# Le projet ISSA : *l'intelligence artificielle au service de la recherche bibliographique*

Porté par trois institutions, ISSA, projet d'indexation automatique des publications d'une archive scientifique ouverte, a été conçu comme un outil d'aide aux recherches bibliographiques complexes.

**Lauréat de l'appel à projet CollEx-Persée[1] en 2020,** le projet ISSA[2] – Indexation Sémantique d'une archive scientifique et Services Associés pour la science ouverte – est porté par trois institutions : le Cirad[3], Inria Sophia Antipolis Méditerranée[4] et IMT Mines Alès[5]. La motivation d'origine, portée par un besoin d'indexation automatique des publications d'une archive scientifique ouverte, s'est rapidement enrichie avec des objectifs plus ambitieux de services de recherche et de visualisation innovants. Les administrateurs d'archives ouvertes gèrent une grande quantité de métadonnées parmi lesquelles les mots-clés qui viennent décrire les publications. Cette indexation est réalisée manuellement la plupart du temps, soit par les déposants eux-mêmes (mots-clés libres en général), soit par des documentalistes spécialistes qui utilisent des descripteurs thématiques ou géographiques issus d'un vocabulaire contrôlé ou d'un thésaurus. Cette activité est exigeante et chronophage, et l'automatisation de l'indexation constitue un besoin clairement identifié par les services d'information scientifique et technique (IST) ou les bibliothèques.

Par ailleurs, ces dernières années, plusieurs évolutions ont radicalement transformé la façon dont les chercheurs et les professionnels en IST interagissent avec la littérature scientifique. En effet, la quantité de publications augmente en flèche, que ce soit dans les revues, les conférences ou par le biais de dépôts de prépublications (par exemple arxiv.org), de sorte qu'il est de plus en plus difficile de trouver des articles correspondants à des critères de recherche parfois très spécifiques.

## LA PLACE DES ARCHIVES OUVERTES DANS L'ÉCOSYSTÈME DE LA LITTÉRATURE SCIENTIFIQUE

Dans ce contexte, les archives scientifiques ouvertes jouent un rôle central pour appuyer les recherches bibliographiques. Cependant, les services de recherche classiques à base de mots-clés proposés nativement par les plateformes ne parviennent souvent pas à saisir la richesse des associations sémantiques entre les articles, de sorte que certaines recherches complexes trouvent difficilement des réponses. Il est donc nécessaire de développer de nouveaux outils qui permettent aux utilisateurs de s'orienter dans cette masse de connaissances. Pour relever ces défis, le projet ISSA, guidé par les objectifs de la science ouverte et s'adossant aux principes FAIR, vise à :

- Fournir un pipeline intégré, générique et réutilisable pour l'analyse et le traitement des articles d'une archive scientifique ouverte
- Traduire le résultat en un index sémantique représenté sous la forme d'un graphe de connaissance RDF[6]
- Développer des services de recherche et de visualisation innovants qui exploitent cet index sémantique pour permettre aux utilisateurs d'explorer les règles d'association thématique, les réseaux de copublications, les articles avec des sujets cooccurrents, etc.

## AGRITROP, CAS D'USAGE DU PROJET ISSA

Pour démontrer la pertinence et l'efficacité de la solution, le projet ISSA s'appuie sur un cas d'usage qui sert de preuve de concept : Agritrop[7], l'archive ouverte institutionnelle du Cirad, contenant plus de 110 000 ressources dont 12 000 articles en libre accès, spécialisée dans les domaines de l'agronomie, de la biodiversité et du développement durable. Le thésaurus multilingue Agrovoc[8], géré par l'Organisation des nations unies pour l'alimentation et l'agriculture, est utilisé pour l'indexation comme vocabulaire de référence spécifique au domaine.

Le processus de construction du graphe de connaissance (ou index sémantique) fait appel à plusieurs techniques d'intelligence artificielle : traitement du langage naturel, ingénierie des connaissances, Web sémantique et données liées. La première étape consiste à récupérer les informations contenues dans l'archive ouverte grâce au protocole OAI-PMH[9]. Dans un premier temps, toutes les métadonnées récupérées sont transformées au format RDF et viennent peupler l'index sémantique : titre, auteurs, résumé, licence, date, langue, identifiants de la publication, lien les PDF en accès libre, etc. Par la suite, les données textuelles des articles telles que le titre, le résumé ou le corps du texte sont traitées afin d'en extraire

[1] www.collexpersee.eu
[2] https://issa.cirad.fr
[3] www.cirad.fr
[4] https://inria.cl/en/centre-inria-sophia-antipolis-mediterranee
[5] https://www.imt-mines-ales.fr
[6] Resource Description Framework : langage de base du Web sémantique développé par le W3C
[7] https://agritrop.cirad.fr
[8] www.fao.org/agrovoc
[9] www.openarchives.org/pmh

automatiquement des descripteurs thématiques et géographiques et des entités nommées, c'est à dire des mentions d'entités reconnaissables dans le texte. Descripteurs et entités nommées sont liés à des bases de connaissance généralistes comme Wikidata[10] et DBpedia[11], géographiques comme GeoNames[12] ou encore à des ressources terminologiques plus spécifiques adaptées à un domaine scientifique donné, par exemple le thésaurus Agrovoc dans le cas d'Agritrop. Ces informations sont transformées en RDF et viennent enrichir à leur tour le graphe de connaissance qui contient alors toutes les informations utiles à la description des publications de l'archive – métadonnées classiques et pour les articles en accès libre, descripteurs thématiques et entités nommées liées. L'ensemble, décrit selon les formats du web sémantique, est naturellement relié au Web des données et interrogeable via un point d'accès SPARQL (langage de requête de données RDF). Les connaissances de milliers de publications produites par des milliers de chercheurs se retrouvent ainsi connectées, publiées sur le Web et interrogeables !

