RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

CULTIVER
PROTÉGER
autrement

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION
Liberté
Égalité
Fraternité

INVESTIR L'AVENIR

AGENCE NATIONALE DE LA RECHERCHE
ANR

Integrating and querying heterogeneous datasets with Askomics

The DeepImpact case study.

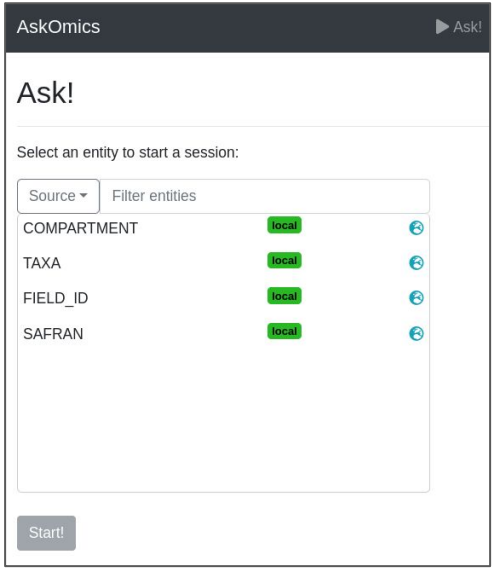
Matéo Boudet, Fabrice Legeai

1. AskOmics: Comment ca marche?
2. Intégration des ontologies dans AskOmics
    a. *L'ontologie FALDO*
    b. *Les ontologies en général*
3. Application au projet *DeepImpact*

AskOmics:

- Intégration de données hétérogènes (génération du RDF)
- Interrogation de données locales et distantes (génération du SPARQL)

Les atouts du web sémantique, sans les difficultés: AskOmics s'en charge!



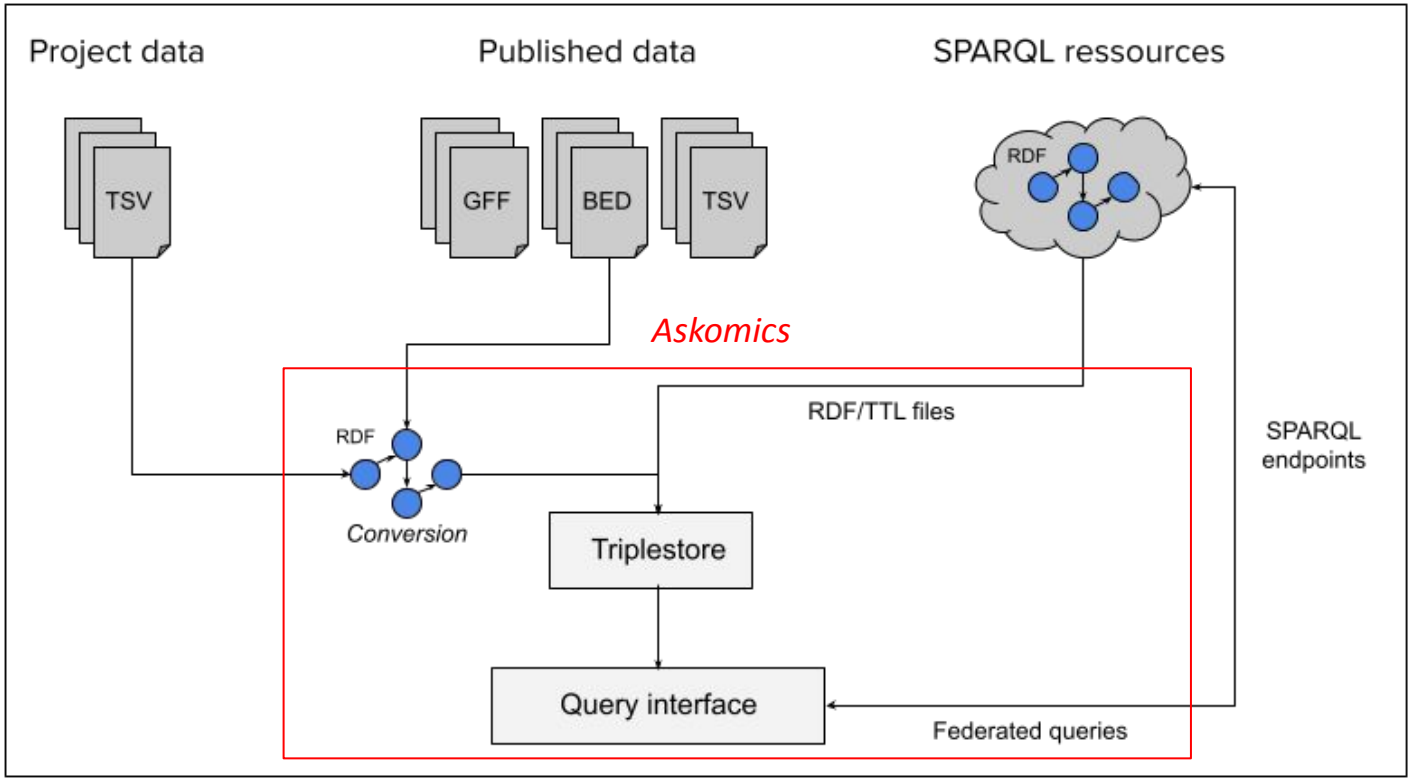
Répondre à des questions comme:

