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

Management et intégration des données du projet *DeepImpact*

Le web sémantique via AskOmics

Victor Mataigne, Matéo Boudet

Intégration et interrogation de données hétérogènes via AskOmics

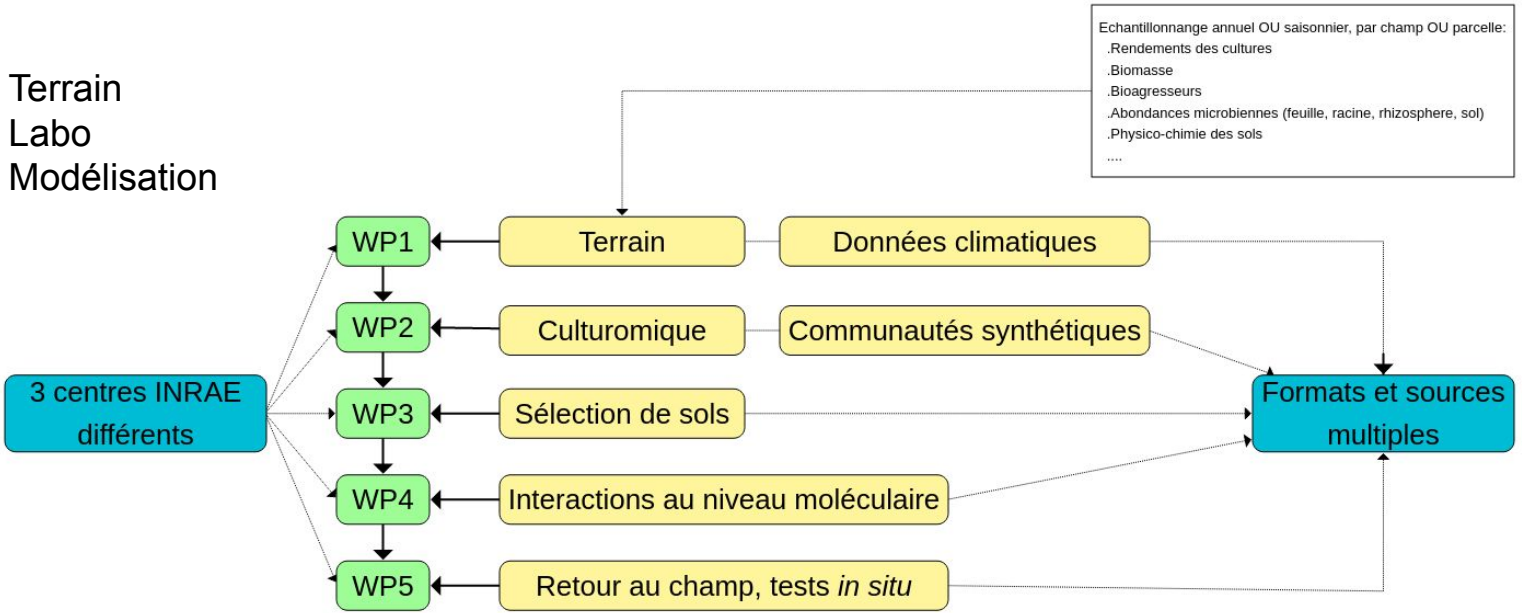
Cas d'étude du projet DeepImpact

DeepImpact - contexte

- Analyse des interactions plante-microbiote pour promouvoir la défense des plantes aux bioagresseurs
- Conception de solutions agroécologiques (i.e. microbiotes synthétiques) contre les stress biotiques pour une meilleure production des plantes cultivées
- Espèces étudiées : blé (*Triticum aestivum*) et colza (*Brassica napus*)
- **200** champs, **3** zones géographiques, **2** ans / **4** saisons d'échantillonnage
- Plusieurs Work Packages successifs (champ →labo→modélisation)

Données : “l’agroécologie du climat au génome”

