When Graphs Become Too Large

FORVM, Semantic Annotation and Multigraph Quotients *at Scale*

Team : David BENABEN, Thomas BELLEMBOIS, Kevin BILLET, Alain BOUCHEREAU, Matéo BOUDET, Cécile CABASSON, Younes DELLERO, Maxime DELMAS, Olivier FILANGI, Clément FRAINAY, Franck GIACOMONI Nicolas GUILHOT, Yann GUITTON, Antoine MAHUL, Guillaume MARTI, Meije MATHE, Nils PAULHE, Sylvain PRIGENT, Faustine SOUC, Florence Vinson, Magalie WEBER

Presented by : Guillaume LAISNEY
guillaume.laisney@inrae.fr

Classic data processing pipeline for a metabolomic study

Chromatography

MS

NMR

Signal treatment / Bioinformatics / Statistics

Biology / Modelisation

Annotation

Identification

Biological interpretation

Extraction Data

Statistical analysis

Data acquisition

Analytical Chemistry

Mapping & visualization

Metabolic DB

Enrichment analysis

tyrosine metabolism

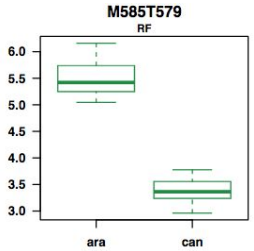
EC 1.3.99.1

...

carboxylic acid

p-value

Metabolite annotations



Brief Bioinform, Volume 14, Issue 6, November 2013, Pages 737–744, https://doi.org/10.1093/bib/bbs055

Source : personnel – Ecole chercheur W4E Bruxelles 2020

The FORVM Project – Research part

Potentialities of Knowledge Representation and Automated Reasoning Methods

Methods to extract associations between compound and biomedical concepts, using literature metadata enrichment

› Bioinformatics. 2021 Nov 5;37(21):3896-3904. doi: 10.1093/bioinformatics/btab627.

FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases

Maxime Delmas [1], Olivier Filangi [2], Nils Paulhe [3], Florence Vinson [1], Christophe Duperier [3], William Garrier [4], Paul-Emeric Saunier [4], Yoann Pitarch [5], Fabien Jourdan [1], Franck Giacomoni [3], Clément Frainay [1]

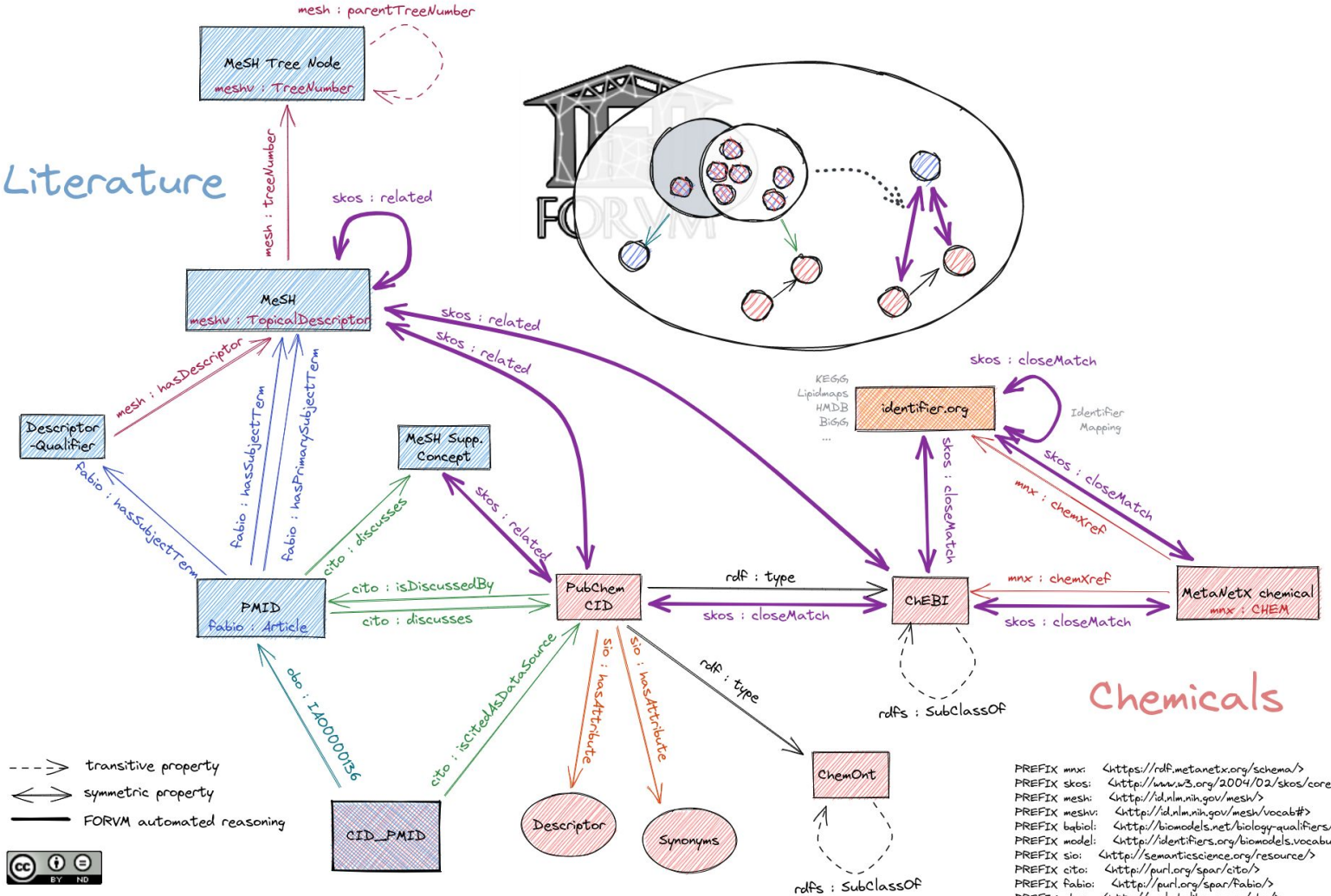
OXFORD (GIGA)n SCIENCE

GigaScience, 2023, 12, 1–13
DOI: 10.1093/gigascience/giad065
Research

Suggesting disease associations for overlooked metabolites using literature from metabolic neighbors

Maxime Delmas [1], Olivier Filangi [2], Christophe Duperier[3], Nils Paulhe [3], Florence Vinson [1,4], Pablo Rodriguez-Mier [1], Franck Giacomoni [3], Fabien Jourdan [1,4] and Clément Frainay [1,*]

[1]Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, 31300 Toulouse, France
[2]IGEPP, INRAE, Institut Agro, Université de Rennes, Domaine de la Motte, 35653 Le Rheu, France
[3]Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France
[4]MetaboHUB-Metatoul, National Infrastructure of Metabolomics and Fluxomics, Toulouse, 31300, France
*Correspondence address. Clément Frainay, INRAE TOXALIM UMR 1331 180 chemin de Tournefeuille - BP.93173 F-31027 TOULOUSE cedex 3, France.
Tel: +33 582066314; E-mail: clement.frainay@inrae.fr


mesh : parentTreeNumber
MeSH Tree Node
meshv : TreeNumber
mesh : treeNumber
Literature
skos : related
MeSH
meshv : TopicalDescriptor
skos : related
skos : related
mesh : hasDescriptor
Descriptor
-Qualifier
fabio : hasSubjectTerm
fabio : hasSubjectTerm
fabio : hasPrimarySubjectTerm
MeSH Supp.
Concept
cito : discusses
skos : related
PMID
fabio : Article
cito : isDiscussedBy
cito : discusses
PubChem
CID
rdf : type
skos : closeMatch
ChEBI
skos : closeMatch
identifier.org
KEGG
Lipidmaps
HMDB
BiGG
...
Identifier
Mapping
skos : closeMatch
skos : closeMatch
mnx : chemXref
mnx : chemXref
skos : closeMatch
MetaNetX chemical
mnx : CHEM
rdfs : SubClassOf
Chemicals
obo : IAO0000136
cito : isCitedAsDataSource
sio : hasAttribute
sio : hasAttribute
rdf : type
ChemOnt
CID_PMID
Descriptor
Synonyms
rdfs : SubClassOf
transitive property
symmetric property
FORVM automated reasoning
PREFIX mnx: <https://rdf.metanetx.org/schema/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX biomodels: <http://biomodels.net/biology-qualifiers/>
PREFIX model: <http://identifiers.org/biomodels.vocabulary#/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX obo: <http://purl.obolibrary.org/obo/>

FORVM Computation

FORVM Computation

9 billion input triples (graph edges)
+ reasoning
+ complex SPARQL queries
+ statistics

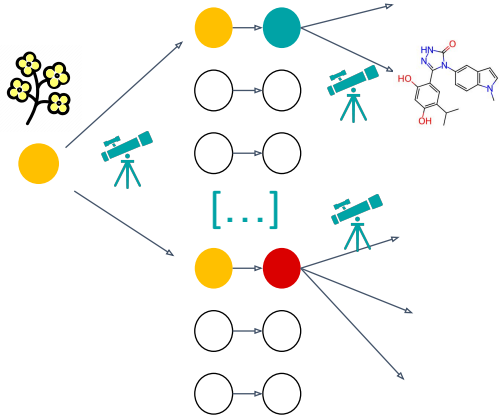
-> requires a large single machine dedicated server + triple store

-> Three weeks computation, punctuated by manual interventions due to errors

Graphs / Scalability / Big Data ?

1 billion triples -> approx. 100 GB
Storing RDF triples is relatively easy !

