# Intégration de données agricoles et environnementales à l'aide des technologies du Web sémantique

Yael Tirlet

Yael TIRLET

Master 2 Bioinformatique et Génomique – Parcours Informatique et Biologie Intégrative

Année 2021/2022

---

# Intégration de données agricoles et environnementales à l'aide des technologies du Web sémantique

---

Maîtres de stage : Matéo BOUDET et Olivier DAMERON

IRISA – Équipe Dyliss – 263 Avenue Général Leclerc 35000 Rennes



Master bioinformatique

UNIVERSITÉ DE RENNES 1

# ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) Yael TIRLET.....
étudiant(e) en Master 2 Bioinformatique et Génomique.....
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 25/05/2022

Signature : 

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

# Remerciements

Je tiens à remercier toute l'équipe Symbiose pour son accueil et la bonne ambiance qui règne au sein de L'IRISA.

Je souhaite remercier particulièrement mes maîtres de stage Matéo Boudet et Olivier Dameron qui m'ont encadrée sur ce stage et m'ont permis de progresser à bien des niveaux, ainsi que

Fabrice Legeai et Kévin Gazengel qui m'ont été d'une aide précieuse durant ce projet.

# Abréviations

B.napus : Brassica napus (colza)

DEG : Differencially Expressed Gene

IRI : International Resource Identifier

M : Malade

NCBITAXON : Ontologie ‘NCBI Taxonomy’

OTU : Operational Taxonomic Unit

P.brassica : Plasmodiophora brassicae

RDF : Ressource Description Framework

S : Saine

T : Tenor (phénotype du colza)

T1 : Temps intermédiaire

T2 : Temps final

VM : Virtual Machine

Y : Yudal (phénotype du colza)

# Sommaire


Introduction 1

I - Matériel et Méthodes 2
- 1- Article et jeu de données original 2
- 2- AskOmics 2
- 3- AskoR 5
- 4- Construction de Machines Virtuelles sur Genouest 5
- 5- Ontologie NCBITAXON 5

II- Résultats 6
- 1- Rassemblement des données à intégrer dans AskOmics 6
  - a- Fichiers d'expression génique 6
  - b- Fichiers Contrast – Context – Condition 6
  - c- Fichiers gff 6
  - d- Fichiers Soil – Dilution 7
  - e- Fichiers de comptage des OTUs 7
  - f- Fichier des caractéristiques physico-chimiques des sols 7
  - g- Fichiers OTUs et Taxons 8
  - h- Récupération de l'ontologie NCBITAXON 10
- 2- Choix des requêtes 11
- 3- Structure des données et peuplement d'AskOmics 12
- 4- Requêtes 14
  - a- Validation 14
  - b- Comparaison des méthodes 14
- 5- Templates et Formulaires 15
- 6- Intégration du NCBITAXON à AskOmics 17

III- Discussion 19
- 1- Correction de fichiers d'entrées de AskoR 19
- 2- Problèmes rencontrés lors de la modification et de la complétion du fichier d'OTU 20
- 3- Pertinence des différents fichiers intégrés 21
- 4- Intégration d'une partie du NCBITAXON 21
- 5- Validation du modèle 22
- 6- Limites 23
- 7- Perspectives 24

Conclusion 25

# Introduction

Les interactions entre les plantes et la faune du sol ne sont aujourd'hui un secret pour personne : Bactéries, Fungi et Algues échangent avec la flore de différentes manières [1]. Ces interactions sont essentielles pour les plantes, notamment pour l'approvisionnement en azote [2]. Il existe des relations privilégiées entre les plantes et certains micro-organismes de leur rhizosphère (partie du sol entourant les racines) grâce à des substances chimiques dégagées par les racines ou produites par les micro-organismes eux-mêmes [3]. De nombreuses études cherchent à comprendre ces interactions et les conséquences que ces micro-organismes et leur diversité peuvent avoir sur la vie des plantes [4-5-6].

Le projet Deep Impact dans lequel le stage s'inscrit vise à comprendre l'impact que la diversité microbienne du sol ainsi que les conditions physico-chimiques et environnementales ont sur la santé et le rendement des cultures de blé et de colza. Il regroupe trois équipes de L'INRAE (Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : équipes de Rennes, Dijon et Toulouse) sur les années 2021/2022 et 2022/2023.

Dans trois régions de France (Occitanie-Toulouse, Bourgogne-Franche Comté et Bretagne-Normandie), des relevés seront faits sur différents champs pour comparer les communautés microbiennes et comprendre les interactions entre celles-ci et les plantes. Seront relevées des données caractérisant le sol : population microbienne (comptage et classification taxonomique des bactéries, champignons), pH, teneur en différents minéraux… mais aussi les différentes espèces constituant l'écosystème comme les plantes adventices et les insectes ravageurs. En plus de cela, des données météorologiques pourront être obtenues grâce à des petites stations météo dans chaque parcelle étudiée. Enfin, afin de comprendre l'impact de l'environnement sur les cultures, des relevés seront effectués à différentes périodes pour quantifier et qualifier le développement et la croissance des plantes (automne, fin hiver et fin printemps) et pour analyser les sols (automne et printemps).

Tout cela va générer de nombreuses données hétérogènes qu'il convient de transformer avant de pouvoir les interroger. Dans ce contexte on se demande alors sous quelle forme il faudra modifier et structurer les données afin qu'elles soient facilement et efficacement manipulables dans le cadre du projet Deep Impact.

Le Web sémantique [7-8-9] pourrait être une solution adaptée à ce genre de données diversifiées et hétérogènes. Le Web sémantique permet de représenter les données sous formes de graphes et de les lier entre elles. En revanche son utilisation nécessite la maîtrise des langages RDF et SPARQL, ce qui est un frein pour des utilisateurs biologistes. De plus les données du projet ne sont pas à ce jour au format du Web sémantique. Le défi est donc d'allier Web sémantique et simplicité d'utilisation pour les biologistes.

**Les objectifs de ce stage sont alors de rassembler et transformer les données à intégrer et de définir les requêtes à faire sur celles-ci. Dans un deuxième temps il faudra choisir une structure pour l'intégration des données et peupler la base RDF sur ce modèle. Enfin, il sera temps d'exécuter les requêtes définies.**

# I - Matériel et méthodes

Le projet Deep Impact étant à ses débuts, nous n'avons pas encore tous les relevés. Nous avons alors décidé de travailler sur des données similaires issues d'autres projets afin de se familiariser avec elles et de déterminer les requêtes que nous aurons besoin de faire, qui influenceront sur le schéma des données. Grâce à la solution logicielle AskOmics (cf I-2), on peut aussi garder en mémoire ces requêtes pour les réutiliser. Les requêtes ainsi faites sur ces données pourront servir de modèle à celles du projet Deep Impact. Nous avons donc travaillé à partir d'un article [10] et nous avons pu récupérer les annexes et les données brutes associées.

## 1- Article et jeu de données originaux

L'article « *Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes* » [10] traite de l'impact des différentes communautés microbiennes du sol (notamment *Plasmodiophora brassicae*) sur la résistance de *Brassica napus* (colza) à une maladie (hernie).

Différentes conditions sont testées pour comprendre les interactions entre le microbiote et les plantes. Deux génotypes différents sont donc comparés : Tenor et Yudal, dans plusieurs conditions. En effet, on compare les plantes infectées et saines, plusieurs milieux de culture (3 milieux plus ou moins riches en micro-organismes) ainsi qu'à deux temps différents (temps intermédiaire et temps final). Les relevés sont faits pour trois répétitions pour chacune des dilutions. Les caractéristiques physico-chimiques des différents sols sont elles aussi connues (mesures faites avant semence).

Ainsi, les données extraites de l'article sont des comptages bruts d'expression de gènes de Brassica napus (2 fichiers d'environ 15Mo et de plus de 100 000 lignes chacun), des données sur les différents sols (Annexe 8 de l'article 10) et des descriptions et comptages des Unités Taxonomiques Opérationnelles (OTUs) (plus de 32000 lignes chacun).

## 2- AskOmics

Le logiciel AskOmics [11] permet à la fois d'intégrer des jeux de données (entre eux, avec des bases de connaissances) en les transformant en RDF et de composer de façon intuitive des requêtes complexes sur ces données, sans nécessité de connaissance du langage SPARQL. Il fournit ainsi aux biologistes et aux expert(e)s une interface pour manipuler des données qui sont représentées grâce au Web Sémantique [7-8-9] :

RDF pour la représentation et l'intégration de jeux de données *, et SPARQL pour leur interrogation ** [12-13]. Cette application Web permet de répondre à un besoin de rassembler des données hétérogènes, contenues dans des fichiers multiples aux formats différents (fichiers tabulés csv et tsv ou formats spécifiques comme GFF). Ces données sont converties dans un langage commun et AskOmics permet alors de créer simplement des requêtes habituellement compliquées.

L'étape d'intégration va permettre la conversion des données au format RDF. Lors de l'intégration des données, on choisit (par tableau) l'entité qui pourra être un point de départ des requêtes. Les autres colonnes sont alors des attributs associés à cette entité de départ (Figure 1 et Figure 2). Ces attributs peuvent être du texte, des nombres, des catégories ou bien un « lien » vers une autre entité de départ (sous la forme 'xxx@YY' avec 'YY' le nom de la deuxième entité de départ) ; ce qui permet de lier les données entre elles et de faire des requêtes sur les liens entre entités. Si le tableau représente des données décrivant une position sur un génome (type QTL par exemple), des catégories 'Reference' (chromosome), 'strand', 'start' et 'end' sont préexistantes.

