Intégration de données phénotypiques, environnementales et de biodiversité à l'aide des technologies du Web Sémantique

Olivier Dameron[1], **Yael Tirlet**[1,2],
Matéo Boudet[2,3], Fabrice Legeai[2]

[1] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[2] INRAe IGEPP, F-35000 Rennes, France
[3] Plateforme GenOuest, IRISA - UMR 6074, F-35000 Rennes, France

2022-11-22

UMR IRISA DyLiSS INRAE GenOuest

Contexte : projet DeepImpact

Le projet DeepImpact (ANR 2021–2027)

INRAE

- Vise à comprendre l'impact de l'environnement et de la faune du sol sur le rendement et la santé des cultures de blé et de colza
- Mesures faites sur différentes parcelles (espèces, sol, météo,. . .)
  - phénotypes
  - environnement
  - biodiversité
- Regroupe 3 équipes de l'INRAE : Rennes, Dijon et Toulouse
  - 10 partenaires ; 5 disciplines
    (agronomie, écologie, pathologie, sciences des plantes, informatique)
  - porteur : Christophe Mougel
  - budget : 3 millions €

Un problème générique : intégrer et interroger

- grande quantité et diversité de données
- connaissances du domaine

Objectif

Définir un schéma de données permettant :

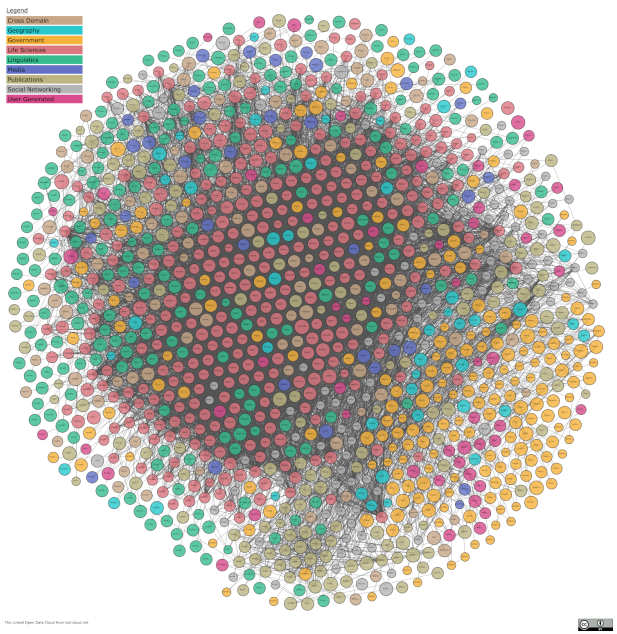
- d'éviter de n-pliquer l'ingénierie des données
- de constituer une banque mutualisée de requêtes
- de combiner les données de différentes études

Approche

- définir le schéma de données + lien avec bases de connaissances
- mettre en œuvre avec Web Sémantique
- constituer une librairie de requêtes réutilisables
- faire l'analyse (évidemment)

État de l'art

Web Sémantique



`https://lod-cloud.net`

- 1700+ bases de données en sciences de la vie, hétérogènes et interdépendantes
- Le Web Sémantique permet de les intégrer et des interroger

Web Sémantique

- (+) solution pertinente et générique
- (-) problèmes
  - données Deep Impact (entre autres) au format tabulé
  - représentation peu accessible pour les biologistes

AskOmics : présentation générale

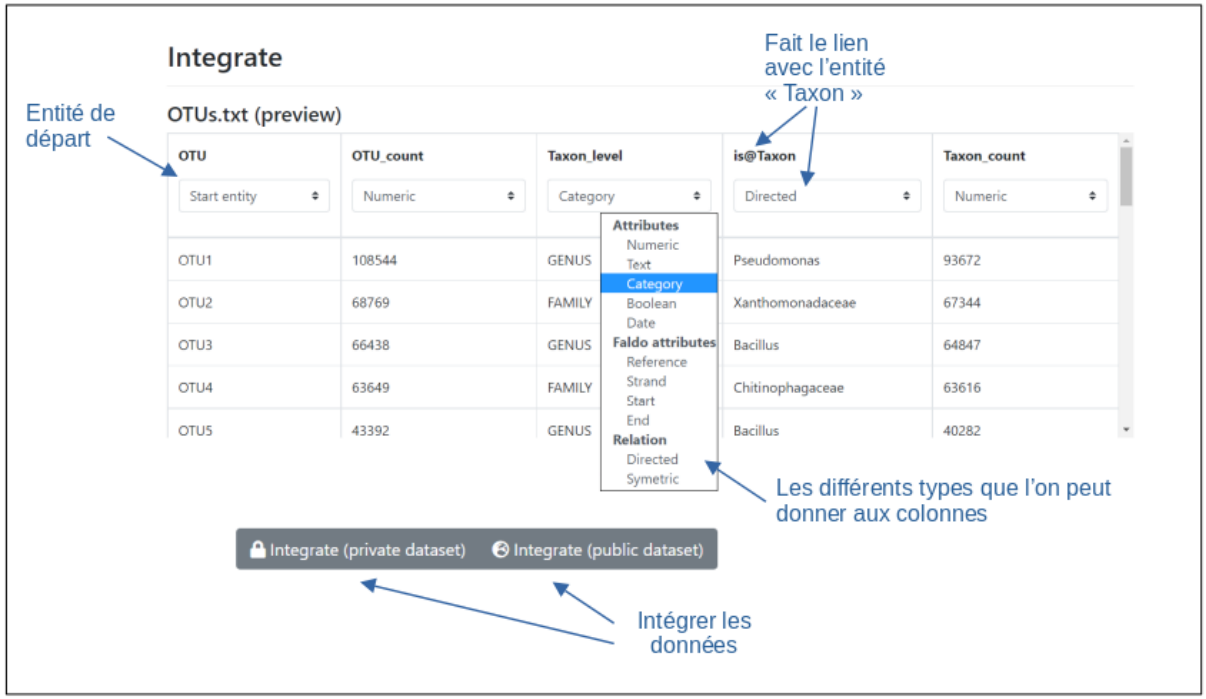
Interface pour manipuler simplement les données grâce au Web sémantique
Développé par l'INRAe et Dyliss

Fonctionnalités

- **Intégrer les données** (entre elles et avec des connaissances)
  - en les convertissant au format RDF
  - en les déployant sous forme d'un endpoint
- **Aider à la composition des requêtes** en générant automatiquement le code SPARQL