The real cost comes from exploration
as exploration often translates into long join pipelines



INRAE

The Metabolomics Semantic Datalake Project

Metabolomics Semantic Datalake

- **Big Data Cluster**
- **Automatised knowledge graph ingestion + update**
- **“Easy” distributed KG exploration**
- **Infrastructure As Code**

OpenStack cloud
Scalable infrastructure

Funded by SaPIs 2021 / ALIMH Department / MetaboHub / DIPSO / CATIs EMPREINTE, BARIC & PROSODIE

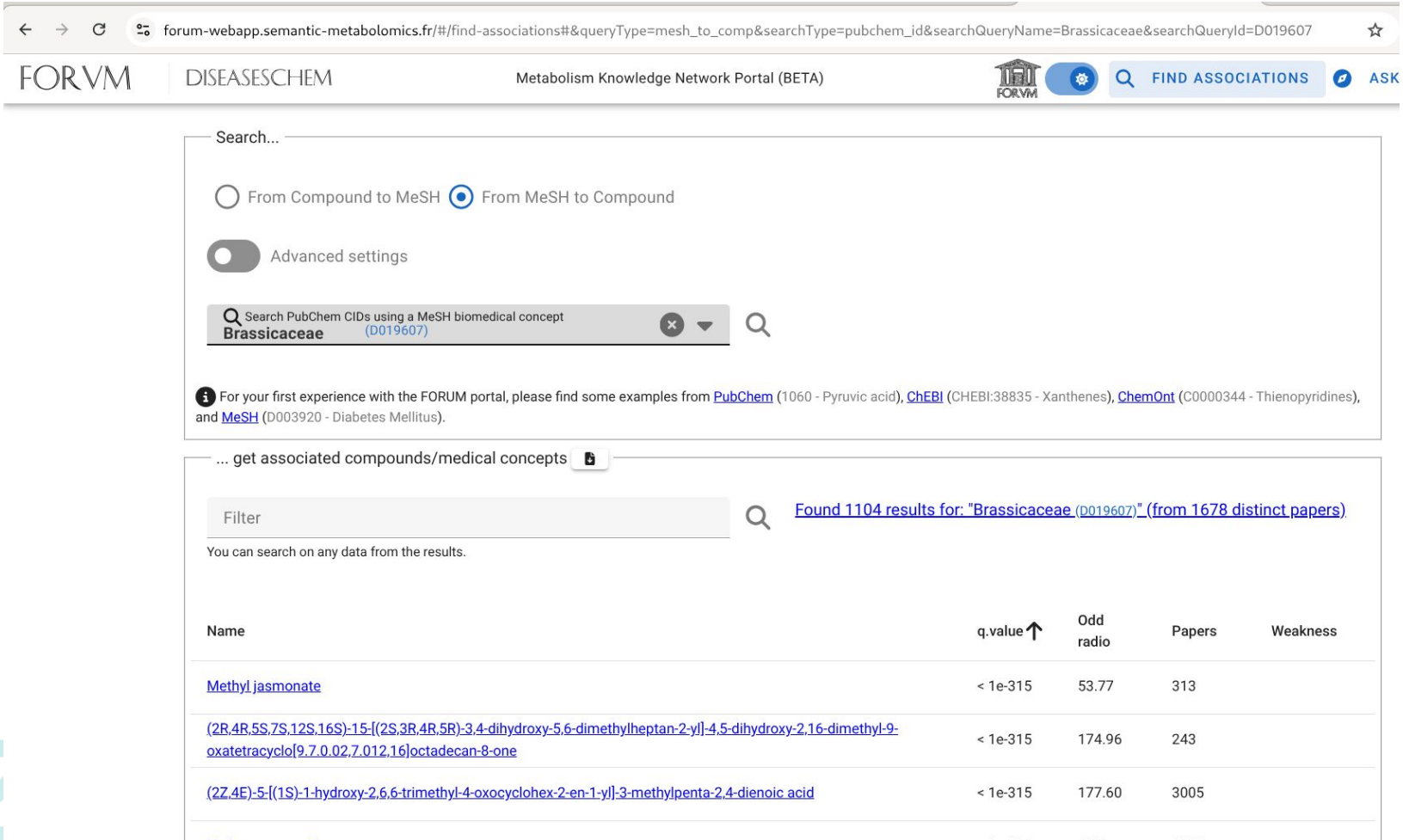
FORVM Graph Distributed Computation

9 billion triples (graph edges)
+ reasoning
+ complex SPARQL queries

-> ~~3 weeks~~ 34h computation, without any manual intervention
(triggered by PubChem releases)

-> New FORVM releases are automatically pushed to an
S3 bucket for publication

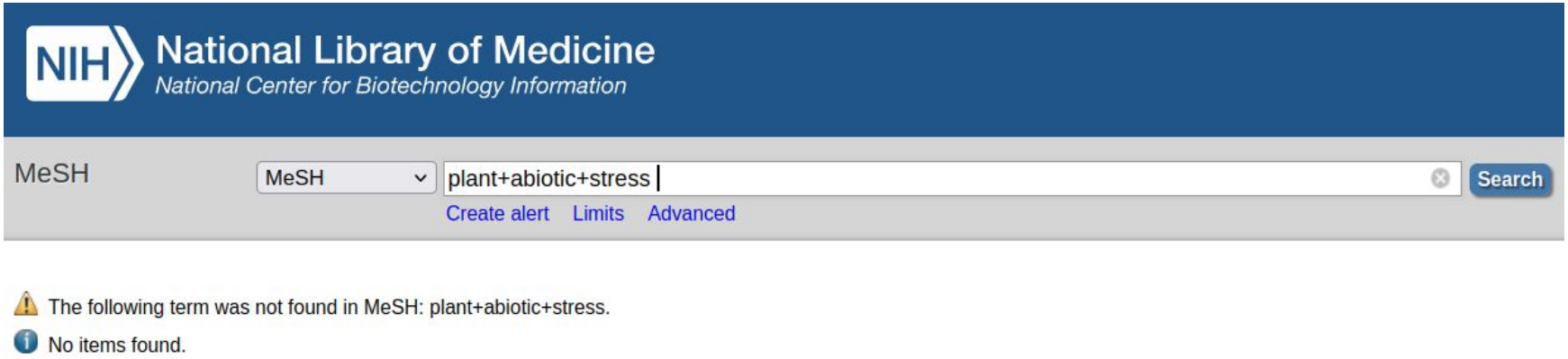
FORVM Graph *Distributed* Computation



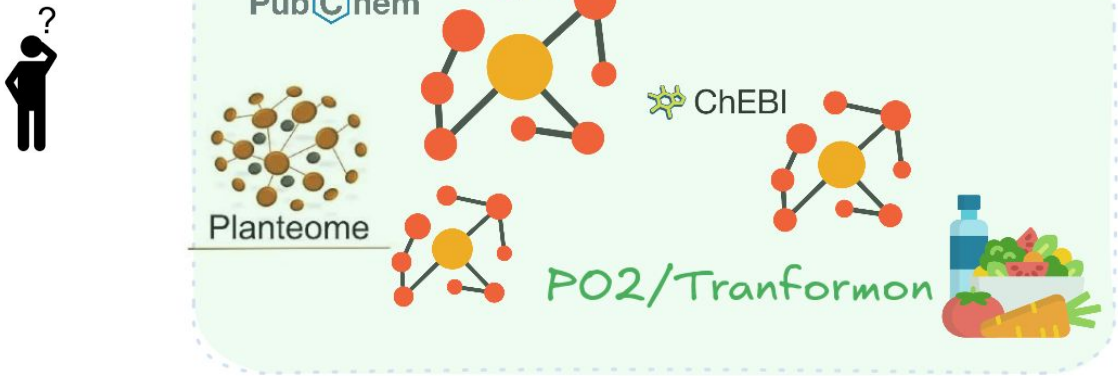
FORVM For Plants ?

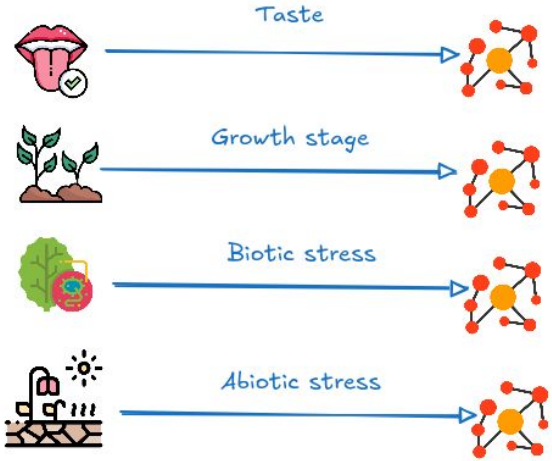
FORVM For Plants ?

- Lack of Annotations (Labelling!) in Scientific Literature on Plant -Omics Studies
- Use of a Domain-Specific Controlled Vocabulary in Biomedicine
  (MeSH : Medical Subject Headings)



Semantic Tagging of Scientific Literature





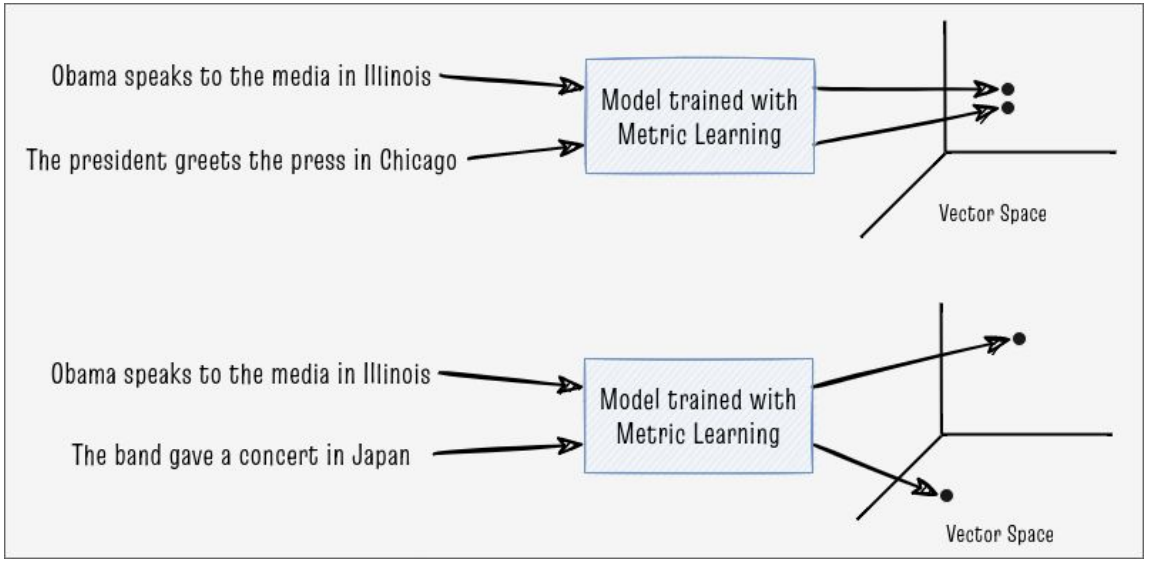
Plant Ontology (PO): Describes plant anatomy, morphology, and developmental stages.
Plant Trait Ontology (TO): Covers the phenotypic characteristics of plants.
Plant Stress Ontology (PSO): Covers biotic and abiotic stresses in plants.