- Terrain
- Labo
- Modélisation



Des données nombreuses, hétérogènes, fortement liées entre elles

Besoin d’une gestion des données

- Concevoir l'infrastructure
  - Ontologie (standardisation, vocabulaire commun et contrôlé)
  - Stockage et accès modèle de données
- Alimenter les données
  - Modèles permettant de suivre l'ontologie
  - Dépôt facilité
  - Vérification automatique
  - Pipelines données brutes → Données à exploiter
- Intégrer et exploiter les données
  - Schéma de données et interface pour construire des requêtes pour récupérer les données à analyser

Gestion des données orientée utilisateur

Infrastructure et ontologie des données

Interlocuteurs multiples (agents de terrain, chercheurs …)

- Choix des types “d’entités” à représenter dans des fichiers distincts (ex: rendement, taxon, champ, localisation, pratiques agricoles…)
- Standardisation des noms et type de variables pour entité → **Création de templates**
- Identifiants uniques pour chaque entité et chaque relevé

Un vocabulaire et des templates que tout le monde peut s’approprier

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

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

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

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

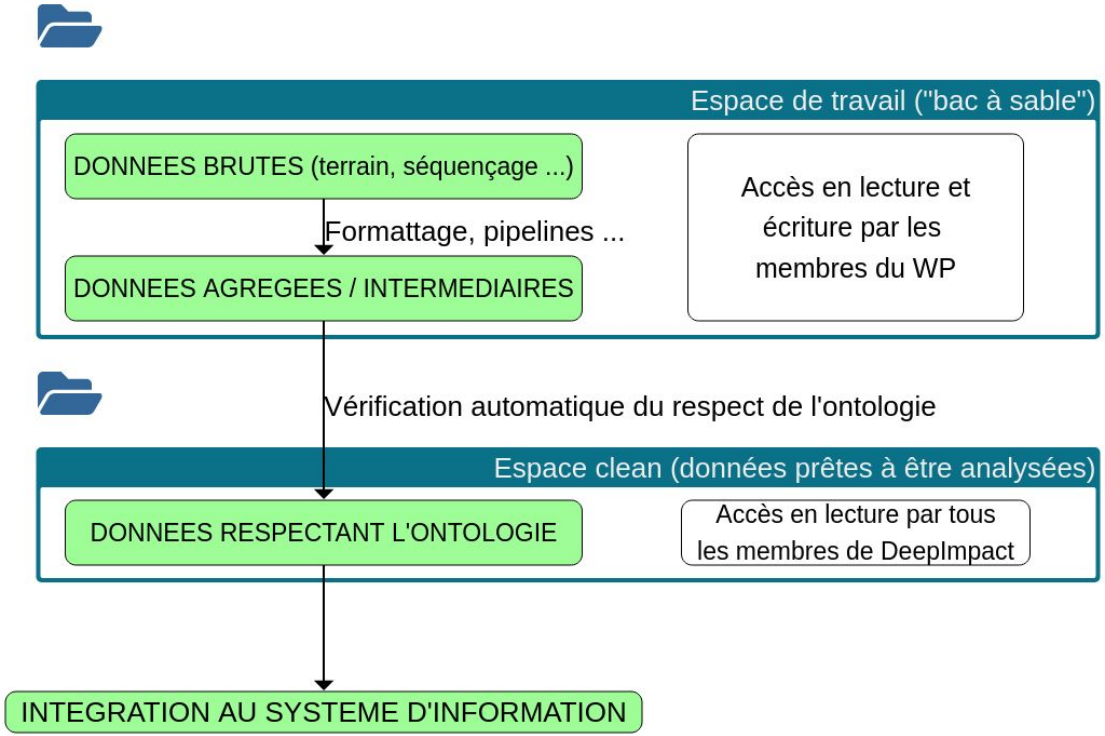
#blast_taxonomy	blast_subject	observation_sum	AF034_Bn_Y21AU_RH_8B_01	AF034_Bn_Y21AU_RH_8B_02	AF034_Bn_Y21AU_RH_8B_03	AF034_Bn_Y21AU_RH_8B_04	AF034_Bn_Y21AU_RH_8B_05
Bacteria;Proteoba	multi-subject	6887	0	0	581	323	0
Bacteria;Proteoba	multi-subject	4899	0	0	0	0	0
Bacteria;Bacteroid	2714719444	4742	0	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