`https://askomics.org`

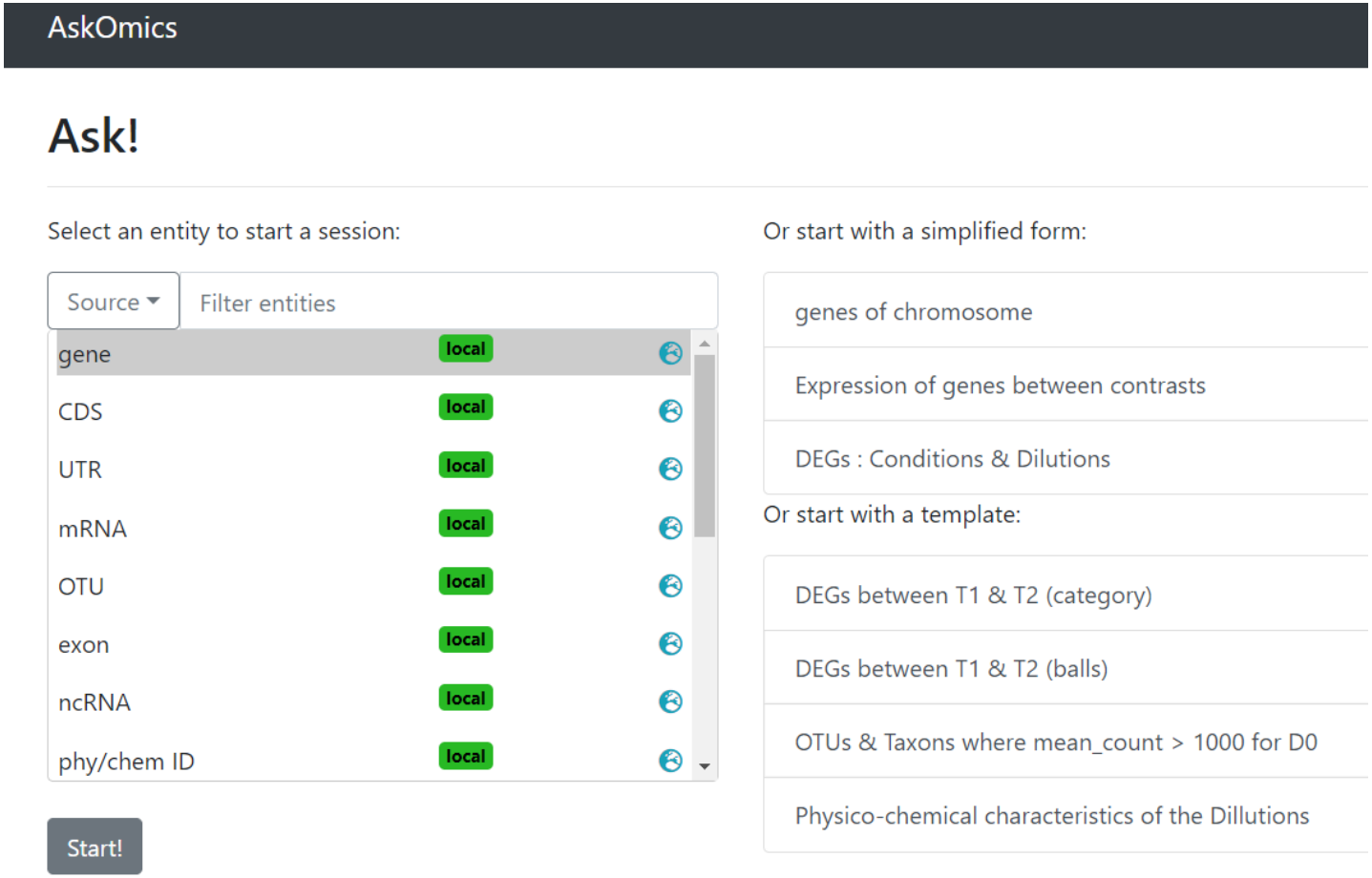
AskOmics (1) : convertir et intégrer les données en RDF



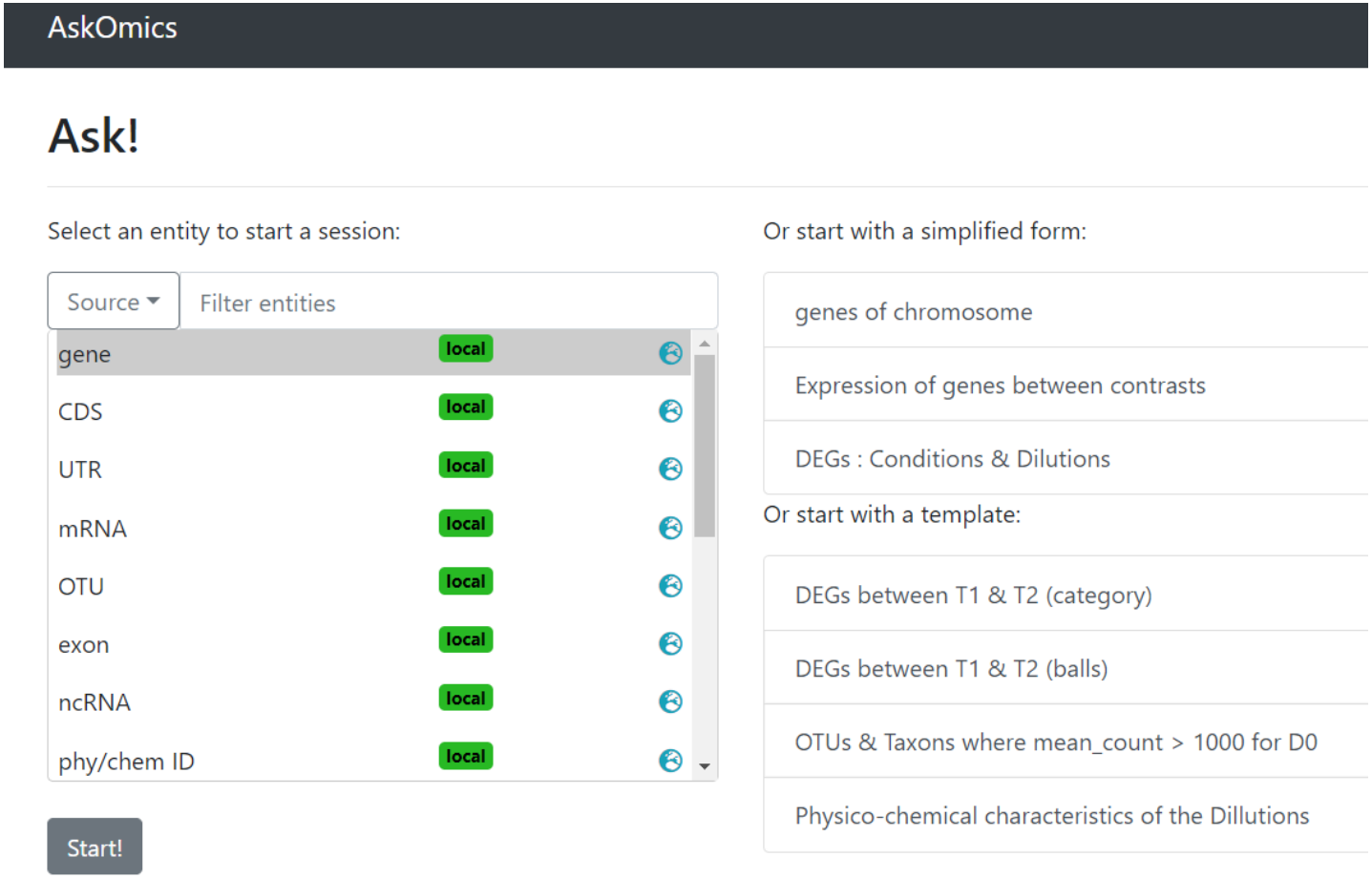
À ce stade :

