Rapport scientifique du Projet ISSA 2

Indexation Sémantique d'une archive scientifique et Services Associés pour la science ouverte

Table des matières

Présentation synthétique du projet

Résumé

Le projet ISSA 2 (Lauréat 2022 CollEx-Persée) fait suite au projet ISSA (Lauréat 2020 CollEx-Persée) qui s'intéressait à l'indexation sémantique des publications scientifiques dans une archive ouverte et à l'exploitation de cette indexation pour développer des services aux chercheurs et aux documentalistes.

S'inscrivant dans la dynamique de la science ouverte et le respect des principes FAIR, le projet ISSA 2 propose une méthode générique pour explorer les données disponibles dans une archive ouverte afin 1) d'en extraire des connaissances nouvelles, 2) d'exploiter ces connaissances avec un objectif bibliométrique et 3) de proposer des services aux chercheurs et aux documentalistes, en particulier en termes de bibliométrie et de recherche d'information. Le projet s'appuie sur l'index sémantique produit dans le cadre du projet ISSA, décrivant les publications d'une archive ouverte : métadonnées, descripteurs et entités nommées mentionnées dans le texte, liés à des bases de connaissances standard. Les méthodes proposées exploitent des techniques de fouille de données ainsi que des techniques de construction, publication et exploitation de graphes de connaissances (Web sémantique et Web des données). L'apport de l'intelligence artificielle doit permettre d'enrichir et élargir les possibilités offertes par des indicateurs bibliométriques par ailleurs bien connus. Deux archives ouvertes serviront de cas d'utilisation tout au long du projet : Agritrop, l'archive ouverte du Cirad, et l'instance HAL de l'UR EuroMov Digital Health in Motion.

Coordination scientifique

- **Anne Toulet**, Cirad, Déléguée à l'information scientifique et à la science ouverte (DiscO), Ingénierie des connaissances et web sémantique.
- **Franck Michel,** Université Côte d'Azur, CNRS, Inria (Wimmics), Ingénieur de recherche CNRS, Ingénierie des connaissances et web sémantique.
- **Andon Tchechmedjiev**, IMT MINES ALÈS (EuroMov Digital Health in Motion, Axe SemTaxM), Enseignant-chercheur, Ingénierie des connaissances et Traitement Automatique des Langues.

Dates du projet : Décembre 2022 – Décembre 2024

Subvention CollEx-Persée : 90 986 €

Recrutements et stages spécifiques au projet : 1 Ingénieur Inria (CDD 10 mois), 2 stages de Master 2 en informatique, 2 stages de master en information scientifique (spécialisation Archives ouvertes et Bibliométrie).

Description générale du projet

Contexte et motivations

Depuis plus de vingt ans, le mouvement de la **science ouverte** vise à rendre accessibles librement les résultats de la recherche scientifique. Les publications scientifiques, les données de la recherche ou encore les codes sources des logiciels se retrouvent au cœur de cette dynamique. Le projet ISSA, lauréat de l'AAP CollEx-Persée 2019-2020, s'inscrivait pleinement dans cette dynamique en proposant d'enrichir et d'exploiter les métadonnées d'un corpus de publications. Dans la continuité, le projet ISSA 2 vise à capitaliser les résultats

du projet ISSA pour élargir ses champs d'application et se concentrer sur la place et le rôle des archives ouvertes dans la science ouverte.

Dans ce contexte et de manière tout à fait imbriquée apparaît un lien avec la **bibliométrie**[1]. En effet, dès lors que l'on s'intéresse aux archives ouvertes et aux corpus de publications scientifiques dans la perspective de services aux utilisateurs, il est naturel d'intégrer ce domaine à nos réflexions. Les archives ouvertes institutionnelles sont centrales pour répondre aux besoins des établissements en termes d'instruments d'aide à la décision, d'évaluation et de pilotage, au travers de la production d'indicateurs *ad hoc*. En particulier, elles font partie intégrante du système d'évaluation coordonné en France par le **Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur** (Hcéres) créé en 2013. Basés sur des métriques « traditionnelles », ces indicateurs ne répondent cependant pas à tous les besoins et l'on observe depuis une dizaine d'années une volonté de faire évoluer les méthodes d'évaluation de la recherche avec l'apparition progressive de métriques alternatives, dont le manifeste **« Altmetrics : a manifesto »** [Priem et al. 2010] a été l'une des premières émergences. Fort de ces constats, le projet ISSA 2 s'attache à démontrer comment prendre en compte ces enjeux forts.

Objectifs du projet

S'inscrivant dans une démarche de science ouverte et le respect des principes FAIR, **le présent projet** propose une **méthode générique** pour explorer les données disponibles dans une archive ouverte afin 1) d'en **extraire des connaissances nouvelles**, 2) d'exploiter ces connaissances avec un **objectif bibliométrique** et 3) de proposer des **services aux chercheurs et aux documentalistes**, en particulier en termes de bibliométrie et de recherche d'information. Il s'appuie sur l'index sémantique produit dans le cadre du projet ISSA, décrivant les publications d'une archive ouverte : métadonnées, descripteurs et entités nommées mentionnées dans le texte, liés à des bases de connaissances standard. Les méthodes proposées exploitent des techniques de **fouille de données** ainsi que des techniques de construction, publication et exploitation de **graphes de connaissances** (**Web sémantique et Web des données**). L'**apport de l'intelligence artificielle** doit permettre d'**enrichir et élargir** les possibilités offertes par des indicateurs bibliométriques par ailleurs bien connus. Deux archives ouvertes serviront de cas d'utilisation tout au long du projet : Agritrop, l'archive ouverte du Cirad, et l'instance HAL de l'UR EuroMov Digital Health in Motion.

Environnement institutionnel et partenariats

Le projet ISSA 2 associe trois partenaires : la **Délégation à l'information scientifique et à la science ouverte (DiscO)** du **Cirad** à Montpellier, l'équipe de recherche **Wimmics** du centre **Inria Sophia Antipolis Méditerranée** et l'Unité de Recherche **EuroMov Digital Health in Motion** (Axe SemTaxM), **IMT Mines Alès**. Ces trois partenaires, aux compétences très complémentaires, ont déjà fait la preuve de leur capacité à travailler ensemble, en particulier pendant le projet initial ISSA lauréat de l'AAP CollEx-Persée 2019-2020.