WP1 : 10 templates totalisant quelques dizaines de variables qui permettent de gérer toutes les données d' échantillonnage relevées par plusieurs agents de terrain

Dépôt et échange des données

- Complexe (et risqué) de travailler par échange de mail
- Besoin d'un espace de travail et de dépôt avec contrôle des accès: Dépôt des données sur **CeSGO** (Owncloud de la plateforme **GenOuest**)
  - Possibilité de lancer les scripts de calcul sur les données
  - Gestion fine des permissions

Organisation du stockage et accès



Petites variations selons les WP:

- WP1 : vérification de l'ontologie des tableaux de données avec *checkcel* (outil maison)
- WP2 : les espaces de travail / clean servent à séparer les données de séquençage bruts et intermédiaires (démultiplexage, logs …) des tableaux d'affiliations et de comptages définitifs

Comment requêter les données standardisées ?

- Quels liens entre quelles entités?
- Comment gérer les différentes périodicités des analyses?
  - Biomasse annuelle des champs
  - Physico-chimie saisonnière des champs
  - Bioagresseurs / adventices par parcelle par saison

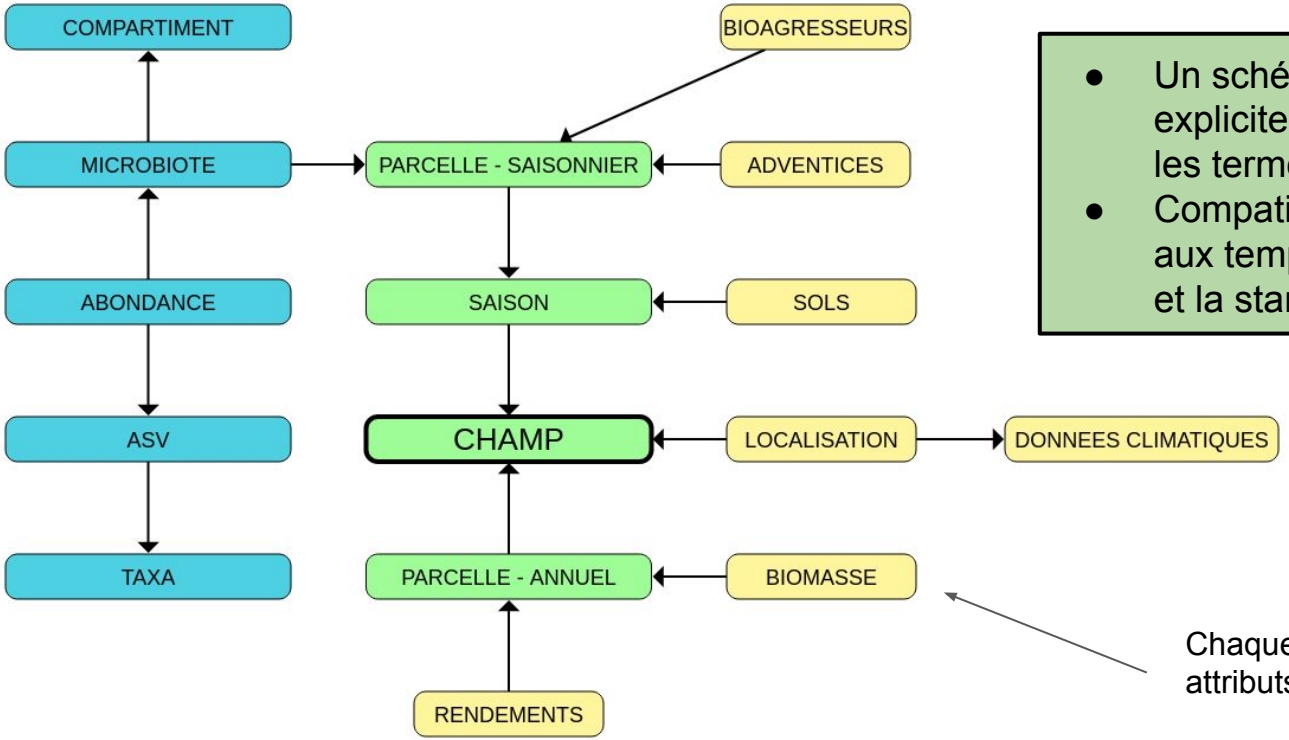
Besoin de mise en place d'une 'abstraction' pour lier les données

