

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

Olivier Dameron¹, **Yael Tirlet**^{1,2},
Matéo Boudet^{2,3}, Fabrice Legeai²

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

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

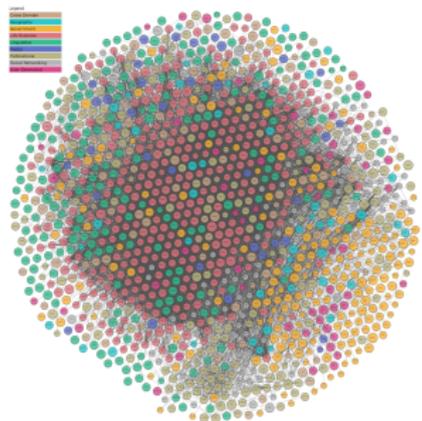
Définir un schéma de données permettant :

- d'éviter de n-plier 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



<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 dépliant 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

Integrate

OTUs.txt (preview)

OTU	OTU_count	Taxon_level	is@Taxon	Taxon_count
Start entity	Numeric	Category	Directed	Numeric
OTU1	108544	GENUS	Pseudomonas	93672
OTU2	68769	FAMILY	Xanthomonadaceae	67344
OTU3	66438	GENUS	Bacillus	64847
OTU4	63649	FAMILY	Chlorophagaceae	63616
OTU5	43392	GENUS	Bacillus	40282

Entité de départ

Fait le lien avec l'entité « Taxon »

Attributes
Numeric
Text
Category
Boolean
Date
Faldo attributes
Reference
Strand
Start
End
Relation
Directed
Symetric

Les différents types que l'on peut donner aux colonnes

Integrate (private dataset) Integrate (public dataset)

Intégrer les données

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

Ask!

Select an entity to start a session:

Source ▾	Filter entities
gene	local 
CDS	local 
UTR	local 
mRNA	local 
OTU	local 
exon	local 
ncRNA	local 
phy/chem ID	local 

Start!

Or start with a simplified form:

genes of chromosome

Expression of genes between contrasts

DEGs : Conditions & Dilutions

Or start with a template:

DEGs between T1 & T2 (category)

DEGs between T1 & T2 (balls)

OTUs & Taxons where mean_count > 1000 for D0

Physico-chemical characteristics of the Dilutions

AskOmics (2.1) : Construction itérative requête SPARQL

AskOmics

Ask!

Select an entity to start a session:

Source ▾	Filter entities
gene	local 
CDS	local 
UTR	local 
mRNA	local 
OTU	local 
exon	local 
ncRNA	local 
phy/chem ID	local 

Start!

Or start with a simplified form:

genes of chromosome

Expression of genes between contrasts

DEGs : Conditions & Dilutions

Or start with a template:

DEGs between T1 & T2 (category)

DEGs between T1 & T2 (balls)