Le **Cirad**, porteur administratif du projet ISSA 2, est l'organisme français de recherche agronomique et de coopération internationale pour le développement durable des régions tropicales et méditerranéennes. Le Cirad a une mission ambitieuse : contribuer à un monde plus durable et à la réalisation des objectifs de développement durable. C'est en rassemblant un large panel de disciplines – des sciences de la vie aux sciences sociales et politiques – que le Cirad fournit des analyses sur les systèmes biologiques, techniques,

[1] La bibliométrie est l'analyse statistique de la production, de la diffusion et de l'impact des publications scientifiques

sociaux et institutionnels. En lien avec son engagement pour une science ouverte, le Cirad conçoit, développe et met à disposition des plateformes de diffusion des connaissances scientifiques qu'il produit ou coproduit avec ses partenaires. Dans cet objectif, la DiscO assure l'acquisition et la diffusion de la production scientifique. Elle accompagne les collectifs du Cirad dans l'analyse d'information, la publication et la valorisation des données de la recherche.

Wimmics est une équipe-projet commune à **Inria** (Sophia Antipolis Méditerranée) et au laboratoire I3S (UMR 7271 Université Côte d'Azur, CNRS). Ses travaux visent à établir des ponts entre deux formes de sémantique qui coexistent sur le Web : d'un côté la sémantique formelle qui inclut la modélisation des connaissances et le raisonnement formel, d'un autre côté la sémantique sociale associée à l'information produite par les humains.

EuroMov Digital Health in Motion est une unité mixte de recherche créée en janvier 2021 entre **IMT Mines Alès**, l'Université de Montpellier et les CHU de Montpellier et de Nîmes. Ses activités se situent à la croisée de plusieurs domaines scientifiques : la science du mouvement humain, la santé et les sciences du numérique. Son axe SemTaxM vise à étudier la sémantique et la taxonomie du mouvement sous l'angle de l'apprentissage automatique multimodal et de techniques de modélisations sémantiques et d'ontologies dans la recherche d'information, le filtrage, la visualisation ou la recommandation, et dispose d'une compétence forte en Traitement Automatique du Langage Naturel (TAL).

Résultats scientifiques

Archives ouvertes et bibliométrie

Pour soutenir les différentes tâches du projet, nous avons étudié l'existant d'une part en termes de bibliométrie (indicateurs, métriques), services associés et perspectives ; d'autre part, en termes de données disponibles dans les deux archives ouvertes qui ont servi de cas d'usage au projet : Agritrop, l'archive ouverte du Cirad, et l'instance HAL de l'UR EuroMov Digital Health in Motion.

Un état de l'art a été produit s'attachant à répondre aux questions suivantes : 1) *Quels sont les critères de mesure courants et les métriques associées ?* 2) *Quels sont les principaux acteurs et services existants autour de ces métriques ?* 3) *Quelles sont les tendances et les perspectives pour les métriques et les services associés ?*

Dans un second temps et pour chacune deux archives ouvertes Agritrop et Hal-Euromov, une étude des schémas de métadonnées, des données disponibles (auteurs, titres, résumés, descripteurs, ODD, etc.), du format et des standards utilisés a été menée afin de permettre la mise en œuvre de nouvelles métriques et cas d'usage.

Le pipeline ISSA

Rappels : le pipeline développé par ISSA

Le premier projet ISSA a permis le développement et la mise en production d'un pipeline (Figure 1) pour l'analyse des documents d'une archive scientifique ouverte. Ce pipeline est conçu pour être **générique** (non lié à un domaine ou une communauté), **réutilisable** (pour traiter n'importe quelle archive) et **extensible** (on peut y ajouter de nouvelles étapes de traitement).

Les métadonnées et le texte intégral des publications sont traités afin d'en extraire des **descripteurs thématiques** et **géographiques**, et des **entités nommées** (EN, mentions de texte). Les descripteurs et les EN

sont liés à trois référentiels sémantiques généralistes : Wikidata, DBpedia et GeoNames. D'autres ressources sémantiques spécifiques au domaine peuvent être intégrées à chaque étape.

Le pipeline intègre et exploite des outils existants suffisamment matures pour une utilisation en production. La figure ci-dessous illustre les étapes du pipeline : (1) Les métadonnées sont extraites de l'archive ouverte, (2) traduites en RDF pour former la base du graphe de connaissances stocké dans une base de données RDF (Virtuoso). (3) Le texte intégral est extrait des PDFs à l'aide de l'outil *Grobid.* (4) Les descripteurs thématiques sont extraits à l'aide du système de classification *Annif*, et les EN sont extraites à l'aide de trois outils : *DBpedia Spotlight* (lien avec DBpedia), *Entity-fishing* (lien avec Wikidata), et *pyclinrec* (lien avec un vocabulaire spécifique au domaine). (5) Les descripteurs et EN sont traduits en un ensemble unifié de données RDF stockées dans la base de données RDF. (6) Une fois publié, le graphe de connaissances peut être exploité pour proposer des services de visualisation et d'exploration.

