilisation de l'apprentissage automatique reste un moyen fiable de pallier le manque de métadonnées ouvertes et riches. Des réseaux d'échanges d'informations peuvent aussi se structurer, comme nous le proposons avec les déclinaisons locales du BSO[13].
Le développement des outils avancés d'enrichissement automatique ne doit donc pas nous faire oublier un objectif essentiel, la poursuite de la mise en place d'un réseau d'infrastructures ouvertes, proposant des métadonnées riches pour rendre à la communauté scientifique le contrôle de l'information qu'elle produit elle-même.

Éric Jeangirard
Data Scientist, Département des outils d'aide à la décision, SIES – MESR
eric.jeangirard@recherche.gouv.fr

[7] https://ror.org

[8] www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte

[9] www.legifrance.gouv.fr/jorf/id/JORFTEXT000045960058

[10] https://science-miner.com

[11] R. Gori, 2022, La Fabrique de nos servitudes

[12] L Chaignon, D Egret; Identifying scientific publications countrywide and measuring their open access: The case of the French Open Science Barometer (BSO). Quantitative Science Studies 2022; doi:10.1162/qss_a_00179

[13] https://barometredelascienceouverte.esr.gouv.fr/about/declinaisons