OTUs & Taxons where mean_count > 1000 for D0

Physico-chemical characteristics of the Dilutions

AskOmics (2.2) : Construction itérative requête SPARQL

Query Builder

Filter links Filter nodes

Show FALDO relations Remove Node

Uri

Label

reference
chrCnn_random
chrA03_random

strand
+
-

ID

Alias

Name

Run & preview Run & save

AskOmics (2.3) : Construction itérative requête SPARQL

Query Builder

Filter links Filter nodes Show FALDO relations

Run & preview

Uri =

Label =

reference
chrCnn_random
chrA03_random

strand
+
-

ID =

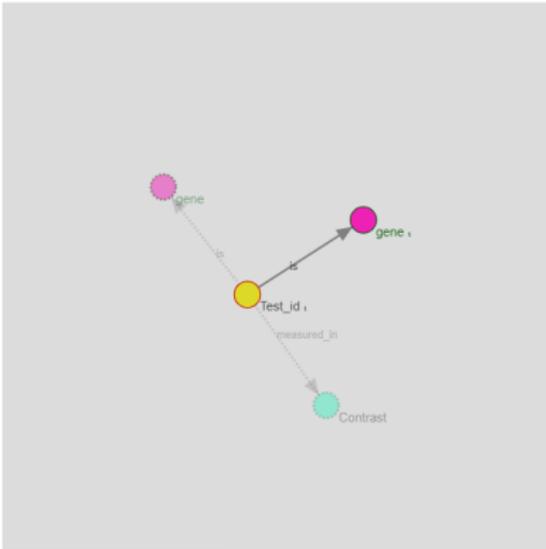
Alias =

Name

AskOmics (2.4) : Construction itérative requête SPARQL

Query Builder

Filter links Filter nodes Show FALDO relations



Run & preview

Uri =

Label =

Expression
T1STD0<T1STD3
T1STD0=T1STD3

Significance
1
0

T1STD0/T1STD0 = +

T1STD3/T1STD3 = +

T1SYD3/T1SYD3

AskOmics (2.5) : Construction itérative requête SPARQL

Query Builder

Filter links Filter nodes Show FALDO relations

Run & preview

Uri =

Label =

Expression
T1SYD6>T2SYD6
T1SYD6=T2SYD6

Significance
1
0

T1STD0/T1STD0 = +

T1STD3/T1STD3 = +

T1SYD3/T1SYD3

AskOmics (3) : Traitement d'une requête SPARQL

(3) SPARQL query is processed by the endpoint

Results

<input type="checkbox"/>	Id	Description	Exec time	Template	Form	Public	Status	Rows	Size	Actions
<input type="checkbox"/>	28	Genes and chromosome whose expression is the same at T1 and T2	17s	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	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

```
1 PREFIX : <http://askomics.org/ont#>
2 PREFIX askmedici <http://askomics.org/interact/>
3 PREFIX dc <http://purl.org/dc/terms/>
4 PREFIX faido <http://biobankdash.org/biobankdash/>
5 PREFIX owl <http://www.w3.org/2002/07/owl#>
6 PREFIX rdfs <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
7 PREFIX rdf <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
8 PREFIX rdfs <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
9 PREFIX xsd <http://www.w3.org/2001/XMLSchema#>
10
11 SELECT DISTINCT ?gene1_Label ?gene1_reference ?test_id1_Label
12 WHERE {
13   ?test_id1_Label <http://askomics.org/data/tx/> ?gene1_ref .
14   ?gene1_ref rdfs:type <http://askomics.org/data/gene1 .
15   ?gene1_ref rdfs:label ?gene1_Label .
16   ?gene1_ref faido:localization faido:reference ?gene1_referencecategory .
17   ?gene1_referencecategory rdfs:label ?gene1_reference .
18 }
```

(1) SPARQL query is generated automatically

gene1_Label	gene1_reference	Test_id1_Label
EnaA02g34680D	chrA02	T1SYD6v1T2SYD6_BnaA02g34680D
EnaA02g09470D	chrA02	T1SYD6v1T2SYD6_BnaA02g09470D
EnaA02g10460D	chrA02	T1SYD6v1T2SYD6_BnaA02g10460D
EnaA02g14050D	chrA02	T1SYD6v1T2SYD6_BnaA02g14050D
EnaA02g15610D	chrA02	T1SYD6v1T2SYD6_BnaA02g15610D
EnaA02g16920D	chrA02	T1SYD6v1T2SYD6_BnaA02g16920D
EnaA02g21750D	chrA02	T1SYD6v1T2SYD6_BnaA02g21750D
EnaA02g24150D	chrA02	T1SYD6v1T2SYD6_BnaA02g24150D
EnaA02g25530D	chrA02	T1SYD6v1T2SYD6_BnaA02g25530D
EnaA02g29010D	chrA02	T1SYD6v1T2SYD6_BnaA02g29010D