- les fichiers tabulés ont été convertis en RDF...
- ... et intégrés dans une seule base...
- ... qui est déployée sous forme d'un SPARQL endpoint
- Vous n'avez plus qu'à écrire vos requêtes SPARQL (enjoy !)

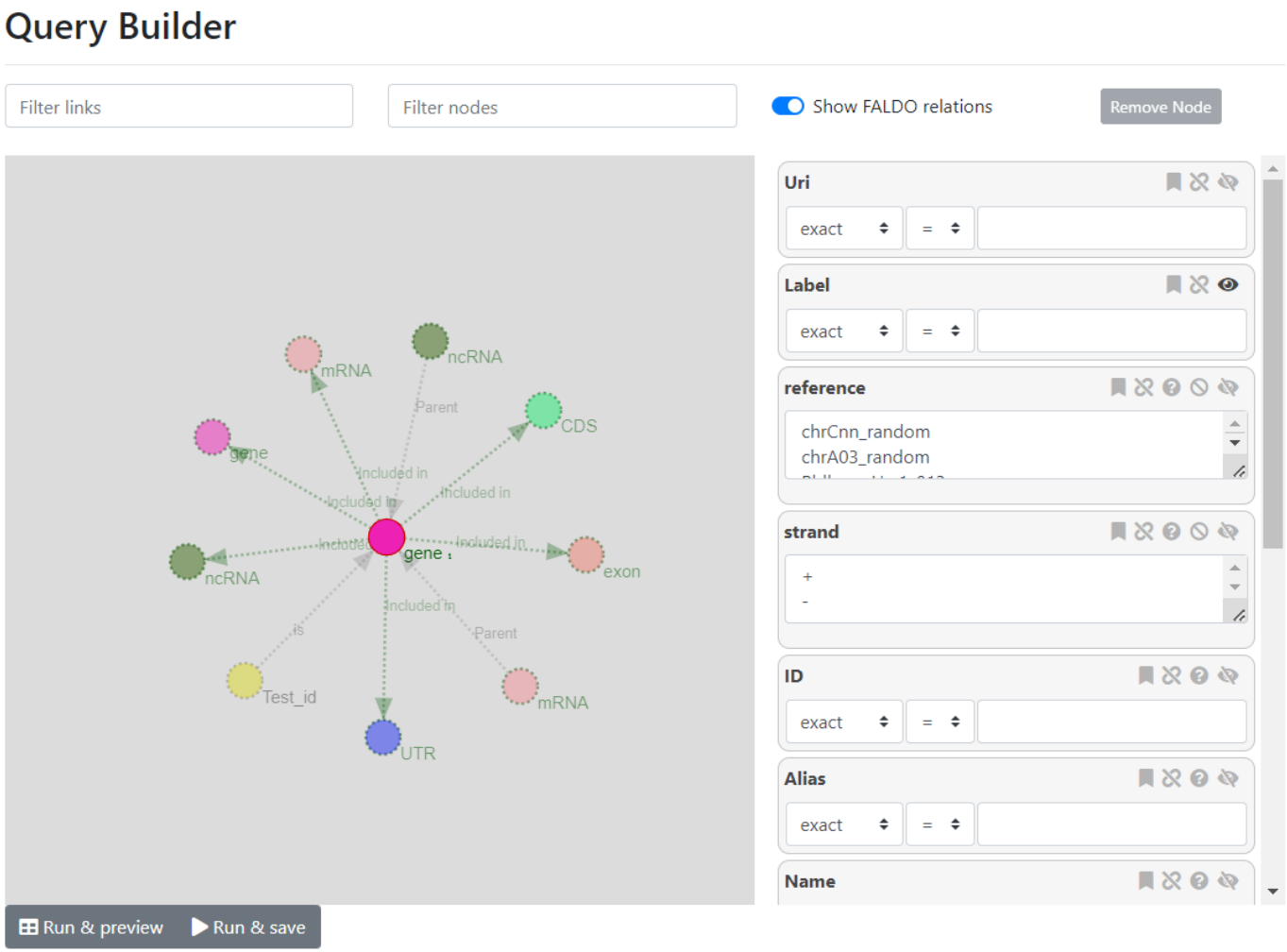
AskOmics (2) : Générer intuitivement des req. SPARQL