"Liste des taux de calcium et de magnésium des sols des champs en agriculture biologique du sud de la France à la 2ème saison d'échantillonnage de la 1ère année"

"Rendements (en biomasse sèche) de tous les champs de l'Est de la France de l'année 1"

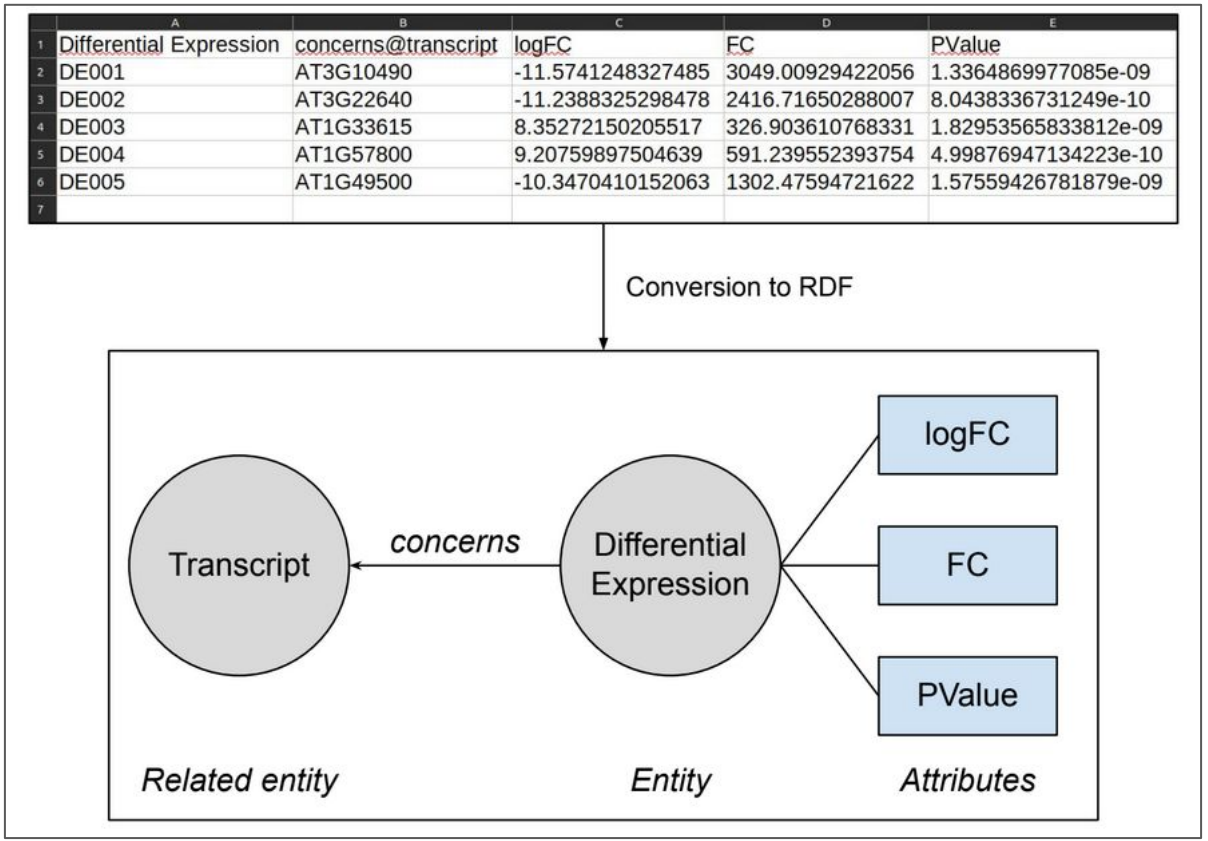
"Liste des taxa trouvés dans les microbiotes de racines des champs de l'ouest de la France en agriculture conventionnelle au 2ème échantillonnage de la 2ème année"

AskOmics: vue d'ensemble



Intégration de données : CSV/TSV

	A	B	C	D	E
1	Differential Expression	concerns@transcript	logFC	FC	PValue
2	DE001	AT3G10490	-11.5741248327485	3049.00929422056	1.3364869977085e-09
3	DE002	AT3G22640	-11.2388325298478	2416.71650288007	8.0438336731249e-10
4	DE003	AT1G33615	8.35272150205517	326.903610768331	1.82953565833812e-09
5	DE004	AT1G57800	9.20759897504639	591.239552393754	4.99876947134223e-10
6	DE005	AT1G49500	-10.3470410152063	1302.47594721622	1.57559426781879e-09
7					



Des données au RDF

Lors de l'intégration des données, AskOmics génère deux graphes:

- L'abstraction du fichier *(basée sur l'en-tête du fichier)*
  - Le type d'entité
    - *askomics:Gene*
  - Les attributs (uri, type)
    - *askomics:Gene* a un attribut *rdfs:label* de type 'String'
  - Les relations
    - *askomics:Gene est lié à askomics:mRNA par la relation askomics:Parent*

