# ISSA : un graphe de connaissances au service de la recherche bibliographique

Anne Toulet*, Franck Michel**, Anna Bobasheva**, Aline Menin **, Sébastien Dupré *, Marie-Claude Deboin *, Marco Winckler **, Andon Tchechmedjiev***

* Cirad Avenue Agropolis 34398 Montpellier Cedex – France
anne.toulet@cirad.fr,
** I3S (Univ. Côte d'Azur, CNRS, Inria) 06900 Sophia Antipolis - France
franck.michel@inria.fr
*** EuroMov Digital Health in Motion (IMT Mines Alès) 30100 Alès - France
andon.tchechmedjiev@mines-ales.fr

**Résumé.** Face à la multiplication des publications scientifiques, les archives scientifiques ouvertes jouent un rôle central pour aider les utilisateurs à effectuer des recherches bibliographiques. Cependant, les services de recherche classiques basés sur des mots-clés ne parviennent pas toujours à apporter des réponses satisfaisantes à certaines recherches complexes. Dans cet article, nous présentons les méthodes, outils et services mis en œuvre dans le cadre du projet ISSA pour répondre à cette problématique. Le projet vise à (1) fournir un pipeline générique, réutilisable et extensible pour l'analyse des documents d'une archive scientifique ouverte, (2) traduire le résultat en un index sémantique représenté sous la forme d'un graphe de connaissances RDF ; (3) développer des services de recherche et de visualisation qui exploitent cet index sémantique. Le projet ISSA s'inscrit dans la dynamique de la science ouverte et s'appuie sur un cas d'usage, Agritrop, l'archive ouverte du Cirad.

## 1 Explorer la littérature scientifique

Ces dernières années, l'accélération du rythme des publications et le "tout numérique" ont radicalement transformé la façon d'interagir avec la littérature scientifique, et les utilisateurs ont dû adapter leurs pratiques. Dans ce contexte, les bases de données bibliographiques, les moteurs de recherche et les archives ouvertes occupent une place de premier plan. Cependant, les services de recherche classiques, reposant généralement sur des correspondances de mots clés ou de noms d'auteurs, ne parviennent souvent pas à saisir la richesse des associations sémantiques entre les articles. Certaines recherches complexes trouvent difficilement des réponses et les résultats parfois peu pertinents obligent l'utilisateur à un filtrage manuel fastidieux. Devant ces difficultés, il est important de proposer aux chercheurs et aux professionnels de l'information des outils permettant de s'y retrouver.

Le présent article est un résumé de l'article publié dans la conférence ISWC 2022 « ISSA : Generic Pipeline, Knowledge Model and Visualization Tools to Help Scientists Search and

Make Sense of a Scientific Archive » (Toulet et al. (2022)). Nous y présentons les méthodes, outils et services mis en œuvre dans le cadre du projet ISSA [1] pour répondre à ces besoins. ISSA vise à (1) fournir un **pipeline générique, réutilisable et extensible pour l'analyse des documents d'une archive scientifique ouverte**, (2) traduire les résultats en un **index sémantique sous la forme d'un graphe de connaissances RDF** ; (3) **développer des services innovants de recherche et de visualisation exploitant cet index**. Orientée vers la généricité et la réutilisabilité, la solution proposée adhère aux principes FAIR (Wilkinson et al. (2016)) et aux recommandations de la science ouverte. Les traitements font appel à diverses techniques d'intelligence artificielle : TALN, ingénierie des connaissances, web sémantique. Les métadonnées et le texte intégral des publications sont traités afin d'en extraire des descripteurs thématiques [2] et des entités nommées. Pour exploiter au mieux la puissance du web sémantique, les descripteurs thématiques et les entités nommées sont liés à des référentiels sémantiques (bases de connaissance, ontologies, thésaurus) tels que Wikidata, DBpedia et GeoNames. Le graphe de connaissances résultant sert de clé de voûte au développement de services de recherche et de visualisation. Afin de démontrer l'efficacité de la solution proposée, le pipeline a été testé et déployé sur une archive ouverte en production : Agritrop [3], l'archive ouverte du Cirad [4].