## PROPOSITION DE SERVICES À VALEUR AJOUTÉE

Le graphe de connaissance sert de clé de voûte au développement d'outils de recherche et de visualisation.

Un premier résultat quasi immédiat est la possibilité de consulter les notices de l'archive ouverte par le biais d'une visualisation enrichie : métadonnées classiques, résumé avec entités nommées surlignées et liens vers les bases de connaissance, affichage des descripteurs obtenus automatiquement, visualisation cartographique des entités nommées géographiques du texte.

Deux autres outils de visualisation permettent d'aider à la résolution de requêtes complexes :

- ARViz extrait et visualise des règles d'association reliant les descripteurs thématiques des articles.

La *Figure 1* illustre comment les concepts mentionnés dans les articles de l'archive peuvent être utilisés pour découvrir et visualiser les règles d'association. Dans l'exemple, les articles mentionnant les concepts Covid-19 et sécurité alimentaire (a) mentionnent fréquemment le concept de pandémie (b).

- LDViz permet quant à lui d'explorer les réseaux sémantiques formés par des entités aussi variées que des descripteurs thématiques, des auteurs, des institutions, etc. En visualisant ces réseaux, LDViz permet aux utilisateurs de résoudre des questions de compétence complexes.

Avec différentes techniques de visualisation, la *Figure 2* montre comment un utilisateur peut rechercher des articles mentionnant le concept de santé ou l'un de ses sous-concepts (a) et (b), découvrir qu'il est souvent mentionné avec le changement climatique (c), et obtenir la liste des publications

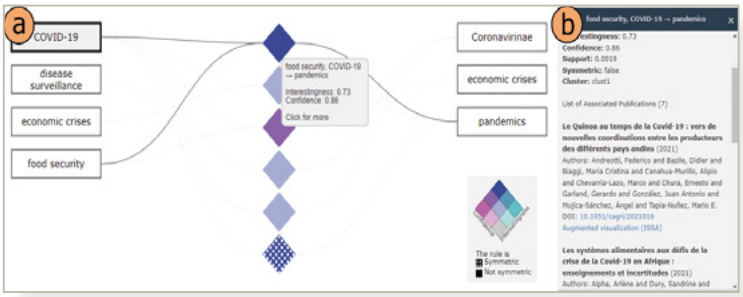


Figure 1 - Recherche par règles d'association

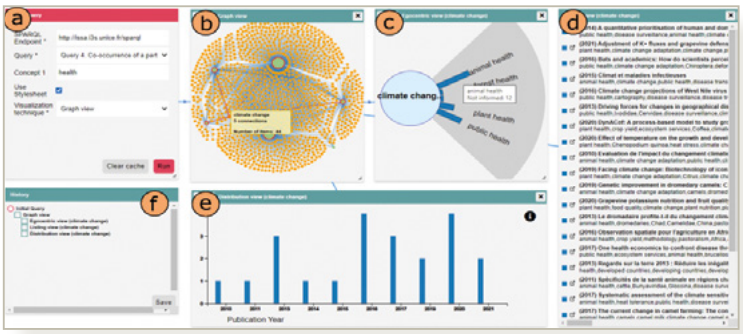


Figure 2 - Recherche par cooccurrence de concepts

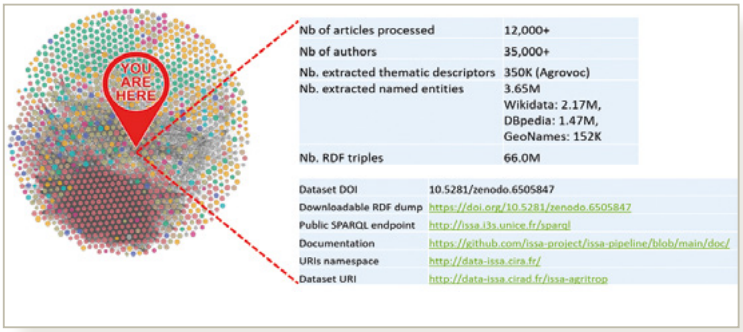


Figure 3 - Le jeu de données ISSA Agritrop dans le LOD

associées (d) et leur répartition dans le temps (e).

S'inscrivant pleinement dans la dynamique de la science ouverte et des données *FAIR*, le travail présenté est rendu disponible sous licence ouverte avec tous les documents d'accompagnement nécessaires pour faciliter sa réutilisation. Le jeu de données ISSA généré dans le cas Agritrop est publié sur le *Linked Open Data Cloud* également sous licence ouverte (*Figure 3*).


Anne Toulet
*Coordinatrice scientifique du projet ISSA pour le Cirad*
anne.toulet@cirad.fr

Franck Michel
*Coordinateur scientifique du projet ISSA pour Inria*
fmichel@i3s.unice.fr

Andon Tchechmedjiev
*Coordinateur scientifique du projet ISSA pour IMT Mines Alès*
andon.tchechmedjiev@mines-ales.fr


[10] www.wikidata.org
[11] www.dbpedia.org
[12] www.geonames.org