Visualisation

- Le graphe de données *(chaque ligne est une entité)*
  - *askomics:GeneA a askomics:Gene*
  - *askomics:GeneA rdfs:label "Gene A"*
  - *askomics:GeneA askomics:derived_from askomics:GeneB*

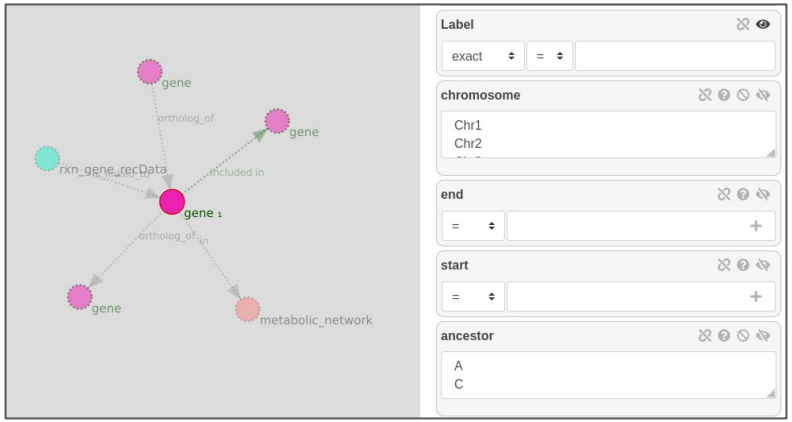
Requête

Des données au RDF: l'abstraction (visualisation)

Pour la visualisation, AskOmics lance une requête sparql, et récupère:

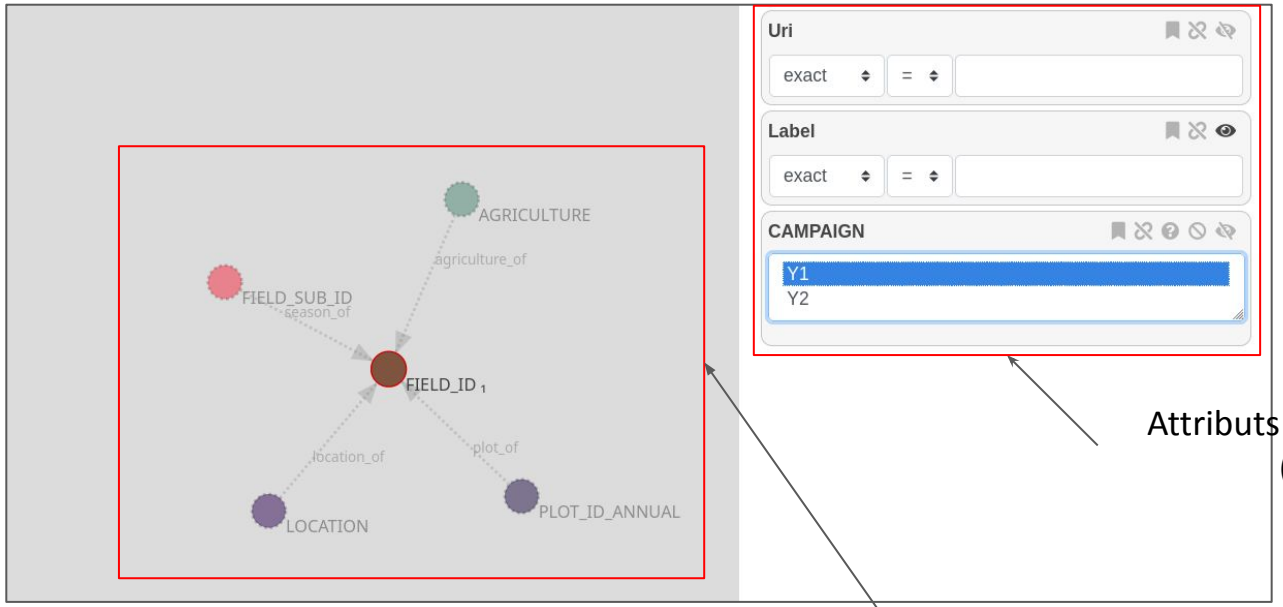
- Toutes les entités
- Toutes les relations
- Tous les attributs
- \+ Restriction aux graphes publiques et ceux de l'utilisateur.

Ensuite, AskOmics construit la visualisation à partir de ces données



AskOmics: Interface de construction de requêtes

“Rendement (en biomasse sèche) de tous les champs de l’Est de la France pour l’année 1”



Attributs du type sélectionné (“Année 1”)