PO2/TransformON, an ontology for data integration on food, feed, bioproducts and biowaste engineering

Semantic Tagging of Scientific Literature

Using Semantic Textual Similarity (STS)

Attention (Vaswani et al., 2017): a mechanism that models dependencies between tokens, allowing models to capture global contexts.



Semantic Tagging of Scientific Literature

Pipeline : Semantic Textual Similarity (STS)

PO:0025235
leaf stomatal pore

Ontological Concept

DEFINITION of « **leaf stomatal pore** » : A phyllome stomatal pore that is part of a leaf stomatal complex

| 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |

0.72

0.12

0.26

Cosine similarity

| 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |

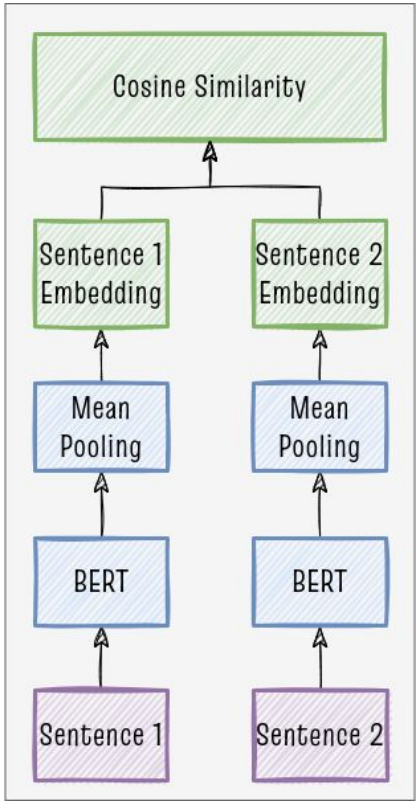
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |

| 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |

Abstract
+ title

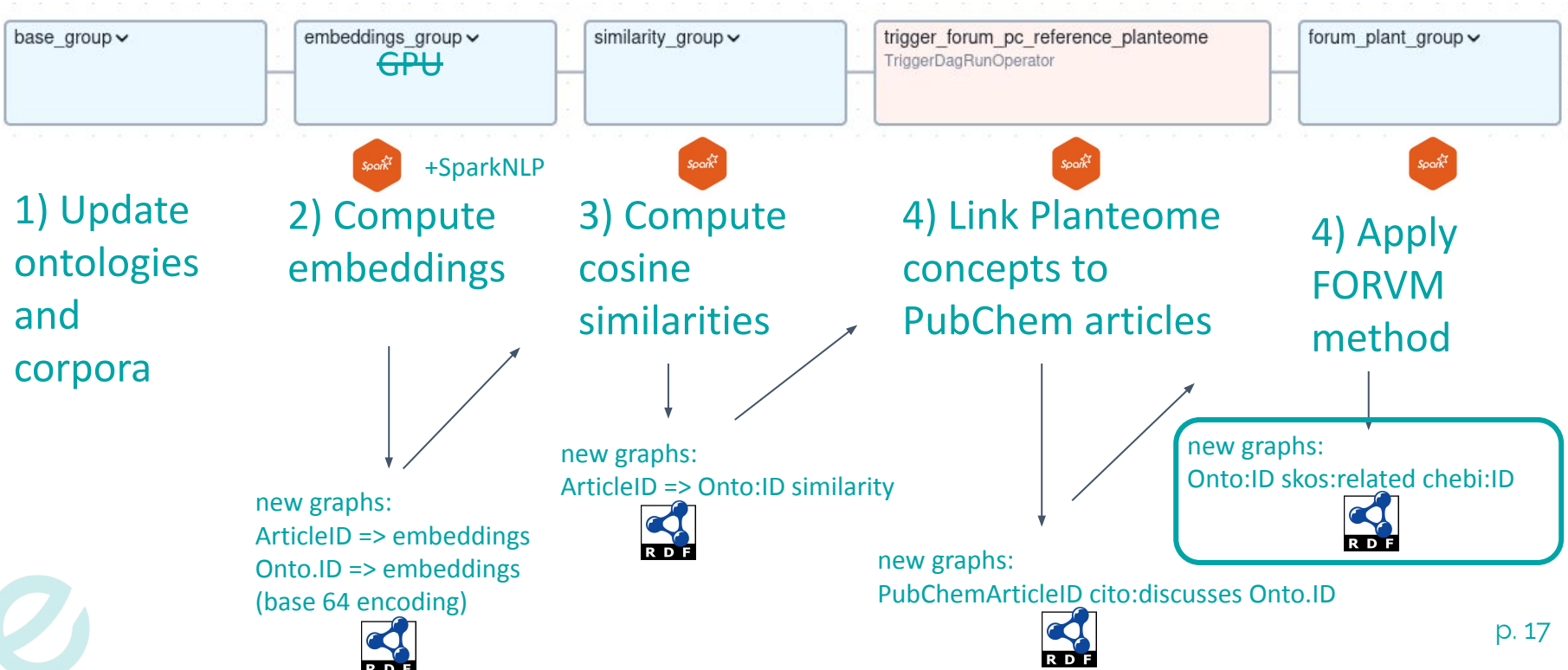
Title : *The structural correlations and the physiological functions of stomatal morphology and leaf structures in C3 annual crops*

Abstract : *A comprehensive study of the correlations between the structural traits and on their relationships with gas exchange parameters may provide some useful information into leaf development and improvement in efficiencies of photosynthetic CO2 fixation and transpirational water loss. In the present study, nine plant materials from eight crop species were pot grown in a growth chamber. ..*

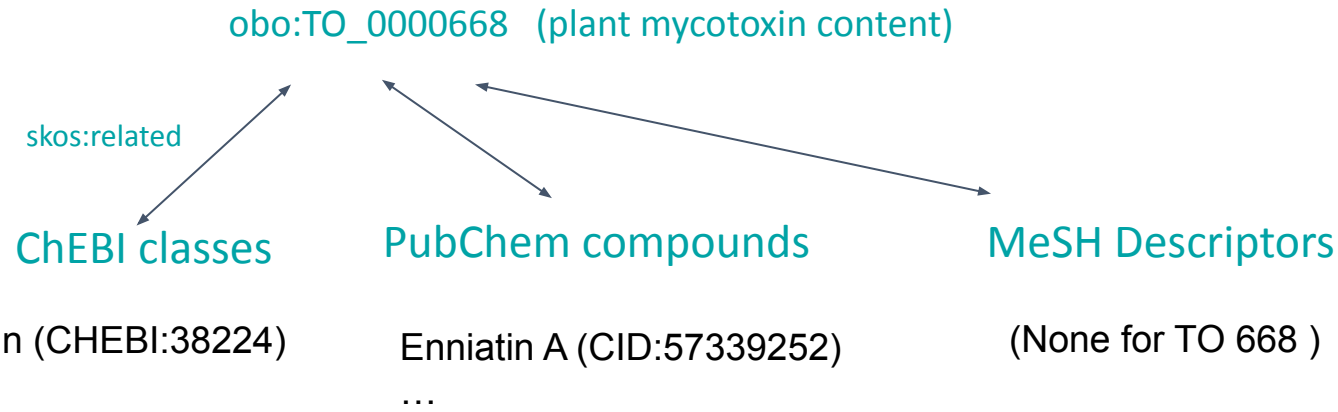


Semantic Tagging of Scientific Literature

Pipeline : Semantic Textual Similarity (STS)



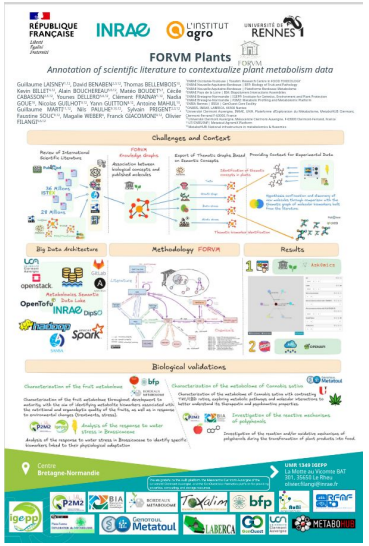
Forum Plants Example Result



Digit-Bio Seed Project (2026-2027)

2026

- Refine semantic textual similarity pipeline and provide performance indicators
- Add new corpora derived from ISTEX dedicated to plants and food



2027

Characterization of the fruit metabolome

Characterization of the fruit metabolome throughout development to maturity, with the aim of identifying metabolite biomarkers associated with the nutritional and organoleptic quality of the fruits, as well as in response to environmental changes (treatments, stress).

Analysis of the response to water stress in Brassicaceae

Analysis of the response to water stress in Brassicaceae to identify specific biomarkers linked to their physiological adaptation

Characterization of the metabolome of Cannabis sativa

Characterization of the metabolome of Cannabis sativa with contrasting THC/CBD ratios, exploring metabolic pathways and molecular interactions to better understand its therapeutic and psychoactive properties.

Investigation of the reactive mechanisms of polyphenols

Investigation of the reaction and/or oxidative mechanisms of polyphenols during the transformation of plant products into food.

