10 ans d'ISIDORE

Pouyllau Stéphane, Minel Jean-Luc, Capelli Laurent, Bunel Mélanie, Sauret Nicolas, Capelli Laurent, Busonera Pauline, Desseigne Adrien, Baude Olivier, Jouguet Hélène

2021/10/19

Mot-clés : isidore, moteur de recherche, outil de découverte, réseau social académique, topic modelling, latent dirichlet analysis, réseaux de neurones

Keywords: search engine, discovery tool, academic social network, isidore, topic modelling, latent dirichlet analysis, neural networks

ISIDORE a 10 ans !

En 2011 paraissait la 1ère version d'ISIDORE. Son 10ème anniversaire est l'occasion de rappeller l'historique du projet et de présenter ses futures grandes lignes en cours de définition dans le cadre du programme "Huma-Num Science Ouverte" [1].

Cette année ISIDORE a franchi les 10 millions de documents et de données indexés, enrichis et catégorisés venant de plus de 9000 bases et entrepôts de données du monde entier. Déjà plus de 2000 chercheur · e · s y possèdent un compte, profitant ainsi des nombreuses fonctionnalités de veille scientifique et de découverte personnalisables.

Au fait, qu'est-ce qu'ISIDORE ?

Initialement, ISIDORE est un moteur de recherche permettant de découvrir et de trouver des publications, des données numériques et profils de chercheur · e · s en sciences humaines et sociales (SHS) venant du monde entier.

Il permet de rechercher dans le texte intégral de plusieurs millions de documents (articles, thèses et mémoires, rapports, jeux de données, pages Web, notices

1. Programme financé par le fond national pour la science ouverte du ministère de la recherche, de l'enseignement-supérieur et de l'innovation.

de bases de données, description de fonds d'archives, etc.), des signalements événements (séminaires, colloques, etc.). De plus, ISIDORE relie entre eux ces millions de documents en les enrichissant à l'aide de concepts scientifiques issus des travaux des communautés de recherche des SHS.

Il est accessible sur le Web sur le portail isidore.science.

Il propose également des fonctionnalités de réseau social scientifique. À ce titre, il entre dans la catégorie des moteurs et assistants de recherche et offre de nombreuses fonctionnalités pour organiser de la veille scientifique.

Principes fondateurs

ISIDORE moissonne, c'est le terme consacré, des entrepôts de données et de publications afin de récupérer des métadonnées textuelles et du texte intégral, les enrichir et les indexer à l'aide d'un moteur de recherche. L'exploitation des métadonnées des documents ainsi que du texte intégral, permet d'analyser ces informations afin de les enrichir en les reliant d'une part à des concepts de référentiels scientifiques (thésaurus, vocabulaires scientifiques, etc.) édités soit par les communautés scientifiques des SHS elles-mêmes, soit par les grandes bibliothèques de recherche internationales, et d'autre part à des identifiants d'auteurs (ORCID, IDRef, IdHAL, VIAF, etc.). Le choix d'utiliser en même temps, de façon alignée, les référentiels issus des grands instruments que sont les bibliothèques nationales et ceux élaborés dans les équipes de recherche SHS permet d'asseoir ISIDORE au cœur des dispositifs socio-techniques de recherche d'information car la matière le composant est créée directement par les communautés de recherche SHS.

ISIDORE procède, et c'est un de ses points forts, à plusieurs types de traitements sémantiques :