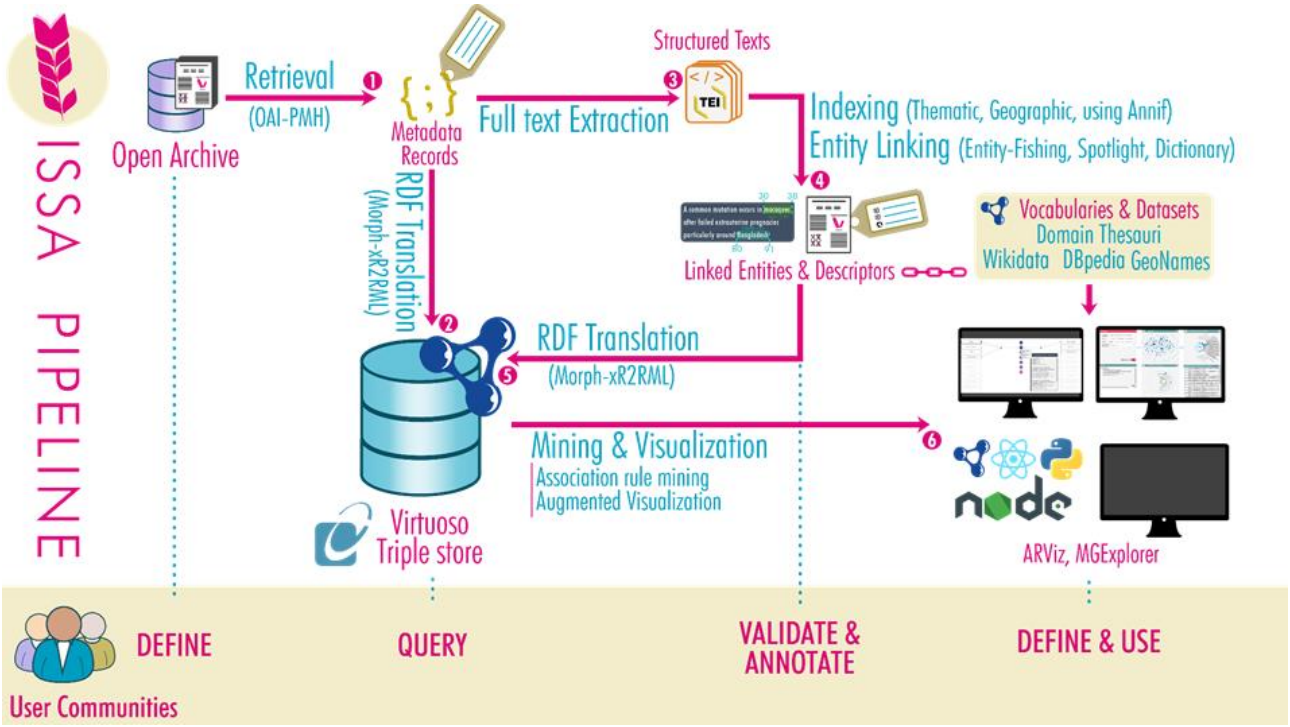

Figure 1 : Le pipeline ISSA

L'ensemble du pipeline ISSA a été déployé sur l'archive institutionnelle du Cirad, Agritrop. Le thésaurus multilingue Agrovoc a été utilisé comme vocabulaire de référence spécifique au domaine.

Application du pipeline à une collection HAL

Dans une première étape, nous avons procédé à une réorganisation du code afin de supporter l'hébergement simultané de plusieurs archives sur le même déploiement. Cela repose sur (1) la mise en commun de tous les outils indépendants de l'archive (Virtuoso, Annif, DBpedia Sportlight, Entity-fishing, pyclinnrec, Morph-xR2RML), et (2) la configuration de chaque environnement uniquement via des paramètres de configuration.

Nous avons ensuite déployé le pipeline pour la collection HAL de l'UR EuroMov Digital Health in Motion (EDHM), accessible comme Agritrop via le protocole OAI-PMH. Cette étape a permis de valider les objectifs de généricité et de réutilisabilité du pipeline.

Extension du pipeline pour les objectifs liés à la bibliométrie

L'analyse du domaine de la bibliométrie, issue du WP2, a permis de dégager plusieurs objectifs pour ISSA 2 :

- Analyser les réseaux de co-publication, co-occurrence et collaboration, non seulement entre auteurs mais en prenant également en compte leurs institutions, ce qui nécessite de connaître les affiliations des auteurs (T1.2) ;
- Analyser la répartition des publications/institutions par Objectif Développement Durable (ODD) (T3.2) ;
- Analyser le positionnement thématique des chercheurs, des institutions (T1.2) ;
- Analyser l'interdisciplinarité en mesurant la diversité des thématiques abordées par publication/auteur/institution, en se basant sur l'indice de diversité Rao-Stirling (T3.2).

Or les données nécessaires à ces analyses (affiliations, ODD, thématiques) ne sont pas présentes dans le graphe de connaissances issu du premier projet ISSA. Ces données sont cependant disponibles dans plusieurs archives bibliographiques comme OpenAlex, Scopus et Web of Science (WoS). Nous avons privilégié OpenAlex pour la taille de son corpus (en constante augmentation), la nature ouverte de son approche, et la transparence des méthodes de classifications utilisées pour les ODD et les thématiques.

Ainsi, nous avons étendu le pipeline en y intégrant la collecte, depuis OpenAlex, des données suivantes pour chaque publication : affiliation des auteurs, classification par ODD, classification thématique.

A partir de ces données, un nouveau composant calcule, pour chaque publication, l'indice de diversité Rao-Stirling. Utilisant les méthodes déjà présentes dans le pipeline, ces données sont transformées en RDF afin d'enrichir le graphe de connaissances.

Le pipeline ainsi étendu a été déployé pour les deux archives Agritrop et HAL, et est depuis exécuté une fois par mois, gardant ainsi le graphe de connaissances à jour avec les dernières publications ajoutées aux archives.

Amélioration du processus d'extraction de règles d'association

Par ailleurs, dans le contexte des T2.1 et T2.2, nous avons exploré différentes stratégies pour ajuster le processus d'extraction des règles d'association afin de détecter en priorité des connaissances dites « nouvelles », c'est-à-dire n'étant pas représentée dans les ontologies de référence du domaine. De telles associations permettraient la découverte d'associations ne faisant pas partie de l'inconscient collectif du domaine concerné. En effet, typiquement des investigations scientifiques dans l'air du temps (par exemple menées par des groupes visibles selon le paradigme dominant) sont connues de tous, mais ne favorisent pas toujours l'innovation scientifique en rupture, qui se trouve souvent dans des signaux faibles [Simonton 1979].

Nous avons ainsi exploré des variations de la méthodologie proposée par [Cadorel et al., 2020], d'une part pour améliorer la pertinence des règles, notamment par des architectures d'apprentissage profond de type auto-encodeur permettant une réduction dimensionnelle plus pertinente comme base du calcul, et d'autre part, nous avons exploré des stratégies d'intégration de représentations vectorielles des ontologies domaine afin de permettre le calcul d'un score de nouveauté pour détecter les connaissances nouvelles.

Bien qu'une évaluation des améliorations proposées permette d'obtenir des scores d'intérêt et de confiance des règles plus élevés, en pratique nous avons observé que les scores de nouveauté ne permettent pas d'obtenir le résultat escompté vis-à-vis des utilisateurs (documentalistes dans nos cas d'usage), mais au contraire mettent en avant des relations jugées comme « évidentes ».