Résultats

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



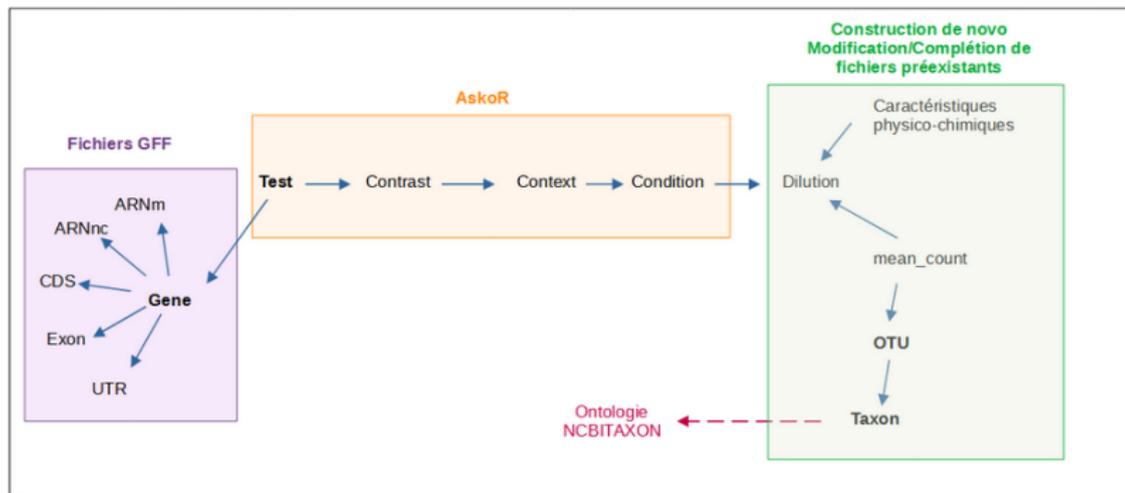
NCBI Taxon

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

The diagram illustrates the contribution of a query bank, showing two contexts (CONTEXTE 1 and CONTEXTE 2) and a query interface (A and B).

CONTEXTE 1 and **CONTEXTE 2** are represented by complex graphs of nodes and edges. In CONTEXTE 1, a green box highlights "Conditions : Yudal - T1 - Saine" and a purple box highlights "Dilution D0/High". In CONTEXTE 2, a green box highlights "Conditions : Yudal - T1 - Saine" and a purple box highlights "Dilution D6/Low".

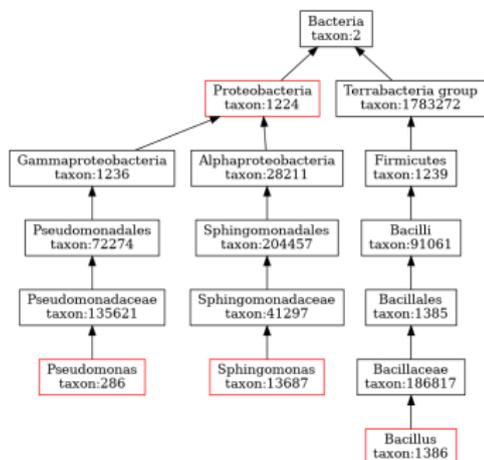
A shows a query interface with fields for URI, Label, Expression, and Significance. The URI field is set to "exact", the Label field is set to "exact", and the Expression field contains the SPARQL query: `111900-111904` and `111900-111904`. The Significance field is set to "1".

B shows a query interface with fields for URI, Label, Expression, and Significance. The URI field is set to "exact", the Label field is set to "exact", and the Expression field contains the SPARQL query: `111900-111904` and `111900-111904`. The Significance field is set to "1".

Contexte 1 :
Yudal
T1
Saine
Dilution D0/High

Contexte 2 :
Yudal
T1
Saine
Dilution D6/Low

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

The screenshot displays the AskOmics interface. On the left, an ontology diagram shows three nodes: a green node labeled 'NCBITAXON₁', a blue node labeled 'OTU₁', and another green node labeled 'NCBITAXON₂'. A red line labeled 'Children of' connects 'NCBITAXON₁' to 'NCBITAXON₂'. A grey line labeled 'Taxon' connects 'NCBITAXON₁' to 'OTU₁'. A green arrow points from the 'NCBITAXON₁' node to the 'Uri' field in the configuration panel. A red arrow points from the 'Children of' relationship to the 'Ontological Relation' dropdown menu.

The configuration panel on the right includes three input fields for 'Uri', 'Label', and 'Taxon rank', each with a dropdown menu set to 'exact' and an equals sign. To the right of these fields are eye icons and a blue annotation: 'Possibilité d'afficher le Label et le rang du taxon'. Below these fields is the 'Ontological Relation' section, which has a 'Search on ...' label and a dropdown menu. The dropdown menu is open, showing options: 'children of', 'children of', 'descendants of', 'parents of', and 'ancestors of'. A blue arrow points to the second 'children of' option with the annotation: 'Possibilité de choisir la recherche à faire dans l'ontologie'.

AskOmics contient des primitives de parcours d'ontologies

Synthèse

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 AskOmic mais peuvent être écrites directement en SPARQL
- Automatisation de l'intégration des fichiers sur AskOmic