- L'annotation ou l'enrichissement sémantique : les termes présents dans les métadonnées des documents, le plus souvent issus des mots-clés auteurs, du titre et du résumé, sont comparés aux entrées des référentiels par le biais d'un algorithme fondé sur une analyse morphologique des termes. Si une équivalence s'effectue entre un terme issu du document et une entrée de l'un des référentiels, alors la ressource sera reliée à ladite entrée du référentiel. Les référentiels sont multilingues et alignés entre eux. Ainsi, l'annotation sémantique est multilingue. Aujourd'hui, ISIDORE enrichit dans trois langues : anglais, français et espagnol permettant de couvrir les bassins linguistiques anglophone, francophone et hispanophone.
- La catégorisation disciplinaire : ISIDORE utilise un classifieur sémantique qui, après avoir été entrainé sur un corpus de référence, catégorise dans les 27 disciplines SHS du référentiel MORESS (aussi utilisé par HAL), tous les documents présents dans ISIDORE. L'entrainement du classifieur est réalisé à l'aide soit d'une catégorisation manuelle réalisée par les chercheurs dans HAL-SHS lors du dépôt de leurs publications, soit

par catégorisation manuelle d'articles confié à nos collègues de l'INIST-CNRS (afin d'équilibrer l'entrainement sur les trois langues d'ISIDORE).
— La détection des auteurs : ISIDORE détecte les auteurs des documents et enrichit la forme auteur (prénom et nom) à l'aide d'identifiants auteurs internationaux (ORCID, VIAF, ISNI) et nationaux (IdHAL, IDRef).

Dans ce cadre, il est tout à fait possible, pour une communauté de recherche des SHS, de proposer des référentiels, des données, des bases de données pouvant enrichir l'ensemble des documents et données présents dans ISIDORE.

Au-delà du portail Web isidore.science, bien connu maintenant des communautés de recherche, ISIDORE propose plusieurs type d'accès :

— À destination des développeurs, des API (moteur de recherche, référentiels, suggestions)
— À destination des chercheur · e · s et du monde des bibliothèques, de la documentation et des archives, une publication des métadonnées enrichies dans le *Linked Open Data* selon les principes du Web sémantique (RDF) et au travers d'un *SPARQL end point*.

Par ailleurs, ISIDORE est utilisé par d'autres portails et sites qui exploitent certaines "briques logicielles" le composant. Dès 2017, dans le cadre d'un hackathon avec les collègues canadiens d'Erudit est né le dispositif *ISIDORE à la demande*. Ce dispositif, modulaire, irrigue aujourd'hui d'autres portails tels que DARIAH (avec par exemple l'expérience du *Parallel Semantic Search Engine*), STYLO ou encore des outils de découverte de bibliothèques.

Aujourd'hui, ISIDORE est un écosystème offrant des outils de recherche pour les chercheur · e · s, les doctorants, les documentalistes, bibliothècaire et archivistes, les développeurs et *data-scientists*.

Un peu d'histoire

Les principales étapes

Conçu en 2009, la réalisation d'ISIDORE a été pilotée par l'équipe du Très grand équipement (TGE) Adonis (Maignien 2011) opéré par le CNRS avec l'assistance par de la société Atos Consulting. ISIDORE a été réalisé entre octobre 2009 et décembre 2010 [2] par les équipes du Centre pour la communication scientifique directe (CCSD) avec l'implication des sociétés Antidot SA, Sword et Mondéca. ISIDORE a été lancé en version "béta" (Pouyllau 2011) le 8 décembre 2010, lors des rencontres de l'infrastructure Adonis à Valpré (France) et en version complète le 4 avril 2011. Il est actuellement développé et exploité par les équipes de l'infrastructure Huma-Num qui améliorent en permanence les différentes parties et briques technologiques qui le composent.

2. ISIDORE a bénéficié du plan de relance gouvernemental de 2009-2010.

Dès 2015, ISIDORE a été repensé une première fois afin de proposer des enrichissements en 3 langues : français, anglais et espagnol en utilisant la puissance du Web sémantique et l'alignement terminologique entre ces 3 langues sur des milliers de concepts. Cela était déjà effectué depuis plusieurs décennies par les bibliothèques nationales et les scientifiques, dans les grands référentiels internationaux comme Rameau (BnF), LCSH (*Library of Congress*, USA), ou BNE (*Biblioteca National de España*), mais aussi dans les référentiels scientifiques produits par les communautés scientifiques SHS internationales.