- "Cohérente" scientifiquement
- Utilisable par des non-experts

Création d'un schéma de données RDF

Intégration : de l'ontologie aux requêtes

Exemple avec le WP1 (terrain)



- Un schéma de données qui explicite tous les liens entre les termes de l'ontologie
- Compatibilité assurée grâce aux templates de données et la standardisation

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

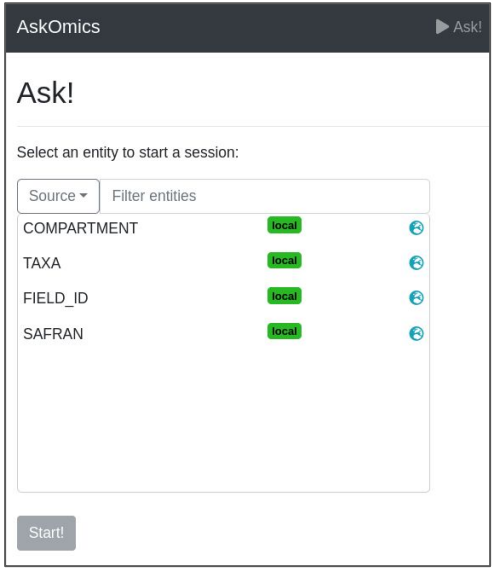
Intégration des données: quel outil?

- Possibilité d'intégrer des données hétérogènes
  - Lien sur les IDs/URIs
- Interface de requête 'simple' pour les membres du projet
- Possibilité de gérer les données 'au fil de l'eau'
  - Interface *adaptée aux données* automatiquement
- Possibilité de créer / modifier / partager des requêtes

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:

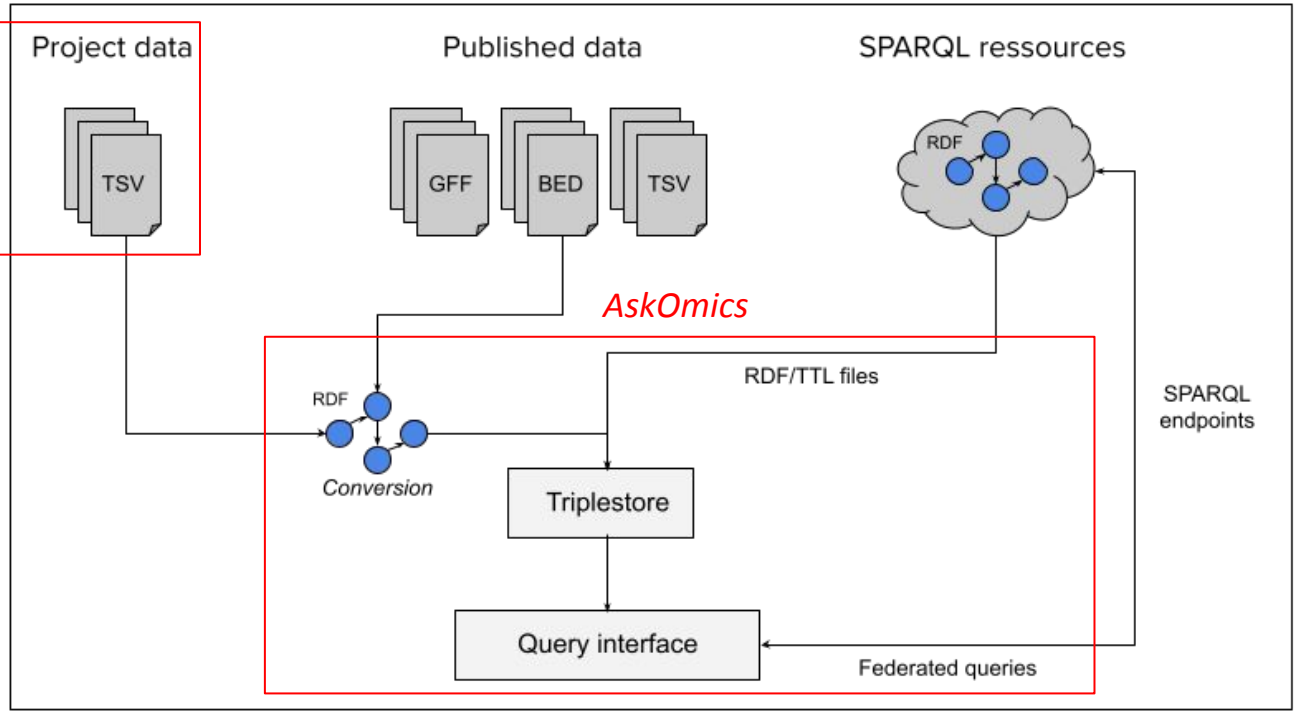
"Liste des taux de calcium et de magnésium des sols des champs en agriculture biologique du sud de la France à la 2ème saison d'échantillonnage de la 1ère année"