Cela s'explique par le fait qu'avant que des relations ne soient intégrées aux ontologies de domaine, il faut qu'un consensus soit établi à l'échelle de la communauté, consensus qui n'émerge que lorsque l'idée en question est déjà bien ancrée dans la culture du domaine. Par ailleurs, les règles d'association sont un outil statistique qui fait ressortir les associations fréquentes, et qui par définition sont peu adaptées à découvrir des innovations de rupture dans la masse de la production scientifique.

Nos préconisations portent sur l'utilisation d'approches d'extraction de relations par apprentissage profond, particulièrement celles d'inférence causale, étant donné qu'elles ne sont pas uniquement limitées aux relations fréquentes du fait des capacités de généralisation avancées de ces modèles pourtant également basés sur l'apprentissage statistique (voir par exemple [Yang et al. 2025 ; Yuan et al. 2021 ; Zhao et al. 2023]). Il convient de noter que la sortie de telles approches serait compatible avec des outils de visualisation tel que ArViz (présenté par la suite) modulo des modifications minimales.

Services de visualisation

Recherche par mots-clés exploitant l'index sémantique (tâche T3.1)

Dans le cadre du premier projet ISSA, nous avons développé plusieurs outils de visualisation à destination des utilisateurs afin de démontrer l'intérêt de l'index sémantique des publications d'une archive ouverte.

Dans la tâche T3.1 du projet ISSA 2, nous avons souhaité accompagner l'utilisateur dans le processus de recherche en développant une interface exploitant conjointement les descripteurs et EN liés à des concepts issus de bases de connaissances, et la richesse sémantique fournie par ces mêmes bases de connaissances et autorisant certaines formes de raisonnement.

Les deux figures suivantes illustrent cette interface. Elle permet à l'utilisateur de sélectionner des concepts liés aux descripteurs (ici, Agrovoc) et/ou aux EN liés (ici, Wikidata). S'appuyant sur ces référentiels, la recherche est naturellement multi-langue et intègre les synonymes des mots saisis par l'utilisateur. Elle implémente en outre plusieurs types de raisonnements :

- Figure 2. Recherche hiérarchique : recherche des publications annotés avec les concepts/EN sélectionnés ou de concepts/EN plus spécifiques. Dans la figure ci-dessous, la recherche du concept Agrovoc "climate change" retourne également les publications liées à "global warming" qui est un sous-concept de "climate change" dans Agrovoc.
- Figure 3. Recherche par termes proches : retourne les publications liées à des termes connexes au terme sélectionné. C'est le cas "climate change adaptation" qui est un concept connexe à "climate change" dans Agrovoc.

Lorsque les bases de connaissances intègrent des données géographiques, la recherche hiérarchique permet en outre d'effectuer une recherche géographique. Dans l'exemple ci-dessous, la recherche par l'EN "Southeast Asia" de Wikidata retourne des publications mentionnant les EN "Thailand", "Indonesia", "Vietnam" etc.

À partir de ces recherches, les métadonnées de chaque publication peuvent être visualisées grâce à la visualisation des notices bibliographiques développée lors du premier projet ISSA.

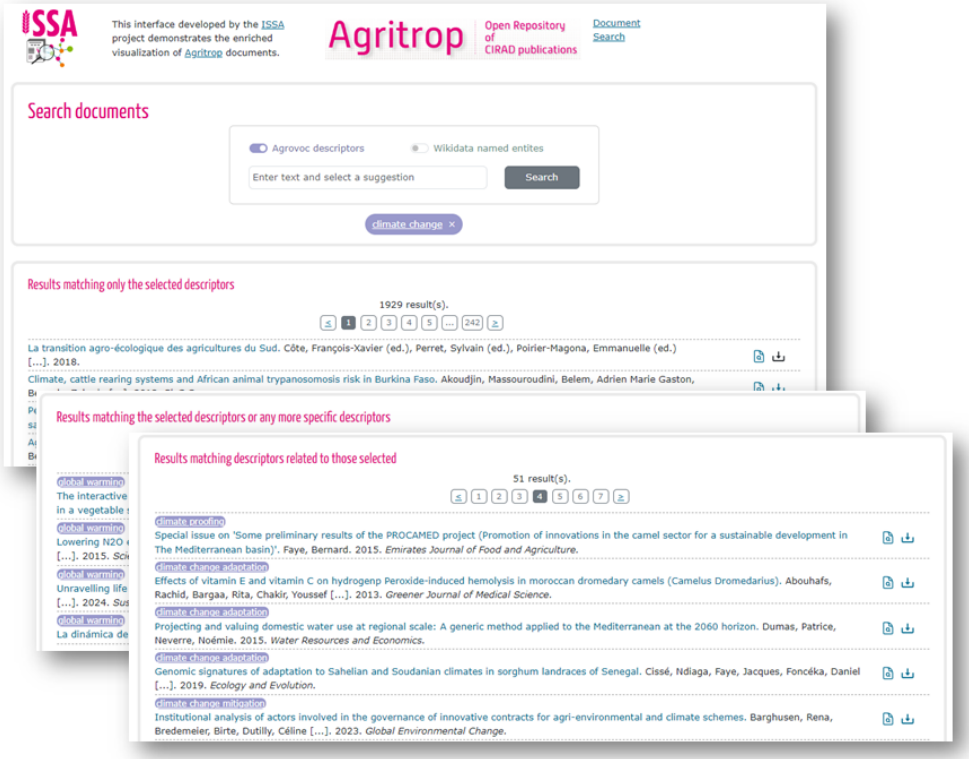

Figure 2 : Un exemple de recherche hiérarchique

Figure 3 : Un exemple de recherche par termes proches

Réseaux de co-publication/co-occurrence (tâche T3.1)

En complément des nouvelles données intégrées au graphe de connaissances sur les affiliations des auteurs, l'outil MGExplorer [Menin et al., 2021] a été étendu avec plusieurs visualisations nouvelles. ()

La visualisation présentée dans la Figure 4 montre les collaborations entre le Cirad et d'autres institutions et/ou unités de recherche.