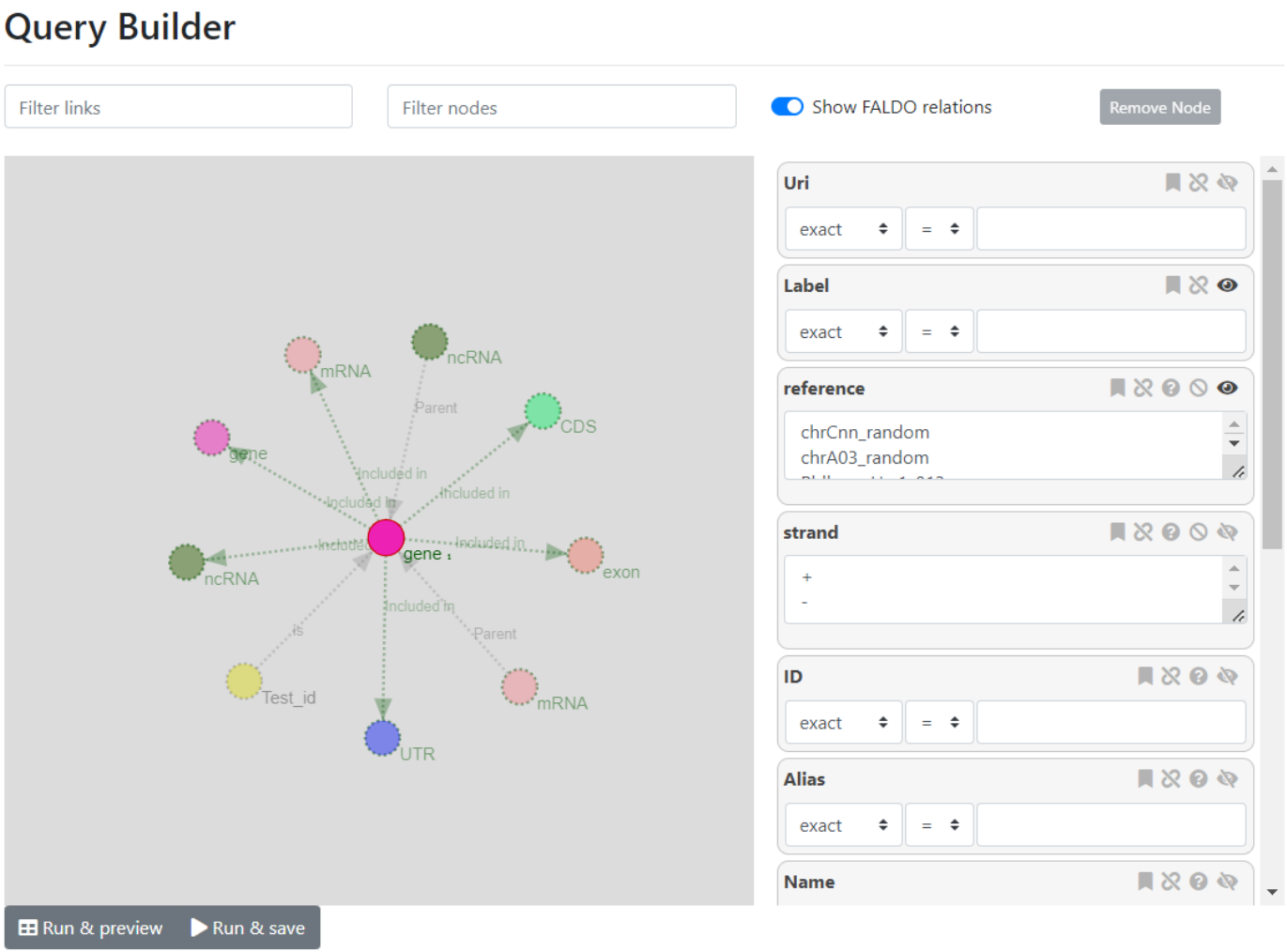
AskOmics (2.1) : Construction itérative requête SPARQL



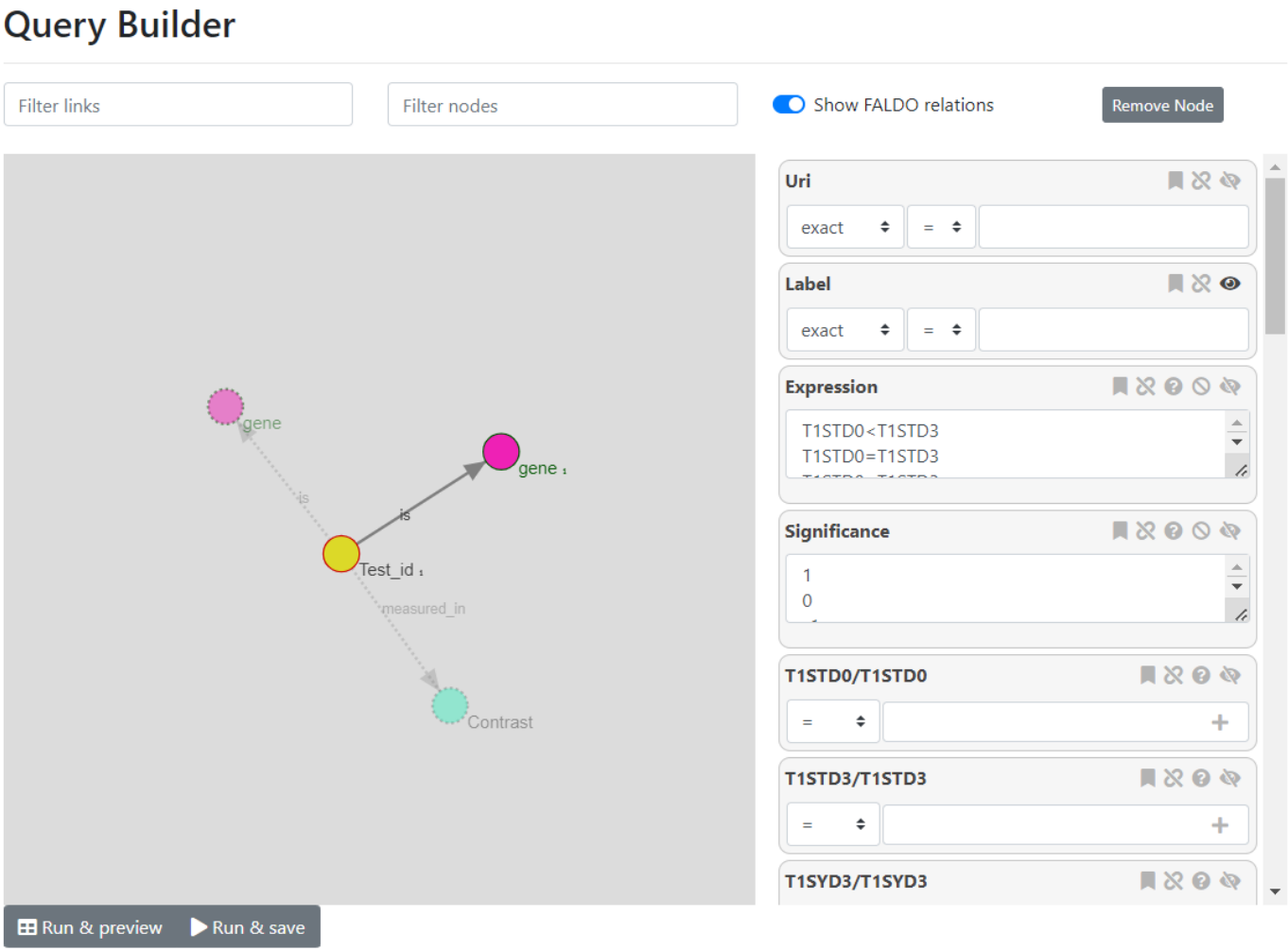
AskOmics (2.2) : Construction itérative requête SPARQL