"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

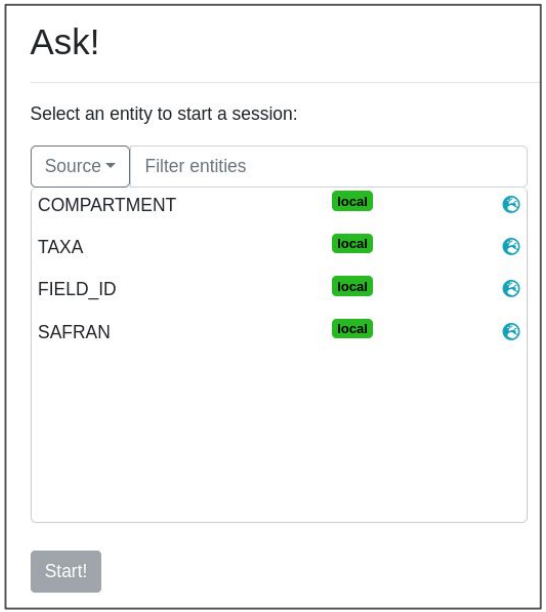
- *Formatées*
- *Nettoyées*
- *Validées*



AskOmics dépend de la '*propreté*' des données: **l'étape de validation est primordiale**

Construction de la requête: point de départ

Choix du type d'entité de départ

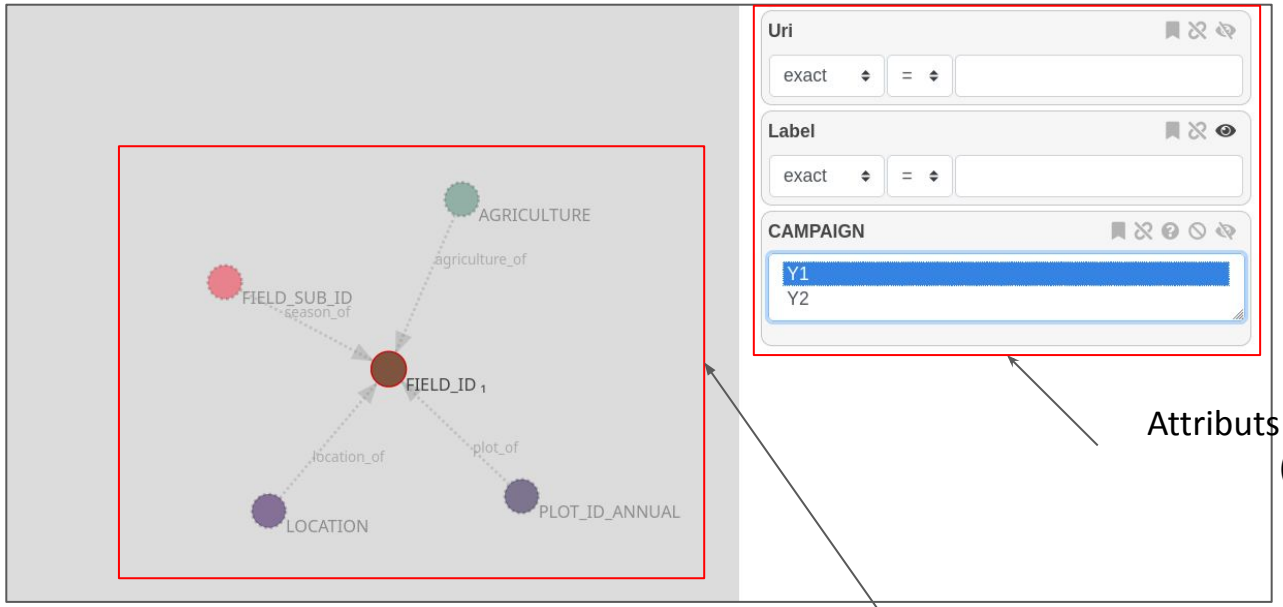