Le reste de ce document est organisé comme suit : la section 2 propose un état de l'art sur les initiatives connexes au projet ISSA. La section 3 décrit le pipeline mis en place pour traiter les documents d'une archive ouverte. La section 4 présente les services de recherche et de visualisation qui exploitent ce graphe. Enfin, la section 5 tire des conclusions et propose des perspectives.

## 2 État de l'art

Il existe une variété de méthodes et d'outils conçus pour traiter le contenu des documents textuels, extraire des connaissances et proposer des services avancés. Des initiatives telles que **Research Data Alliance** (RDA [5]), **Go Fair** [6] ou **European Open Science Cloud** (Budroni et al. (2019)), ont jeté les bases de la mise en œuvre des principes FAIR pour la science ouverte. Le projet **OpenMinted** [7] visait à créer une infrastructure européenne générique de type "Software as a Service" pour l'exploration de textes, basée sur une architecture modulaire. Après 5 ans de développement, le projet n'a pas réussi à fournir un prototype entièrement fonctionnel, se contentant de poser les composants fondamentaux de l'infrastructure. Le projet connexe **Visa TM** [8] devait être le composant central d'extraction de connaissances, intégrant des thésaurus et des ontologies de nombreux domaines, mais il n'est parvenu qu'à une intégration très préliminaire. À l'inverse, le projet ISSA adopte une approche plus modeste mais ciblée et décentralisée, en proposant un pipeline générique adaptable à de multiples domaines,

1. https://issa.cirad.fr/
2. Les descripteurs thématiques sont des mots-clés liés à des vocabulaires de référence, des thésaurus ou des ontologies, qui caractérisent un article dans son ensemble.
3. https ://agritrop.cirad.fr/
4. Centre de coopération internationale en recherche agronomique pour le développement https://www.cirad.fr/
5. RD Alliance project website. https://www.rd-alliance.org/
6. https://www.go-fair.org/
7. http://openminted.eu/
8. https://www.ouvrirlascience.fr/projet-visa-tm/

basé sur l'intégration d'outils existants robustes, et déployable par chaque communauté. ISSA met également l'accent sur l'utilisation de données ouvertes liées, du Web sémantique et des principes FAIR, qui sont absents d'OpenMinted. L'infrastructure **ISTEX**, qui devait être le fournisseur de corpus pour OpenMinted (Kettani et al. (2018)), a des objectifs liés à ISSA dans la mesure où elle vise à constituer des corpus de publications scientifiques et à fournir aux communautés de recherche des outils pour explorer des sous-ensembles pertinents de ces corpus. Cependant, l'objectif principal est de permettre la création et le téléchargement de sous-ensembles de corpus selon des critères très précis, d'extraire la terminologie et de fournir une visualisation descriptive des résultats grâce à l'outil LODEX (Benedetti et al. (2015)). Les aspects indexation et graphe de connaissance consolidé du projet ISSA sont absents. Le projet **Covid-On-The-Web** (Michel et al. (2020)) est le plus récent et est celui qui a le plus de points communs avec ISSA. Il fournit aux chercheurs des moyens d'accéder, extraire et interroger des connaissances à partir de la littérature relative à la famille des coronavirus, en construisant et en exploitant un graphe de connaissances décrivant les concepts et les arguments extraits de plus de 100 000 articles scientifiques. Mais il ne s'agit pas d'un pipeline réutilisable. En résumé, le **projet ISSA possède un atout majeur** absent de toutes ces initiatives : proposer un pipeline intégré et générique, facile à déployer et à personnaliser.