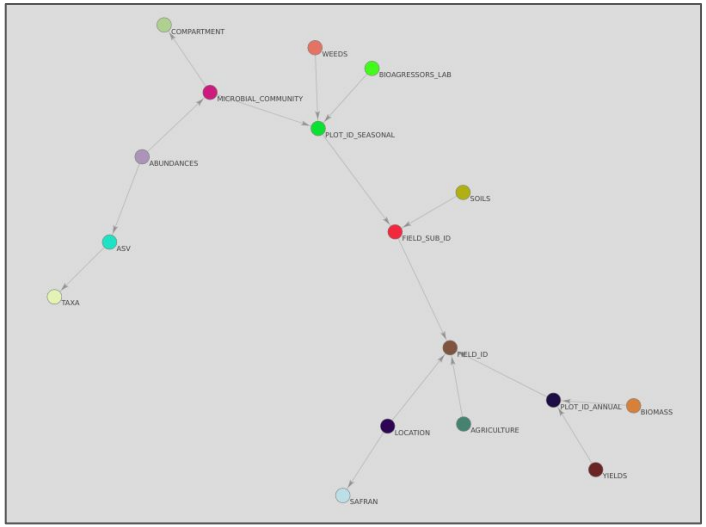
Entités liées au type sélectionné

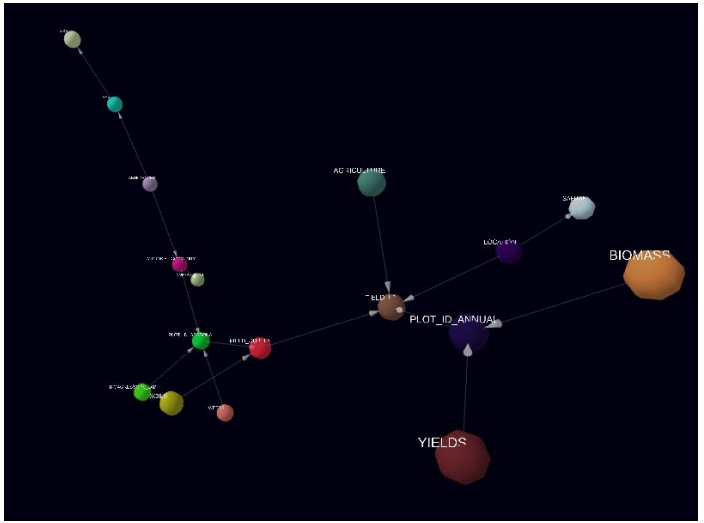
“Tous les champs de l’année 1”

Construction itérative / progressive des requêtes d’entité en entité en spécifiant les attributs

AskOmics: Quoi de neuf?

- Visualisation de l’abstraction complète (2D/3D)
- Meilleur support pour les ontologies
- Notion de distance entre deux attributs dans les requêtes
  - Ex: ‘Tous les gènes qui commencent à moins de 200 bases de la fin de AT3G104090’
  - ‘Toutes les entités dont l’attribut A est > à l’attribut B’
- Amélioration générale de l’expérience utilisateur





Les ontologies: FALDO

<FALDO is the Feature Annotation Location Description Ontology. It is a simple ontology to describe sequence feature positions and regions>

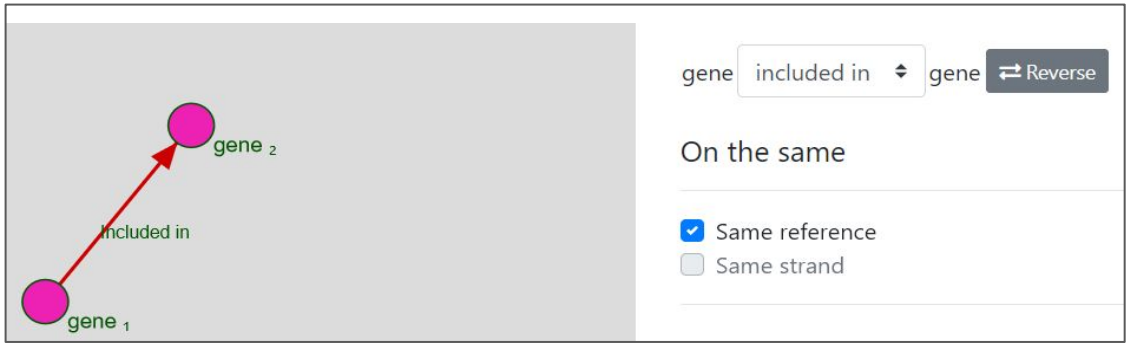
Si l'entité décrit une position sur le génome *(reference, start, end, strand)*, AskOmics utilise l'ontologie FALDO pour la génération du RDF.

askomics:DE001	*faldo:location*	*_:location*
_:location	*rdf:type*	*faldo:region*
_:location	*faldo:begin*	*_:begin*
_:location	*faldo:end*	*_:end*
_:begin	*rdf:type*	*faldo:ExactPosition*
_:begin	*faldo:position*	*"1"*
_:begin	*faldo:reference*	*"C1"*
_:end	*rdf:type*	*faldo:ExactPosition*
_:end	*faldo:position*	*"100"*
_:end	*faldo:reference*	*"C1"*

Les ontologies: FALDO

Avantage de l'utilisation de l'ontologie:

- Toutes les entités 'positionnelles' sont décrites de la même façon
- Possibilité de proposer des requêtes 'avancées' entre ces entités




```
?gene1_uri rdf:type <http://askomics.org/data/gene> .
?gene1_uri faldo:location/faldo:end/faldo:position ?gene1_end .
?gene1_uri faldo:location/faldo:begin/faldo:position ?gene1_start .
?gene23_uri rdf:type <http://askomics.org/data/gene> .
?gene23_uri faldo:location/faldo:end/faldo:position ?gene23_end .
?gene23_uri faldo:location/faldo:begin/faldo:position ?gene23_start .

FILTER (?gene1_start > ?gene23_start && ?gene1_end < ?gene23_end) .
```

Mais: Extrêmement lent!