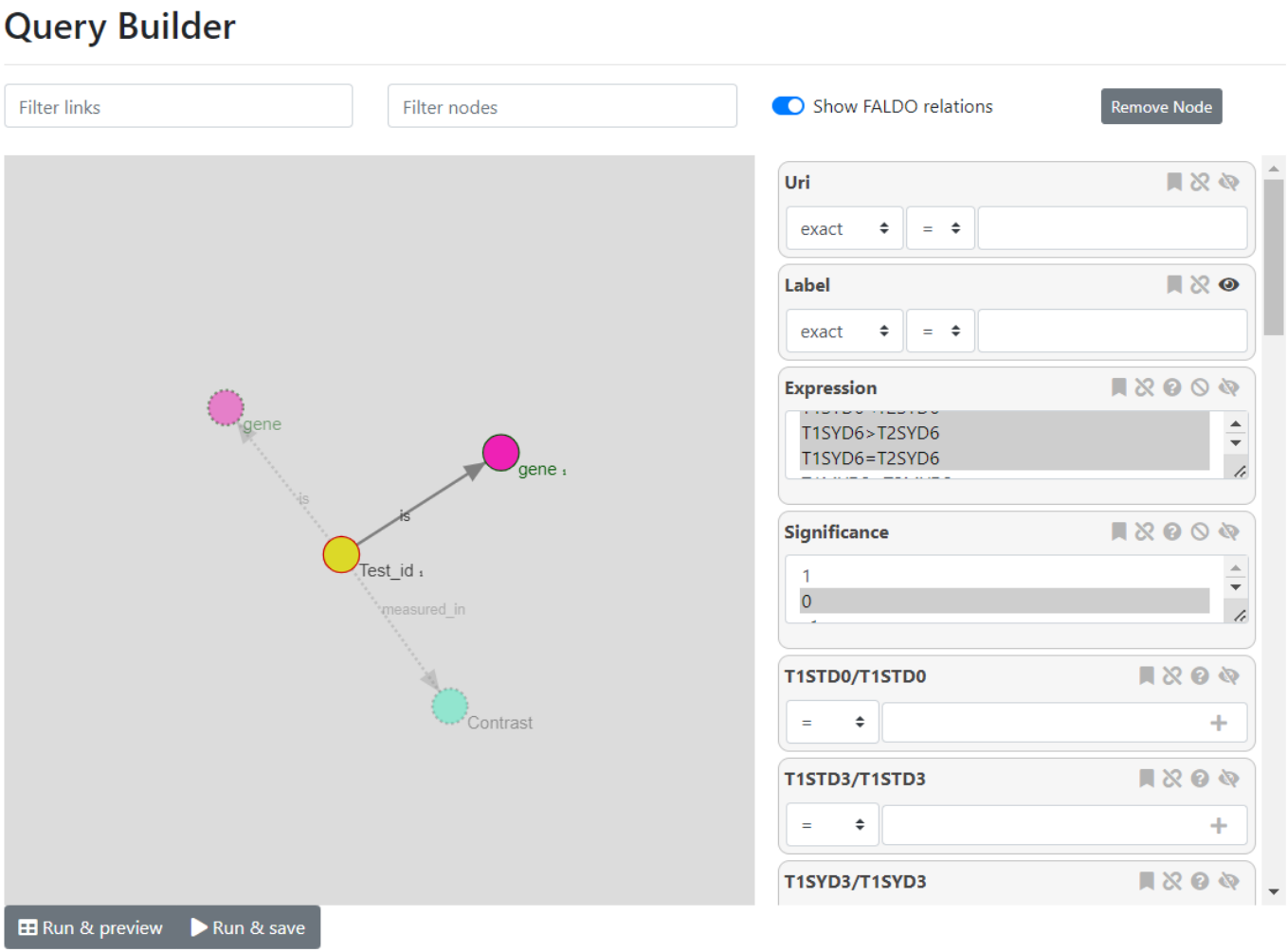
AskOmics (2.3) : Construction itérative requête SPARQL



AskOmics (2.4) : Construction itérative requête SPARQL



AskOmics (2.5) : Construction itérative requête SPARQL



AskOmics (3) : Traitement d'une requête SPARQL

(3) SPARQL query is processed by the endpoint

Results

Id	Description	Exec time	Template	Form	Public	Status	Rows	Size	Actions
28	Genes and chromosome whose expression is the same at T1 and T2	17s				Success	568 121	33.0 MB	Preview Download Form Redo Sparql

(2) SPARQL query is sent to the endpoint

(4) query results are sent back to the user

SPARQL query

```
PREFIX : <http://askomics.org/data/>
PREFIX askomics: <http://askomics.org/internal/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX faldo: <http://biohackathon.org/resource/faldo/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?gene1_Label ?gene1_reference ?Test_id1_Label
WHERE {
    ?Test_id15_uri <http://askomics.org/data/is> ?gene1_uri .
    ?gene1_uri rdf:type <http://askomics.org/data/gene> .
    ?gene1_uri rdfs:label ?gene1_Label .
    ?gene1_uri faldo:location/faldo:begin/faldo:reference ?gene1_referenceCategory .
    ?gene1_referenceCategory rdfs:label ?gene1_reference .
}
```