En 2018, une importante mise à jour des interfaces a été réalisée afin d'internationaliser ISIDORE grâce à l'implémentation d'une interface multilingue et d'ajouter de nombreuses fonctionnalités destinées aux utilisateurs. Conçue avec l'aide directe de chercheur · e · s SHS (historien · ne · s, sociologues, géographes, linguistes, etc.) et des designers de la société Atelier Universel, cette nouvelle interface, fondée sur les principes du réseau social entre chercheur · e · s et d'outil de veille scientifique, a permis de créer un nouveau site Web, isidore.science qui s'inscrit aujourd'hui dans une diffusion internationale comme, par exemple, dans les catalogues de l'*European Open Science Cloud* et *Re3Data*.

Dix années d'apprentissage

En 2009, la conception et la réalisation d'ISIDORE reposaient sur quelques hypothèses et convictions des fondateurs(Poupeau 2016).

Il s'agissait, dans un premier temps, d'offrir un portail d'accès unifié à l'ensemble des publications scientifiques et données produites par la communauté des chercheur · e · s et des enseignants-chercheur · e · s en SHS. En effet, en 2009, rechercher une publication nécessitait d'accéder à une dizaine de portails (HAL-SHS, Persée, Cairn, rien que pour la france, etc.), chacun offrant une interface de recherche spécifique.

Il s'agissait ensuite d'exploiter les possibilités offertes par l'utilisation de thésaurus disciplinaires, pour offrir des fonctionnalités de recherche plus puissantes que celles offertes à cette époque par les moteurs de recherche existants (Google Scholar notamment). C'est ce qui est appelé enrichissement sémantique dans le langage "isidorien" (cf. ci-dessus). Cette exploitation se concrétisait notamment par l'affichage dans l'interface Web des termes génériques proposés par les différents thésaurus.

Il y avait également une volonté d'exploiter les possibilités offertes par l'apprentissage supervisé (*machine learning*) pour catégoriser automatiquement les documents. Cette technique d'apprentissage exige une base d'entrainement (le *training set*) composée de plusieurs centaines d'articles pour chaque catégorie afin de construire le classifieur (principalement à l'aide de HAL-SHS et de collaborations avec l'INIST-CNRS à partir de 2015 pour l'anglais et l'espagnol).

Enfin, il s'agissait de contribuer à l'interopérabilité entre les différents entrepôts

de publications scientifiques en explicitant le modèles de données (l'ontologie) et en le formalisant dans les langages du Web sémantique (RDF, RDFS, SKOS, OWL).

Ces convictions et hypothèses ont été confrontées, en dix ans, à la critique implacable des usages. Il est apparu, au fil des discussions avec des panels d'utilisateurs, que l'affichage des enrichissements montrait des limites et pouvait être source d'erreurs. L'enrichissement sémantique et la catégorisation automatique offrent des outils d'aide à la recherche particulièrement puissants, mais leur visualisation devait être repensée, notamment en exploitant mieux le multilinguisme, premier enseignement qui a été mis en œuvre dans la version actuelle de ISIDORE.

Il est apparu que l'évolution rapide des thèmes de recherche nécessite des mises à jour très régulières des référentiels émanant des communautés de recherche SHS et des mises à jour de la base d'entrainement du classifieur. Ces tâches, extrêmement chronophages, n'ont pas pu être réalisées régulièrement. Un comptage via une requête SPARQL nous indique au 19 juillet 2021 que sur 3 381 711 documents de type article présents dans ISIDORE, seuls 2 215 744 documents de type article sont catégorisés avec une confiance de plus de 70%, soit presque 35% d'articles non catégorisés. Ce deuxième enseignement nous invite à mener une réflexion sur les nouvelles technologies disponibles pour surmonter cette difficulté.