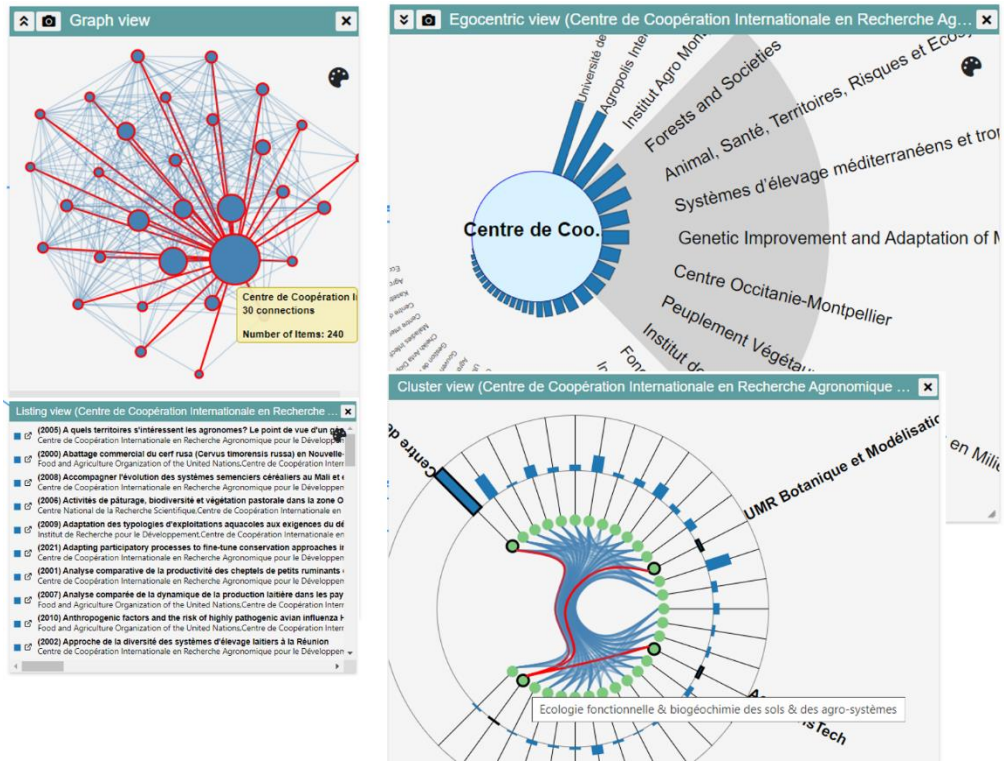

Figure 4 : Collaborations entre le Cirad et d'autres institutions et/ou unités de recherche

La visualisation présentée dans la Figure 5 montre les relations entre les auteurs/autrices et les thématiques qu'ils et elles abordent dans leurs publications : en haut à droite, soit les thématiques sur lesquelles travaille un ou une chercheuse particulière ; en bas à droite, les scientifiques qui travaillent sur une thématique particulière (bas, droite).

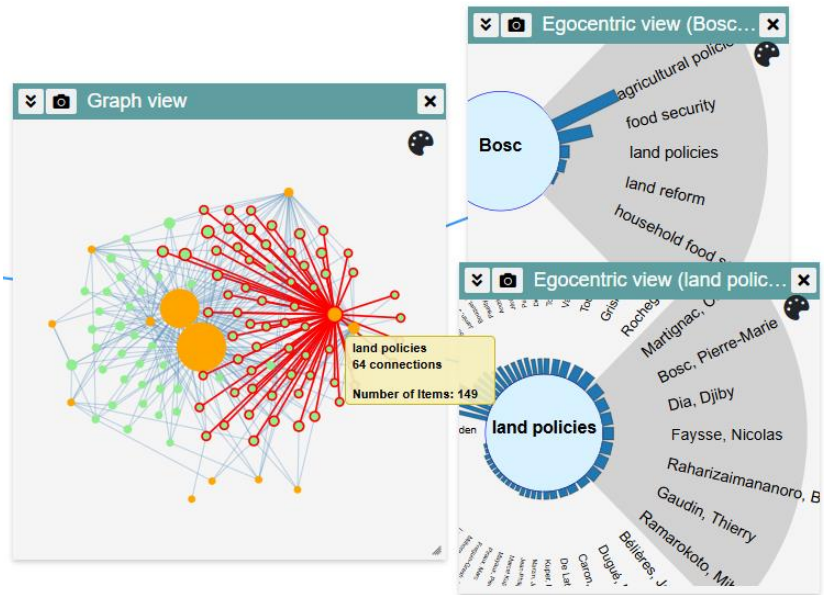

Figure 5 : Relations entre les auteurs/autrices et les thématiques abordées dans les publications

Indicateurs bibliométriques et évolution temporelle (tâches 1.2, 3.2)

Les visualisations précédentes utilisant MGExplorer permettent d'analyser des co-publications/co-occurrences et leur évolution temporelle deux à deux, mais se prêtent mal à l'analyse des réseaux plus complexes.

Ainsi, pour explorer des réseaux plus complexes et analyser leur dynamique temporelle, nous avons étendu et adapté l'outil Muvin (MUltidimensional VIsualization of Networks) [Menin et al., 2022]. Dans Muvin, l'axe des abscisses représente la chronologie en années, et l'axe des ordonnées représente les entités qui collaborent, ces nœuds pouvant être des auteurs, institutions, thématiques, etc.

Notons que le terme générique de "collaboration" se comprend bien dans le cas d'auteurs et d'institutions, mais n'est pas adapté dans d'autres cas comme celui des thématiques pour lesquelles on parlera plus facilement de co-occurrence. De plus, chaque ligne peut combiner deux types de représentation :

- un diagramme de nœuds représentant les objets de la collaboration : typiquement, les nœuds sont des publications représentées dans le temps (axe x) et par auteurs, institutions, thématiques etc. (axe y) ;
- un diagramme de flux (streamgraph) représenté comme des "vagues" de couleur qui permet d'ajouter une dimension supplémentaire

La Figure 6 montre l'application de Muvin aux auteurs et leurs publications. L'exploration suit une approche incrémentale dans laquelle l'utilisateur choisit un ou une autrice pour commencer l'exploration. Les streamgraphs fournissent un aperçu de sa production au fil du temps en termes de quantité (hauteur de la vague) et de type de production (couleur de la vague), ce dernier correspondant aux articles de conférence (orange), articles de revues (vert), livres (jaune), etc. Chaque publication est représentée par un nœud gris dont la taille encode le nombre de collaborateurs.

Nous avons exploré l'utilisation de Muvin à d'autres type d'indicateurs. Dans la Figure 7, les streamgraphs représentent les 6 principaux ODD dans les publications impliquant le Cirad et AgroParisTech.