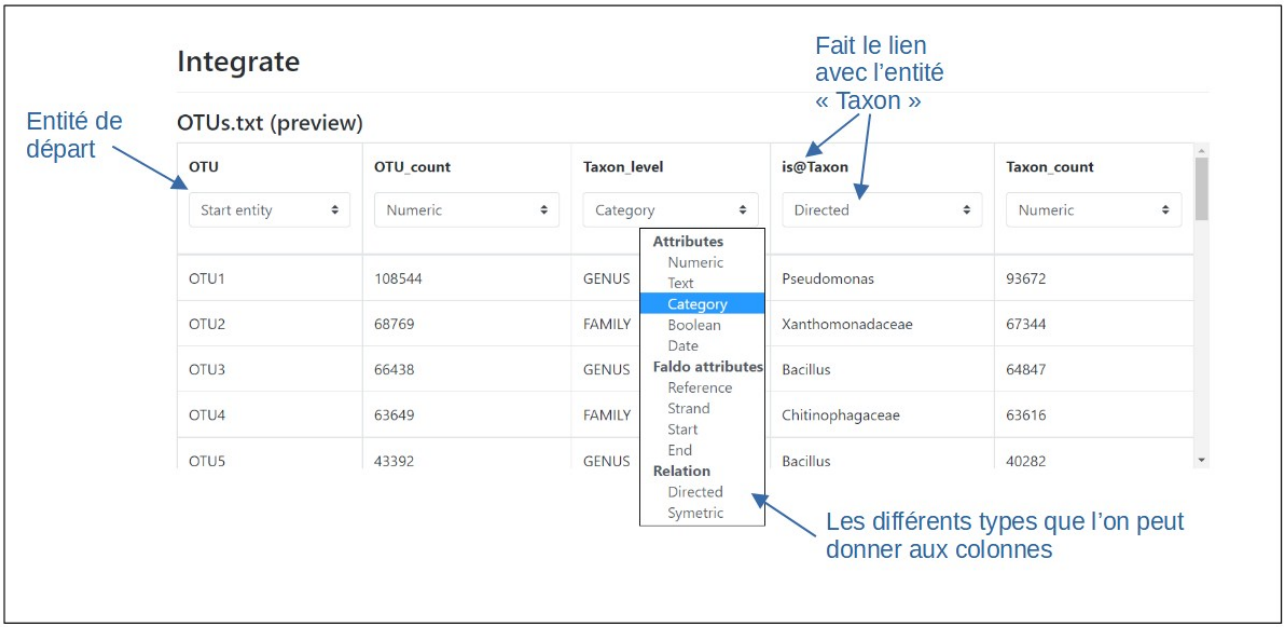


**Figure 1: Capture d'écran du menu d'intégration d'un fichier tabulé sur AskOmics**

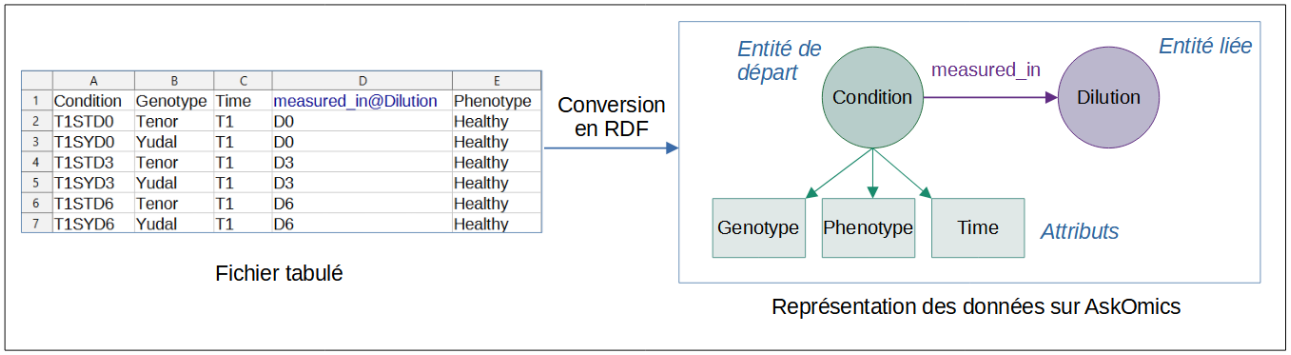


**Figure 2: Résultat de la conversion d'un fichier en RDF par le logiciel AskOmics**

* http://www.w3.org/RDF/

** http://www.w3.org/TR/sparql11-overview/

Après l'intégration de toutes les données, celles-ci sont stockées sous forme de graphes dans la base RDF. La structure des données issue des entêtes des fichiers tabulés servira à la création de requêtes. L'utilisateur peut soit écrire directement des requêtes SPARQL pour interroger les données RDF, soit utiliser l'outil d'aide à la composition de requêtes. Il s'agit d'une interface en 'clique-boutons' pour que n'importe quel utilisateur n'ayant jamais fait de SPARQL puisse s'en sortir facilement.

Les entités de départ (par exemple 'Gene', 'Contrast'…) sont représentées par des boules (Figure 3). Lorsque l'on clique sur une boule apparaissent alors les différents attributs associés à l'entité. On peut ainsi choisir les colonnes à afficher, appliquer des filtres (choisir une catégorie en particulier, restreindre une colonne numérique aux entités dont ce chiffre ne dépasse pas un certain seuil…). Si des liens existent entre des entités de départ, il y aura une flèche entre les deux boules correspondantes. En suivant cette flèche on peut donc compléter la requête. En effet il est possible de composer les requêtes de façon itérative : « Afficher tous les tests », « Afficher tous les tests et les gènes associés », « Afficher tous les tests et les gènes associés qui sont situés sur le chromosome A10 »...

On peut décider de lancer un aperçu de la requête pour voir quelques résultats et la forme qu'ils auront ; ce qui peut permettre de vérifier que la requête correspond bien à ce que l'on veut. Si c'est le cas on peut alors cliquer sur 'Run & save' ; la requête (qui peut être longue ou générer un nombre important de réponses) sera lancée sur toutes les données ; et les résultats ainsi que la requête effectuée seront sauvegardés. Les requêtes peuvent être enregistrées pour être réutilisées plus tard (notamment sur un autre jeu de données ayant la même structure), partagées avec d'autres utilisateurs ou servir de modèle en laissant l'utilisateur choisir certains paramètres.

C'est sur cet outil que nous avons choisi d'intégrer les données pour les regrouper et les structurer.

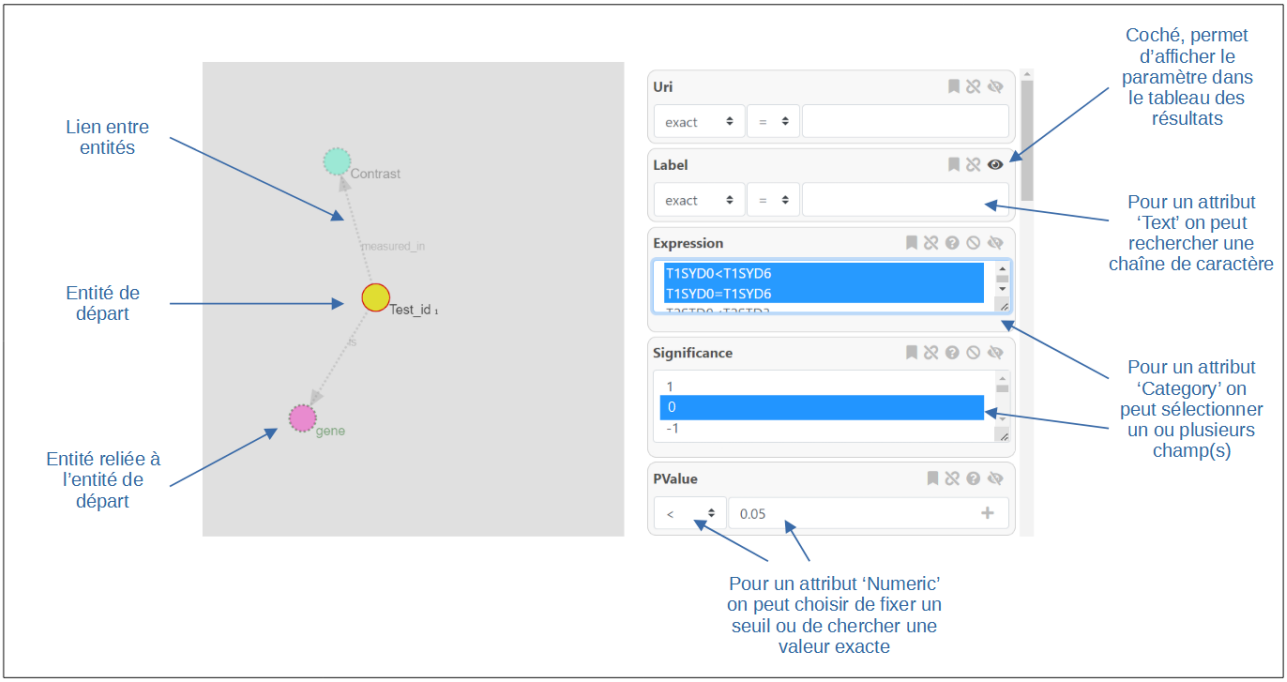


**Figure 3: Capture d'écran de la construction d'une requête sur AskOmics**

## 3 - AskoR

Lors de la structuration et complétion des données, nous avons utilisé AskoR [14] sur nos comptages bruts afin de récupérer des données utilisables sur AskOmics.

AskoR est un pipeline développé sur R qui permet de faire des traitements statistiques sur des données d'expression génétique et de transcriptomique, puis de formater le résultat pour pouvoir l'importer dans AskOmics.

A partir de données brutes (comptages), on va pouvoir faire différents tests statistiques, divers graphiques et des tables de données. Sont inclues dans le programme des étapes de normalisation et de validation des données. AskoR offre à l'utilisateur une utilisation personnalisée car le code peut être facilement remodelé pour correspondre aux différents tests et/ou graphiques voulus.

Le code renvoie aussi une série de fichiers intégrables dans AskOmics : des tableaux d'expression génétique différentielle et des fichiers 'Contrast', 'Context' et 'Condition'. Ce sont ces différents fichiers qui nous serviront par la suite.

## 4 – Construction de Machines Virtuelles sur Genouest

Afin d'utiliser AskOmics, je l'ai déployé sur des Machines Virtuelles (VMs) sur la plateforme de bio-informatique Genouest*. Cela permet d'avoir accès à des ressources matérielles plus conséquentes, pour faciliter l'utilisation d'AskOmics.

## 5- Ontologie NCBITAXON

Une ontologie [15] est un ensemble structuré de termes qui sont liés les uns aux autres par des assertions. L'ontologie NCBITAXON correspond à une classification des espèces décrites dans les bases de données publiques. J'ai utilisé le NCBITAXON premièrement via l'EBI ** afin de récolter des données sur les OTUs. J'ai plus tard récupéré l'ontologie sous format RDF depuis Bioportal [16-17] afin de pouvoir l'interroger en SPARQL et ajouter une partie de l'ontologie sur AskOmics.

---

* GenOuest bioinformatics core facility (https://www.genouest.org)

** https://www.ebi.ac.uk/

# II - Résultats

## 1 – Rassemblement des données à intégrer dans AskOmics

Nous avons décidé d'intégrer une grande quantité de fichiers (55 fichiers, presque 600 Mo de données) sur AskOmics afin de regrouper les données. Ces fichiers ont été crées soit via AskoR, soit de novo, soit à partir de fichiers préexistants. J'ai intégré les fichiers suivants :

### a - Fichiers d'expression génique

Les 40 fichiers d'expression différentielle des gènes de *Brassica napus* couvrant les différents contrastes ont été obtenus à partir des données brutes via AskoR. Ils ont tous la même structure. Ils contiennent 11 colonnes différentes correspondant à l'identifiant du Test, au contraste, à l'expression génique dans chacun des deux contextes, au gène concerné, à la p-value, au FC…

### b - Fichiers Contrast – Context – Condition

Les fichiers Contrast, Context et Condition sont eux aussi créés par AskoR. Ils sont liés ensemble et indissociables : un contraste s'écrit sous la forme Context1vsContext2 et, de la même façon, un contexte correspond à un ensemble de conditions.