L'interopérabilité, via les API et le Sparql endpoint, n'a pas donné lieu à des exploitations croisées avec d'autres dépôts, en particulier à l'échelle européenne, alors que d'autres institutions, comme par exemple la BnF, ou encore des projets tels que Biblissima, semblent avoir mieux valorisé ces fonctionnalités.

Les principes du FAIR sont venus conforter les choix initiaux (le F, A et I) mais aussi questionner la séparation entre ISIDORE et NAKALA, ce dernier étant en charge de conserver les données de la recherche (le R). Aujourd'hui, il convient de repenser les liens entre ISIDORE et NAKALA : c'est le cœur du programme Huma-Num Science Ouverte actuellement en cours de réalisation.

Vers demain...

La multiplication des moteurs de recherche et de découverte depuis le début des années 2000 a complexifié (Pouyllau S. 2020), le paysage de la recherche d'information (Chaudiron 2008) pour les chercheur · e · s. ISIDORE, par son inclusion en 2013 dans l'écosystème d'Huma-Num, profite cependant, autour de lui, de la mise en œuvre d'une intégration de services numériques (authentification centralisée, entrepôts de données créés avec NAKALA et moissonnables, etc.). Mais si les interfaces Web ont fleuri, les fonctionalités de recherche, qui s'appuyent sur la qualité des données indexées, sur la capacité du dispositif sociotechnique à les analyser, les regrouper, les trier et les relier, sont restées le plus souvent dans leurs états initiaux. La création de chaînes de traitement en *Machine Learning* et

surtout leurs exploitation, utilisant comme ISIDORE un apprentissage entrainé par référentiel, nécessitent des cycles de vie rapides et souvent complexes sur le plan des choix et de la validation scientifique des référentiels d'enrichissement. Cependant, le développement de nouvelles technologies, comme l'apprentissage profond (*Deep Learning*) ou le (*Topic Modelling*) ouvrent de nouveaux horizons.

ISIDORE dans l'éco-système d'Huma-Num : de l'amélioration continue au programme HNSO

Depuis sa création, ISIDORE a évolué régulièrement et plus particulièrement au niveau de ses chaines d'enrichissements sémantiques et ses briques de traitement de l'information. Dès 2015, avec l'ajout de l'annotation, de la catégorisation et des enrichissements en anglais et espagnol, les chaines de traitement furent largement mises à jour, reprises et retravaillées. Si l'année 2018 fut consacrée à l'évolution des interfaces et à l'ajout de fonctionnalités de réseau social scientifique, depuis 2019, les équipes d'Huma-Num travaillent sur plusieurs améliorations qui sont en cours de réalisation. Nous proposons ici d'en présenter deux, qui sont en cours de finalisation.

Gérer ses données dans ISIDORE et l'arrivée du temps réel dans l'indexation des données

Créé avec un nombre limité de sources à moissonner et à indexer, ISIDORE s'est fortement développé en 10 ans et la question de mise à jour en temps réel s'est très vite posée. Les moyens limités du début, les difficultés pour stabiliser l'équipe et d'autres choix tel que le multilinguisme dans l'annotation et la catégorisation ont retardé fortement l'instruction de cette fonctionalité. Depuis quelques mois, l'équipe d'Huma-Num, et en particulier son pôle ACCES, qui opère les développements du code et des briques qui composent ISIDORE, avec comme partenaire industriel Antidot, ont travaillé sur la mise en œuvre d'un moissonnage à périodicité rapide : la mise à jour d'ISIDORE se fera quotidiennement et non plus mensuellement. Le cahier des charges est complexe car ISIDORE ne se limitant pas à l'indexation de métadonnées et de texte intégral, c'est toute la chaine de traitement qu'il a fallu reprendre. Dans cette évolution d'ISIDORE, une nouvelle base de données, contenant les sources — c'est à dire les bases de données à collecter et moissonner (en OAI-PMH ou à l'aide du couple Sitemap/XML et expression en RDFa de métadonnées) a été imaginé par le pôle ACCES. Elle donnera la possibilité aux fournisseurs de données d'éditer leurs bases, d'ajouter des points de collecte, etc.

