

Génération et utilisation de graphes de connaissances pour aider à la contextualisation des données de métabolomique

Olivier Filangi^{1,2}, Guillaume Laisney^{1,2}, David Benaben^{4,5}, Alain Bouchereau^{1,2}, Matéo Boudet^{1,3}, Cécile Cabasson^{4,5}, Younes Dello^{1,2}, Yann Guitton⁶, Guillaume Marti⁸, Meije Mathé⁹, Nils Paulhe⁷, Sylvain Prigent^{4,5}, Faustine Souc⁷, Magalie Weber¹⁰, Clément Fraiseney⁹ and Franck Giacomoni⁷

¹ Institute for Genetics, Environment and Plant Protection (IGEPP), National Research Institute for Agriculture, Food and Environment (INRAE), Institut Agro, Université Rennes, 35650 Le Rheu, France

²Metabolic Profiling and Metabolomic Platform (P2M2), Biopolymers Interactions Assemblies, Institute for Genetics, Environment and Plant Protection, 35650 Le Rheu, France

³INRIA, IRISA, GenOuest Core Facility, Beaulieu Campus, Rennes, France

⁴UMR Biology of Fruit and Pathology, University of Bordeaux, INRAE, Bordeaux, France

⁵Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, Villenave d'Ornon, France

⁶Oniris, INRAE, LABERCA, Nantes, France

⁷University Clermont Auvergne, INRAE, UNH, Metabolism Exploration Platform, MetaboHUB Clermont, Clermont-Ferrand, France

⁸Metatoul-AgromiX Platform, LRSV, University of Toulouse, CNRS, UT3, INP Toulouse

⁹Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse 31300, France

¹⁰INRAE, UR BIA, F-44316, Nantes, France

Résumé

Cet article présente une e-infrastructure de graphes de connaissances, le "Metabolomic Semantic Data Lake", pour contextualiser les données métabolomiques. Elle relie des concepts biologiques et annote la littérature scientifique grâce à des méthodes issues de l'intelligence artificielle, palliant le manque d'annotation dans certains domaines. Basée sur des technologies Big Data et le Web Sémantique, elle offre une API pour un accès facile aux données. Les graphes FORVM ChemDisease et FORVM Plants permettent d'analyser des voies toxicologiques et d'identifier des biomarqueurs végétaux. L'objectif est d'améliorer l'intégration et l'interprétation des données en sciences de la vie.

Mots-clés

Métabolomique, web sémantique, IA, graphe de connaissance, données massives, lac de données.

Abstract

This article introduces a knowledge graph e-infrastructure, the Metabolomic Semantic Data Lake, for contextualizing metabolomics data. It links biological concepts and annotated scientific literature using AI, addressing annotation gaps. Based on Big Data technologies and the Semantic Web, it offers an API for easy data access. FORVM ChemDisease and FORVM Plants graphs enable the analysis of toxicological pathways and the identification of plant bio-

markers. The aim is to enhance data integration and interpretation in life sciences.

Keywords

Metabolomics, semantic web, IA, knowledge graph, big data, data-lake.

Contextualisation des données

Introduction

Les avancées technologiques dans l'acquisition de données en biologie ont profondément transformé les modalités de la collecte d'informations. En effet, l'instrumentation utilisée en phénotypage moléculaire permet avec une précision toujours plus accrue de la mesure de produire des volumes de données toujours plus importants. Pour la bioinformatique, ces flux croissants nécessitent le développement de nouvelles approches adaptables à grande échelle en traitement et analyse de données. L'intégration de technologies informatiques dans ces approches permet à la communauté scientifique d'aller plus loin dans l'interprétation des résultats des études multimodales. Dans ce contexte, la contextualisation des données "omiques" ou "comment lier ces résultats expérimentaux générés à haut débit aux connaissances scientifiques disponibles dans les publications et revues spécialisées?" est devenue une étape clé dans la compréhension et l'intégration de ces données. Notamment, de par le nombre de publications scientifiques augmentant