Par exemple le contraste 'T2MTD0vsT2MYD0' se décompose en les deux contextes 'T2MTD0' et 'T2MYD0' dont les conditions sont respectivement : 'Temps final' (T2) 'Malade' (M) 'Tenor' (T) 'Dilution D0' (D0) et 'Temps final' (T2) 'Malade' (M) 'Yudal' (Y) 'Dilution D0' (D0). Ce sont ces fichiers qui vont permettre de faire des comparaisons entre les différentes conditions et les lier aux données d'expression génétique.

### c - Fichiers gff

J'ai aussi intégré deux fichiers gff de *Brassica napus*. Un fichier gff contient des informations relatives aux gènes, ARN, exons… On y retrouve notamment des séquences et leur position dans le génome : début, fin, brin, phase.

### d - Fichiers Soil – Dilution

Dans les fichiers de contraste, les répétitions de chaque sol ne sont pas mentionnées. On a simplement la dilution avec laquelle on les a préparés. Le fichier de conditions renvoie donc à une entité Dilution qui indique si le sol est hautement (D0/'High'), moyennement (D3/'Medium') ou faiblement (D6/'Low') riche en micro-organismes.

En revanche, dans les fichiers de comptage des OTU (cf II-1-e), les comptages bruts ont été faits dans chacune des trois répétitions des trois dilutions différentes. Ainsi, on a créé une entité Soil (Sol) qui prend en compte la répétition et qui est aussi liée à l'entité Dilution. Il est important de remarquer que la répétition A pour la dilution D0 n'a rien à voir avec la répétition A pour la dilution D3 par exemple. L'entité Soil est très peu utilisée dans les requêtes que j'ai faites mais permet de garder les informations de répétition.

### e - Fichiers de comptage des OTUs

Toutes les OTUs ont été comptabilisées dans les trois répétitions de chaque dilution.
Le fichier de comptage a été remodelé par un script Python que j'ai écrit pour qu'il soit lisible par le logiciel et facilement manipulable avec les autres données.

Pour ne pas s'embêter avec les répétition qui ne sont pas beaucoup utilisées, j'ai aussi écrit un script qui calcule la moyenne par dilution pour chacune des OTU. On a alors le choix de regarder dans le détail des répétitions ou uniquement la moyenne. C'est ce fichier de moyennes que j'ai utilisé principalement pour les requêtes. Il est directement lié à l'entité Dilution.

### f - Fichier des caractéristiques physico-chimiques des sols

Enfin j'ai ajouté un tableau donnant les caractéristiques physico-chimiques des trois sols (pH, teneurs en minéraux et matière organique…). Cette table correspond à l'annexe numéro 8 de l'article utilisé ; je l'ai simplement légèrement modifiée pour convenir au logiciel.

## g- Fichiers OTUs et Taxons

Le fichier brut décrivant chacune des OTUs n'était pas du tout pratique à lire et encore moins à manipuler. Chaque OTU était décomposée en Phylums, Ordres, Familles… Dans chaque colonne on trouvait alors les noms de plusieurs taxons correspondants avec leur comptage respectif entre parenthèse (Figure 4). Pour simplifier la lecture de ce fichier, j'ai écrit un script en Python.

J'ai alors décidé de garder idéalement (pour chaque OTU) le taxon avec le compte maximum pour le Genre (c'est à dire le taxon majoritaire le plus précis). Le taxon est retenu uniquement si son compte est supérieur à 70 % du compte total des taxons de l'OTU et s'il est connu (on ne retient pas les 'Unknown').

Si ces conditions ne sont pas remplies, on remonte d'un niveau (ex : on remonte de Genus à Family). Si on finit par remonter jusqu'au Phylum, on garde le taxon majoritaire, quelque soit la proportion. Dans ce cas là, si le taxon majoritaire est 'Unknown' le programme python change son nom par 'Unidentified'. Après ce premier traitement où, pour chaque OTU, on garde un taxon, son niveau, son compte et le compte total des taxons, on va faire une requête sur l'ontologie NCBITAXON via l'API de l'EBI.

Cela m'a permis de récupérer, pour chaque taxon, le label correspondant dans l'ontologie et l'IRI associé. L'IRI pourra me permettre plus tard de faire des requêtes SPARQL depuis AskOmics sur l'ontologie NCBITAXON. Mais sur le moment, l'IRI m'a servi à faire une deuxième requête pour trouver les synonymes associés au taxon. Tous ces paramètres sont alors stockés dans un second fichier pour chaque taxon distinct.
Les fichiers ainsi formés peuvent alors être intégrés dans AskOmics.

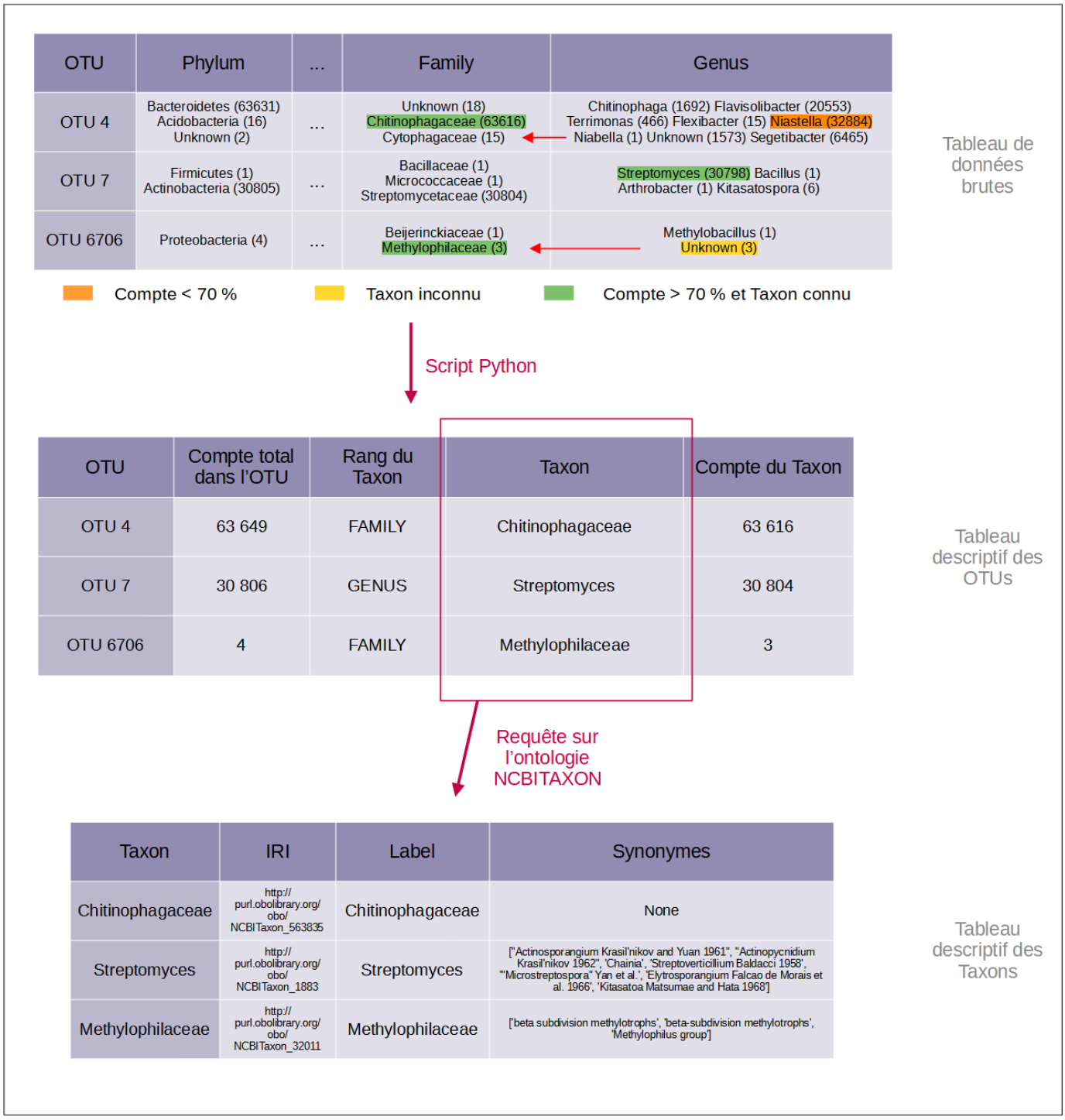


**Figure 4: Processus de création des fichiers de description des OTUs et des Taxons**

### h – Récupération de l'ontologie NCBITAXON

Afin d'avoir plus d'informations sur les Taxons présents dans chacune des OTUs, nous avons voulu récupérer une partie de l'ontologie NCBITAXON qui pourrait être intégrée sur AskOmics. Pour cela j'ai récupéré l'ontologie sur Bioportal au format ttl. Il s'agit d'un fichier de 1,02 Gigaoctets contenant 1 983 907 classes et jusqu'à 10 relations par classe.

De plus, à partir de notre fichier de Taxons, j'ai extrait les IRIs, en enlevant l'IRI correspondant à 'Unidentified' et les champs vides (certains Taxons n'avaient pas d'IRI).

A partir de ces environ 850 IRIs de Taxons distincts, j'ai construit une requête SPARQL visant à trouver l'ancêtre ou les ancêtres commun(s) le(s) plus bas. Pour lancer cette requête, j'ai utilisé le serveur SPARQL Fuseki* sur lequel j'ai chargé le fichier RDF du NCBITAXON.

Après plusieurs essais infructueux et quelques heures de recherches, je me suis rendu compte que, parmi nos Taxons, il y avait bien 838 Bactéries mais aussi 4 Eucaryotes (*Bosea yervamora, Labrys fuzhouensis, Lamprocystis* et *Eremococcus turbinatus*), 5 Archées (*Thaumarchaeota, Euryarchaeota, Methanosarcinaceae, Methanosarcina* et *Ignisphaera*) et 2 Taxons qui ne semblaient pas appartenir à l'ontologie (*Bacillus* et *Candidatus Liberibacter*). En effet, pour ces deux derniers Taxons, la requête python sur l'EBI avait renvoyé un IRI qui n'était pas reconnu lors des requêtes sur l'ontologie NCBITAXON. Des recherches à la main sur le site de l'EBI et celui de Bioportal m'ont permis de voir que *Bacillus* et *Candidatus Liberibacter* appartenaient bien à l'ontologie. Pour le premier, la requête python ne renvoyait pourtant pas le bon IRI. Pour le deuxième, son Label était '*Liberibacter*' ; '*Candidatus Liberibacter*' correspondant au synonyme associé. Les requêtes pour ces deux taxons renvoyaient donc à des sous-classes qui n'apparaissaient pas dans l'ontologie téléchargée. J'ai alors récupéré les bons IRIs à la main et les ai rajoutées dans la liste des IRIs.