FALDO: Les triplets en plus

(Comparaison entité / entité) + (triplets multiples pour définir une position dans faldo) =

temps de calcul excessif

Deux moyens pour réduire ce temps de calcul:

- Création de triplets **‘directs’** Entité <-> Valeur
  - Ex: askomics:DE001 askomics:faldoStart “1”
- Séparation du génome en ‘**blocs**’ (ex: blocs de 10000 bases) et ajout de ces blocs en ‘attribut’ de l’entité
  - Ex: askomics:DE001 askomics:includeIn “block1”
  - Possibilité de rajouter l’info sur les chromosomes & strand également

Lors de la génération de la requête, on commence par demander que les entités aient un bloc en commun -> Accélération de la requête

FALDO: Les triplets en plus

Sur ~200 000 gènes:

```
?gene1_uri rdf:type <http://askomics.org/data/gene> .
?gene1_uri faldo:location/faldo:end/faldo:position ?gene1_end .
?gene1_uri faldo:location/faldo:begin/faldo:position ?gene1_start .
?gene23_uri rdf:type <http://askomics.org/data/gene> .
?gene23_uri faldo:location/faldo:end/faldo:position ?gene23_end .
?gene23_uri faldo:location/faldo:begin/faldo:position ?gene23_start .

FILTER (?gene1_start > ?gene23_start && ?gene1_end < ?gene23_end) .
```

10+ minutes

```
?gene1_uri askomics:includeInReference ?block_1_23 .
?gene23_uri askomics:includeInReference ?block_1_23 .
?gene1_uri rdf:type <http://askomics.org/data/gene> .
?gene1_uri faldo:location/faldo:end/faldo:position ?gene1_end .
?gene1_uri faldo:location/faldo:begin/faldo:position ?gene1_start .
?gene23_uri rdf:type <http://askomics.org/data/gene> .
?gene23_uri faldo:location/faldo:end/faldo:position ?gene23_end .
?gene23_uri faldo:location/faldo:begin/faldo:position ?gene23_start .

FILTER (?gene1_start > ?gene23_start && ?gene1_end < ?gene23_end) .
```

14 secondes

Les ontologies: en général

FALDO est la seule ontologie ‘par défaut’ présente dans AskOmics

Mais: possibilité d’intégrer d’autres ontologies selon les données du projet

- *NCBITAXON*
- *Autres (Crop Ontology…)*

-> Deux possibilités pour l’intégration de l’ontologie

- Intégration directement des données AskOmics
- Requêtes fédérées sur un endpoint distant (uniquement l’abstraction en local)

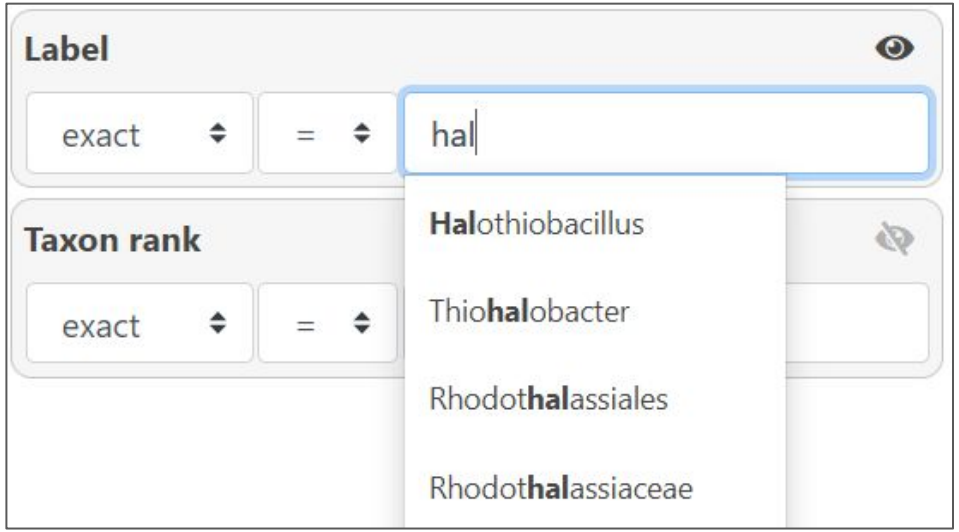
Pour l’intégration des données,

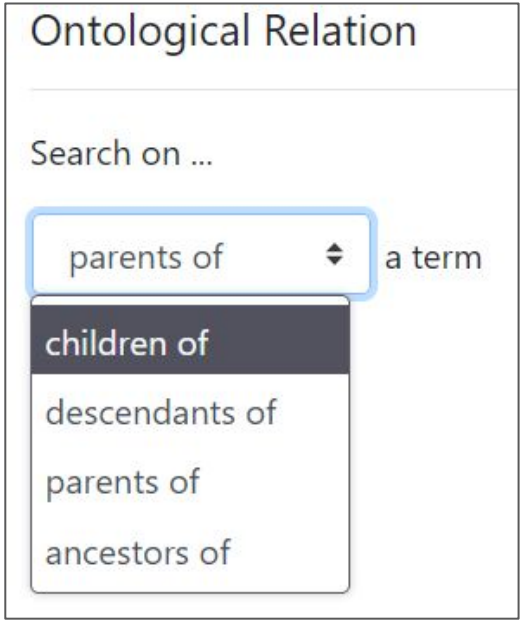
- Utilisation de l’URI du terme
- Spécification du type de colonne ‘Ontologie’

POC en cours: Intégration du thésaurus INRAE dans AskOmics

Les ontologies: en général

- Les données utilisent des termes d'une ontologie
- Possibilité de faire une requête avec auto-complétion sur le label
- Possibilité de faire une requête sur ascendants / descendants du terme
    - Ex: Toutes les entités dont l'attribut 'Type' descend de 'Field'





Les ontologies: pourquoi?