considérablement chaque année, la construction d'une bibliographie exhaustive et régulièrement mise à jour est rendue particulièrement difficile et chronophage. Pour relever ce défi, l'étiquetage sémantique des articles scientifiques offre une solution facilitant la recherche bibliographique et ciblant avec précision les thématiques liées aux objets d'étude (taxonomie d'espèces, environnement, conditions expérimentales, ...). Des initiatives telles que celles portées par le "National Institutes of Health" (NIH) ont permis le développement d'ontologies biomédicales, utilisées pour annoter la littérature scientifique. Ainsi, l'ensemble du corpus des Archives des revues biomédicales et des sciences de la vie (NIH PubMed) est annoté régulièrement par des workflows automatisés, sous la supervision d'équipes expertes dans le domaine biomédical. Cependant, ces avancées et les efforts de standardisation restent dépendants des disciplines scientifiques. Ainsi, en science des plantes, par exemple, des approches similaires sont encore trop peu développées. Une autre problématique réside dans la pertinence des annotations : certaines conclusions ou annotations peuvent s'avérer erronées ou obsolètes dans certains articles. Ces limitations sont souvent attribuables à la qualité variable de la production scientifique ou aux erreurs introduites par les processus d'annotation automatisés. Dans ce contexte, nous présentons une e-infrastructure de gestion et d'intégration de données pour la métabolomique permettant de mettre à disposition une fédération de données aux collaborateurs de l'infrastructure nationale MetaboHUB. Son architecture informatique garantit i) l'interopérabilité entre les données métabolomiques et les autres silos et ressources « omiques », ii) la traçabilité des données et des traitements et iii) d'automatiser l'enrichissement des métadonnées des données et la production de bases de connaissances (à partir de publications, bases de métabolites, banques de signatures spectrales). Le travail décrit en suivant s'intéresse plus spécifiquement aux graphes de connaissances du projet FORVM, utiles à la contextualisation des données issues d'expérimentations. Les méthodes développées dans la cadre de ce projet nous permettent d'établir des liens fiables entre des concepts biologiques provenant d'ontologies variées et d'annoter des corpus volumineux de revues scientifiques. La production régulière de ces graphes est rendue possible grâce à une infrastructure informatique "Big Data", dédiée au traitement de données massives au format 'RDF' (Resource Description Framework). Cette infrastructure intègre également des techniques d'intelligence artificielle issues de la famille des modèles de langage de grande taille (LLM), permettant l'annotation automatique de la littérature scientifique. Une interface de programmation d'application (API) a été également développée afin de faciliter l'exploitation de ces graphes. Elle permet notamment à des applications tierces d'exploiter ces connaissances produites en continu notamment au travers de la mise à disposition d'outils et de portails web servant à la contextualisation des données en métabolomique. L'utilité et la plus-value de cette e-infrastructure sont également abordées au travers de cas d'usage illustrant une variété d'applications de ce workflow à des expériences

menées par différentes plate-formes d'analyses métabolomiques de l'infrastructure nationale MetaboHUB.

Étiquetage sémantique de la littérature scientifique

Malgré les avancées significatives dans les sciences biologiques, tous les domaines n'ont pas progressé de manière équitable. Le secteur de la santé humaine bénéficie de ressources spécialisées et d'outils très robustes, tels que les MeSH (Medical Subject Headings)[1], la base de données PubMed et l'identification des substances chimiques pour indexer les sujets des articles. Ces éléments facilitent l'accès aux connaissances scientifiques publiées [9]. En revanche, l'étude des végétaux accuse un certain retard. Des initiatives comme Planteome proposent des ontologies spécialisées (Plant Ontology, Trait Ontology, Plant Stress Ontology) pour standardiser les concepts liés aux plantes [2], mais leur utilisation se limite principalement à l'annotation d'objets biologiques. L'étiquetage sémantique de la littérature scientifique à l'aide d'ontologies pourrait permettre au domaine végétal de développer des outils comparables aux standards établis dans le secteur biomédical. Aujourd'hui, la démocratisation des grands modèles de langage offre une opportunité unique pour dépasser cette limitation. Ces modèles permettent d'analyser massivement les textes grâce à l'augmentation constante des publications scientifiques accessibles en téléchargement via des plateformes telles que [PubMed](#) ou [ISTEX](#), qui constituent le socle de la bibliothèque scientifique numérique nationale. La méthode d'étiquetage sémantique de ces publications scientifiques est basée sur l'utilisation du modèle SBERT (Sentence-BERT) qui est une adaptation du modèle BERT ("Bidirectional Encoder Representations from Transformers"), spécialement conçue pour générer des représentations vectorielles (embeddings) de phrases entières et permettant de comparer efficacement leur similarité sémantique. Il est ainsi possible de comparer le contenu d'un résumé et les définitions des concepts d'une ontologie. En combinant un calcul de similarité avec un seuil prédéfini, il devient alors possible d'automatiser l'étiquetage d'un nombre de textes scientifiques.