Enfin, dans la Figure 8, les streamgraphs représentent la valeur de l'indice de diversité de Rao-Stirling pour trois URM du Cirad ayant des publications en commun. La valeur de l'indice, comprise entre 0 et 1, a été discrétisée par des intervalles [0,0-0,4[, [0,4-0,6[, [0,6-0,8[, etc. Remarquons qu'ici les nœuds représentant les publications sont masqués afin de ne pas surcharger inutilement la visualisation.

Ces utilisations variées de Muvin montrent la puissance et la flexibilité de cet outil, et sa capacité à produire des visualisations multi-dimensionnelles. La durée du projet ne nous a cependant pas permis de réaliser une évaluation formelle en mettant l'outil dans les mains des documentalistes. Si les premiers retours sont positifs, ils font aussi apparaître des pistes d'amélioration, qu'il s'agisse de la variété des indicateurs bibliométriques et leur visualisation, ou de l'ergonomie et l'utilisabilité de l'outil lui-même. Néanmoins, cette expérimentation a permis de mettre la lumière sur des perspectives possibles pour l'évolution du domaine de la bibliométrie.

Figure 6 : Utilisation de Muvin pour visualiser les collaborations entre auteur·rice·s

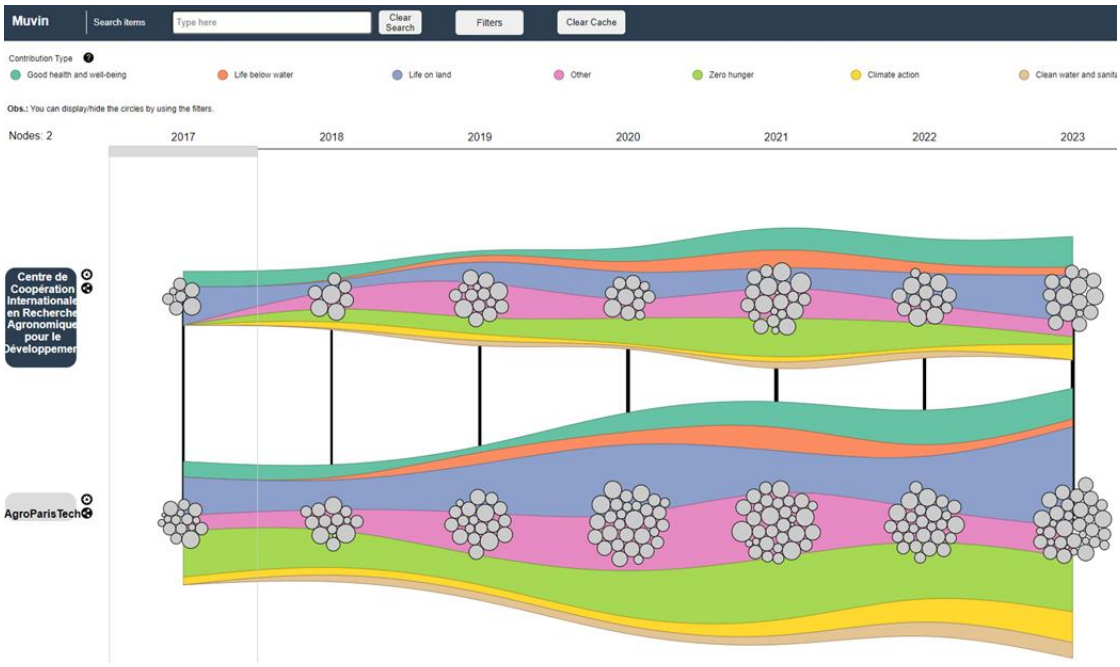

Figure 7 : Six principaux ODD dans les publications impliquant le Cirad et AgroParisTech

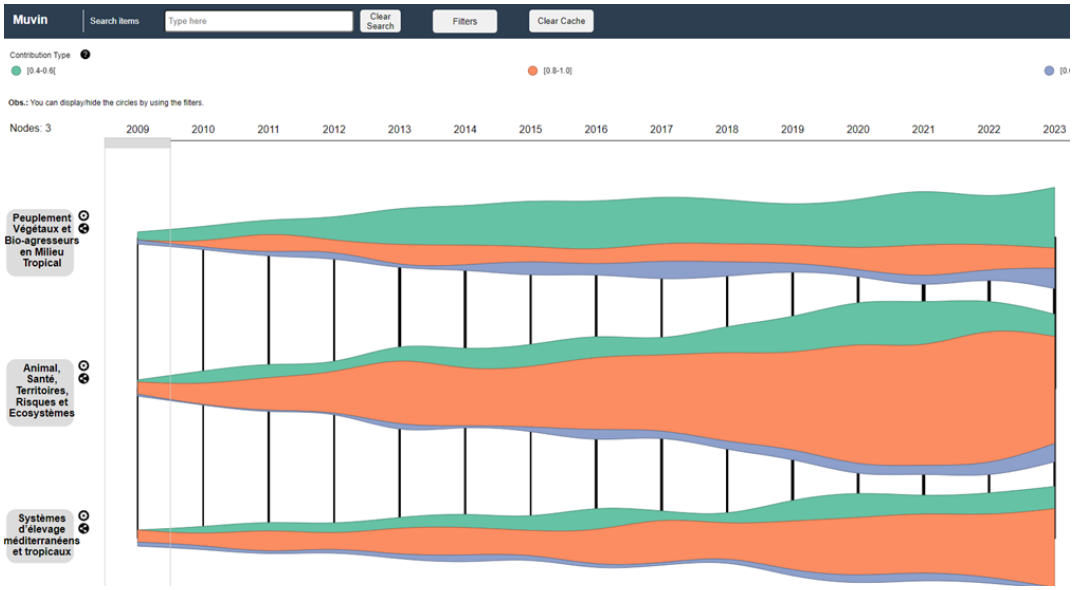

Figure 8. Comparaison de l'indice de Rao-Stirling pour 3 UMR du Cirad ayant des collaborations

ARViz

ARViz est un outil générique conçu pour l'exploration de règles d'association. Une règle d'association "A ⍰ B" se comprend de la façon suivante : si un document possède le descripteur thématique A (l'antécédent), alors il possède le descripteur thématique B (le conséquent) avec une certaine probabilité.