Mieux connaître les auteur · e · s pour mieux découvrir leurs travaux et nouer des collaborations

Une autre amélioration en cours est la détection des auteur · e · s des documents et le traitement effectué dessus. L'identification des auteur · e · s est une fonction-

nalité majeure depuis qu'ISIDORE a été doté, en 2018, d'une couche de réseau social académique. En effet, suivre des auteur · e · s, découvrir les revues dans lesquelles ils/elles publient, les travaux qu'ils/elles déposent dans des archives ouvertes, etc. passe par la fiabilisation de la détection des auteur · e · s et par les capacités d'ISIDORE à, par exemple, désambiguïser les homonymes, associer les bons identifiants (IdRef, Wikidata, BnF, ORCID, VIAF, ISNI, etc.) aux bonnes personnes. Au-delà, il s'agit avec ces matériaux de proposer de nouvelles fonctionnalités comme la suggestion de mise en contact entre des personnes travaillant sur des sujets proches mais appartenant à des pays, ou des écoles de pensées différentes ou encore la détection de thématique de recherches émergentes ou d'experts d'un domaine très spécialisé (Pouyllau S. 2020). C'est avec le partenaire industriel Antidot, qu'a été mis au point une nouvelle chaine de traitement, qui permet d'analyser les formes auteurs, par utilisation de techniques en *machine learning*. Il s'agit d'un travail difficile, méticuleux, qui nécessitera aussi l'inclusion de tests avec des panels de chercheurs pour améliorer en continu, sur le long terme, ce type de dispositif.

ISIDORE au cœur d'Huma-Num, les travaux menés dans le cadre du programme "HNSO"

Plus largement, Huma-Num s'est engagée, avec le programme "Huma-Num Science Ouverte" financé par le fond national pour la science ouverte, dans le développement d'un vaste programme d'amélioration continue des deux briques qui en constitue son cœur : ISIDORE et NAKALA. Dans ce cadre, un certains nombre de ponts ont été lancés entre ISIDORE et NAKALA, en particulier dans l'utilisation des référentiels et thésaurus fournis par les communautés SHS internationales. Utiliser les référentiels d'ISIDORE pour aider les chercheur · e · s à indexer plus efficacement leurs jeux de données dans NAKALA et utiliser le texte intégral des jeux de données déposés dans NAKALA pour améliorer la classification des documents dans ISIDORE, sont au cœur des travaux menés en ce moment et jusqu'en 2024 dans le cadre du programme "Huma-Num Science Ouverte". Au-delà des interconnexions, qui relèvent des aspects informatiques du développement, c'est aussi l'ensemble de la chaine documentaire qui doit être repensée en particulier au niveau du cycle de vie (mises à jour régulières, etc.) des référentiels scientifiques et de l'ontologie des types d'ISIDORE. Ce travail est engagé dans le cadre de ce programme. Avec le développement conjoint d'HumanID, le "hub" d'authentification et d'accès aux services Web proposé par Huma-Num, c'est vers une intégration de services en mode "sans couture" pour les chercheur · e · s qui est en cours de création à Huma-Num et qui préfigure ce qui sera sans doute au centre du vaste projet *COMMONS*, développé en coopération avec les infrastructures OpenEdition et Métopes.

Au-delà des briques logicielles et des méthodes, HNSO s'intéresse, pour l'avenir d'ISIDORE, aux évolutions des pratiques de recherche d'information des chercheur · e · s. C'est l'une des missions du HN Lab dont les activités, gravitant autour de 4 axes de recherche, alimentent en réflexions et en études le

programme HNSO et donc le développement d'ISIDORE.