Une preuve de concept pour relier métabolites et maladies humaines

Le projet FORVM a consisté en la construction d'un graphe de connaissances innovant qui fédère, à grande échelle, des bases de données en sciences de la vie, ainsi que des corpus issus de la littérature scientifique, afin d'établir des associations entre métabolites et maladies d'origine humaine. En s'appuyant sur les technologies du web sémantique et les techniques avancées d'inférence ontologique, les développements effectués dans FORVM ont permis d'identifier des relations implicites entre concepts biomédicaux et chimiques, facilitant ainsi l'interprétation des signatures métaboliques [5, 4]. Ce graphe comprend pour sa version actuelle plus de 9 milliards de faits et 9 millions de relations inférées. Un portail web (<https://forum-webapp.semantic-metabolomics.fr>) offre aux chercheurs un accès simplifié via une interface intuitive et un point d'accès avancé

permettant d'adresser des requêtes complexes au format SPARQL. En combinant analyses statistiques et exploration sémantique, FORVM aide à relier certains métabolites peu étudiés à des maladies. Ce graphe constitue ainsi un outil majeur pour renforcer l'intégration des données métabolomiques dans les études sur la santé humaine. Ce projet est inscrit dans la feuille de route scientifique de l'infrastructure nationale MetaboHUB et intégré dans le cadre de ses activités bioinformatiques.

Mise en place d'une infrastructure de production avec le "Metabolomic Semantic Data-Lake"

Le "Metabolomic Semantic Data-Lake" est une infrastructure orientée "Big Data" et implémentant un cluster Spark/Hadoop hébergé au centre de données INRAE de Toulouse. Elle met en œuvre une suite de workflows automatisés permettant l'ingestion fiable et régulière de données au format RDF issues de plus d'une vingtaine de bases bio-informatiques utiles au domaine de la métabolomique. Ces workflows s'appuient sur la technologie Apache Airflow et sur le développement de tâches en Python afin d'assurer leur maintenance, ainsi que la mise à jour automatique et continue des données [8]. L'infrastructure repose également sur une API Scala conçue pour simplifier l'exploration et l'utilisation des graphes RDF. Cette API utilise le framework SANSa (Semantic Analytics Stack), un moteur Big Data spécialisé dans le traitement massif des données RDF [10]. Grâce à SANSa, il a été possible d'exploiter sémantiquement les données stockées dans les graphes tout en facilitant leur interrogation, que ce soit en mode interactif (console Spark) ou en mode batch (soumission automatisée). Actuellement, cette infrastructure permet également la génération automatique des nouvelles versions des graphes, en s'appuyant sur la méthodologie FORVM, ce qui assure une mise à jour très régulière des informations de contextualisation de données pour les plateformes de métabolomique partenaires du projet.

Mise à disposition des données contextuelles

Les graphes de connaissances FORVM, développés au sein de notre infrastructure « Metabolomic Semantic Data-Lake » permettent de contextualiser des études sur des modèles biologiques, allant de l'humain aux cultivars. Dans ces graphes, les entités sont reliées par une unique relation nommée « forvm-related », dont la validité repose sur plusieurs tests statistiques. Cette approche garantit la fiabilité des liens établis entre les différentes entités. La richesse des graphes FORVM réside dans le type RDF (`rdf:type`) des entités impliquées dans la relation « forvm-related ». Le graphe FORVM ChemDisease propose ainsi des associations entre des bases de données telles que PUBCHEM et MeSH, CHEBI et MeSH, CHEMONT et MeSH, ainsi qu'entre différents termes MeSH. De manière analogue, le graphe FORVM Plants relie des entités issues de PUBCHEM, CHEBI, CHEMONT et PLANTEOME, permettant ainsi de couvrir un large spectre d'interactions biologiques et métaboliques. Pour chaque relation « forvm-related »,

