Metabolomic Semantic Datalake : A Scalable Approach to Managing Metabolomics Semantic Resources

Guillaume Laisney, D. Benaben, Christophe Duperier, M. Boudet, Franck Giacomoni, O. Filangi

Metabolomic Semantic Datalake : A Scalable Approach to Managing Metabolomics Semantic Resources

G. LAISNEY[1,2], D. BENABEN[3,4], C. DUPERIER[5], M. BOUDET[2,6], F. GIACOMONI[5] and O. FILANGI[1,2]

Corresponding author : guillaume.laisney@inrae.fr

1 Metabolic Profiling and Metabolomic Platform (P2M2), BIA-IGEPP, Le Rheu, France, **2** INRAE, UMR1349 Institute for Genetics, Environment and Plant Protection (IGEPP), Institut Agro, Université Rennes, Le Rheu, France, **3** INRAE, Univ. Bordeaux, UMR BFP, Villenave d'Ornon, France, **4** Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, Villenave d'Ornon, France, **5** Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand, France, **6** INRIA, IRISA, GenOuest Core Facility, Campus de Beaulieu, Rennes, France

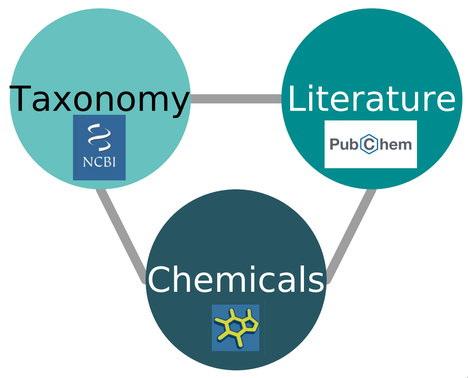
CONTEXT

RDF knowledge graphs enable interoperability; databases published in this format can be combined and explored as a single entity. However, querying these massive assembled graphs requires a powerful and scalable computing architecture.

EXAMPLE :

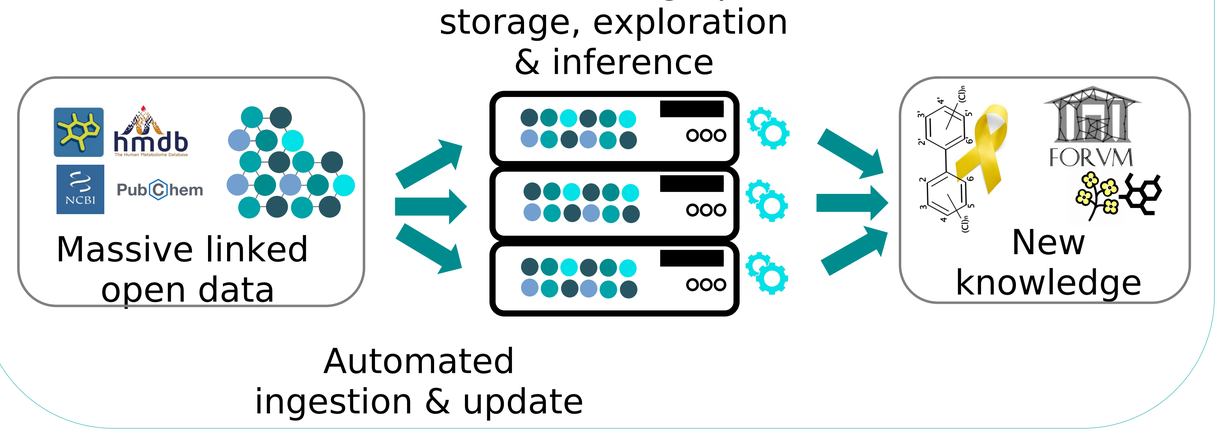
NCBI Taxonomy.
PubChem Literature.
ChEBI molecular entities.

These graphs can be explored as one single database, but they hold more than 3.5×10^9 edges.



GOAL

The goal of the "Metabolomic Semantic Datalake" (MSD) is to assess a scalable distributed processing architecture designed to explore large-scale knowledge graphs.



RESULTS

The Spark/Hadoop Big Data cluster at the INRAE data center in Toulouse can be securely accessed through an encrypted connection. Continuous Integration and Deployment (CI/CD) ensure smooth operation and enable easy scaling.

A suite of workflows has been created, allowing for knowledge graphs ingestion into a distributed Hadoop file system (HDFS) using Apache Airflow and a custom Python library.

A Scala library offers exploration capabilities for the extensive graph collection and simplifies their loading and usage within the Spark and SANSA[1] framework.

Distributed metadata extraction techniques improve comprehension of semantic data available in the data lake and facilitate retrieval of graphs based on specific properties or vocabularies used.

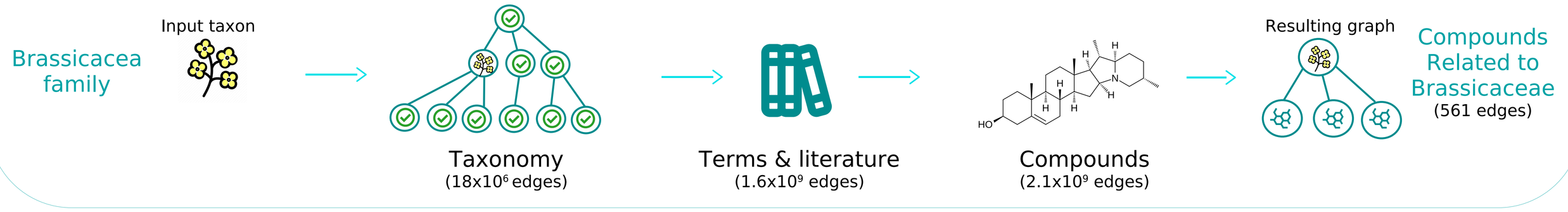
Code is available at :
https://github.com/eMetaboHUB/MSDDBM

EXAMPLE APPLICATION

An example use case is the extraction from literature of compounds related to a given taxon, such as a species or family.

Our distributed program begins by searching the NCBI taxonomy: it moves up a specified number of ranks above the target taxon, then down to identify all related taxa. For example, we currently look for relatives of the Brassicaceae family by searching the Brassicales order. Next, it explores the literature for articles mentioning these taxa in their titles. The metadata from these articles leads to potential metabolites of the target taxon and can be added to a dereplication database.

Apache Airflow runs this task automatically after each update of the source graphs and sends an alert when new results are found.



ONGOING WORK

The FORUM[2] knowledge graph, comprising nine billion triplets, will stay up-to-date through an automated workflow triggered by source data updates.

A thesis is currently underway at the Laboratory of Food Toxicology of the UMR TOXALIM in Toulouse, whose topic is the creation and exploitation of a knowledge graph to elucidate links between persistent organic pollutants and endometriosis. This work will leverage the opportunities offered by the Metabolomic Semantic Datalake.

[1] Lehmann, J. et al. Distributed Semantic Analytics using the SANSA Stack. Proceedings of 16th International Semantic Web Conference - Resources Track (ISWC'2017), 147–155 (2017).
[2] Delmas, M. et al. FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. Bioinformatics 37, 3896–3904 (2021).