Après avoir mis de coté les Archées et les Eucaryotes, j'ai lancé une requête sur les Bactéries afin de trouver leur plus petit ancêtre commun. On trouve alors '*Bacteria*' ; en effet plusieurs de nos taxons sont déjà au rang Phylum.

Dans un deuxième temps, j'ai créé une deuxième requête SPARQL qui m'a permis de récupérer un arbre depuis les taxons jusqu'à '*Bacteria*' et qui recense tous les rdfs:subClassOf, les types, les rangs et les labels de chaque taxon. Ce fichier complété contient alors 1202 taxons distincts. La hiérarchie de ce graphe RDF a été schématisée dans la Figure 5. La Figure 6 illustre quant à elle la structure simplifiée du graphe sur uniquement quelques taxons.

---

* https://jena.apache.org/documentation/fuseki2/

## 2 – Choix des requêtes

Afin de valider l'utilisation d'AskOmics pour interroger ce type de données, j'ai parcouru l'article à la recherche de résultats reproductibles. N'ayant que les données brutes complètes chez *B.napus* mais pas chez *P.brassica*, seuls les résultats chez *B.napus* peuvent alors être reproduits.
J'ai alors choisi de reproduire, entre autres, ces différentes requêtes :

- DEGs entre D0 et D3 à T2 pour les plantes Tenor malades
- Comparaison entre T1 et T2 des DEGs D0vsD3 et D0vsD6 pour les plantes Yudal saines
- Comparaison pour les plantes Tenor malades entre T1 et T2 des DEGs D0vsD3 et D0vsD3
- …

J'ai aussi choisi différentes requêtes-tests à effectuer afin de vérifier la fonctionnalité de la structure et la bonne liaison des données entre elles.

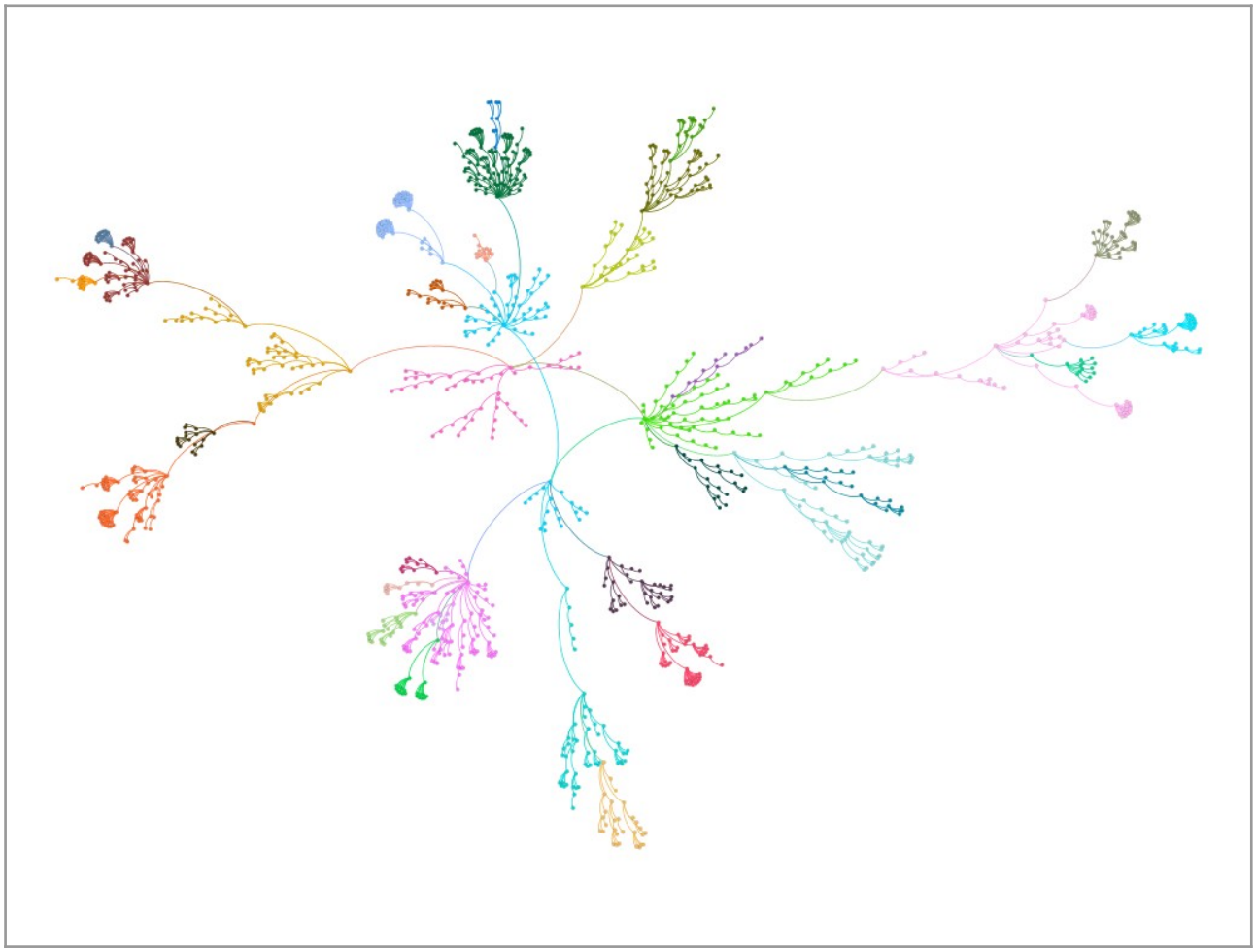

**Figure 5: Graphe de la hiérarchie des 850 taxons de l'étude dans la partie extraite du NCBITAXON**

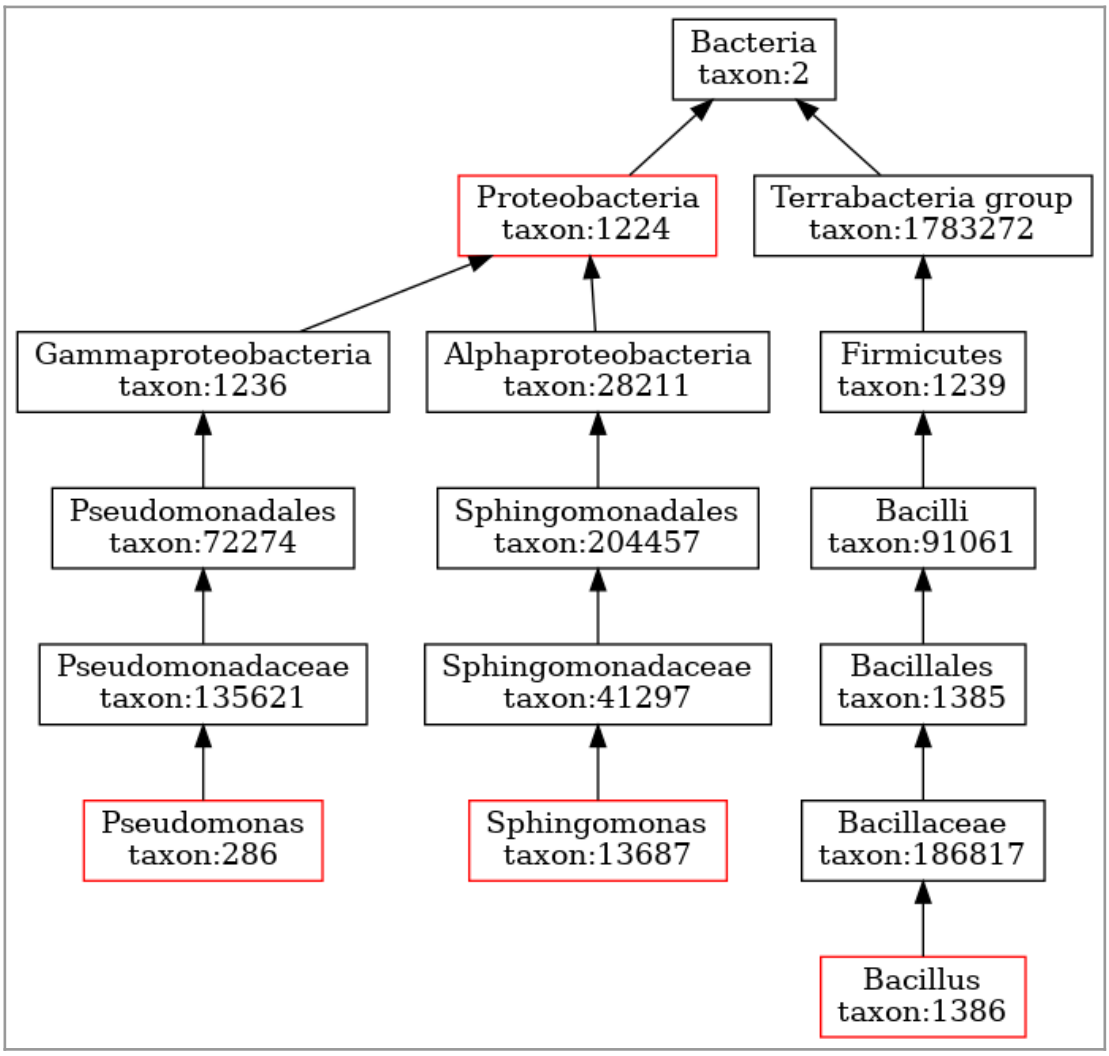


**Figure 6: Graphe de la hiérarchie des taxons – Exemple des 4 Taxons les plus présents dans les OTUs** (encadrés en rouge)

## 3 – Structure des données et peuplement d'AskOmics

La Figure 7 représente la structure souhaitée des données sur AskOmics. Les fichiers intégrés représentent alors 56 806 920 triplets au format RDF sur la plateforme.