des informations détaillées sont fournies à l'aide de l'ontologie STATO (STATistics Ontology). Ces informations comprennent la cooccurrence, c'est-à-dire la fréquence à laquelle les deux entités sont mentionnées ensemble dans la littérature scientifique, ainsi que des mesures statistiques telles que la p-value et la q-value issues du test exact de Fisher, le ratio de cotes pour quantifier l'intensité de l'association, la p-value ajustée du test χ^2 de Yates, et le seuil de l'indice de fragilité, qui indique le nombre minimal d'articles à retirer pour que l'association ne soit plus statistiquement significative. Grâce à cette structuration, les graphes FORVM permettent une exploration des interactions métaboliques et biologiques, en s'appuyant sur des ontologies et des référentiels. À titre d'exemple, dans le graphe FORVM ChemDisease, les termes MeSH "Fluoresceins" (D005452) et « Xanthenes » (CHEBI :38835) sont statistiquement associés et formalisés par une relation RDF "forvm-related". La cooccurrence de citation des deux termes atteint 15151 (Données 2025). L'analyse statistique par le test exact de Fisher (stato :0000073) donne une q.value (obi :0001442) de $4,26 \times 10^{-290}$ et un ratio de cotes (stato :0000182) de 22,11, ce qui indique une association extrêmement significative entre ces deux concepts. Un autre exemple issu du même graphe concerne le concept MeSH « TGF-beta Superfamily Proteins » (D055411), relié au concept MeSH « Osteoporosis, Postmenopausal » (D015663), illustrant la capacité du graphe à formaliser des liens pertinents entre entités biomédicales. Dans le graphe FORVM Plants, une association statistique similaire est observée entre le terme Planteome « whole plant fruit ripening stage » (PO :0007010) et le composé « Ascorbic acid » (CHEBI :22652). Avec les objectifs de garantir l'interopérabilité applicative de nos ressources, d'offrir des accès simplifiés aux scientifiques mais également pour intégrer ces graphes dans des applications tierces, un outil basé sur le standard OpenAPI est en cours de développement. Ce middleware permettra de rendre accessible une sélection des informations FORVM, dans un format standardisé, facilitant leur intégration par exemple dans des workflows de traitement haut débit comme sous la plateforme Workflow4Metabolomics [7] ou dans de nouveaux portails web dédiés à des familles spécifiques de maladies.

Résultats préliminaires

L'infrastructure FORVM permet aujourd'hui de mettre à disposition de la communauté scientifique un panel de ressources publiquement accessibles. Il comprend notamment un workflow complet de génération et d'exploitation de graphes (Figure 1) et plusieurs graphes de connaissances aidant à la contextualisation des données métabolomiques dans les domaines de la santé et des sciences des plantes. Le graphe de connaissances FORVM ChemDisease, consolidé depuis 2021, contient des millions d'associations entre composés chimiques et concepts biomédicaux, offrant une structure robuste pour explorer les relations complexes entre les composés chimiques, les changements métaboliques et les problèmes de santé. Ce graphe constitue un outil puissant pour analyser les mécanismes

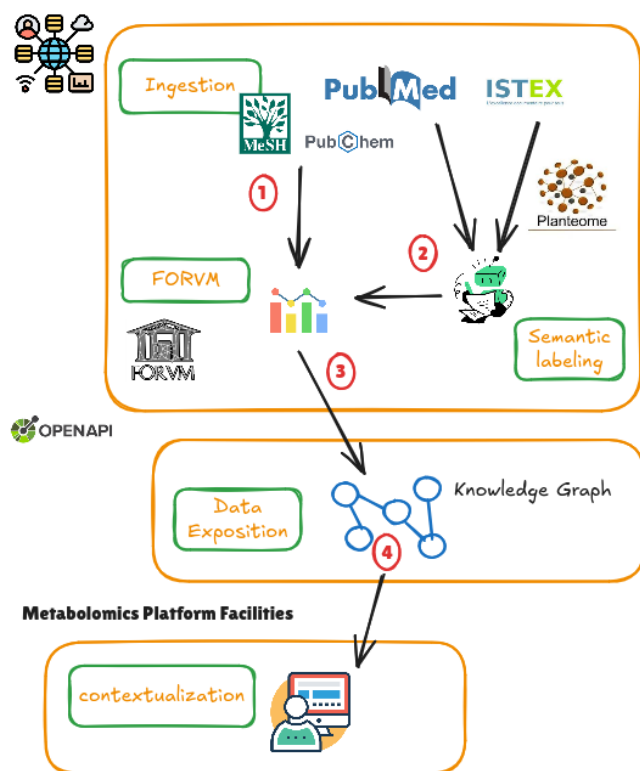


FIGURE 1 – Workflow de génération et d'exploitation des graphes de connaissances FORVM