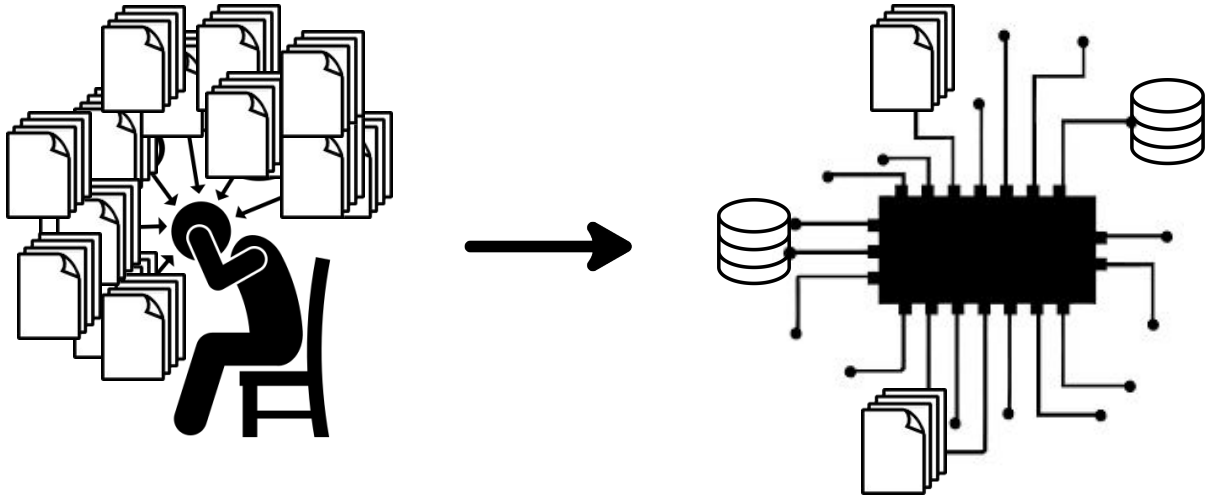
MetabolinkAI WP 2.1
Decentralized Knowledge Representation and Management

anr

Swiss National Science Foundation

The Challenge

metabolomics data interpretation = primary bottleneck



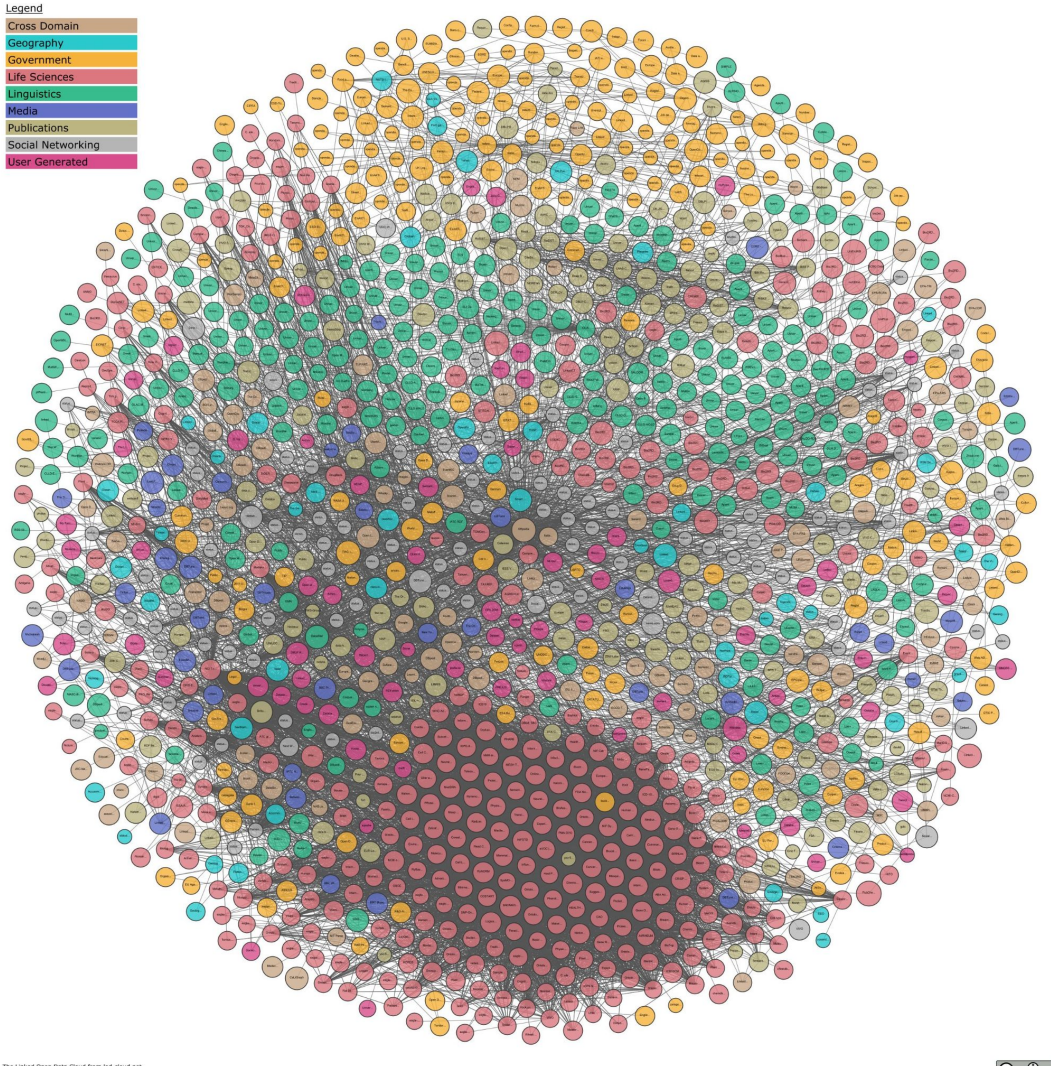
Objective: an AI assistant backed by a “Metabolomics Knowledge Hub”

WP2 Sub-Challenge

Charting the L.O.D (Linked Open Data)

At web scale, WebSem resources often becomes data without a manual:

- What is it about? (scope/domain)
- How is it modeled? (missing schema/structural description)
- How does it connect? (few explicit links to external graphs)



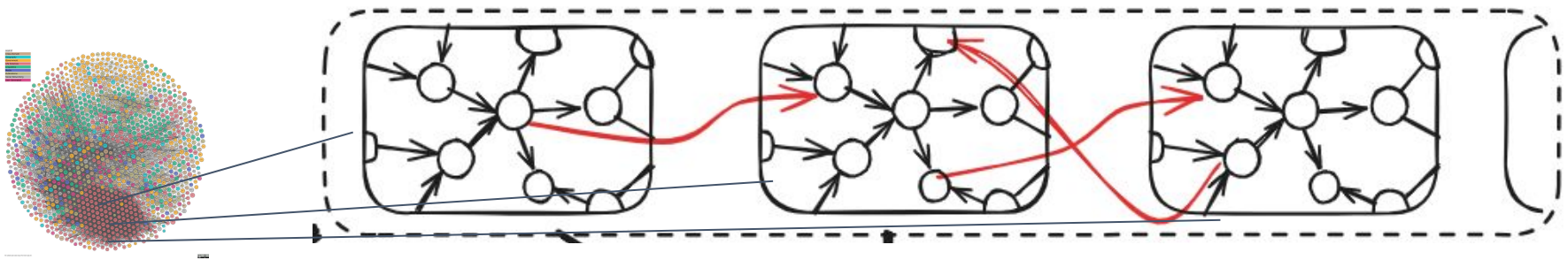
INRAE

Queryability Across Many Graphs

Given a set of accessible graphs :

-> we want to infer what questions are answerable,

-> ideally, machines should be able to suggest and compose queries



INRAE

Summarising RDF Graphs : Existing Approaches

DistLODStats

DistLODStats: Distributed Computation of RDF Dataset Statistics

Gezim Sejdiu[1], Ivan Ermilov[2], Jens Lehmann[1,3] and Mohamed Nadjib Mami[1,3]

[1] Smart Data Analytics, University of Bonn, Germany
sejdiu@cs.uni-bonn.de, jens.lehmann@cs.uni-bonn.de, mami@cs.uni-bonn.de
[2] Department of Computer Science, University of Leipzig, 04109 Leipzig, Germany
iermilov@informatik.uni-leipzig.de
[3] Fraunhofer IAIS, Germany
jens.lehmann@iais.fraunhofer.de, mohamed.nadjib.mami@iais.fraunhofer.de

Resource type	Software Framework
Website	http://sansa-stack.net/distlodstats/
Permanent URL	https://doi.org/10.6084/m9.figshare.6080711

DistLODStats

- Very useful,
- scalable (!)
- We've tested it on 60+ graphs,
- we had to debug it though…
- Summaries lack structural information

```
<http://stats.lod2.eu/rdf/void/?source=/data/pubchem_reference_v2025-08-02>

void:classes 1;
void:classPartition [
void:class <http://rdf.ncbi.nlm.nih.gov/pubchem/vocabulary#Reference>;
void:triples 42397590;
];
void:entities  78860615;
void:distinctSubjects  42399021;
void:distinctObjects  36461571;
void:properties 23;
void:propertyPartition [
void:property <http://purl.org/spar/fabio/hasSubjectTerm>;
void:triples 268288707;
],[
void:property <http://purl.org/dc/terms/creator>;
void:triples 194673970;
],[
void:property <http://purl.org/spar/fabio/hasPrimarySubjectTerm>;
void:triples 117750347;
],[
void:property <http://rdf.ncbi.nlm.nih.gov/pubchem/vocabulary#discussesAsDerivedByTextMining>;
void:triples 108175139;
], …
```

INRAE

https://forge.inrae.fr/mss/metabolomics-semantic-stack/-/blob/main/apps/msdTools/src/main/scala/fr/inrae/msdtools/MsdRDFStatistics.scala

RDFQuotient

RDF graph summarization for first-sight structure discovery

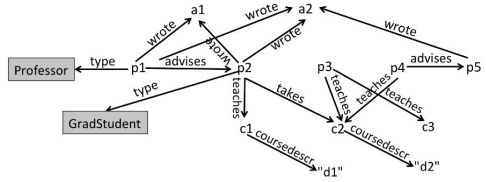
François Goasdoué, Pawel Guzewicz, Ioana Manolescu

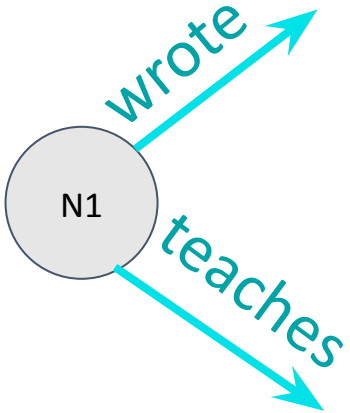
François Goasdoué, Pawel Guzewicz, Ioana Manolescu. RDF graph summarization for first-sight structure discovery. The VLDB Journal, 2020, 29 (5), pp.1191-1218. 10.1007/s00778-020-00611-y . hal-02530206v2

https://rdfquotient.inria.fr/

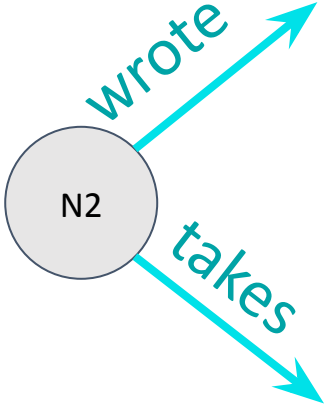
INRAE

RDFQuotient's Central Feature
Transitive Cooccurrence of Properties

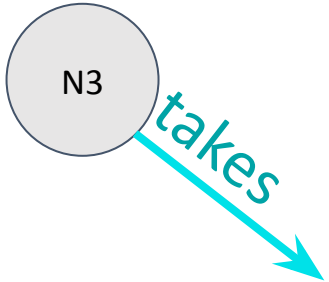




(Professor ?)