(1) SPARQL query is generated automagically

gene1_Label	gene1_reference	Test_id1_Label
BnaA02g34680D	chrA02	T1SYD6vsT2SYD6_BnaA02g34680D
BnaA02g09470D	chrA02	T1SYD6vsT2SYD6_BnaA02g09470D
BnaA02g10460D	chrA02	T1SYD6vsT2SYD6_BnaA02g10460D
BnaA02g14050D	chrA02	T1SYD6vsT2SYD6_BnaA02g14050D
BnaA02g15610D	chrA02	T1SYD6vsT2SYD6_BnaA02g15610D
BnaA02g16920D	chrA02	T1SYD6vsT2SYD6_BnaA02g16920D
BnaA02g21750D	chrA02	T1SYD6vsT2SYD6_BnaA02g21750D
BnaA02g24150D	chrA02	T1SYD6vsT2SYD6_BnaA02g24150D
BnaA02g25530D	chrA02	T1SYD6vsT2SYD6_BnaA02g25530D
BnaA02g29010D	chrA02	T1SYD6vsT2SYD6_BnaA02g29010D

10 1 2 3 >

Résultats

Validation : article et jeu de données

Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes
Daval, Stéphanie, et al. *Microbial Biotechnology*, 2020
`https://doi.org/10.1111/1751-7915.13634`

- Mesures d'expression génique pour différentes conditions
- Trois dilutions différentes pour les sols, trois répétitions par dilution
- Données :
  - comptages bruts d'expression de gènes de *Brassica napus*
  - comptages bruts des OTUs pour chaque réplicats de chaque sol
  - fichier de description des OTU
  - données sur les différents sols. . .

Données et requêtes similaires à ce qu'on pourra avoir avec DeepImpact
Idéal pour la validation

Contribution1 : Rassemblement, Modification et Création des données

	Nature	Taille	R	M	C
40	Expr génique	10Mb	X		X
2x3	Contraste - contexte - conditions		X		X
2	GFF	76+70Mb	X		
1	Comptage OTU	2Mb		X	
1	Caract physico-chimique sols			X	
1	OTU et taxons	1Mb		X	X

- 55 Fichiers
- 600 Mb de données

Modification et complétion des données OTUs

OTU	Phylum	...	Family	Genus
OTU 4	Bacteroidetes (63631) Acidobacteria (16) Unknown (2)	...	Unknown (18) Chitinophagaceae (63616) Cytophagaceae (15)	Chitinophaga (1692) Flavisolibacter (20553) Terrimonas (466) Flexibacter (15) Niastella (32884) Niabella (1) Unknown (1573) Segetibacter (6465)
OTU 7	Firmicutes (1) Actinobacteria (30805)	...	Bacillaceae (1) Micrococcaceae (1) Streptomycetaceae (30804)	Streptomyces (30798) Bacillus (1) Arthrobacter (1) Kitasatospora (6)
OTU 6706	Proteobacteria (4)	...	Beijerinckiaceae (1) Methylophilaceae (3)	Methylobacillus (1) Unknown (3)

Compte < 70 % Taxon inconnu Compte > 70 % et Taxon connu

Pour chaque OTU, on cherche à déterminer (le plus précisément possible) l'espèce majoritaire représentative

Modification et complétion des données OTUs

OTU	Compte total dans l'OTU	Rang du Taxon	Taxon	Compte du Taxon
OTU 4	63 649	FAMILY	Chitinophagaceae	63 616
OTU 7	30 806	GENUS	Streptomyces	30 804
OTU 6706	4	FAMILY	Methylophilaceae	3

Tableau descriptif des OTUs

Résultat du pré-traitement

Modification et complétion des données OTUs

OTU	Compte total dans l'OTU	Rang du Taxon	Taxon	Compte du Taxon
OTU 4	63 649	FAMILY	Chitinophagaceae	63 616
OTU 7	30 806	GENUS	Streptomyces	30 804
OTU 6706	4	FAMILY	Methylophilaceae	3