Vers de nouvelles méthodes pour l'enrichissement sémantique des contenus : les réseaux de neurones ou le *Topic Modelling*

En 10 ans, les techniques de classifications et d'annotations sémantiques ont largement progressé. Deux approches permettent d'envisager une catégorisation plus efficiente et moins chronophage en temps de traitement. Tout d'abord, les récents progrès réalisés ces cinq dernières années dans le développement d'architectures de *Deep Learning* fondées sur les réseaux de neurones (CNN, LSTM, Tranformers) et la construction de ressources langagières à large couverture (BERT et DocBERT (Adhikari A. 2019), FLAUBERT, CAMEMBERT, etc) permettent d'obtenir des catégorisations avec un score F1 proche des 91%. Néanmoins, ces approches nécessitent toujours des corpus d'entrainement qui peuvent être complexes a produire. Une autre approche, le *Topic Modelling* ou *Latent Dirichlet Analysis* (LDA), qui repose sur la construction d'un modèle probabiliste (Blei D. 2001) ne nécessite pas de corpus d'entrainement. En revanche, les catégories affectées aux articles (les *topics*) ne peuvent pas être prédéfinies. L'avantages de ces méthodes — pour un outil ancré dans des communautés de recherche pouvant travailler sur des terrains et objets très différent — est de mieux épouser les *topics* produits par les écoles de pensée. De nos jours, l'émergence de ces *topics*, que l'on peut voir comme des "concepts à potentielle courte vie", est accélérée par la diffusion très rapide d'articles et d'ouvrages en libre accès. Ces *topics* pourraient être rapprochés, reliés même, aux référentiels situés produits par les communautés, aux référentiels des bibliothèques nationales déjà présents depuis 2010 dans ISIDORE, voir à utiliser des définitions scientifiques venant de dispositifs coopératifs tel que wikidata, base de données d'articles en texte intégral, etc. pour être désambiguïsés, validés, etc. par des intelligences artificielles entrainées par des humains (chercheur · e · s, professionnel · le · s de l'information et de la documentation, *data-scientists*, etc.). Dans ce cadre, la création du HN Lab au sein d'Huma-Num permet de préparer le terrain en testant des hypothèses de travail, très en amont, en faisant de la recherche et des preuves de concept. Ainsi, des travaux menés par le HN Lab sur les revues *Intermédialités* et *Études françaises* en partenariat avec l'entreprise DSI Group pourront alimenter ces réflexions [3]. Comme on le voit, sur plan des chaînes d'annotation sémantique, plusieurs voies d'amélioration s'ouvrent pour les 10 prochaines années d'ISIDORE.

Références

Adhikari A., Tang R. et Lin J., Ram A. 2019. « DocBERT : BERT for Document Classification ». https://arxiv.org/abs/1904.08398.

3. Ce travail s'incrit dans le cadre du programme de recherche Revue2.0 mené à l'Université de Montréal. Voir le log du HN Lab pour plus d'information.

Blei D., Ng A. et Jordan M. 2001. « Latent Dirichlet Allocation ». *The Journal of Machine Learning Research* 3 (janvier) :601-8.

Chaudiron, M., S. et Ihadjadene. 2008. « Quelles analyses de l'usage des moteurs de recherche ». *Questions de communication* 14. https://doi.org/10.4000/questionsdecommunication.604.

Maignien, Yannick. 2011. « ISIDORE, de l'interconnexion de données à l'intégration de services ». https://archivesic.ccsd.cnrs.fr/sic_00593320.

Poupeau, Gautier. 2016. « Bilan de 15 ans de réflexion sur la gestion des données numériques ». http://www.lespetitescases.net/bilan-reflexion-sur-la-gestion-des-donnees-numeriques.

Pouyllau, Stéphane. 2011. « ISIDORE : une plateforme de recherche de documents et d'information pour les Sciences Humaines et Sociales ». https://archivesic.ccsd.cnrs.fr/sic_00605642.

Pouyllau S., Capelli L. et MINEL J-L., Bunel M. 2020. « "We" : a Proposal for the TRIPLE platform », septembre. Zenodo. https://doi.org/10.5281/zenodo.4059099.