Parce que c'est FAIR, mais aussi:

FALDO:

- Je veux: "Les gènes inclus dans un QTL spécifique"

NCBITAXON:

- Je veux: "Les environnements où on trouve en abondance des bactéries du genre Pseudomonas"

Thesaurus INRAE:

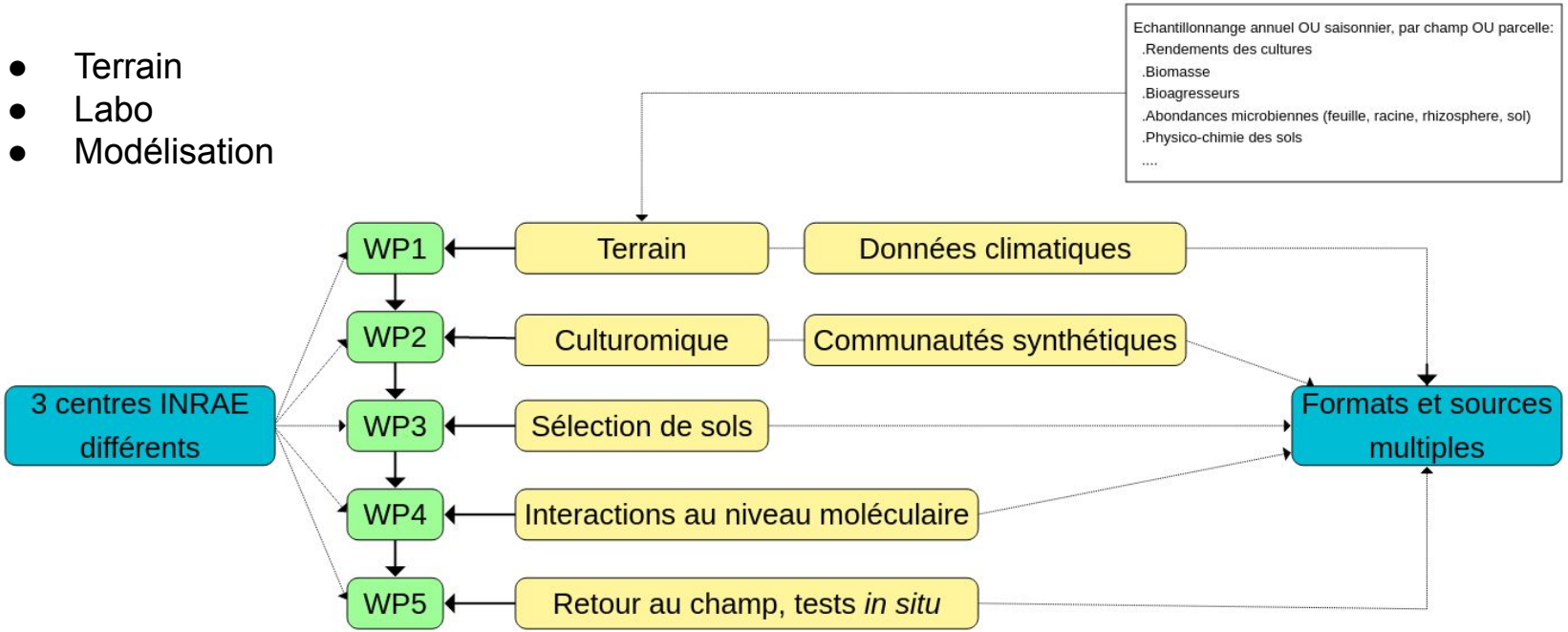
- Je veux: Toutes les entités dont l'attribut *type* 'descend' du terme '*crop protection'*

Et DeepImpact, alors? Contexte:

- Analyse des interactions plante-microbiote pour promouvoir la défense des plantes aux bioagresseurs
- Recherche de solutions agroécologiques contre les stress biotiques
- Espèces étudiées : blé et colza
- **200** champs, **3** zones géographiques, **2** ans / **4** saisons d'échantillonnage
- Plusieurs Work Packages successifs (champ →labo→modélisation -> champ)

Données : “l’agroécologie du climat au génome”

- Terrain
- Labo
- Modélisation



Des données nombreuses, hétérogènes, fortement liées entre elles

Intégration : des données aux requêtes

Exemple avec le WP1 (terrain)

