COMMUNIQUÉ DE PRESSE

Paris, le 31 août 2017

# YAGO, l'une des 1ères grande bases publiques de connaissance, révèle son code et s'ouvre aux contributions des développeurs

**Yago constitue l'une des premières grandes bases de connaissance publiques. Développée par 7 scientifiques du Max Planck Institute for Informatics (Sarrebruck, Allemagne) et de Télécom ParisTech (Paris, France), la base est aujourd'hui le fruit d'une collaboration entre ces deux institutions et Ambiverse, une spin-off du Max Planck. Yago inclut des informations de sources variées dans un format lisible par l'ordinateur. Le recours à l'intelligence artificielle a permis d'élargir le champ de ses usages en améliorant efficacité et mode de fonctionnement pour diverses applications. Désormais, cette très grande base de données d'accès gratuit devient open source et rend public son code source, afin d'amplifier encore son développement collaboratif. En août 2017, les chercheurs de l'équipe ont aussi gagné pour leurs travaux un prix du *Artificial Intelligence journal*, le plus important journal dans le domaine de l'intelligence artificielle.**

Quiconque est coutumier des recherches sur Internet a pu faire le constat suivant : les mots ont presque tous plusieurs significations. Selon le sens qu'on a en tête, on va s'attendre à différents résultats de recherche. Les moteurs de recherche dits « intelligents » résolvent ce problème en faisant appel à des bases de connaissance. Celles-ci comprennent des films, des entreprises, des personnalités, des produits et beaucoup d'autres informations. Développée par les scientifiques du Max Planck Institute for Informatics et de Télécom ParisTech, Yago a dès ses débuts, en 2007, été mise à disposition gratuitement.

« Si vous entrez dans Google le terme "Thales" par exemple, cela ne représente qu'une succession de lettres pour le moteur de recherche » explique le professeur Gerhard Weikum, Directeur scientifique au Max Planck Institute for Informatics à Sarrebruck. « Une base de connaissance va permettre de relier cet ensemble de lettres à plusieurs significations possibles, comme ici pour notre exemple à "Thales Group", la multinationale française d'équipements électronique, ou à "Thales de Milet", le philosophe et savant grec. » De nos jours, il est difficile d'imaginer les moteurs de recherche se passer de ce genre de connaissances de fond, de contexte. D'ailleurs, c'est grâce à l'utilisation des bases de connaissance que Google peut afficher en résultats de cette requête à la fois les cours en bourse, le logo et le PDG de Thales Group, en plus des résultats de recherche classique.

Les projets de recherche académique ont été pionniers dans ce domaine et ont permis aux bases de connaissance de voir le jour : parmi eux Yago tout particulièrement et, un peu plus tard, DBpedia[i]. A l'origine sujet de thèse de Fabian Suchanek au Max Planck Institute, Yago est aujourd'hui le fruit d'une collaboration avec Télécom ParisTech (où Fabian Suchanek est devenu professeur) et Ambiverse, une spin-off du Max Planck. Yago intègre les informations de Wikipédia et d'autres sources, afin de leur donner de « l'intelligence » et de les contextualiser. Par exemple, le système sait ainsi placer le siège de Thales Group à Neuilly-sur-Seine et le lieu de naissance de Thales de Milet en Turquie.

Beaucoup d'applications relevant de divers secteurs industriels font appel à l'intelligence artificielle afin d'améliorer leur efficacité et, surtout, leur mode de fonctionnement. Ainsi, Yago a vu le champ de ses usages s'élargir. Les applications intelligentes peuvent, avec l'aide de Yago, effectuer des recherches dans plusieurs langues tout en répertoriant des faits à la fois spatialement et temporellement, ce qui rend possible la requête suivante : « Recherche tous les scientifiques ayant vécu au XXe, nés dans la région du Grand Paris et ayant reçu un prix Nobel ». Par exemple Primal.com, start-up canadienne, est une application dans laquelle Yago aide des entreprises à mieux comprendre les intérêts de leurs clients et à formuler des recommandations de contenus et de produits qui vont répondre à leurs besoins spécifiques. L'utilisation la plus connue de Yago ces dernières années fut lorsqu'IBM l'intégra au système d'intelligence artificielle Watson, qui remporta le jeu télévisé « Jeopardy! » en 2011.

Mais Ambiverse, spin-off du Max Plack Institute, va plus loin : elle s'est servie de Yago pour analyser les fameux « Panama Papers ». Il a fallu seulement quelques heures à Ambiverse pour découvrir de nouveaux éléments sur les propriétaires des comptes au Panama, une tâche qui aurait autrement requis un effort manuel considérable.

Une telle analyse fut rendue possible par le fait que Yago catégorise tout individu dans une structure sémantique. Jusqu'à présent, les ordinateurs ont stocké de grandes quantités de données, sans être capables cependant de les classer et a fortiori de les comprendre. La structure de Yago change la donne en permettant à l'ordinateur de pouvoir distinguer, par exemple, « Gerd Müller », le champion du monde de football en 1974, de « Gerd Müller », le ministre allemand de la Coopération économique et du Développement. Johannes Hoffart, Directeur général d'Ambiverse, ajoute : « Yago attribue des personnes à des contextes, il est donc facile de déterminer si plus d'athlètes ou plus de politiciens possèdent des comptes au Panama ». De telles structures étaient auparavant élaborées manuellement, ce qui constitue en fait une tâche très complexe, en termes de production comme de vérification.