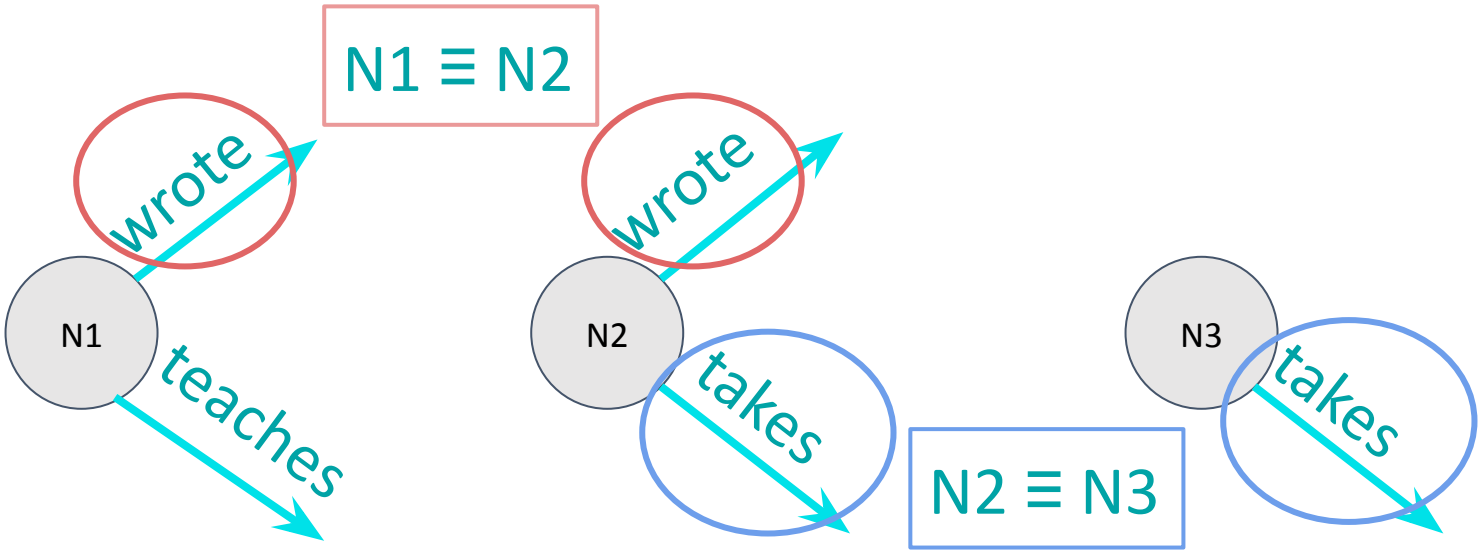
(GradStudent ?)



(Student ?)

INRAE

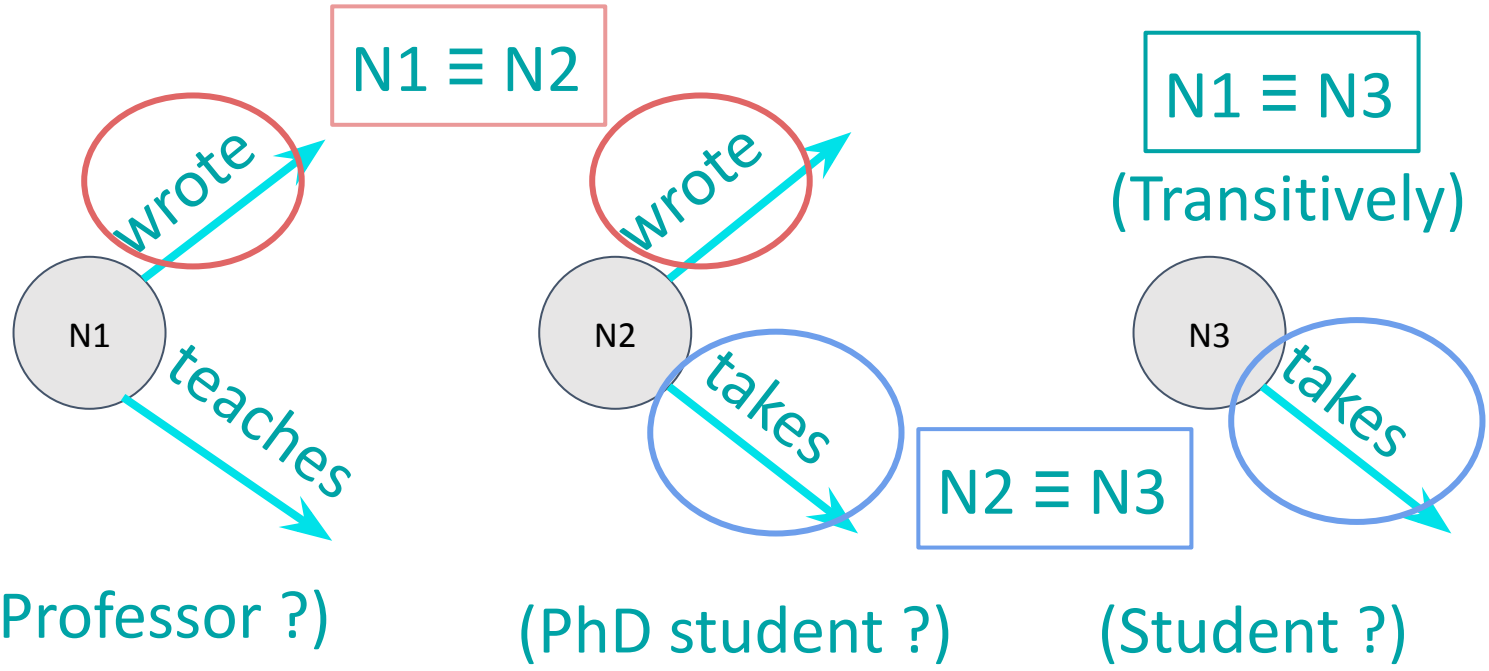
RDFQuotient's Central Feature
Transitive Cooccurrence of Properties



(Professor ?) (GradStudent ?) (Student ?)