biologiques sous-jacents et identifier des cibles potentielles dans le domaine de la santé humaine. Le graphe de connaissances expérimental (2025) *FORVM Plants* associe les composés chimiques et les termes des ontologies "Plantome", qui décrivent l'anatomie, la morphologie et le développement des plantes ainsi que leurs traits biochimiques, morphologiques, de croissance, de qualité, de stress et de rendement. En liant ces informations aux composés chimiques, ce graphe propose une approche innovante pour l'identification et la contextualisation de biomarqueurs végétaux dans diverses conceptions expérimentales. Il facilite ainsi la compréhension des interactions complexes entre les plantes et leur environnement chimique. Les associations contenues dans ces deux graphes peuvent être interrogées via des portails web dédiés. Le portail *FORVM ChemDisease* (<https://forum-webapp.semantic-metabolomics.fr>) est une application spécifique qui a été développée dans le cadre de la publication de la méthodologie *FORVM* [5]. Le portail *FORVM Plants* est également accessible publiquement (<https://askomics-forum-plants-192-168-100-118.vm.openstack.genouest.org/>). Il repose sur le système d'information AskOmics [6], qui permet une intégration rapide des données RDF grâce à ses interfaces intuitives. Ce portail offre à l'utilisateur, un éditeur visuel de requêtes et de consultation des résultats, adapté à des utilisateurs novices sur les technologies du web sémantique. Enfin, ces portails donnent accès à des "end points SPARQL" (<https://forum.semantic-metabolomics.fr/sparql/> et <https://askomics-forum-plants-192-168-100-118.vm.openstack.genouest.org/sparql>). Ces "endpoints" permettent d'exploiter pleinement les données disponibles grâce à des requêtes personnalisées, ouvrant ainsi la voie à l'exploration de questions scientifiques plus complexes.

Applications expérimentales

Le graphe de connaissances *FORVM ChemDisease* permet le développement d'un pipeline pour simplifier l'analyse des voies toxicologiques pour les études biologiques. Une application notable de cette méthodologie est le projet *Endoxomics*, qui vise à identifier les événements clés reliant les polluants organiques persistants (POPs) à la physiopathologie de l'endométriose [11]. Le graphe de connaissances *FORVM Plants*, quant à lui, permet d'explorer les relations complexes entre données biologiques et chimiques dans trois projets. Le premier se concentre sur la réponse au stress hydrique chez les Brassicacées, en identifiant des biomarqueurs spécifiques liés à leur adaptation physiologique. Le deuxième projet vise à caractériser le métabolome de *Cannabis sativa* présentant des ratio THC/CBD contrastés, en examinant les voies métaboliques et les interactions moléculaires pour mieux comprendre ses propriétés thérapeutiques et psychoactives [12]. Enfin, le troisième projet étudie les mécanismes de régulation du métabolisme redox au cours du développement et de la maturation des fruits de tomate [3], avec pour objectif d'identifier des biomarqueurs clés associés à la qualité et à la santé des fruits.

Remerciements

Nous remercions l'infrastructure MetaboHUB financée par l'Agence Nationale de la Recherche dans le cadre du programme France 2030 (MetaboHUB ANR-11-INBS-0010; MetEx+ ANR-21-ESRE-0035; MetaboHUB (JVCE) ANR-24-INBS-0012) pour le financement d'un contrat de 18 mois pour un ingénieur d'étude, ainsi que pour les moyens humains et matériels mobilisés dans le développement du projet. Nous exprimons également notre gratitude envers INRAE et son Schéma Directeur du Numérique, le département ALIMH, à l'unité Ingenium et à la DiPSO pour les financements initiaux du projet *Metabolomic Semantic Datalake*, ainsi qu'aux CATI *eMPREInTE*, *PRODOSIE* et *BARIC* pour leur rôle essentiel dans la réalisation, le maintien et l'évolution de l'infrastructure matérielle du cluster "Big Data" utilisé dans ce projet.

Références

- [1] Barbara Bushman, David Anderson, and Gang Fu. Transforming the Medical Subject Headings into Linked Data : Creating the Authorized Version of MeSH in RDF. *Journal of Library Metadata*, 15(3-4) :157–176, 2015.
- [2] Laurel Cooper, Austin Meier, Marie-Angélique Laporte, Justin L. Elser, Chris Mungall, Brandon T. Sinn, Dario Cavaliere, Seth Carbon, Nathan A. Dunn, Barry Smith, Botong Qu, Justin Preece, Eugene Zhang, Sinisa Todorovic, Georgios Gkoutos, John H. Doonan,