Dans la Figure 9, 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 cet exemple, la règle survolée indique que les antécédents *climate change mitigation* et *tropical forests* ont comme conséquent carbon sequestration avec une confiance maximale.

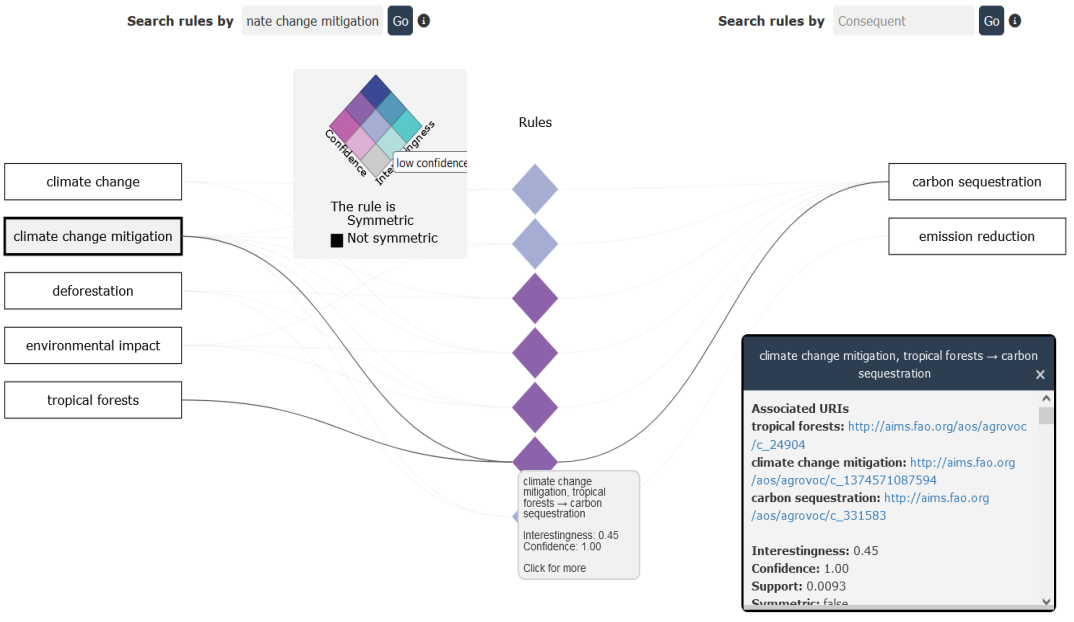

Figure 9. ARViz, visualisation et exploration de règles d'association entre descripteurs thématiques

Diffusion et valorisation des résultats

Dissémination

Un effort important du projet a porté sur l'aspect **pédagogie/retour vers les communautés** pour une **montée en compétence des jeunes** et le transfert de la méthodologie vers d'autres établissements.

Au cours du projet, 2 stages de Master 2 en informatique et 2 stages de master en information scientifique (spécialisation Archives ouvertes et Bibliométrie) ont été proposés.

Un **atelier** de fin de projet s'est déroulé au grand amphithéâtre du Cirad le 10 septembre 2025. Cette journée a permis de présenter le projet, les résultats obtenus et la manière dont ils peuvent exploités par d'autres.

Communication

Différents canaux et différentes formes de communication ont été utilisés pour diffuser les informations concernant le projet :

- La page projet CollEx-Persée qui propose une fiche synthétique du projet ainsi que des pointeurs vers des informations en lien avec le projet
- Le site Web du projet ISSA ; on y trouve les informations importantes, les annonces d'évènements, les pointeurs vers des informations en lien, etc.
- Des actualités sur nos différents sites institutionnels

Code source et composants logiciels

Le code du pipeline ISSA et des composants logiciels sont identifiés par un DOI et disponibles sur les dépôts GitHub du projet sous licence **Apache 2.0** (open-source et libre). Le Tableau 1 précise ces éléments.

Projet	DOI	Dépôt
ISSA pipeline v2.1	10.5281/zenodo.14230463	https://github.com/issa-project/issa-pipeline https://hal.science/hal-04807540
Arviz et fouille des règles d'association	10.5281/zenodo.6511786 10.5281/zenodo.6511146	https://github.com/Wimmics/arviz https://github.com/Wimmics/association-rules-mining
MGExplorer	10.5281/zenodo.6511782	https://github.com/Wimmics/ldviz
Recherche par mots-clés et visualisation des notices bibliographiques v1.1	10.5281/zenodo.10450607 10.5281/zenodo.10450633	https://github.com/issa-project/web-visualization https://github.com/issa-project/web-backend

Tableau 1. Code et composants logiciels

Dataset ISSA Agritrop et documentation

Le jeu de données produit dans le cadre du cas d'usage Agritrop est disponible sous forme d'un dump RDF téléchargeable et via un point d'accès SPARQL public. Il est identifié par un DOI et placé sous licence *Open Data Commons*. La documentation publiée dans le cadre du projet ISSA est en accès ouvert sous licence Creative Commons Attribution 4.0 International. Ces informations sont résumées dans le Tableau 2.

DOI du jeu de données ISSA Agritrop v2.0	10.5281/zenodo.10381606
Licence	Open Data Commons 1.0
Nombre de publications Nombre de publications avec texte Nombre d'auteurs	42,000+ 14,900+ 46,000+
RDF dump téléchargeable	https://doi.org/10.5281/zenodo.10381606
Point d'accès SPARQL	http://data-issa.cirad.fr/sparql
Entités nommées extraites	Entités nommées : 9.45M (Wikidata : 2.62M, DBpedia : 2.1M, Agrovoc : 3.84M, GeoNames : 250K) Descripteurs thématiques : 380K
Nombre de triplets RDF	167M

Tableau 2. Éléments d'information sur le jeu de données produit

Publications en lien avec le projet

Les différents articles mentionnés ci-dessous sont en accès ouvert sous licence Creative Commons Attribution 4.0 International (conférences internationales ISWC 2022 et TDWG 2022) ou sous licence CC BY-ND 2.0 (article publié dans le numéro 107 de la revue « Arabesques »). Un article est à paraître et sera également publié en accès libre (Conférence EGC 2023).