# 3 De l'archive ouverte au graphe de connaissances

Le pipeline ISSA exploite des outils existants pour analyser et indexer les documents d'une archive scientifique ouverte, en établissant des liens entre les articles et des ressources du web de données, et en respectant les standards du web sémantique. La Fig. 1 décrit ce pipeline : (1) Les métadonnées sont extraites de l'archive ouverte, (2) traduites en RDF et stockées dans une base de données RDF Virtuoso. (3) Le texte intégral est extrait des PDFs et, pour chaque article, (4) les descripteurs thématiques et les entités nommées sont extraits du texte et liés à Wikidata, DBpedia, GeoNames et éventuellement à des thésaurus spécifiques du domaine. (5) Les descripteurs et les entités sont traduits en un ensemble unifié de données RDF stockées dans le serveur Virtuoso avec les enregistrements de métadonnées et (6) le graphe de connaissances est exploité pour proposer des services de visualisation et d'exploration.

La transformation en RDF est faite à l'aide du logiciel Morph-xR2RML [9] qui décrit le mapping des données sources vers un modèle RDF qui s'appuie sur des vocabulaires couramment utilisés : Dublin Core Metadata, the FRBR-aligned Bibliographic Ontology (FaBiO), the Bibliographic Ontology (BibO), FOAF, Schema.org, the Web Annotation Vocabulary et PROV-O. Une description complète de la représentation RDF ainsi que des exemples sont fournis dans le dépôt Github du pipeline [10].

**Traitement des métadonnées**. De nombreuses archives ouvertes implémentent nativement le protocole OAI-PMH qui permet de moissonner les métadonnées des documents qu'elles contiennent. Le pipeline ISSA est livré avec un connecteur compatible avec le protocole OAI-PMH. À défaut, les métadonnées peuvent être obtenues à l'aide de diverses interfaces, généralement une API REST et cette étape nécessitera des adaptations mineures du pipeline ISSA : (1) l'écriture d'un connecteur pour s'adapter aux spécificités de l'API de l'archive, et (2) un

9. https://github.com/frmichel/morph-xr2rml/
10. https://github.com/issa-project/issa-pipeline/blob/main/doc/

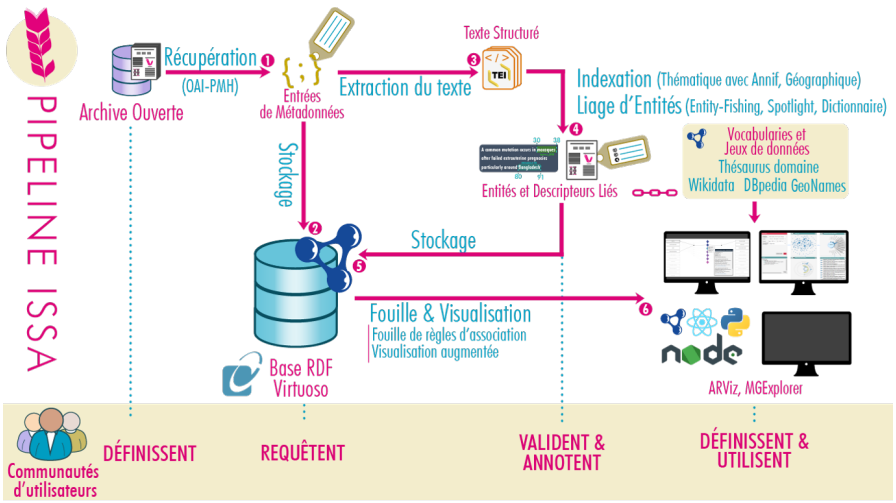


FIG. 1 – *Pipeline ISSA : ressources, services et applications.*

ajustement des mappings permettant de passer du schéma de métadonnées source vers le modèle RDF cible.

**Classification textuelle des articles**. Les descripteurs thématiques sont des mots-clés (généralement moins de dix) qui caractérisent un article dans son ensemble et qui sont liés à un vocabulaire normalisé. Dans certaines institutions, les documentalistes indexent manuellement les documents à l'aide de descripteurs, ce qui permet d'obtenir des annotations précises mais prend beaucoup de temps. S'il existe un corpus suffisamment important indexé avec un vocabulaire de domaine, il est possible d'entraîner un modèle de classification supervisée pour attribuer automatiquement des descripteurs thématiques aux publications. Le pipeline ISSA comprend un tel système de classification grâce à l'intégration de l'outil Annif (Suominen (2019)) développé par la Bibliothèque nationale de Finlande. Annif ne propose pas de nouvelle méthode en soi mais fournit un cadre et une API pour intégrer les modèles et outils d'apprentissage automatique existants.