- Dennis W. Stevenson, Elizabeth Arnaud, and Pankaj Jaiswal. The Planteome database : an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, 46(D1) :D1168–D1180, January 2018.
- [3] Guillaume Decros, Thomas Dussarrat, Pierre Baldet, Cédric Cassan, Cécile Cabasson, Martine Dieuaide-Noubhani, Alice Destailleur, Amélie Flandin, Sylvain Prigent, Kentaro Mori, Sophie Colombié, Joana Jorly, Yves Gibon, Bertrand Beauvoit, and Pierre Pétriaccq. Enzyme-based kinetic modelling of ASC–GSH cycle during tomato fruit development reveals the importance of reducing power and ROS availability. *New Phytologist*, 240(1) :242–257, 2023. _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.19160>.
- [4] Maxime Delmas, Olivier Filangi, Christophe Duperier, Nils Paulhe, Florence Vinson, Pablo Rodriguez-Mier, Franck Giacomoni, Fabien Jourdan, and Clément Frainay. Suggesting disease associations for overlooked metabolites using literature from metabolic neighbors. *GigaScience*, 12 :giad065, January 2023.
- [5] Maxime Delmas, Olivier Filangi, Nils Paulhe, Florence Vinson, Christophe Duperier, William Garrier, Paul-Emeric Saunier, Yoann Pitarch, Fabien Jourdan, Franck Giacomoni, and Clément Frainay. FORUM : building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics*, 37(21) :3896–3904, November 2021.
- [6] Xavier Garnier, Anthony Bretaudeau, Olivier Filangi, Fabrice Legeai, Anne Siegel, and Olivier Dameron. AskOmics, a web tool to integrate and query biological data using semantic web technologies. In *JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques*, page 1, Lille, France, July 2017.
- [7] Franck Giacomoni, Gildas Le Corguillé, Misharl Monsoor, Marion Landi, Pierre Pericard, Mélanie Pétera, Christophe Duperier, Marie Tremblay-Franco, Jean-François Martin, Daniel Jacob, Sophie Goulitquer, Etienne A. Thévenot, and Christophe Caron. Workflow4Metabolomics : a collaborative research infrastructure for computational metabolomics. *Bioinformatics*, 31(9) :1493–1495, May 2015.
- [8] Guillaume Laisney, David Benaben, Christophe Duperier, Matéo Boudet, Franck Giacomoni, and Olivier Filangi. Metabolomic Semantic Datalake : A Scalable Approach to Managing Metabolomics Semantic Resources. In *JOBIM 2024 - Journées Ouvertes en Biologie, Informatique et Mathématiques*, page 1, June 2024.
- [9] Robert Leaman, Rezarta Islamaj, Virginia Adams, Mohammed A Alliheedi, João Rafael Almeida, Rui Antunes, Robert Bevan, Yung-Chun Chang, Arslan Erdengasileng, Matthew Hodgskiss, Ryuki Ida, Hyunjae Kim, Keqiao Li, Robert E Mercer, Lukrécia Mertová, Ghadeer Mobasher, Hoo-Chang Shin, Mujeen Sung, Tomoki Tsujimura, Wen-Chao Yeh, and Zhiyong Lu. Chemical identification and indexing in full-text articles : an overview of the NLM-Chem track at BioCreative VII. *Database*, 2023 :baad005, January 2023.
- [10] Jens Lehmann, Gezim Sejdiu, Lorenz Bühmann, Patrick Westphal, Claus Stadler, Ivan Ermilov, Simon Bin, Nilesh Chakraborty, Muhammad Saleem, Axel-Cyrille Ngomo Ngonga, and Hajira Jabeen. Distributed Semantic Analytics using the SANS Stack. In *Proceedings of 16th International Semantic Web Conference - Resources Track (ISWC'2017)*, pages 147–155. Springer, 2017.
- [11] Meije Mathé, Guillaume Laisney, Olivier Filangi, Maxime Delmas, Christophe Duperier, German Cano Sancho, Fabien Jourdan, Franck Giacomoni, and Clément Frainay. Extracting targeted sub networks from FORUM for toxicological insight. In *JOBIM 2024 - Journées Ouvertes en Biologie, Informatique et Mathématiques*, June 2024.
- [12] Pedro G. Vásquez-Ocmín, Guillaume Marti, Maxime Bonhomme, Fabienne Mathis, Sylvie Fournier, Stéphane Bertani, and Alexandre Maciuk. Cannabinoids vs. whole metabolome : Relevance of cannabinomics in analyzing cannabis varieties. *Analytica Chimica Acta*, 1184 :339020, 2021.