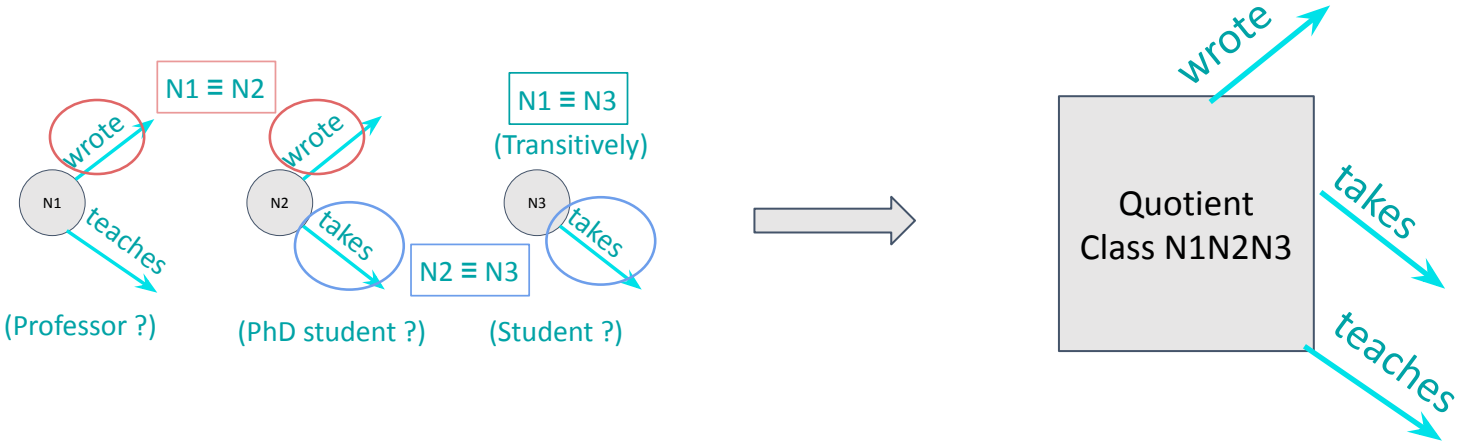
INRAE

RDFQuotient's Central Feature
Transitive Cooccurrence of Properties

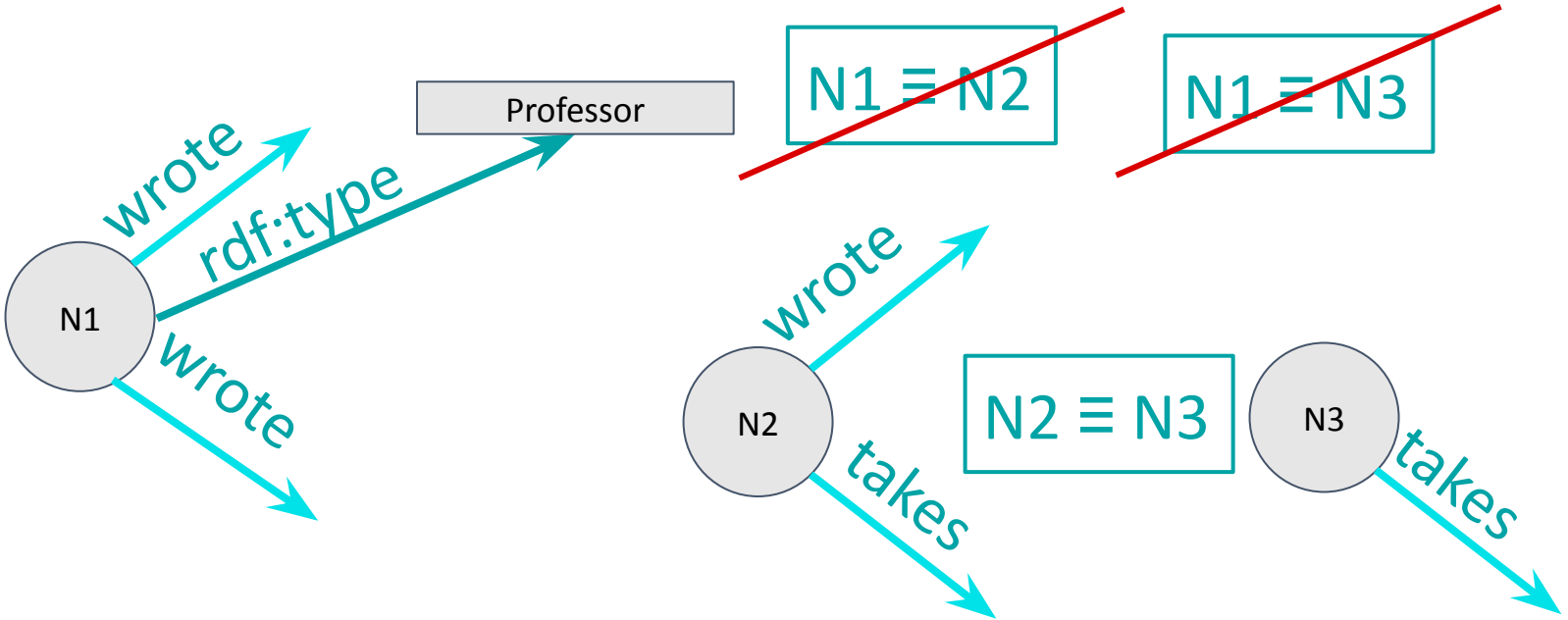


INRAE

RDFQuotient's Central Feature
Transitive Cooccurrence of Properties

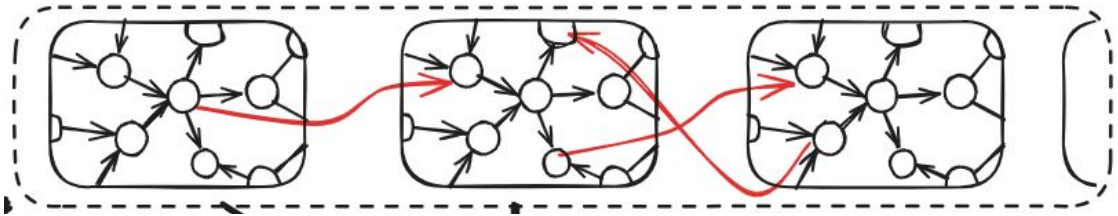


RDFQuotient “**Types first, Then Data**” Approach
-> use and prioritise types if they exist



RDFQuotient's limits

- Not usable with large graphs (not scalable)
- Partly based on a *non* open-source system (OntoSQL/RDFDB)
- Single graph approach -> no links between graphs



4.4E09 triples 8.8E06 triples 6.8E08 triples

INRAE

A Multigraph, Scalable Quotient Project

A Scalable Approach

Open-source, distributed, structural graph summary tool

-> Based on RDFQuotient's transitive cooccurrence of properties & "type then data" strategy

-> Compatible with large graphs (for metabolomics!)

-> Every step of the quotient has to be **scalable**

GraphFrames

INRAE

https://forge.inrae.fr/mss/metabolomics-semantic-stack/-/tree/main/apps/quotient

Simple Multigraph Approach with Indexes

Quotient indexes link graph nodes to their quotient classes, e.g. :

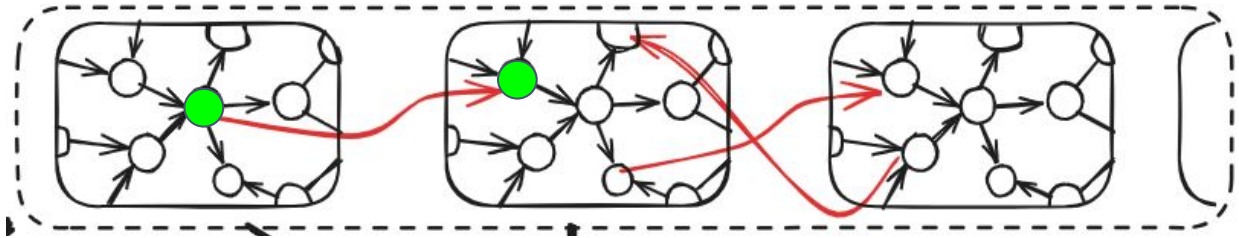
In Gene Ontology's index :

taxon:9397 :hasQuotientClass :NRMdLIF860k .

skos:closeMatch !!!

In NCBI Taxonomy's index :

taxon:9397 :hasQuotientClass :QSdSnR6EdQI .



INRAE

Enabling a Multigraph Structural Overview

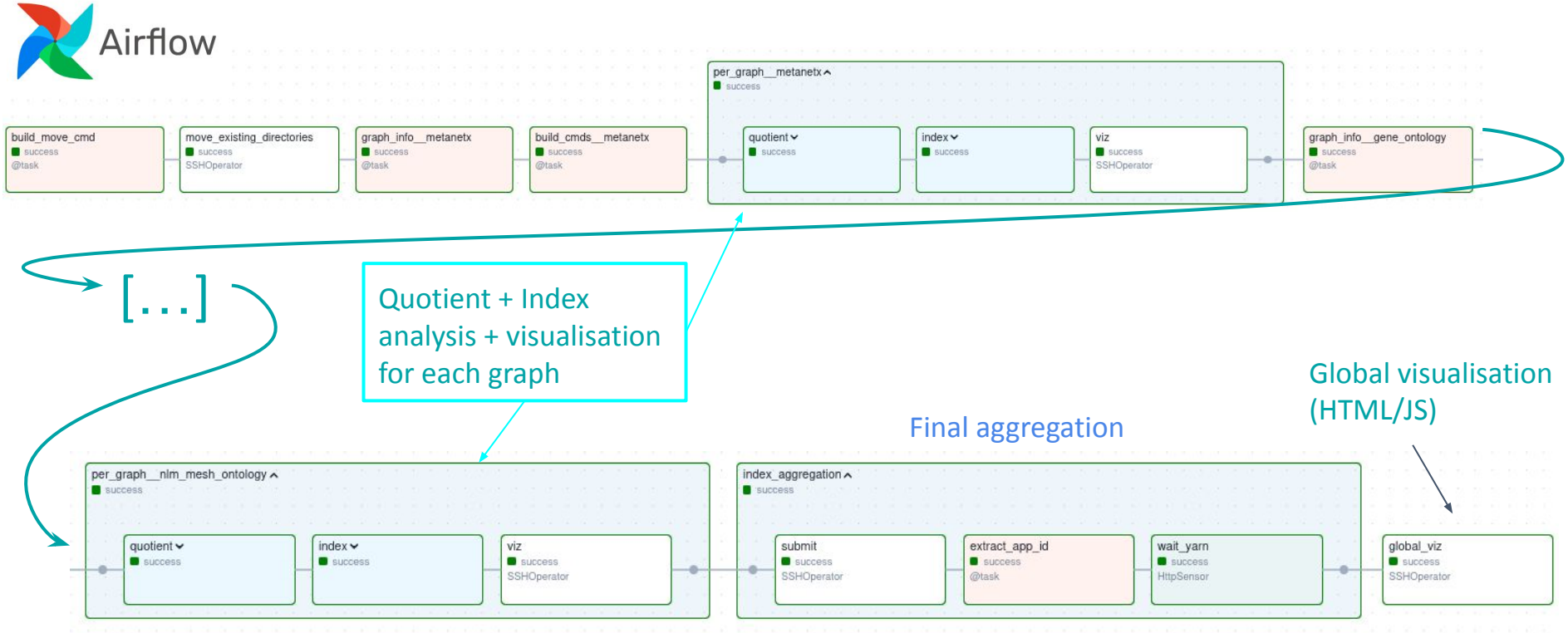
Basic multigraph mode :

1) load a set of graphs as one graph
2) compute a global quotient

Enhanced multigraph mode :

1) compute [quotient + index] for each graph
2) aggregate quotients (using indexes to compute equivalences)

Dedicated Multigraph Dynamic Workflow



INRAE https://forge.inrae.fr/mss/metabolomics-semantic-stack/-/blob/main/workflowManagement/dags/multi_quotient_dag.py