**Extraction et liaison d'entités nommées**. Le pipeline ISSA s'appuie sur trois outils pour identifier, désambiguïser et lier les entités nommées (EN) à partir des articles (titre, résumé et corps) : DBpedia Spotlight (Daiber et al. (2013)) annote les textes en huit langues différentes avec des entités DBpedia ; Entity-fishingé [11] identifie et désambiguïse les EN avec Wikidata ; et l'outil pyclinrec [12] qui permet d'obtenir une annotation par projection sur dictionnaire. Pour chaque article, le pipeline fait appel à chacun des trois outils et traduit leurs sorties respectives en une représentation RDF. Une étape supplémentaire de post-traitement identifie spécifiquement les entités géographiques en recherchant les correspondances entre GeoNames et les concepts Wikidata correspondants.

**Code source, jeu de données, documentation**. L'ensemble du pipeline ISSA a été déployé sur l'archive institutionnelle du Cirad Agritrop, spécialisée dans les domaines de l'agro-

11. https://github.com/kermitt2/entity-fishing
12. https://github.com/twktheainur/pyclinrec

nomie, la biodiversité et le développement durable, qui contient plus de 110 000 références dont 12 000 articles en accès ouvert. Dans ce contexte, le thésaurus multilingue Agrovoc a été utilisé comme vocabulaire de référence spécifique au domaine. Les différentes étapes sont intégrées dans un pipeline complet entièrement décrit dans Toulet et al. (2022). La documentation et les différents composants sont disponibles sur le site GitHub du projet [13] sous licence Apache 2.0 (open-source et libre), et sont identifiés par un DOI qui garantit leur disponibilité à long terme. Le jeu de données associé, ISSA Agritrop, est disponible sous la forme d'un dump RDF téléchargeable, identifié par un DOI et accessible via un point d'accès SPARQL public.

# 4 Services de visualisation et d'exploration

**Notices bibliographiques enrichies**. Le rôle premier d'une archive ouverte est de fournir un accès aux métadonnées des documents. Le prototype ISSA propose une vue enrichie de ces notices pour chaque document. Au-delà du simple affichage des métadonnées, ce service permet de visualiser le résumé de l'article dont les EN sont surlignées et liées aux bases de connaissances Wikidata, DBpedia, GeoNames et Agrovoc dans notre cas d'usage. Les descripteurs thématiques extraits automatiquement sont également affichés, ainsi qu'une visualisation cartographique des lieux mentionnés dans l'article.

**Outils de visualisation avancée**. Des outils de visualisation spécifiques permettent d'exploiter l'index sémantique. Le premier est ARViz, un outil générique conçu pour l'exploration de règles d'association ; le second, MGExplorer, permet de répondre à d'autres besoins d'exploration et de résoudre des questions de compétences complexes (voir Menin et al. (2021a,b)). Nous proposons ci-dessous deux exemples dans le cadre du cas d'utilisation Agritrop.

*Extraction et visualisation de règles d'association avec ARViz*. La Fig. 2 illustre comment les concepts mentionnés dans les articles de l'archive peuvent être utilisés pour découvrir et visualiser des règles d'association. Dans cet exemple, l'outil fournit une représentation intuitive des éléments impliqués dans les règles (Fig. 2a) : les concepts antécédents et conséquents sont représentés à gauche et à droite de la figure respectivement, tandis que des losanges représentent les règles d'association. Leur couleur indique l'intérêt et la confiance des règles mentionnées. Dans notre cas, COVID-19 est l'antécédent, il est associé à trois concepts conséquents : la famille des Coronavirinae, type de virus à l'origine de la maladie, et pandémies ; plus surprenant, les crises économiques. Pour ce dernier, les publications correspondant à cette règle concernent la résilience du secteur alimentaire et la réponse agricole à la crise du COVID-19.