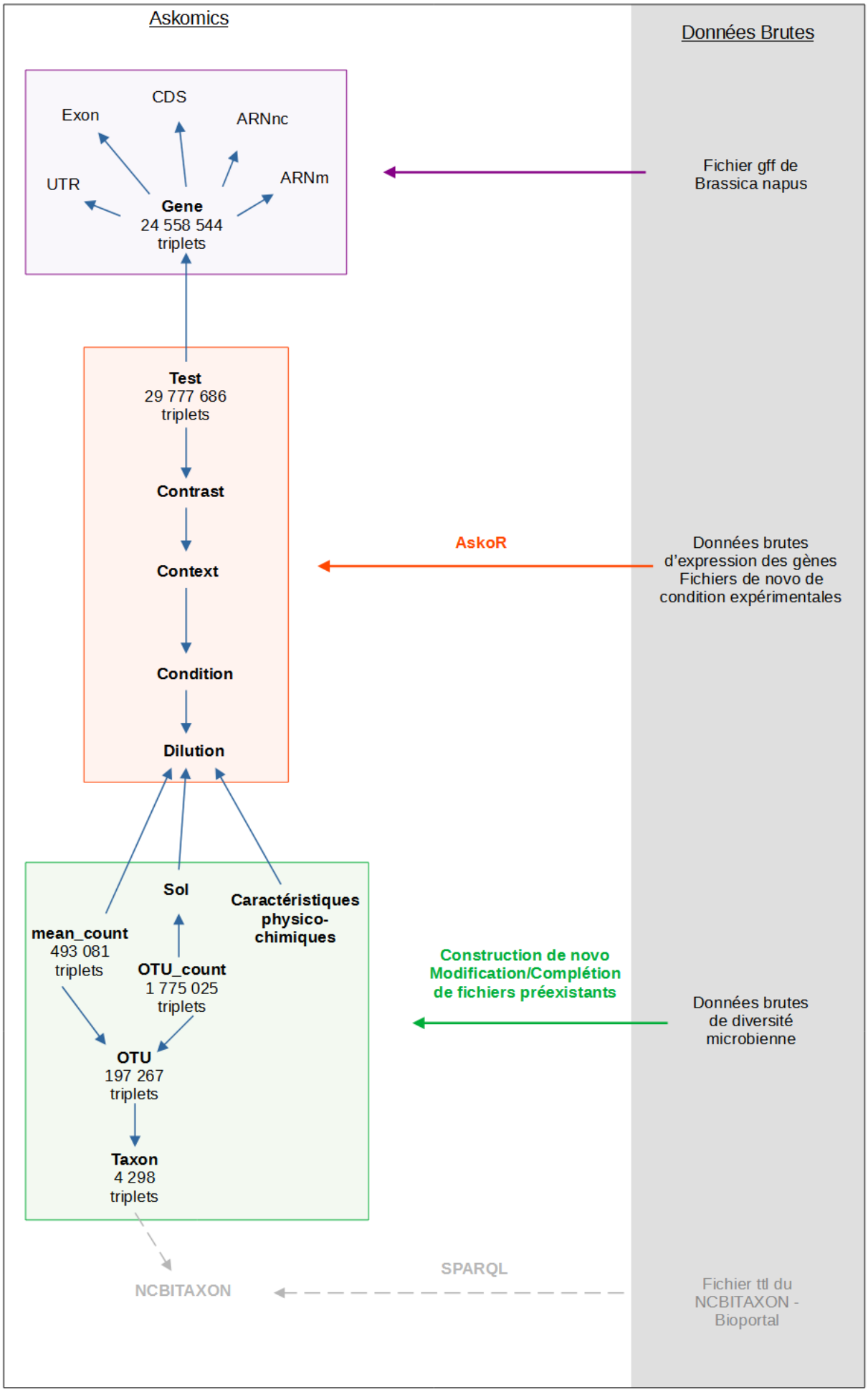


**Figure 7: Structure des données sur AskOmics :** En gris clair l'ajout prévu de l'ontologie

## 4 - Requêtes

### a – Validation

Tous les résultats donnés par AskOmics sont identiques à ceux de l'article, à une exception près : l'article indique 64 DEGs pour le contraste T1MYD0 vs T1MYD6 et 23 pour T2MYD0 vs T2MYD6 alors que je trouve respectivement 63 et 24 DEGs grâce à la plateforme. Il est possible que ce soit une simple erreur de report de données.

En tout, environ une cinquantaine de requêtes différentes ont été faites, que ce soit pour tester l'intégration des données ou pour reproduire les résultats de l'article.

### b - Comparaison des méthodes

J'avais premièrement choisi le type « Text » pour la colonne « Expression » lors de l'intégration des données d'expression différentielle des gènes. J'ai ensuite construit une deuxième VM (Machine Virtuelle) où la colonne « Expression » a été intégrée sous forme de catégories. J'ai alors comparé les manières de construire des requêtes (Figure 8) et les temps d'exécution de celles-ci (Tableau 1) entre les deux méthodes.

On remarque alors que la deuxième méthode, avec le type « Category », renvoie les résultats beaucoup plus rapidement qu'avec le type « Text ». On peut aussi noter que, plus il y a de paramètres, de chemins à suivre, plus le temps d'exécution est long, parfois jusqu'à déclencher un timeout ; même pour la deuxième VM qui est plus performante (Tableau 1).

Par ailleurs, utiliser les catégories permet aussi de ne pas se tromper entre le contexte 1 et le contexte 2 lors de la création de requêtes et/ou de ne pas faire de requêtes avec des contextes ou des contrastes inexistants.

En revanche, certaines requêtes demandent de passer par le fichier de dilution pour relier les OTUs et les fichiers Tests (qui contiennent les données d'expression différentielle). Cela n'empêche pas d'utiliser uniquement les catégories (sans choisir les paramètres dans les boules) mais on est quand même obligé de passer par les entités Contrast, Context et Condition ; on perd donc sûrement un peu de temps.

| Requête | Résultat (nombre de lignes) | Temps pour la VM 1 : attribut 'Text' | Temps pour la VM 2 : attribut 'Category' |
|---|---|---|---|
| Tests dont le contraste est T1STD0vsT1SYD0 | 13 020 | 2 min 4 sec | < 1 sec |
| Tests comparant T1 et T2 | 744 366 | 3 min 26 sec | 28 sec |
| Gènes du chromosome A10 différemment exprimés entre D0 et D6 | 264 | 4 min 41 sec | 2 sec |
| Tests Tenor vs Tenor<br>Signifiance = -1<br>Pvalue < 0,01<br>Gène sur le brin + | 21 910 | 28 min | 1 min 24 sec |
| Tests T1STD0vsT1SYD0 et OTUs/Taxons dont OTUmean > 1500 pour D0 | 185 139 | FAILED (> 12 h) | 8 h 56 min 20 sec |

**Tableau 1**: **Tableau comparatif du temps d'exécution des requêtes en fonction du type choisi**

## 5 - Templates et Formulaires

Sur AskOmics, il est possible d'exécuter à nouveau des requêtes, en allant dans le menu 'Results' pour les relancer. Mais il est aussi possible d'enregistrer ces requêtes sous la forme de Formulaires ou de Templates qui apparaîtront sur le menu de départ des requêtes (Figure 9A).

Ces requêtes permettent une utilisation simple et rapide d'AskOmics.
J'ai choisi de créer à la fois des Formulaires et des Templates, les deux ayant des avantages et des inconvénients.

Utiliser un Formulaire (Figure 9B) permet de ne pas sortir des sentiers battus : l'utilisateur ne peut modifier que certains paramètres de la requête modèle. Ainsi il est bien encadré mais n'a aucune flexibilité.

Au contraire, un Template (Figure 9C) permet de modifier la requête modèle à souhait : rajouter des entités, des paramètres, en enlever… L'utilisateur est beaucoup plus flexible mais il est aussi beaucoup plus facile de s'éloigner de la requête de départ. Il y a bien plus d'informations, parfois inutiles dans le cadre de la requête, qui peuvent perdre l'utilisateur.

Dans un projet où les requêtes sont toutes très similaires, comme par exemple afficher les DEGs entre deux contrastes, il sera alors plus approprié d'utiliser des Formulaires.

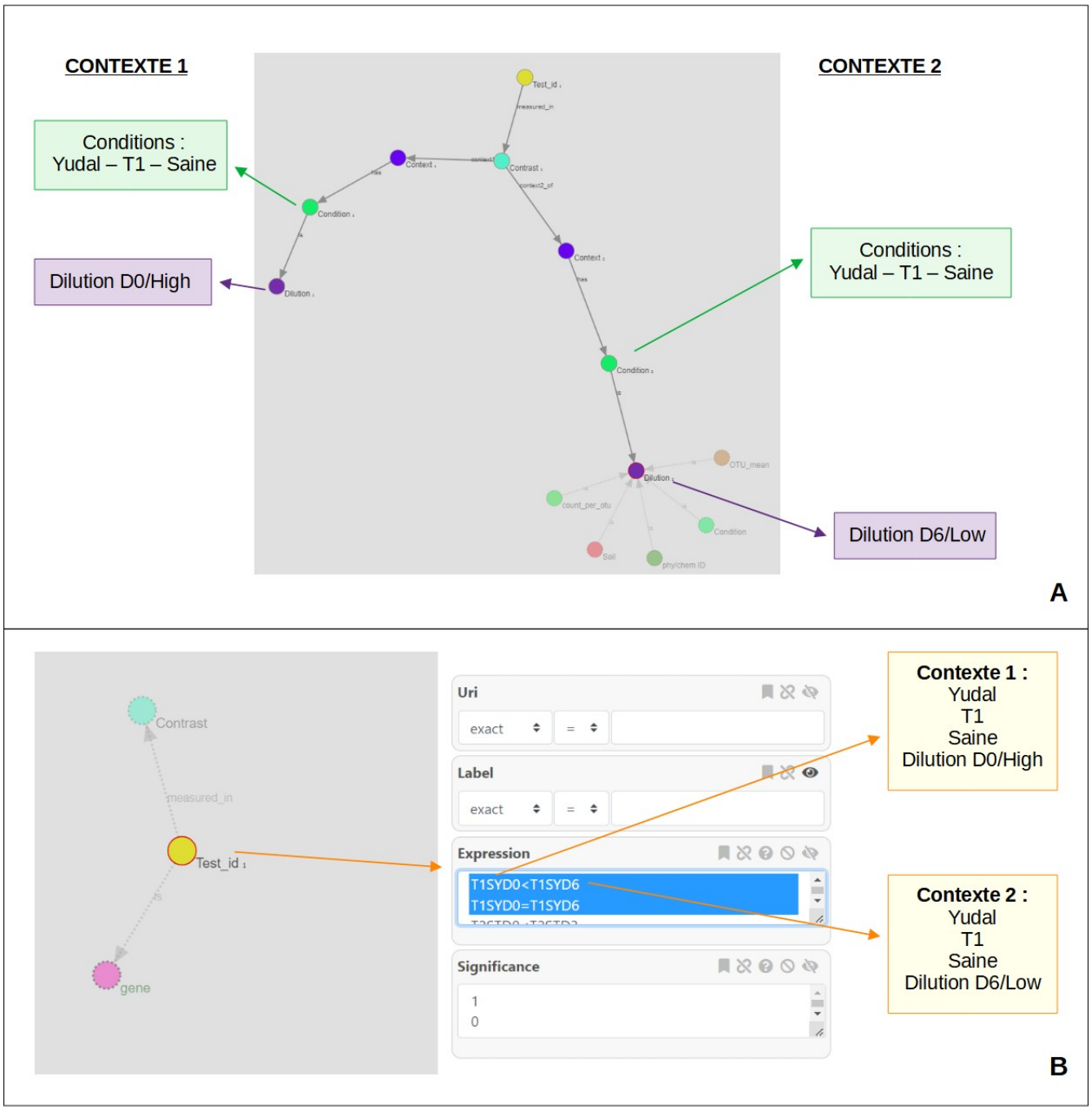


**Figure 8: Schéma comparatif des méthodes de construction de requête** : A- Utilisation du type 'Text', obligation de descendre jusqu'aux entités Dilution pour décrire le contraste. B- Utilisation du type 'Category', on peut choisir le bon contraste au niveau de l'entité de départ.