Machine Readable Output

Quotients are provided as RDF graphs

-> We had to implement a tiny OWL vocabulary

10 properties, enables quotient modeling

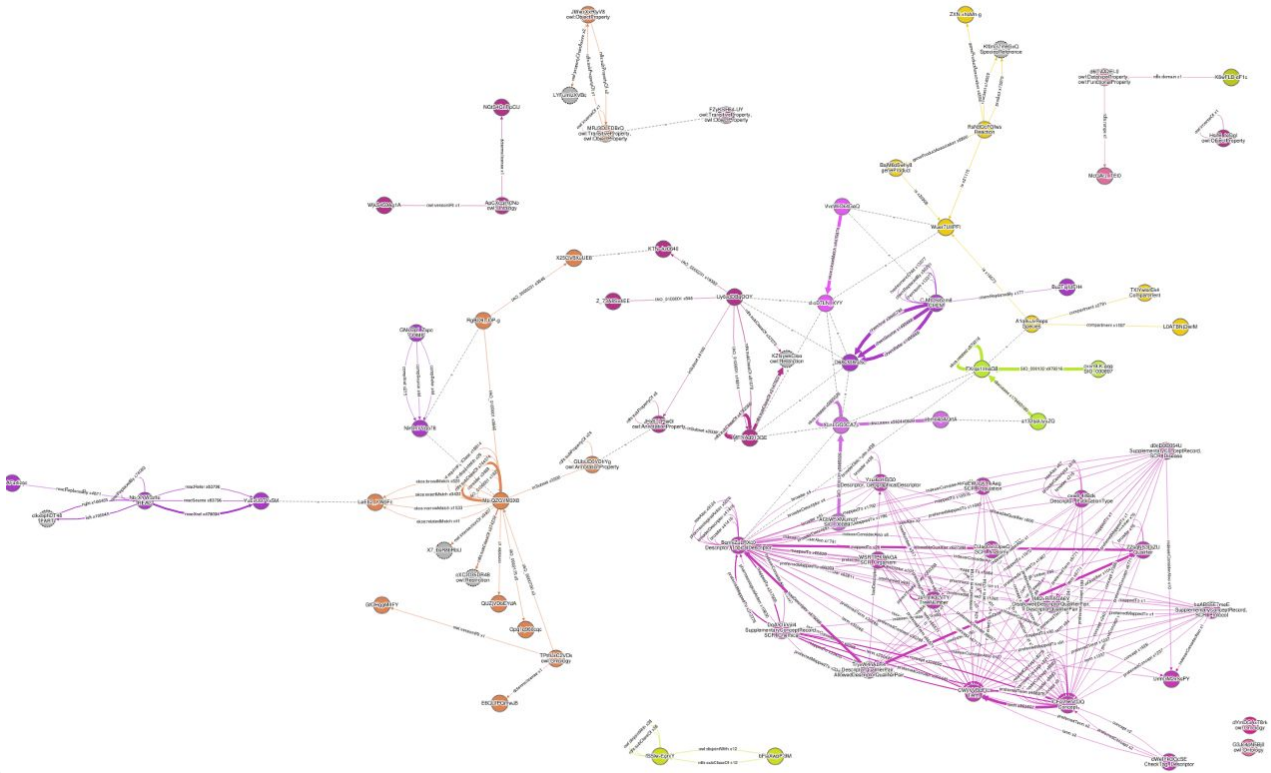
Why ? Example :

~~:quotientClass1 pubchem:inchikey :quotientClass2~~

:quotientClass2 is not an InChIKey !

(and a quotient class may represent a high diversity of nodes)

Multigraph HTML/JS Quotient Visualisation



http://purl.obolibrary.org/obo/go.owl

https://www.metanetx.org/ftp/latest/MNXref.ttl

https://forvm.semantics.metabohub.fr/forum_chebi_mesh/20260213

https://forvm.semantics.metabohub.fr/forum_mappings/20260213

https://nlmpubs.nlm.nih.gov/projects/mesh/rdf/mesh.nt

https://ftp.ebi.ac.uk/pub/databases/chebi/ontology/chebi.owl.gz

https://nlmpubs.nlm.nih.gov/projects/mesh/vocabulary_0.9.3.ttl

https://forvm.semantics.metabohub.fr/forum_vocabulary/20260224

https://github.com/SysBioChalmers/Human-GEM/blob/main/model/Human-GEM.xml

https://forvm.semantics.metabohub.fr/forum_chebi_mesh/20260213

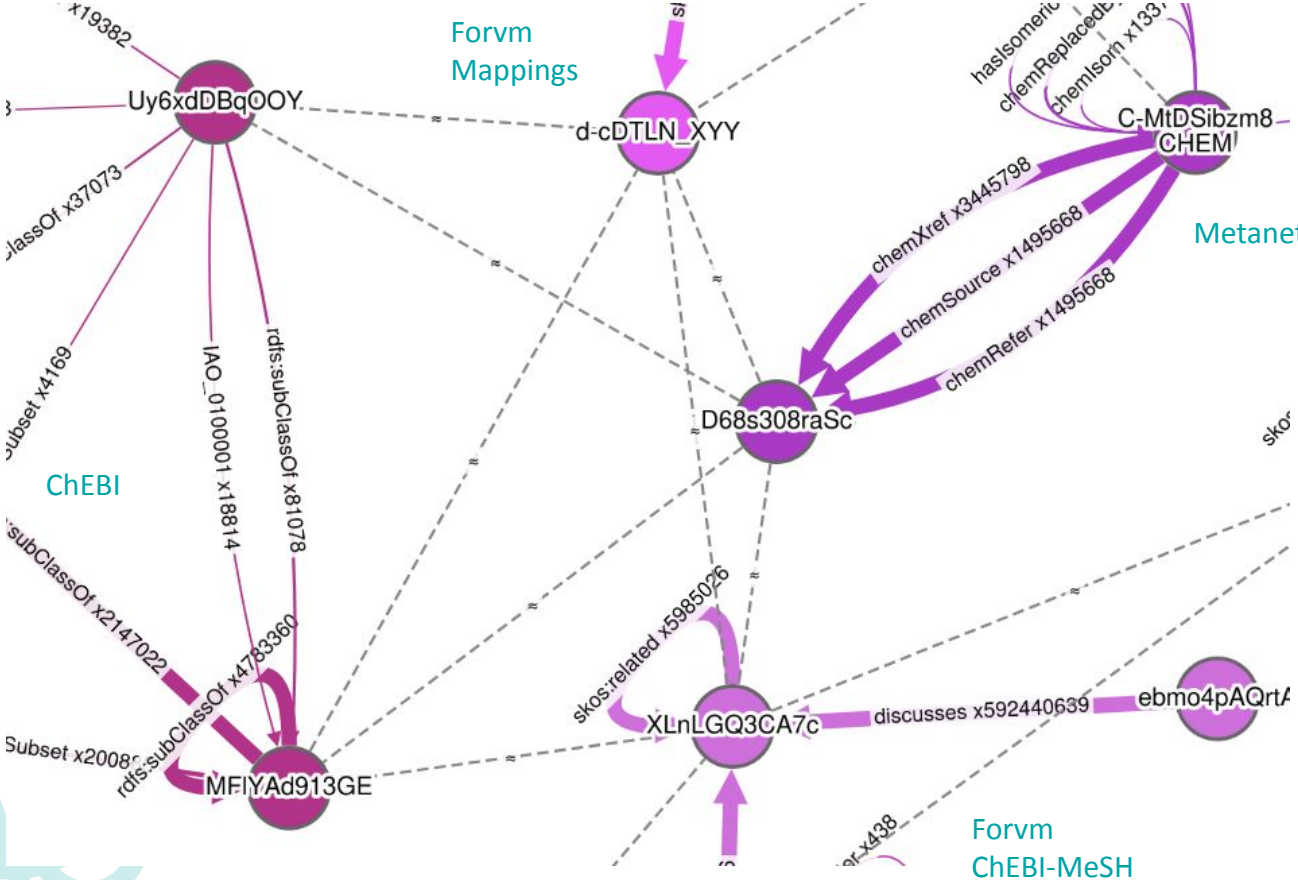
https://forvm.semantics.metabohub.fr/forum_vocabulary/20260224

https://nlmpubs.nlm.nih.gov/projects/mesh/rdf/mesh.nt

https://nlmpubs.nlm.nih.gov/projects/mesh/vocabulary_0.9.3.ttl

https://forvm.semantics.metabohub.fr/

Multigraph HTML/JS Quotient Visualisation



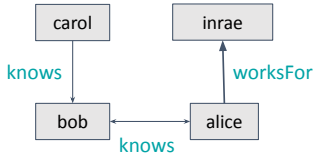
Example.html

- http://purl.obolibrary.org/obo/go.owl
- https://www.metanetx.org/ftp/latest/MNXref.ttl
- https://forvm.semantics.metabohub.fr/forum_chebi_mesh/20260213
- https://forvm.semantics.metabohub.fr/forum_mappings/20260213
- https://nlmpubs.nlm.nih.gov/projects/mesh/rdf/mesh.nt
- https://ftp.ebi.ac.uk/pub/databases/chebi/ontology/chebi.owl.gz
- https://nlmpubs.nlm.nih.gov/projects/mesh/vocabulary_0.9.3.ttl

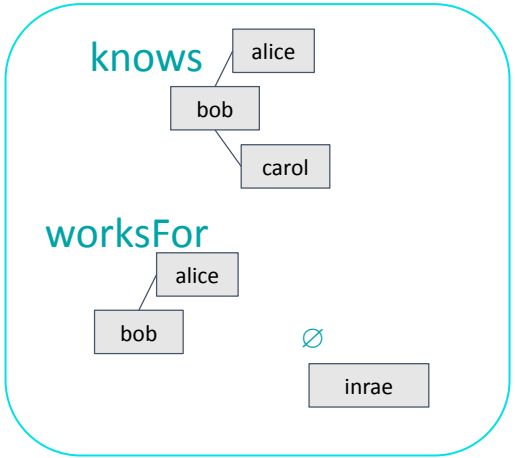
Algorithm / Problems / Solutions

Two-Step Distributed Clique Building

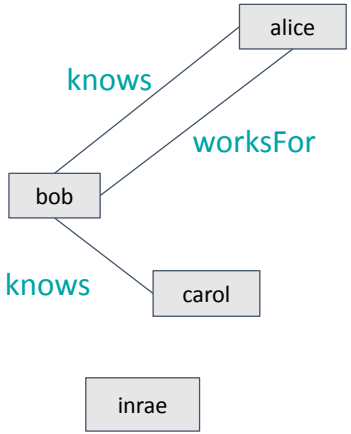
Transitive Propagation Using GraphFrames



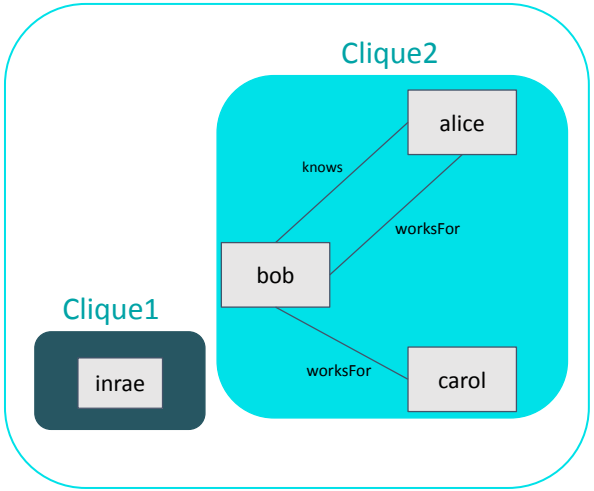
1) Form one star per property* (here : outgoing properties)



-> new Graph (from stars)