Tableau descriptif des OTUs

Les taxons font référence à une base de connaissances

Modification et complétion des données OTUs

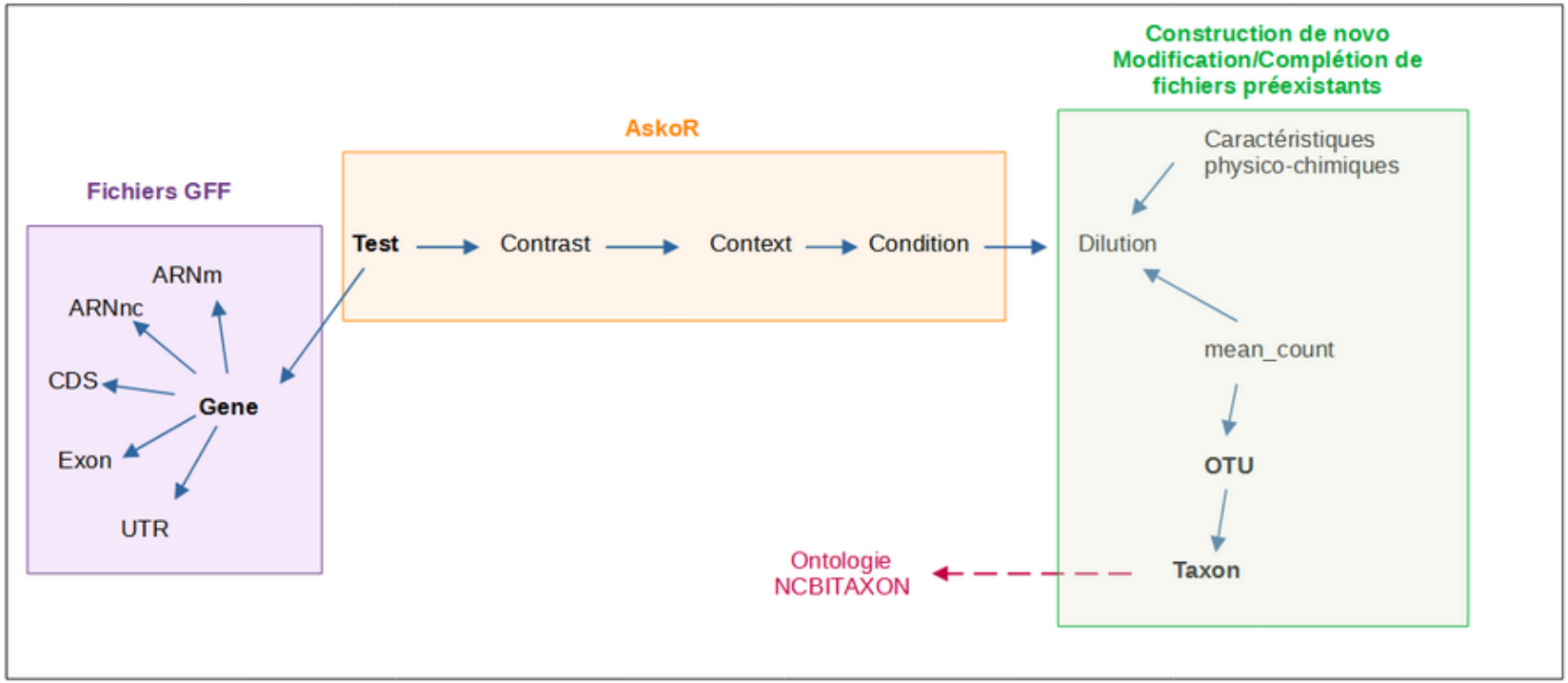
NCBI Taxon

Taxon	Identifiant	Label	Synonymes
Chitinophagaceae	http://purl.obolibrary.org/obo/NCBITaxon_563835	Chitinophagaceae	None
Streptomyces	http://purl.obolibrary.org/obo/NCBITaxon_1883	Streptomyces	["Actinosporangium Krasil'nikov and Yuan 1961", "Actinopycnidium Krasil'nikov 1962", 'Chainia', 'Streptoverticillium Baldacci 1958', ""Microstreptospora" Yan et al.', 'Elytrosporangium Falcao de Morais et al. 1966', 'Kitasatoa Matsumae and Hata 1968']
Methylophilaceae	http://purl.obolibrary.org/obo/NCBITaxon_32011	Methylophilaceae	['beta subdivision methylotrophs', 'beta-subdivision methylotrophs', 'Methylophilus group']

Tableau descriptif des Taxons

Résultat du mapping avec l'ontologie NCBI Taxon

Contribution 2 : Structure de données pour l'intégration



Principe du schéma d'intégration de données
et utilisation d'AskoR pour le contrôle qualité

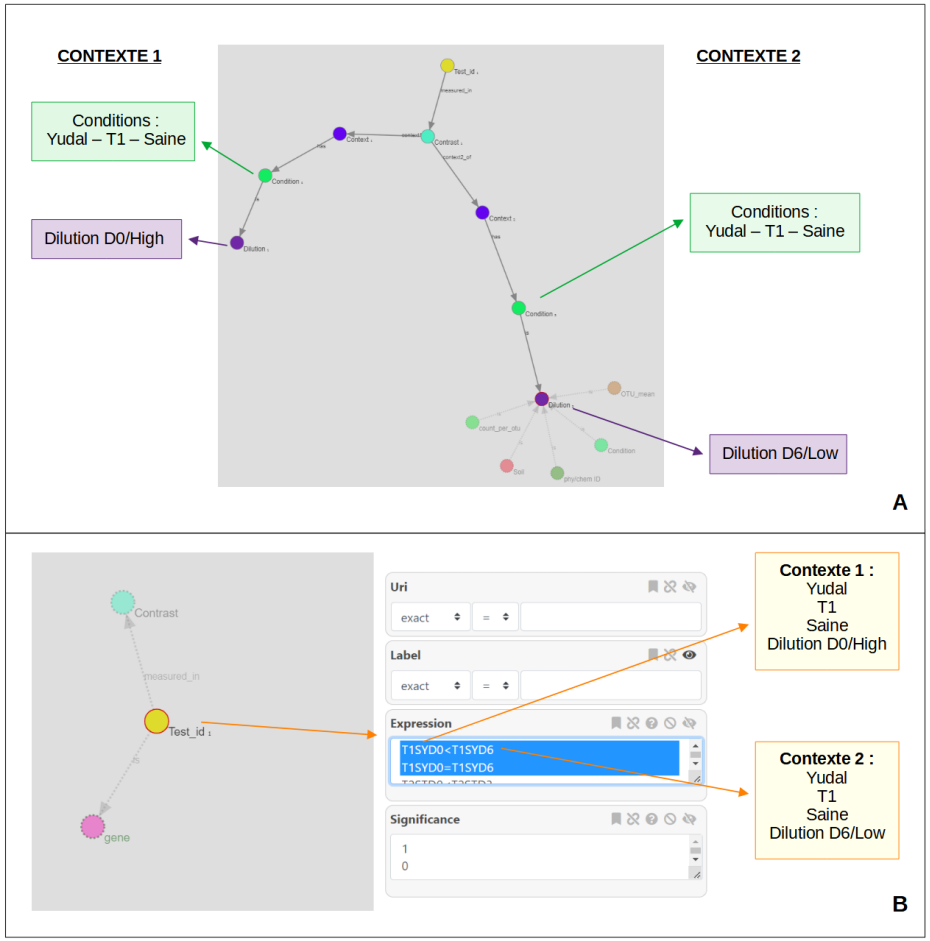
Contribution 3 : Peuplement de la base AskOmics

- 2 480 000 Tests
- 113 851 Gènes
- 32 870 OTUs
- 855 Taxons

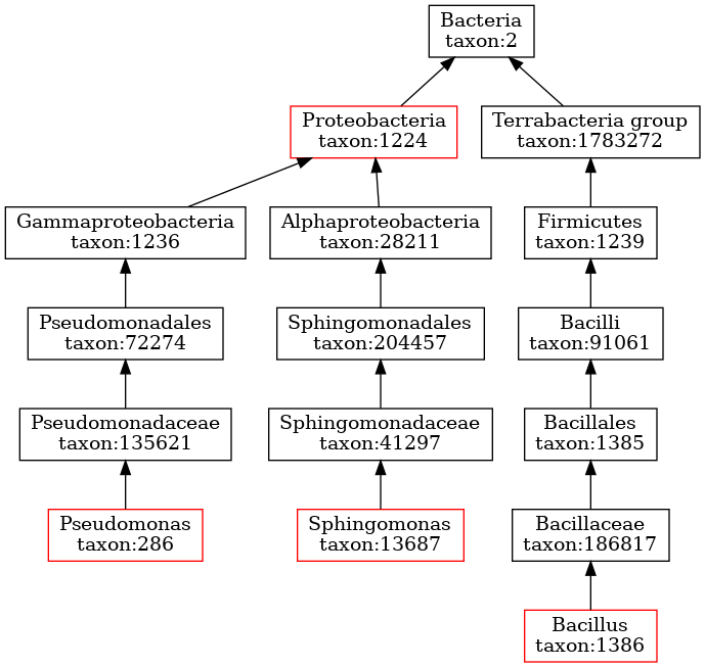
- 56 806 920 triplets au format RDF
- Base RDF déployée sur Genocloud dans une instance d'AskOmics
- `https://gitlab.inria.fr/gogepp_team/stage-yael`