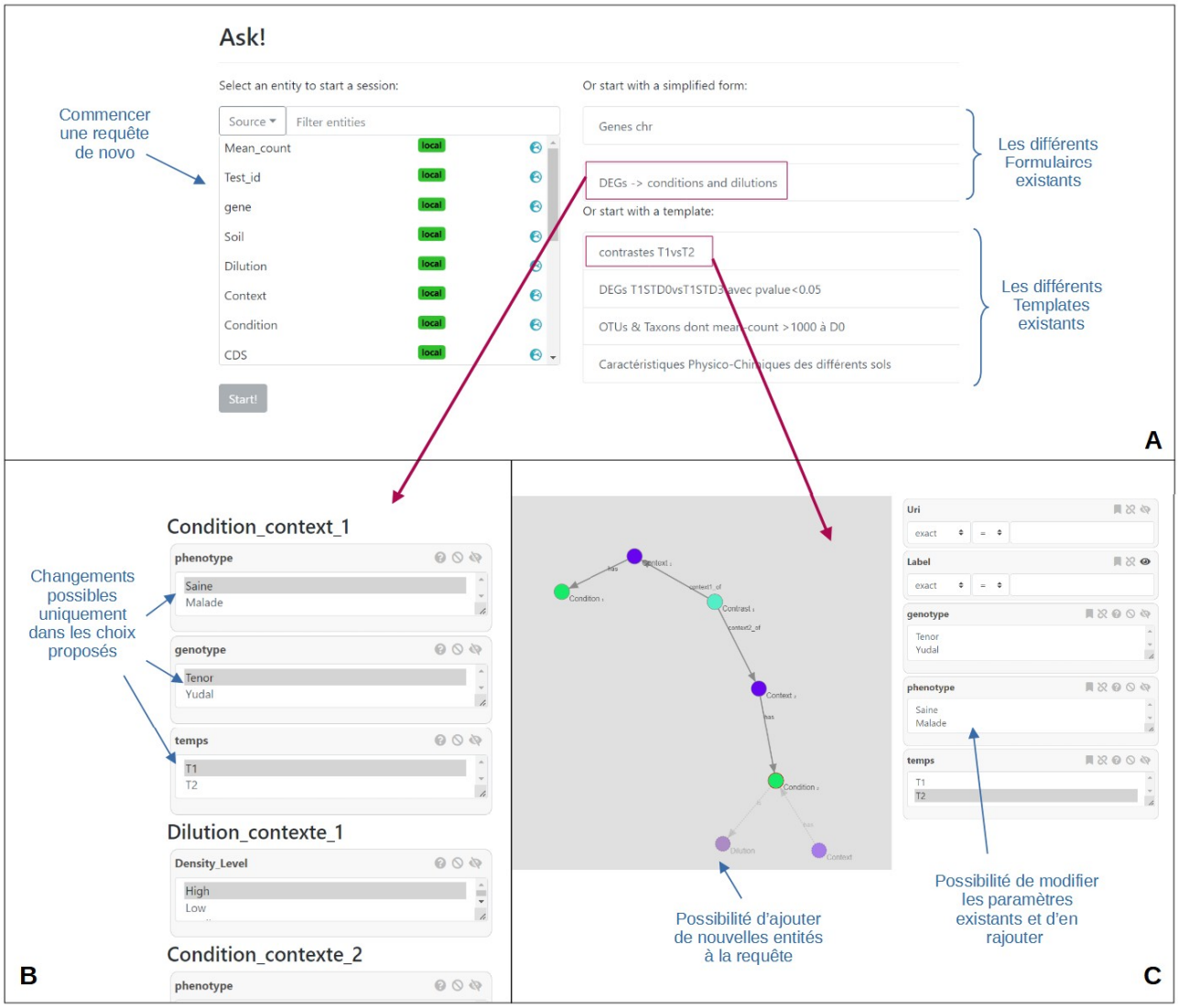


**Figure 9: Utilisation des Forms et Templates** : A- Capture d'écran du menu de départ des requêtes. B- Utilisation d'un Formulaire. C- Utilisation d'un Template.

## 6 - Intégration du NCBITAXON à AskOmics

J'ai testé l'intégration du NCBITAXON à notre structure de données sur une troisième VM.

Le graphe en RDF d'une sous-partie du NCBITAXON contient 4 triplets par taxon :

- rdfs:subClassOf pour afficher des relations de parenté
- rdf:type pour déclarer le taxon comme une classe
- skos:prefLabel pour afficher le Label du taxon
- taxon:RANK pour afficher le rang (genus, family, order…) du taxon.

Avant d'intégrer le fichier sur AskOmics, il a fallu rajouter, pour chacun des taxons, un triplet pour déclarer que le taxon appartient à l'ontologie NCBITAXON. Il a aussi fallu modifier les skos:prefLabel en rdfs:label pour qu'ils soient reconnus par AskOmics. Enfin nous avons ajouté des préfixes et des triplets pour définir le NCBITAXON comme une ontologie appartenant à AskOmics.

J'ai dû aussi remanier les fichiers d'OTUs et de taxons. Le deuxième fichier n'est plus utile car toutes les informations concernant les taxons sont décrites dans l'ontologie. Nous avons donc supprimé le fichier concernant les Taxons. Le fichier des OTUs a quant à lui été modifié : la colonne donnant le nom des taxons a été remplacée par une colonne d'IRIs liés à l'ontologie et la colonne donnant le rang du taxon a été enlevée. J'ai alors rajouté directement dans le fichier des OTUs la colonne des synonymes des taxons, afin de ne pas perdre cette information (nous n'avons pas ajouté les synonymes en créant le graphe RDF de l'ontologie).

L'ajout du NCBITAXON sur AskOmics permet alors d'interroger l'ontologie depuis chaque OTU. On peut ainsi afficher des informations concernant le taxon mais aussi remonter dans la hiérarchie des taxons. En cliquant sur la flèche, on peut choisir de chercher les descendants ou bien les ancêtres du taxon concerné (Figure 10).

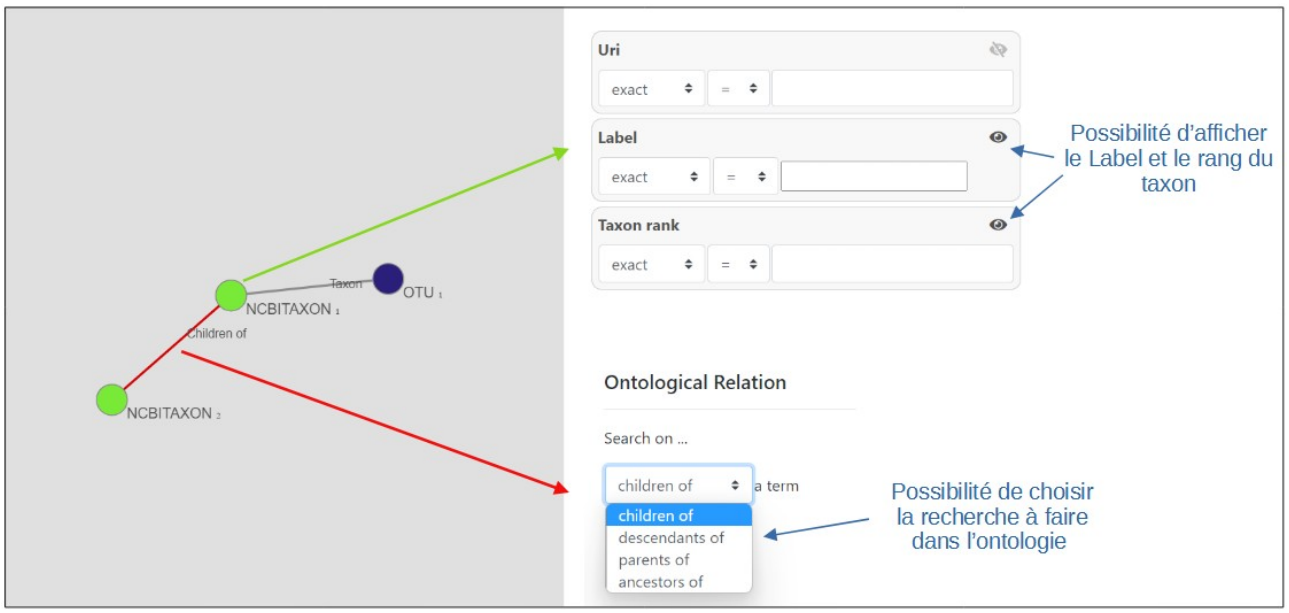


**Figure 10: Utilisation de l'ontologie NCBITAXON sur AskOmics**

# III - Discussion

## 1 - Correction des fichiers d'entrées de AskoR

Lors de l'utilisation d' AskoR sur les données brutes, j'ai été confrontée à plusieurs problèmes. Premièrement les fichiers présentaient des espaces et des tabulations en trop qui étaient peu visibles et qui empêchaient le bon fonctionnement du programme. Après plusieurs corrections j'ai pu obtenir un fichier sans erreurs mais un problème subsistait. En effet, la colonne indiquant les répétitions faites sur les sols (répétitions A, B, C pour chaque dilution) n'était pas prise en compte par AskoR. Pour palier à ce problème, nous avons donc supprimé cette colonne, nous permettant ainsi d'obtenir des résultats. Nous perdons l'information de la répétition mais cela n'a pas de conséquences importantes ; d'autant plus que l'article ne s'appuie pas sur les répétitions et ne les compare pas entre elles.

Lors de la première utilisation d'AskoR je n'avais qu'une petite partie des données brutes mais tous les fichiers annexes importants pour lancer le programme. Lorsque j'ai récupéré le reste des comptages bruts, il m'a fallu créer à la main les fichiers annexes. Cela m'a pris plusieurs heures de temps et d'attention pour ne pas insérer d'erreur (en effet, en premier lieu, une erreur de typographie avait engendré une grosse différence entre les résultats attendus et ceux obtenus et il a fallu revérifier chacun des fichiers) mais m'a permis de choisir les contrastes que je souhaitais étudier par la suite.