*Aide à la résolution de requêtes complexes avec LDViz*. Dans l'exemple qui suit, nous nous intéressons à l'initiative One Health Mackenzie et Jeggo (2019); Lerner et Berg (2015) qui vise à unifier les thèmes de santé publique, animale et environnementale pour mieux comprendre le développement des pandémies et la propagation des maladies émergentes. La Fig. 3 montre comment aider les utilisateurs à rechercher des articles mentionnant le concept de santé ou l'un de ses sous-concepts (a et b), à découvrir qu'il est souvent co-mentionné avec le changement climatique (c), et à obtenir la liste des publications connexes (d) et leur répartition dans le temps (e).

13. `https://github.com/issa-project/`

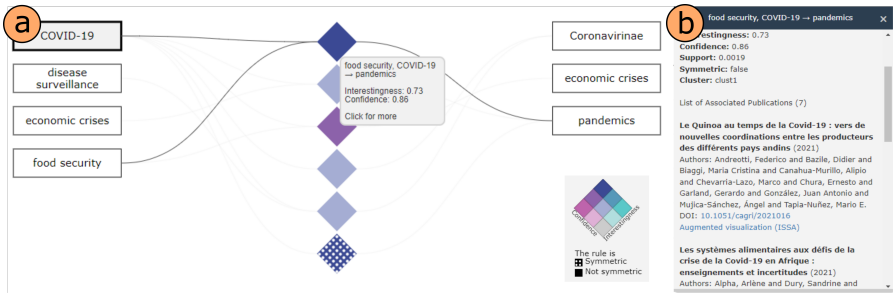


FIG. 2 – *Exploration visuelle (a) des règles d'association impliquant le concept COVID-19 à l'aide d'ARViz et (b) des publications mentionnant les concepts COVID-19, crises économiques et pandémies.*

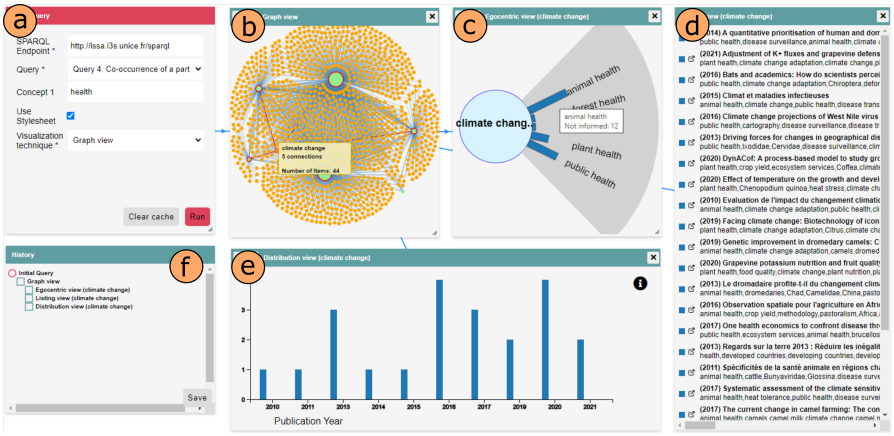


FIG. 3 – *Exploration visuelle de la relation entre les concepts de santé et de changement climatique à l'aide de LDViz.*

# 5 Conclusion et perspectives

Dans cet article, nous avons présenté les méthodes et outils mis en œuvre dans le projet ISSA pour optimiser les recherches bibliographiques. En nous appuyant sur des outils existants et robustes, nous avons conçu un pipeline générique, réutilisable et extensible pour l'analyse et le traitement d'articles provenant d'une archive scientifique ouverte, afin de produire un index sémantique sous la forme d'un graphe de connaissances RDF public. Nous avons développé des services innovants de recherche et de visualisation qui exploitent cet index sémantique pour permettre aux chercheurs, aux décideurs ou aux professionnels de l'information scientifique d'explorer des règles d'association thématiques, des réseaux de co-publications, des réseaux d'articles avec des sujets co-occurrents, etc. Nos expérimentations avec des documentalistes du Cirad ont montré la capacité de ces services à fournir des réponses aux questions de compétences soumises par les chercheurs. À court et moyen termes, nous prévoyons de poursuivre ce travail de plusieurs façons. Tout d'abord, en menant des activités de diffusion afin que d'autres communautés puissent s'emparer du pipeline ISSA en l'adaptant à leurs propres besoins. D'autre part, en enrichissant notre offre de services, notamment en matière de bibliométrie et de recherche d'information.