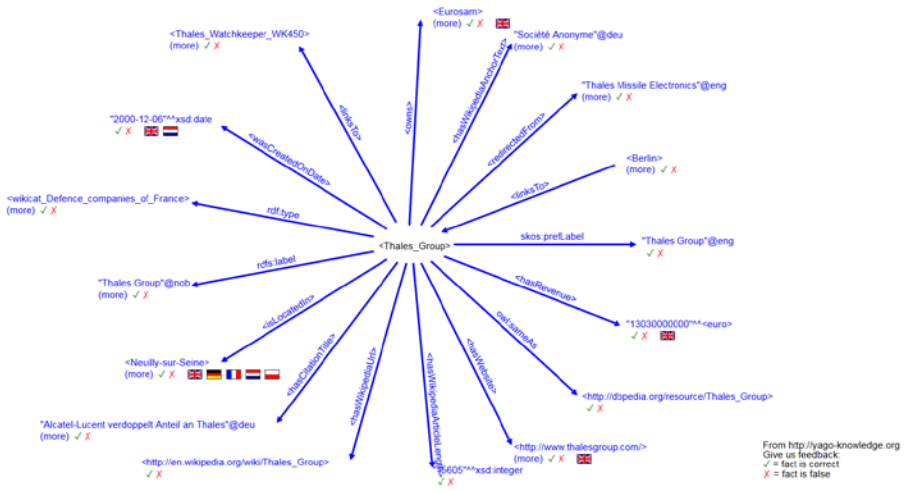
La procédure mise au point pour Yago permet de contourner astucieusement cette tâche fastidieuse. La base de connaissance va de manière systématique puiser dans le réservoir de connaissances de Wikipédia. Non seulement l'information indiquant si telle personne est un(e) athlète ou un(e) politicien(ne), mais également la ou les relations existant entre les deux, vont se présenter dans un format lisible par la machine. Par conséquent, le lien « se trouve à » connecte Thales Group à Neuilly-sur-Seine. Comme chaque page Wikipédia est une entité dans Yago, les chercheurs peuvent fournir à la base de connaissance près de 17 millions d'entités et 150 millions de relations entre celles-ci.

Les chercheurs révèlent à présent le code source de leur base de connaissance sur la plateforme GitHub, sous la licence open source GNU GPL v3. Cette licence d'utilisation du logiciel garantit à chacun le droit d'utiliser, d'étudier, de modifier et de partager le code programme protégé. « La communauté des développeurs va disposer d'une base de connaissance de haute qualité, » conclut Fabian Suchanek, chercheur à l'origine du projet. « Nous espérons non seulement voir apparaître de nouveaux usages pour Yago, mais nous attendons avec un vif intérêt les contributions des développeurs. »

Pour en savoir plus sur le projet et télécharger le code source :

www.yago-knowledge.org

*Grâce à la génération automatique d'entités et des liens entre elles/et des relations qui les lient, les ordinateurs sont en mesure de répondre à des requêtes complexes.*




À PROPOS DE TÉLÉCOM PARISTECH - www.telecom-paristech.fr

Télécom ParisTech forme à innover et entreprendre dans un monde devenu numérique. Ses enseignements et sa recherche intègrent toutes les disciplines des sciences et technologies de l'information et de la communication avec un ancrage sociétal fort, leur permettant de relever les défis majeurs du 21e siècle. Ses cursus diplôment ingénieurs, docteurs et professionnels et attirent 55 % d'étudiants étrangers. Sa recherche présente une expertise internationale, originale et pluridisciplinaire, sur six axes stratégiques : Big Data, Très Grands Réseaux & Systèmes, Confiance Numérique, Design-Interaction-Perception, Modélisation pour le Numérique, Innovation Numérique. École de l'IMT (Institut Mines-Télécom), Télécom ParisTech est membre fondateur du réseau ParisTech et se positionne comme le collège de l'innovation par le numérique de Paris-Saclay, dont l'ambition est de devenir l'un des premiers pôles d'innovation mondiaux.

À PROPOS DE L'INSTITUT MAX PLANCK POUR L'INFORMATIQUE - www.mpi-inf.mpg.de

L'Institut Max Planck pour l'Informatique, situé à Sarrebruck en Allemagne, a été fondé en 1990. Sa mission est le développement des fondations pour les systèmes informatiques efficaces et robustes, y compris les algorithmes ainsi que les applications. L'institut possède 5 départements : Algorithms

and Complexity, Computer Vision and Multimodal Computing, Computational Biology and Applied Algorithmics, Computer Graphics, et Databases and Information Systems. L'institut comprend 200 scientifiques, dont 120 étudiants en thèse.

L'institut appartient à la Société Max Planck, qui comprend 85 instituts dédiés aux sciences fondamentales. La société est réputée pour sa liste de 33 lauréats de prix Nobel et se place parmi le top 5 des instituts de recherche au monde en termes de production scientifique.

## CONTACTS PRESSE


Dominique Célier
01 45 81 75 17 • 06 87 11 95 90
dominique.celier@telecom-paristech.fr

Stéphane Menegaldo
01 45 81 70 95 • 07 82 75 17 30
stephane.menegaldo@telecom-paristech.fr


---

[i] Autre base de connaissance, DBpedia est un projet d'exploration et d'extraction automatiques des données de Wikipédia pour en proposer une version structurée au format du web sémantique.