De plus, comme indiqué dans l'article, une des répétitions n'a donné aucun résultat pour le contexte T2MTD6 : la plante devait être infectée mais ne présentait aucun symptôme. A la place des comptages on trouve alors une colonne remplie de 'NA' pour cette répétition. Malheureusement, la colonne n'est pas détectée par AskoR et empêche le tableau d'être lu correctement. Pour avoir des résultats en sortie de programme, j'ai alors remplacé les 'NA' par des '0'. Le fichier de comptages bruts est alors bien utilisé mais, la norme étant 3 répétitions par contexte, les données sortantes sur le contexte T2MTD6 sont probablement faussées. Par chance, cela n'a pas affecté les différents résultats que l'on obtient ; ils sont bien identiques à ceux de l'article, même lorsque le contexte est concerné par la requête.

J'ai aussi été confrontée à un autre problème : lors du traitement des comptages bruts des plantes malades, le fichier correspondant au dernier contraste n'était pas créé. Il s'est avéré qu'il s'agissait d'un problème de P-value qui s'annulait. Nous avons donc ajouté dans le code d'AskoR une valeur ajoutée très petite, qui empêche la P-value de s'annuler sans pour autant modifier les valeurs des autres P-values. Cette modification a permis de récupérer le dernier fichier d'expression différentielle et de relever un problème dans le code d'AskoR auquel d'autres utilisateurs pourraient être confrontés.

## 2 - Problèmes rencontrés lors de la modification et de la complétion du fichier d'OTU

La création du fichier concernant les OTUs n'a pas été sans difficultés non plus. Au début, je choisissais le taxon de valeur maximale dans la catégorie Genus sans restriction : pas de seuil minimum ni de montée de niveau si le taxon n'était pas connu. Il y avait alors beaucoup d'OTUs 'Unidentified' (environ 2000 OTUs sur presque 33000 soit un peu plus de 6%). En choisissant de remonter d'un niveau (ou plus), j'arrivais parfois à me débarrasser des 'Unidentified'. Cela a permis de réduire la proportion des OTUs inconnues à moins de 1 % (environ 300 OTUs).

Par ailleurs, au début, la requête sur l'EBI pour récupérer des informations sur les taxons ne renvoyait pas d'IRI pour environ 3000 OTUs (50 taxons distincts). Pour certains il a suffit de remplacer « _ » dans les taxons en deux mots par une espace pour trouver un IRI correspondant. Pour les autres, il s'est avéré que la requête cherchait uniquement les réponses exactes. Ainsi, pour *Bacillus*, on ne trouvait pas d'IRI qui correspondait exactement. En effet, lorsque l'on cherche '*Bacillus*' sur le NCBITAXON via l'EBI, on trouve '*Bacillus sp. Bacillus M9*', '*Bacillus sp. Bacillus M12*'… Mais pas simplement '*Bacillus*'. La requête renvoyait alors 'None'. Ainsi, en changeant 'exact = True' par 'exact = False' nous avons résolu le problème. Au lieu de renvoyer 'None', la requête renvoyait l'IRI du premier résultat contenant '*Bacillus*'. Après ce changement on ne retrouve que 71 OTUs sans IRI (soit 0,2% des OTUs). En effet, la requête sur ces Taxons trouve des résultats mais aucun ne contenant le nom du Taxon ; elle renvoie alors 'None'.

De plus, j'ai fait face à un autre problème : nous voulions récupérer les synonymes des taxons pour chacune des OTUs. Or, en faisant uniquement la requête sur NCBITAXON je n'y avais pas accès. J'ai alors programmé une autre requête qui, à partir de l'IRI du taxon, me permet d'avoir accès à beaucoup plus d'informations. Ainsi j'avais accès à une catégorie 'synonyms'. Malheureusement, cette catégorie était systématiquement vide. En cherchant, j'ai finalement trouvé des catégories 'has_related_synonym' et 'has_exact_synonym' qui apparaissaient parfois dans les annotations des taxons. J'ai donc choisi de garder les synonymes trouvés dans ces catégories. Cela m'a permis de remplir la colonne synonymes pour environ un tiers des OTUs.

Enfin, une dernière modification a été apportée. Au départ, un seul fichier de sortie regroupait les informations sur les OTUs et les Taxons. Passer de un seul fichier contenant OTUs et Taxons à deux fichiers distincts a permis de réduire le nombre taxons à chercher sur l'ontologie NCBITAXON (puisqu'on ne récupère que les taxons distincts), ce qui fait gagner de la place et du temps de calcul. En effet, le programme python lançait au départ plus de 32 000 requêtes sur l'EBI pour retrouver les IRI puis à nouveau 32 000 requêtes pour les synonymes. Ce programme initial tournait pendant une dizaine d'heures. En créant deux fichiers, le deuxième ne recensant que les Taxons distincts, on a réduit le nombre de requêtes à environ 1700 (deux fois 850) ; le programme renvoyait alors les fichiers en une heure. Cela a aussi diminué le nombre de triplets stockés sur AskOmics.

## 3 - Pertinence des différents fichiers intégrés

L'article s'appuyant sur l'expression différentielle des gènes dans les différentes conditions mises en œuvre, il est bien évidement nécessaire d'intégrer les fichiers d'expression génique. j'ai trouvé important d'ajouter les fichiers gff permettant d'avoir plus d'indications (chromosome, brin, position..) sur les gènes mentionnés. Le deuxième gros 'bloc' de fichiers concerne les conditions et plus particulièrement les sols et leur diversité en terme de micro-organismes. C'est pourquoi j'ai intégré le fichier descriptif des OTUs et le fichier de leur comptage selon les sols. De plus, lorsque nous aurons les données du projet Deep Impact, il y aura sûrement des données ressemblantes. Les fichiers intégrés pourront alors nous servir de modèle. De la même façon, j'ai choisi d'intégrer les caractéristiques physico-chimiques des sols car c'est le genre de données avec lesquelles nous pourrions être confrontés par la suite. Pour être plus proche du projet Deep Impact, il aurait été intéressant d'intégrer des données de phénotypage et de rendement. Or, l'article sur lequel nous nous sommes basé ne présentait pas ce genre de données. Nous nous sommes donc contentés des données d'expression génétique.

Intégrer tous ces fichiers et créer toutes ces entités peuvent rendre les requêtes plus compliquées à faire mais aussi à comprendre. En revanche, c'est ce qui permet de faire le lien entre le microbiote et le fonctionnement différentiel des plantes.

## 4 – Intégration d'une partie du NCBITAXON

Le remodelage du fichier des OTUs (cf II-6), même s'il permet l'intégration de l'ontologie, nous fait perdre des informations sur certaines OTUs. En effet, lors des requêtes python sur l'EBI, certains taxons n'ont pas d'IRI qui leur correspond. Lorsque l'on remplace les Labels par les IRIs, les OTUs dont les taxons n'ont pas d'IRI se retrouvent avec 'None' dans cette colonne et on perd le nom du taxon. Impossible alors de savoir quel taxon est représentatif de l'OTU (puisque sans IRI, pas de synonyme(s) non plus).

L'intégration définitive du NCBITAXON à notre structure de données demande encore à réfléchir. Est ce que l'ajout global d'information compense la perte de toute information concernant certaines OTUs ? On peut se demander s'il n'est pas plus simple de supprimer directement toutes les OTUs ne possédant pas d'IRI.

On peut aussi essayer de réfléchir à d'autres possibilités. On pourrait par exemple stocker le Label dans les synonymes lorsque le taxon n'a pas d'IRI.

## 5 – Validation du modèle

Nous avons voulu reproduire les résultats de l'article afin de valider notre intégration de données. Or en essayant de le faire, je me suis rendu compte que des données manquaient : nous avions un seul jeu de comptages bruts correspondant bien aux différents sols pour Tenor et Yudal mais seul le temps T2 est représenté. De plus, nous ne savions pas s'il s'agissait des comptages sur les organismes sains ou infectés. Ces données incomplètes renvoyaient alors des résultats complètement différents de ceux de l'article. J'ai alors récupéré le reste des données et recommencé l'intégration de celles-ci sur AskoR puis AskOmics.

Après ce contretemps j'ai alors pu faire des tests sur AskOmics, qui ont été concluants : les résultats étant identiques à ceux données par l'article et l'interface étant facile à manipuler, **l'utilisation de la plateforme AskOmics est pertinente et validée**. La création de Formulaires et/ou de Templates rend d'autant plus accessible l'utilisation de la plateforme.

Nous avons convenu qu'il était plus avantageux d'intégrer le paramètre 'Expression' sous la forme de catégorie (cf Tableau 1), d'autant plus que cette méthode n'empêche pas l'utilisation des entités Contrast, Context et Condition puisqu'elles sont conservées dans l'arborescence des données.

En ce qui concerne le comptage des OTUs, il est plus pertinent d'utiliser la moyenne par Dilution plutôt que les comptages pour chaque répétition. Les répétitions n'étant utilisées nulle part ailleurs, il s'agit ici d'une information inutile et qui, de ce fait, rend plus difficile l'utilisation des comptages bruts. De plus, choisir le fichier des moyennes, permet de passer par moins d'entités et donc de diminuer le temps d'exécution des requêtes. Les entités Count_dil (comptages par dilution et par répétition) et Soil (description des sols comme étant une dilution et une répétition) pourraient donc être retirées de la structure des données.

En revanche, il aurait été intéressant d'ajouter plus de données au modèle. En effet, nous n'avons pas pris en compte l'annotation des gènes concernés par les tests par exemple. Cela aurait pu être intéressant de les intégrer au modèle, pour comprendre d'un point de vue biologique les conséquences de la maladie et des micro-organismes du sol sur la plante. Ce genre d'information sera sûrement important pour la suite du projet Deep Impact et il conviendra de prendre en compte les annotations de gènes.

Ensuite, les données brutes d'expression différentielle comprenaient deux fichiers : un pour les plantes malades et un pour les plantes saines. Ainsi, je n'ai pas pu ajouter de contraste dans AskoR qui compare directement les phénotypes Sains et Malades. Fusionner les deux fichiers de départ pourrait être une solution pour permettre ce genre de comparaisons.

## 6 - Limites

AskOmics semble être une très bonne solution pour une partie de l'interrogation des données. Certains résultats en revanche sont plus difficiles voir impossibles à reproduire avec la plateforme.
A cela s'ajoute un problème : si on inverse le contexte 1 et le contexte 2, on obtient un contraste qui n'existe pas et donc on ne récupère aucun résultat. Heureusement, ce problème est largement atténué par l'utilisation du type 'Category' (cf II – 4 – b) ; même si la vigilance reste de mise.

De plus, comme mentionné plus tôt, lors du traitement des OTUs, nous avons vu que certains Taxons ne font pas partie des Bactéries. Or, lors de séquençage 16S, on n'attend pas d'Archées ni d'Eucaryotes. Il s'agit en fait sûrement de gènes mitochondriaux de ces espèces qui ont été séquencés ; ils sont considérés comme des contaminants. Afin de s'en débarrasser, on pourrait les éliminer dès la construction des fichiers d'OTUs et de Taxons, soit en éliminant l'OTU si le Taxon choisi n'est pas une Bactérie, soit en décidant de choisir un autre Taxon, même s'il n'est pas majoritaire dans l'OTU.