**Remerciements.** Les travaux présentés dans cet article ont été réalisés dans le cadre du projet ISSA, lauréat 2020 de l'appel à projet financé par le GIS CollEx-Persée [14].

# Références

Benedetti, F., S. Bergamaschi, et L. Po (2015). Lodex : A tool for visual querying linked open data.

Budroni, P., J. Claude-Burgelman, et M. Schouppe (2019). Architectures of knowledge : The european open science cloud. *ABI Technik 39*(2), 130–141.

Daiber, J., M. Jakob, C. Hokamp, et P. N. Mendes (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pp. 121–124.

Kettani, F., S. Schneider, S. Aubin, R. Bossy, C. François, C. Jonquet, A. Tchechmedjiev, A. Toulet, et C. Nédellec (2018). Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTEX, un exemple de service de Text et Data Mining pour les scientifiques français. In S. Ranwez (Ed.), *IC : Ingénierie des Connaissances*, Nancy, France, pp. 247–249.

Lerner, H. et C. Berg (2015). The concept of health in one health and some practical implications for research and education : what is one health ? *Infection ecology & epidemiology 5*, 25300.

Mackenzie et Jeggo (2019). The one health approach—why is it so important ? *Tropical Medicine and Infectious Disease 4*, 88.

Menin, A., L. Cadorel, A. G. B. Tettamanzi, A. Giboin, F. Gandon, et M. Winckler (2021a). ARViz : Interactive Visualization of Association Rules for RDF Data Exploration. In *IV*

14. https://www.collexpersee.eu/projet/issa/

*2021 - 25th International Conference Information Visualisation*, Volume 25, Melbourne / Virtual, Australia, pp. 13–20.

Menin, A., R. Cava, C. M. Dal Sasso Freitas, O. Corby, et M. Winckler (2021b). Towards a Visual Approach for Representing Analytical Provenance in Exploration Processes. In *IV 2021 - 25th International Conference Information Visualisation*, Volume 25, Melbourne / Virtual, Australia, pp. 21–28.

Michel, F., F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby, R. Gazzotti, A. Giboin, S. Marro, T. Mayer, M. Simon, S. Villata, et M. Winckler (2020). Covid-on-the-Web : Knowledge Graph and Services to Advance COVID-19 Research. In *ISWC 2020 - 19th International Semantic Web Conference*, Athens / Virtual, Greece.

Suominen, O. (2019). Annif : DIY automated subject indexing using multiple algorithms. *LIBER Quarterly 29*(1), 1–25.

Toulet, A., F. Michel, A. Bobasheva, A. Menin, S. Dupré, M.-C. Deboin, M. Winckler, et A. Tchechmedjiev (2022). Issa : Generic pipeline, knowledge model and visualization tools to help scientists search and make sense of a scientific archive. In *The Semantic Web – ISWC 2022*, Cham, pp. 660–677. Springer International Publishing.

Wilkinson, M., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, et B. Mons (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*.

## Summary

Faced with the proliferation of scientific publications, scientific archives play a central role in helping users carry out bibliographic research. However, traditional keyword-based search services often fail to grasp the richness of the semantic associations between articles. In this paper, we present the methods, tools and services implemented in the ISSA project to tackle these issues. The project aims to (1) provide a generic, reusable pipeline for the analysis and processing of articles of an open scientific archive, (2) translate the result into a semantic index stored and represented as an RDF knowledge graph; (3) develop innovative search and visualization services that exploit this semantic index. The ISSA project is based on a use case that serves as proof of concept: Agritrop, CIRAD's open archive. Fully in line with the open science and FAIR dynamics, this work is available under an open license.