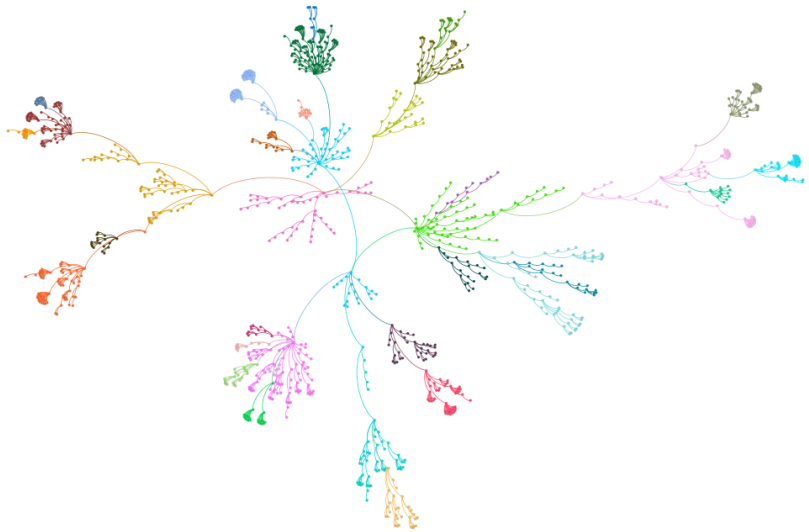
Contribution 4 : banque de requêtes

- 50 requêtes SPARQL
  - validation
  - contrôle qualité
  - évaluation perf.
- mises à disposition
  - templates
  - formulaires



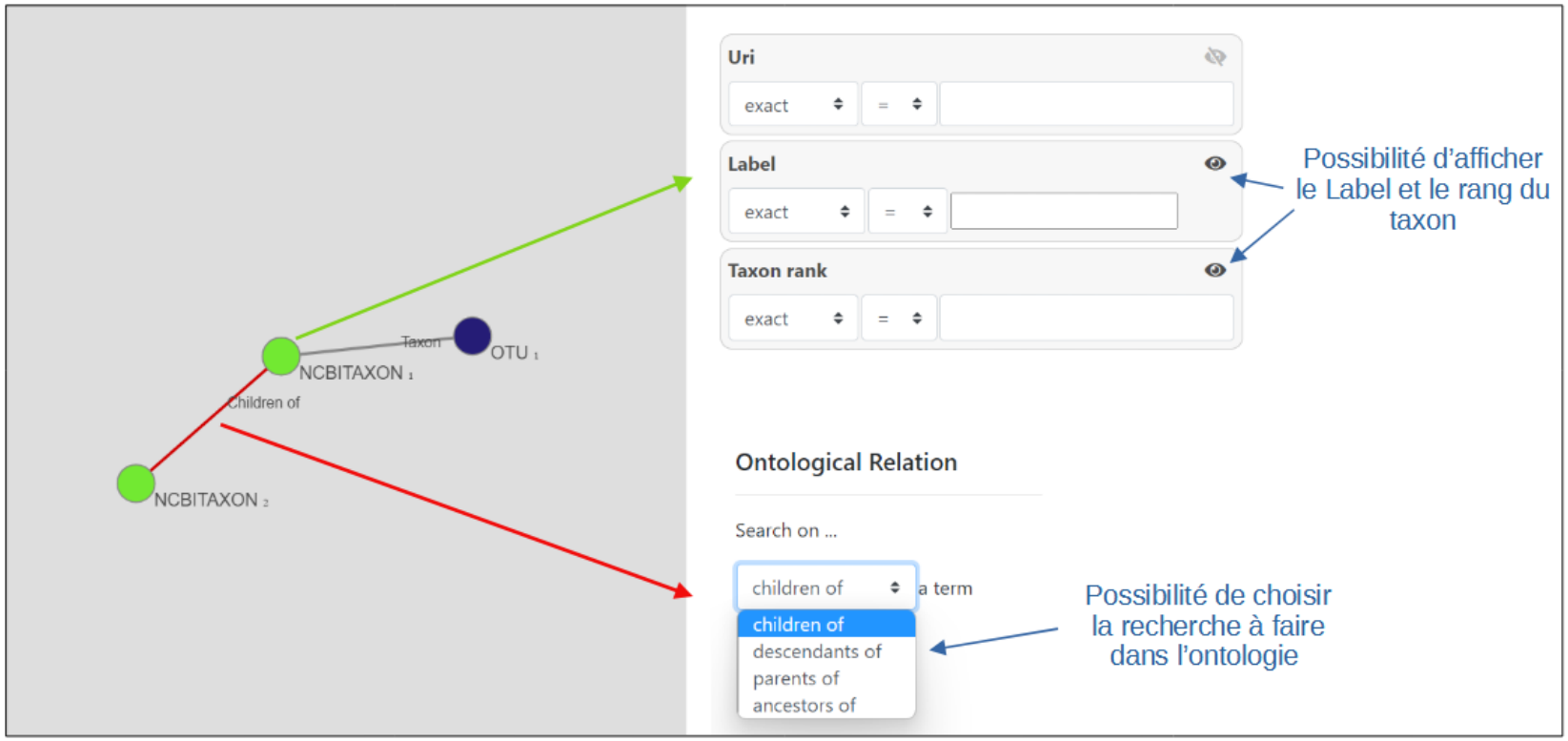
Contribution 5.1 : Récupération d'une partie de l'ontologie NCBI TAXON





- 855 taxons extraits des fichiers OTUs
- 1202 avec la hiérarchie

Contribution 5.2 : Intégration ontologie dans AskOmics



AskOmics contient des primitives de parcours d'ontologies

Synthèse

Conclusion

Contributions

1. Rassembler et modifier les fichiers
2. Définir la structure de données
3. Peupler la base avec les fichiers
4. Créer une librairie de requêtes
5. Intégrer l'ontologie NCBI TAXON

Perspectives

- Certaines requêtes ne sont pas possible à reproduire sur AskOmics mais peuvent être écrites directement en SPARQL
- Automatisation de l'intégration des fichiers sur AskOmics