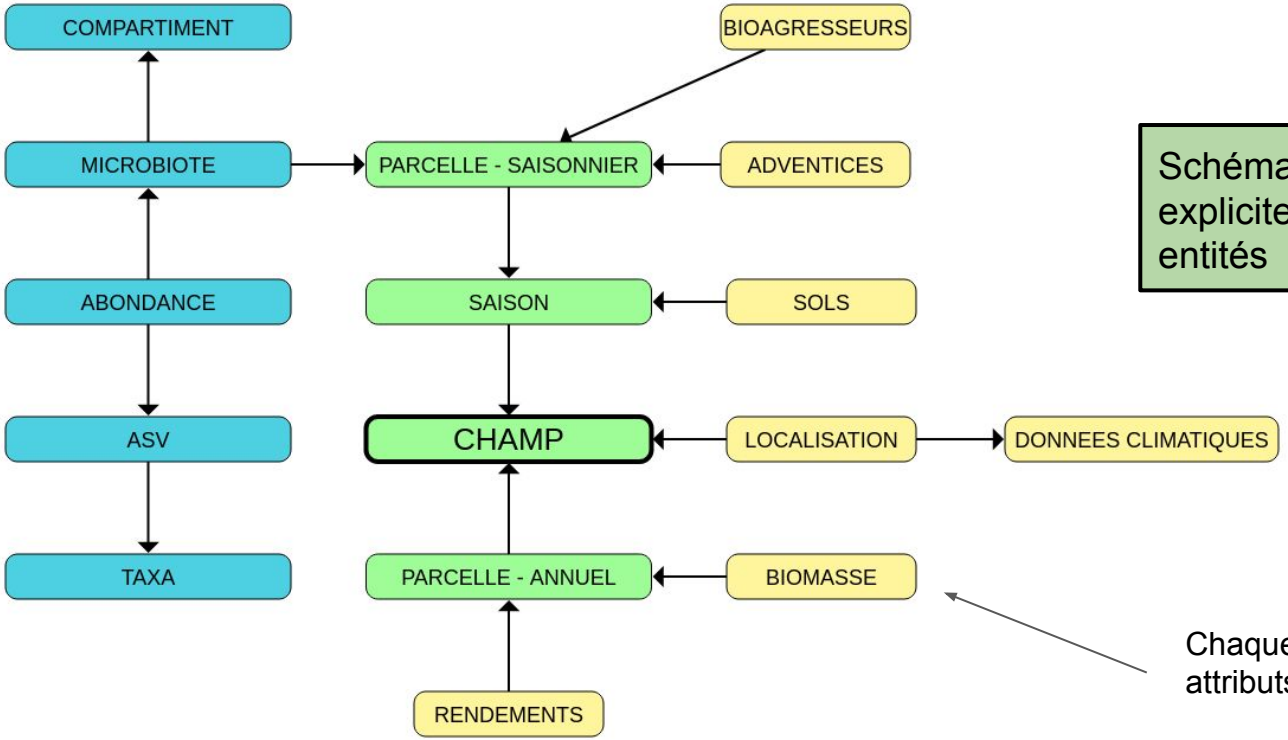


Schéma de données qui explicite les liens entre les entités

Chaque entité à ses propres attributs (i.e. variables)

DeepImpact: les fichiers

Terrain

FIELD_SUB_ID	PLOT_ID	DATE	OPERATOR	WEED_SPECIES	DENSITY_CLASS	PHENOLOGY_STAGE
AF001-Bn-Y1-S1	PA	05/11/2021	PLG, UK, SD	CHEAL	3+	D
AF001-Bn-Y1-S1	PA	05/11/2021	PLG, UK, SD	STEME	3+	C
AF001-Bn-Y1-S1	PA	05/11/2021	PLG, UK, SD	VERPE	P	D
AF001-Bn-Y1-S1	PA	05/11/2021	PLG, UK, SD	POAAN	2	B
AF001-Bn-Y1-S1	PA	05/11/2021	PLG, UK, SD	LAMPU	P	B
AF001-Bn-Y1-S1	PB	05/11/2021	PLG, UK, SD	CHEAL	4	D
AF001-Bn-Y1-S1	PB	05/11/2021	PLG, UK, SD	STEME	4	C
AF001-Bn-Y1-S1	PB	05/11/2021	PLG, UK, SD	VERPE	P	C
AF001-Bn-Y1-S1	PC	05/11/2021	PLG, UK, SD	CHEAL	4	D
AF001-Bn-Y1-S1	PC	05/11/2021	PLG, UK, SD	STEME	4	C
AF001-Bn-Y1-S1	PD	05/11/2021	PLG, UK, SD	CHEAL	4	D
AF001-Bn-Y1-S1	PD	05/11/2021	PLG, UK, SD	STEME	4	C

FIELD_SUB_ID	COARSE_SAND	FINE_SAND	COARSE_SILT	FINE_SILT	CLAY
AF001-Bn-Y1-S1	171	182	253	239	155
AF002-Bn-Y1-S1	17	142	472	236	133
AF003-Bn-Y1-S1	50	112	446	252	140
AF004-Bn-Y1-S1	276	96	258	227	143
AF005-Bn-Y1-S1	218	138	232	248	165
AF006-Bn-Y1-S1	205	95	263	266	171
AF007-Bn-Y1-S1	107	70	219	402	201
AF008-Bn-Y1-S1	55	110	367	313	155
AF009-Bn-Y1-S1	95	106	344	241	214
AF010-Bn-Y1-S1	126	161	237	242	234