Liste détaillée des publications :

- Anne Toulet, Franck Michel, Anna Bobasheva, Aline Menin, Sébastien Dupré, Marie-Claude Deboin, Marco Winckler, Andon Tchechmedjiev. *ISSA: Generic Pipeline, Knowledge Model and Visualization tools to Help Scientists Search and Make Sense of a Scientific Archive*. ISWC 2022 - 21st International Semantic Web Conference, Oct 2022, Hangzhou, China, https://doi.org/10.1007/978-3-031-19433-7_38.
- Franck Michel, Anne Toulet, Anna Bobasheva, Marie-Claude Deboin, Sébastien Dupré, Aline Menin, Marco Winckler, Andon Tchechmedjiev. *Semantic Indexing of Open Scientific Literature to Help Users Discover and Navigate through Publications Networks*. TDWG 2022 - The Biodiversity Information Standards annual conference, Oct 2022, Sofia, Bulgaria, https://doi.org/10.3897/biss.6.93640
- Anne Toulet, Franck Michel et Andon Tchechmedjiev. *Le projet ISSA : l'intelligence artificielle au service de la recherche bibliographique*. Arabesques 107, p. 6-7, 2022. https://dx.doi.org/10.35562/arabesques.3046

– Anne Toulet, Franck Michel, Anna Bobasheva, Aline Menin, Sébastien Dupré, Marie-Claude Deboin, Marco Winckler, Andon Tchechmedjiev. *ISSA, un graphe de connaissances au service de la recherche bibliographique*. EGC 2023 - 23ème conférence francophone sur l'extraction et la gestion des connaissances. 16-20 janvier 2023, Lyon, France

Bibliographie

[Ahmadi et al. 2020] N. Ahmadi, V-P. Huynh, V. Meduri, S. Ortona, P. Papotti. Mining Expressive Rules in Knowledge Graphs. J. Data and Information Quality, 12(2). (2020) https://doi.org/10.1145/3371315

[Cadorel et al., 2020] Lucie Cadorel, Andrea G. B. Tettamanzi. Mining RDF Data of COVID-19 Scientific Literature for Interesting Association Rules. WI-IAT'20 - IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Dec 2020, Melbourne, Australia.

[Cadorel et al., 2020] Lucie Cadorel, Andrea G. B. Tettamanzi. Mining RDF Data of COVID-19 Scientific Literature for Interesting Association Rules. WI-IAT'20 - IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Dec 2020, Melbourne, Australia.

[Gandon et al, 2012] F. Gandon, O. Corby, and C. Faron-Zucker, Le Web sémantique: Comment lier les données et les schémas sur le Web ? Dunod, 2012.

[Gregorio 2019] Stéphanie Gregorio, Alain Collignon, François Parmentier, Nicolas Thouvenin. LODEX : des données structurées au web sémantique. Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019), Jan 2019, Metz, France.

[Kettani et al., 2018] Kettani, F., Schneider, S., Aubin, S., Bossy, R., François, C., Jonquet, C., Tchechmedjiev, A., Toulet, A., and Nédellec, C. Projet VisaTM : l'interconnexion OpenMinTeD – AgroPortal – ISTEX, un exemple de service de Text et Data Mining pour les scientifiques français. Ingénierie des Connaissances, 2018, Nancy, France. pp.247-249, 2018

[Menin et al., 2021] Menin, A., Cava, R., Dal Sasso Freitas, C.M., Corby, O., Winckler, M.Towards a Visual Approach for Representing Analytical Provenance in Exploration Processes. In: IV 2021 - 25th International Conference Information Visualisation. 2021 25th International Conference Information Visualisation (IV), vol. 25, pp. 21–28. Melbourne / Virtual, Australia (Jul 2021), https://hal.archives-ouvertes.fr/hal-03292172

[Menin et al., 2022] Aline Menin, Michel Buffa, Maroua Tikat, Benjamin Molinet, Guillaume Pelerin, et al.. Incremental and multimodal visualization of discographies: exploring the WASABI music knowledge base. WAC 2022 - Web Audio Conference 2022, Jul 2022, Cannes, France. (10.5281/zenodo.6767530). (hal-03748134)

[Menin-Cadorel et al 2021] Aline Menin, Lucie Cadorel, Andrea Tettamanzi, Alain Giboin, Fabien Gandon, Marco Winckler. ARViz: Interactive Visualization of Association Rules for RDF Data Exploration. The 25th International Conference Information Visualisation, 5 - 9, July 2021, University of Technology Sydney , Sydney, Australia.

[Michel et al 2020] Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, et al.. Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research. ISWC 2020 - 19th International Semantic Web Conference, Nov 2020, Athens / Virtual, Greece.

[Priem et al. 2010] J. Priem, D. Taraborelli, P. Groth, C. Neylon (2010), Altmetrics: A manifesto, 26 October 2010. Journal of Personality and Social Psychology, 37(9), 1603–1616. https://doi.org/10.1037/0022-3514.37.9.1603

[Simonton 1979] Simonton, D. K. (1979). Multiple discovery and invention: Zeitgeist, genius, or chance?.

[Tennant et al., 2019] Tennant J., Beamer J. E., Bosman J., Brembs B., Chung N. C., Clement G., Crick T., Dugan J., Dunning A., Eccles D., Enkhbayar A., Graziotin D., Harding R., Havemann J., Katz D. S., Khanal K., Kjaer J. N., Koder T., Macklin P., Madan C. R., Masuzzo P., Matthias L., Mayer K., Nichols D., Papadopoulou E., Pasquier T., Ross-Hellauer T., Schulte-Mecklenbeck M., Sholler D., Steiner T., Szczesny P. & Turner A. (2019). Foundations for Open Scholarship Strategy Development. BITSS <https://osf.io/b4v8p>.

[Wilkinson et al., 2016] Wilkinson M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018

[Yang et al. 2025] Yang, R., Zhu J., Man J., Liu H., Fang L., Zhou Y. GS-KGC: A generative subgraph-based framework for knowledge graph completion with large language models. (2025) Information Fusion, Volume 117, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2024.102868.

[Yuan et Eldardiry 2021] C. Yuan, & H. Eldardiry. Unsupervised Relation Extraction: A Variational Autoencoder Approach. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 1929–1938 (2021). https://aclanthology.org/2021.emnlp-main.147

[Zhao et al. 2023] Zhao B., Wang S., Chi L., Li Q., Liu, X., and Geng, J. (2023) Causal Discovery via Causal Star Graphs. ACM Trans. Knowl. Discov. Data 17, 7, Article 98 (August 2023), 24 pages. https://doi.org/10.1145/3586997