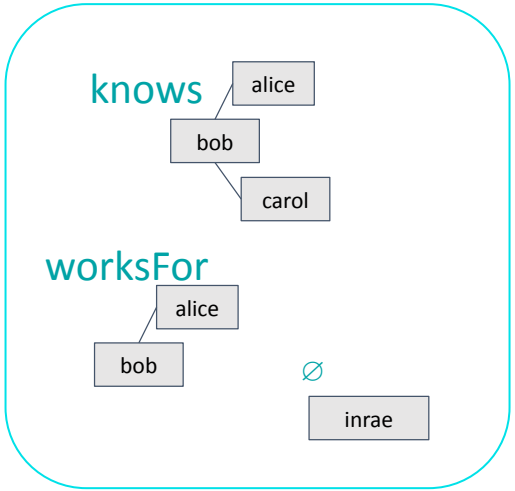
2) Compute connected components



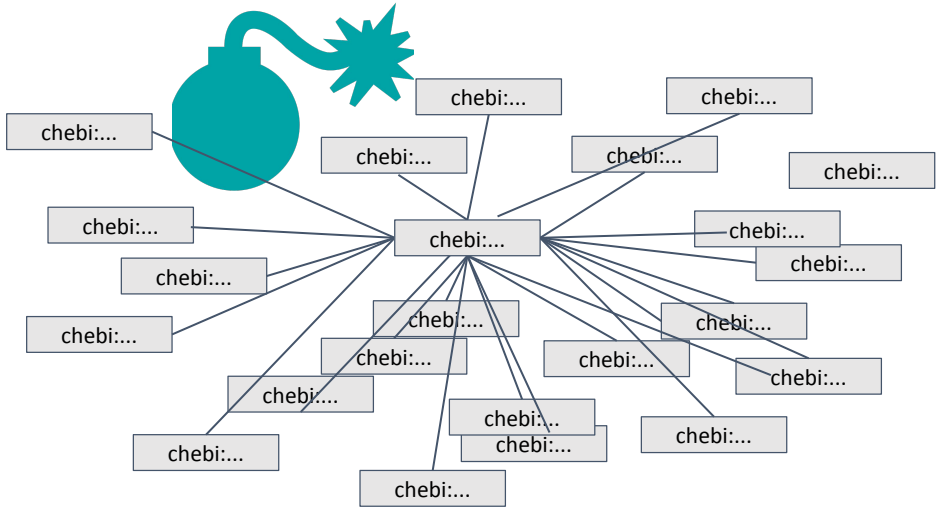
*** incoming / outgoing / both, depending on the algorithm flavour**

Two-Step Distributed Algorithm : Scalability ?

1) Form one star per property



… may lead to skew data (e.g. rdfs:subClassOf Star)



-> All expensive joins and aggregations are salted

Memory Efficiency / Node Dictionary

During quotient computation, we use a graph node dictionary to reduce memory usage and enable DataFrame-level optimisations.

Identifier (kind)	Jena Node
0 (resource)	<http://id.nlm.nih.gov/mesh/C540250>
1 (resource)	<http://id.nlm.nih.gov/mesh/vocab#active>
2 (literal)	true

A New Reasoner

New OWL Horst Reasoner Implementation

We had to implement a new distributed reasoner to support inference using large and complex schemata

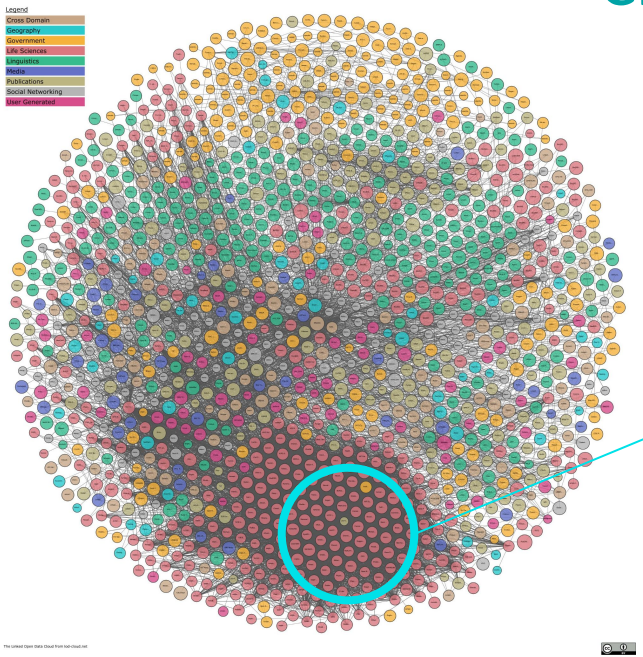
- Needed before quotient computation (and also needed by FORVM)
- Based on Spark DataFrame/Dataset APIs :
  - Query optimization (Catalyst) -> faster jobs
  - Better execution engine (Tungsten / whole-stage code generation)
  - Better I/O optimisation (column pruning, filtered file scan)
  - However, DataFrames are only loosely typed, so we had to use a few techniques to approximate strong typing
- Major improvement : removes the schema-broadcast bottleneck in the previous reasoner + fixes a critical bug in the transitive-closure loop condition

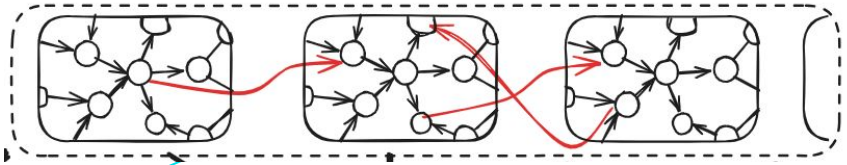
INRAE

https://forge.inrae.fr/mss/metabolomics-semantic-stack/-/tree/main/apps/reasoners

Next Steps

Chart (a tiny subset of) the LOD !

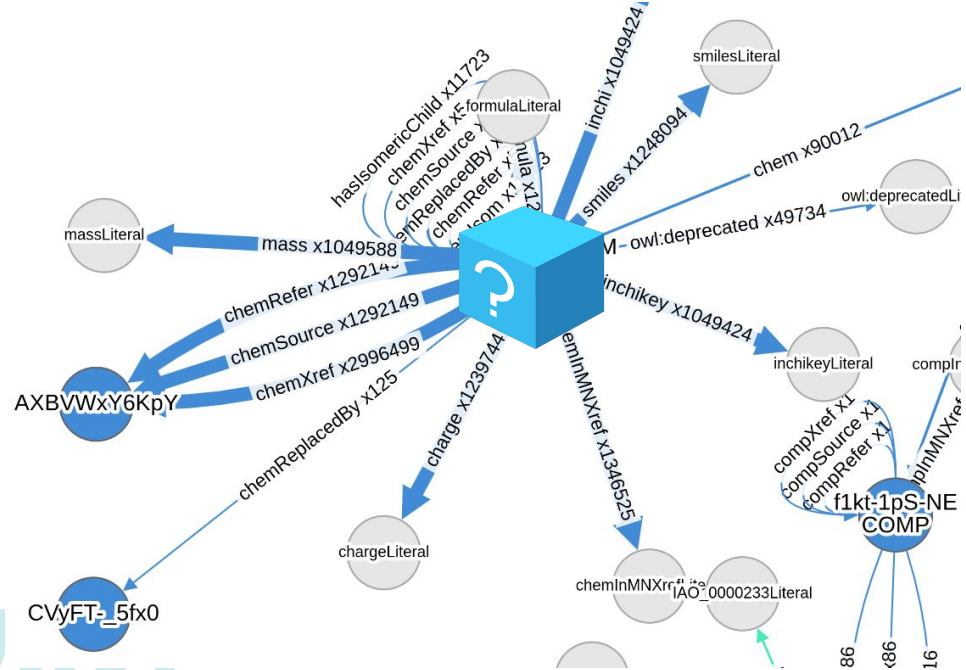




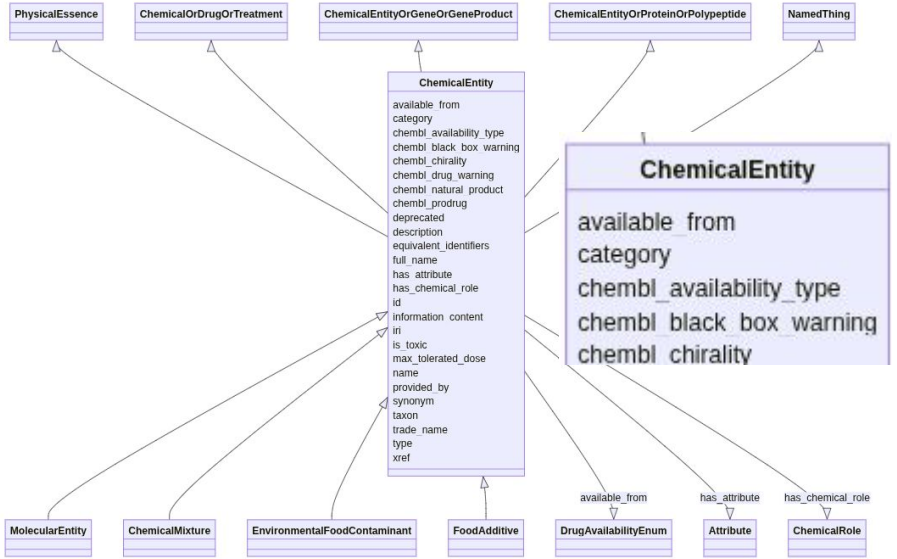
Next Steps

Rule-based assignment of quotient classes to Biolink Model ?

Quotient Class



Biolink Model Class



Project code :
https://forge.inrae.fr/mss/metabolomics-semantic-stack/

MSD core + MSD tools
FORVM (distributed implementation)
FORVM Plants
Workflow Management
Reasoners
Quotient + tools

Thank you !

Team : David BENABEN, Thomas BELLEMBOIS, Kevin BILLET, Alain BOUCHEREAU, Matéo BOUDET, Cécile CABASSON, Younes DELLERO, Maxime DELMAS, Olivier FILANGI, Clément FRAINAY, Franck GIACOMONI Nicolas GUILHOT, Yann GUITTON, Antoine MAHUL, Guillaume MARTI, Meije MATHE, Nils PAULHE, Sylvain PRIGENT, Faustine SOUC, Florence Vinson, Magalie WEBER

Presented by : Guillaume LAISNEY
guillaume.laisney@inrae.fr