Localisation

FIELD_ID	LATITUDE	LONGITUDE	REGION	SAFRAN
AF001-Bn-Y1	48.010587	-1.650226	WEST	2762
AF002-Bn-Y1	48.189962	-2.156227	WEST	2501
AF003-Bn-Y1	48.186577	-1.949645	WEST	2503
AF004-Bn-Y1	47.904364	-2.534406	WEST	2885
AF005-Bn-Y1	47.7100555	-2.6239451	WEST	3271
AF006-Bn-Y1	47.662399	-2.277299	WEST	3400
AF007-Bn-Y1	47.916373	-2.052324	WEST	2890

Climat

jour	mois	an	numero_de_maille	lambx	lamby	prenei_q	prelig_q	pe_q	t_q	tinf_h_q
1	1	2021	2318	7960	23770	0	0	0	0.6	-0.3
2	1	2021	2318	7960	23770	0	0	-0.1	-0.4	-1.3
3	1	2021	2318	7960	23770	0.1	0	0.1	0.1	-0.2
4	1	2021	2318	7960	23770	0.3	0	0.4	0.6	-0.2
5	1	2021	2318	7960	23770	0	0	0.1	0.6	0.3

Abondances microbiennes

#blast_taxonomy	blast_subject	observation_sum	AF034_Bn_Y21AU_RH_8B_01	AF034_Bn_Y21AU_RH_8B_02	AF034_Bn_Y21AU_RH_8B_03	AF034_Bn_Y21AU_RH_8B_04	AF034_Bn_Y21AU_RH_8B_05
Bacteria;Proteoba	multi-subject	6887	0	0	581	323	0
Bacteria;Proteoba	multi-subject	4899	0	0	0	0	0
Bacteria;Bacteroi	2714719444	4742	0	0	0	0	0
Bacteria;Proteoba	multi-subject	4660	0	2016	0	0	0
Bacteria;Proteoba	2644906880	3371	0	0	0	0	0
Bacteria;Proteoba	2598739733	3187	0	0	0	0	0
Bacteria;Proteoba	2649017914	2734	0	0	0	0	0

- Préparation des templates (vocabulaire contrôlé) en amont
- Validation automatique lors du dépôt des données
- Formatage + Intégration (TBA: *automatique*) des données dans AskOmics

-> Interface de requête + mise à disposition de l'endpoint SPARQL

AskOmics & DeepImpact : quelles requêtes?

FIELD_ID1_Label	SOILS69_MAGNESIUM	SOILS69_CALCIUM
AF093-Ta-Y1	2.61	25.04
AF099-Ta-Y1	10.91	46.19

"Liste des taux de calcium et de magnésium des sols des champs en agriculture biologique du sud de la France à la 2ème saison d'échantillonnage de la 1ère année"

"Rendements (en biomasse sèche) de tous les champs de l'Est de la France de l'année 1"

"Liste des taxa trouvés dans les microbiotes de racines des champs de l'ouest de la France en agriculture conventionnelle au 2ème échantillonnage de la 2ème année"

FIELD_ID1_Label	YIELDS47_DRY_WEIGHT_1000_SEEDS	YIELDS47_DRY_WEIGHT
AF016-Bn-Y1	3.58	97
AF016-Bn-Y1	3.22	592.4299999999999
AF016-Bn-Y1	3.94	157.95
AF016-Bn-Y1	3.93	166.4
AF017-Bn-Y1	3.48	312.22
AF017-Bn-Y1	3.71	361.76
AF017-Bn-Y1	4.1	110.85
AF017-Bn-Y1	2.79	361.66
AF018-Bn-Y1	3.55	354.2
AF018-Bn-Y1	3.31	338.37

TAXA1_Label	TAXA1_CLASS	TAXA1_KINGDOM	TAXA1_GENUS	TAXA1_ORDER	TAXA1_PHYLUM
TAX-3	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-1	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-2	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-4	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-5	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-6	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-7	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria
TAX-8	Gammaproteobacteria	Bacteria	Pseudomonas	Pseudomonadales	Proteobacteria

AskOmics & DeepImpact : Intégrer, ou ne pas intégrer ?

- Beaucoup de données & de types de données
  - Volumétrie assez faible = faible nombre de triplets RDF
  - Intégration ‘tel quel’: Données ‘askomics’ == ‘Données chercheurs’
- Mais aussi: données ‘difficiles’ à intégrer *(abondance)*
  - Beaucoup de triplets générés
  - Formatage nécessaire des fichiers: Données ‘askomics’ !== ‘Données chercheurs’

Deux possibilités:

- Intégration complète:
  - Complet, mais perte en performance & schéma de données complexe
- Intégration partielle (liens / chemins) vers les fichiers
  - Pas de filtrage possible, mais récupération des données ‘brutes’

-> Dépend des besoins du projet

Population1_Label	Sequence1_Read1
BO_F_ETRE_W_A	1fcf18ec-de4d-4942-8cd3-41be048d0e8c

Remerciements

- Les membres du projet DeepImpact
- Les (multiples) contributeurs au développement d'AskOmics *(depuis 2015)*
- La plateforme GenOuest pour l'accès aux ressources informatiques

Des questions?