On peut soulever un léger biais dans la récupération des IRIs du NCBITAXON. En effet, l'IRI récupéré sur l'EBI n'est pas identique aux IRIs du fichier turtle de l'ontologie NCBITAXON de Bioportal. La numérotation des taxons était correcte (sauf exceptions mentionnées plus haut) mais la première partie de l'IRI était différente. Nous avons donc du y apporter quelques modifications avant de pouvoir faire des requêtes SPARQL dessus. Il faudrait automatiser cette modification dès le chargement des IRIs ou bien, si cela est possible, lancer la requête python directement sur Bioportal.

Un problème sur le serveur Genouest a engendré une suppression de la première VM, sur laquelle j'avais fait presque une centaines de requêtes (parfois la même requête plusieurs fois, je ne peux donc pas donner un nombre exact de requêtes différentes). C'est aussi sur cette VM que j'avais enregistré les Formulaires et les Templates. Même si les informations importantes avaient été sauvegardées, certaines requêtes sont perdues et j'ai dû refaire tous les Formulaires et Templates sur la deuxième VM.

## 7 – Perspectives

Le travail a été fait sur le fichier de séquençage 16S et donc sur les Bactéries mais nous avons aussi les données de séquençage 18S. Ainsi, nous pourrions lier les données avec les Fungi et, de même que pour les Bactéries, récupérer la partie du NCBITAXON correspondante.

Pour la fin du stage nous envisageons de travailler plus en profondeur à l'intégration de la partie extraite de l'ontologie NCBITAXON sur AskOmics et de la lier à nos données pour pouvoir faire des requêtes dessus.

Je vais aussi adapter les codes python qui m'ont permis de modifier et/ou créer des fichiers de données afin qu'ils soient utilisables simplement par n'importe quel utilisateur en donnant au programme le fichier de données brutes.

Enfin, les OTUs ne sont pas la seule manière de recenser les Bactéries et certains biologistes préfèrent utiliser les ASVs (Amplicon Sequence Variant). Adapter la structure des données dans AskOmics pour qu'elle convienne à la fois aux OTUs et aux ASVs pourrait être une perspective.

# Conclusion

Afin de répondre à notre interrogation de départ qui était de savoir comment transformer et agencer les données pour les interroger facilement, nous avons décidé de travailler avec la plateforme AskOmics et donc d'adapter les données afin de pouvoir les y intégrer.

Mes contributions durant ce stage ont donc été de :

1. Construire un jeu de données en me basant sur les données brutes de l'article « *Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes* [10]» et en étendant le modèle d'AskoR (modèle déjà adapté à AskOmics).
2. Définir un schéma d'intégration de ces données en RDF et convertir les données originales pour correspondre à ce schéma.
3. Créer une fonction d'extraction d'une partie du NCBITAXON (SPARQL) et adapter le fichier (python) à la plateforme.
4. Intégrer toutes ces données sur AskOmics.
5. Faire des requêtes sur le jeu de données. J'ai aussi créé des modèles de requêtes sous la forme de Templates et de Formulaires réutilisables par d'autres utilisateurs afin de simplifier l'interrogation de leur propre jeu de données.

Cette librairie de requêtes a permis de montrer que la structure de données choisie est pertinente et permet bien de reproduire les résultats de l'article. Ainsi, nous pouvons valider la structuration et la conversion des données ainsi faites.

De plus, les tests que j'ai effectués ont permis de pointer du doigt un certain nombre de bugs présents dans la solution AskOmics (relations « fantômes », lenteurs de calcul…) qui ont été corrigés. Ces modifications ont aussi permis de diminuer drastiquement le temps d'exécution des requêtes. Enfin, avec les colonnes (noms et types) que j'ai choisi pour chaque fichier, il a été possible de faire un script d'automatisation de l'import des données.

Le modèle de données et les requêtes correspondantes ainsi créés sur AskOmics pourront être utiles car réutilisés et complétés dans le cadre de l'interrogation des données du projet Deep Impact ou d'autres projets de biologie. Les scripts python et le script SPARQL ont été déposés sur Gitlab et pourront eux aussi être réutilisés.

# Bibliographie


1. Tkacz A, Poole P. The plant microbiome: The dark and dirty secrets of plant growth. Plants, People, Planet. 2021;3:124–129. https://doi.org/10.1002/ ppp3.10167
2. Babalola OO, Fadiji AE, Enagbonma BJ, Alori ET, Ayilara MS and Ayangbenro AS (2020) The Nexus Between Plant and Plant Microbiome: Revelation of the Networking Strategies. Front. Microbiol. 11:548037. doi: 10.3389/fmicb.2020.548037
3. Berendsen, Roeland L., et al. "The Rhizosphere Microbiome and Plant Health." *Trends in Plant Science*, vol. 17, no. 8, 2012, pp. 478–86, https://doi.org/10.1016/j.tplants.2012.04.001.
4. Berg G, Kusstatscher P, Abdelfattah A, Cernava T and Smalla K (2021) Microbiome Modulation—Toward a Better Understanding of Plant Microbiome Response to Microbial Inoculants. Front. Microbiol. 12:650610. doi: 10.3389/fmicb.2021.650610
5. Stefan L, Hartmann M, Engbersen N, Six J and Schöb C (2021) Positive Effects of Crop Diversity on Productivity Driven by Changes in Soil Microbial Composition. Front. Microbiol. 12:660749. doi: 10.3389/fmicb.2021.660749
6. Deng, Siwen, et al. "A Plant Growth-Promoting Microbial Soil Amendment Dynamically Alters the Strawberry Root Bacterial Microbiome." *Scientific Reports*, vol. 9, no. 1, 2019, p. 17677, https://doi.org/10.1038/s41598-019-53623-2.
7. Berners-Lee, T., Hendler, J., Lassila, O. (2001), The semantic web, Scientific American, 284(5), 34–43. http://www.sciam.com/2001/0501issue/0501berners-lee.html.
8. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Scott Marshall, M., Ogbuji, C., Rees, J., Stephens, S., Wong, G. T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.-H. (2007), Advancing translational research with the semantic web, BMC Bioinformatics, 8(3)
9. Schulz, S., Balkanyi, L., Cornet, R., Bodenreider, O. (2013), From concept representations to ontologies: A paradigm shift in health informatics?, Healthcare informatics research, 19(4), 235–242.
10. Daval, Stéphanie, et al. "Soil Microbiota Influences Clubroot Disease by Modulating *Plasmodiophora Brassicae* and *Brassica Napus* Transcriptomes." *Microbial Biotechnology*, vol. 13, no. 5, 2020, pp. 1648–72, https://doi.org/10.1111/1751-7915.13634.
11. Charles Bettembourg, Olivier Dameron, Anthony Bretaudeau, Fabrice Legeai. AskOmics : Intégration et interrogation de réseaux de régulation génomique et post-génomique. IN OVIVE (INtégration de sources/masses de données hétérogènes et Ontologies, dans le

domaine des sciences du VIVant et de l'Environnement), Institut National de Recherche en Informatique et en Automatique (INRIA). Rennes, FRA., Jun 2015, Rennes, France. pp.7. ffhal-01184903

12. Pérez, J., Arenas, M., Gutierrez, C. (2009), Semantics and complexity of sparql, ACM Trans. Database Syst., 34(3), 16:1–16:45. URL: http://doi.acm.org/10.1145/1567274.1567278
13. Salvadores, M., Horridge, M., Alexander, P. R., Fergerson, R. W., Musen, M. A., Noy, N. F. (2012), Using SPARQL to query Bioportal ontologies and metadata, in Proceedings of the International Semantic Web Conference ISWC 2012, vol. 7650 of Lecture Notes in Computer Science, pp. 180–195
14. Alves-Carvalho, S.; Gazengel, K.; Masanelli, S.; Bretaudeau, A.; Robin, S.; Daval, S.; Legeai, F. AskoR, a R Package for Easy RNA-Seq Data Analysis. Proceedings 2021, 1, 0. https://doi.org/10.3390/IECE-10646
15. Bard, J. B. L., Rhee, S. Y. (2004), Ontologies in biology: design, applications andfuture challenges, Nature reviews. Genetics, 5(3), 213–222.
16. Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., Musen, M. A. (2009), Bioportal: ontologies and integrated data resources at the click of a mouse, Nucleic acids research, 37(Web Server issue), W170–W173.
17. Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., Musen, M. A. (2011), BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic acids research, 39(Web Server issue), W541–W545.

Yael TIRLET Année 2021/2022

Master 2 Bioinformatique et Génomique – Parcours Informatique et Biologie Intégrative

---

Résumé :

Dans le cadre du projet Deep Impact qui cherche à étudier le rendement des cultures de blé et de colza en fonction de la biodiversité du sol mais aussi des conditions environnementales, de nombreuses données vont être générées. Ce stage vise à trouver une manière pratique et efficace pour interroger ce genre de données, afin de simplifier le travail des biologistes.

Nous nous sommes appuyés sur les résultats et les données d'un article traitant de la diversité microbienne et de la santé des plantes pour réfléchir à un schéma de données idéal.

Nous avons aussi récupéré une partie de l'ontologie Ncbi Taxonomy afin de la lier à nos données.

Après avoir défini une structure de données et avoir construit le jeu de données correspondant, en transformant les fichiers originaux, en les complétant et parfois même en les créant entièrement, nous avons intégré ces fichiers sur la plateforme AskOmics.

C'est sur cette plateforme que nous avons pu interroger les données de façon simple. Les différentes requêtes effectuées ont montré que l'utilisation d'AskOmics était idéale, aussi bien au niveau de la simplicité d'utilisation que de la justesse des résultats obtenus.

Nous avons alors créé une librairie de requêtes ainsi que des requêtes modèles, réutilisables facilement. Toutes ces requêtes pourront servir lors d'autres projets comprenant des données agricoles, comme le projet Deep Impact.

**Agriculture, Diversité microbienne, Web Sémantique, Structure de données, Ontologie**

Abstract :

The Deep Impact project is studying wheat and rapeseed crop yield depending on soil biodiversity and environmental conditions. This will generate lots of data. This internship aims to find a practical and efficient way to question this data. This will simplify biologists' work.

We have relied on data and results from an article, which was about microbial diversity and plant health to think of an ideal data scheme.

We also retrieved a part of the Ncbi Taxonomy ontology, to link it to our data.

After defining a data structure and constructing the corresponding dataset by modifying the originals files, by completing them or even by creating them entirely, we integrated these files on the AskOmics platform.

This is on this platform that we were able to simply question our data. The different queries that we made showed that using AskOmics was ideal, both for the simplicity of the use as for the accuracy of the results we obtained.

Then we created a library of queries and query models that we can easily reuse. All these queries can be used in other projects with agricultural data, as the Deep Impact project.

**Agriculture, Microbial Diversity, Semantic Web, Dataset structure, Ontology**