"Les attributs X des entités Y qui valident la condition Z"

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

AskOmics: Interface de construction de requêtes

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



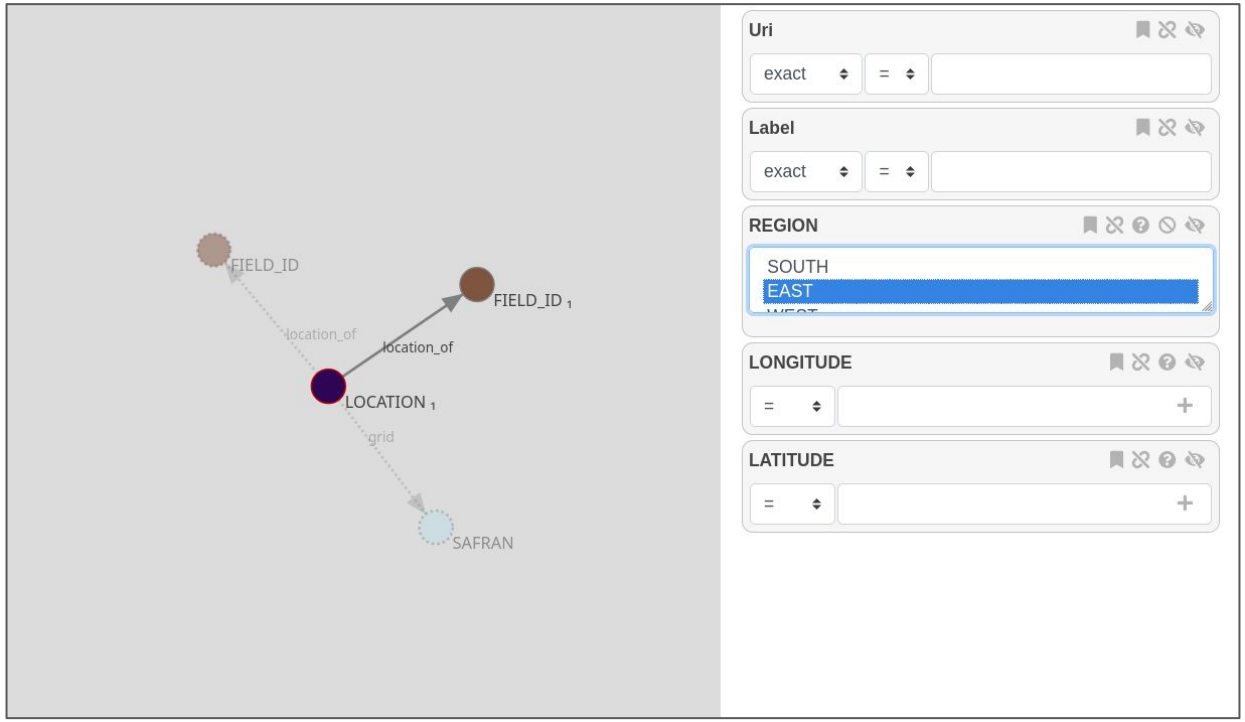
Attributs du type sélectionné (“Année 1”)

Entités liées au type sélectionné

“Tous les champs de l’année 1”

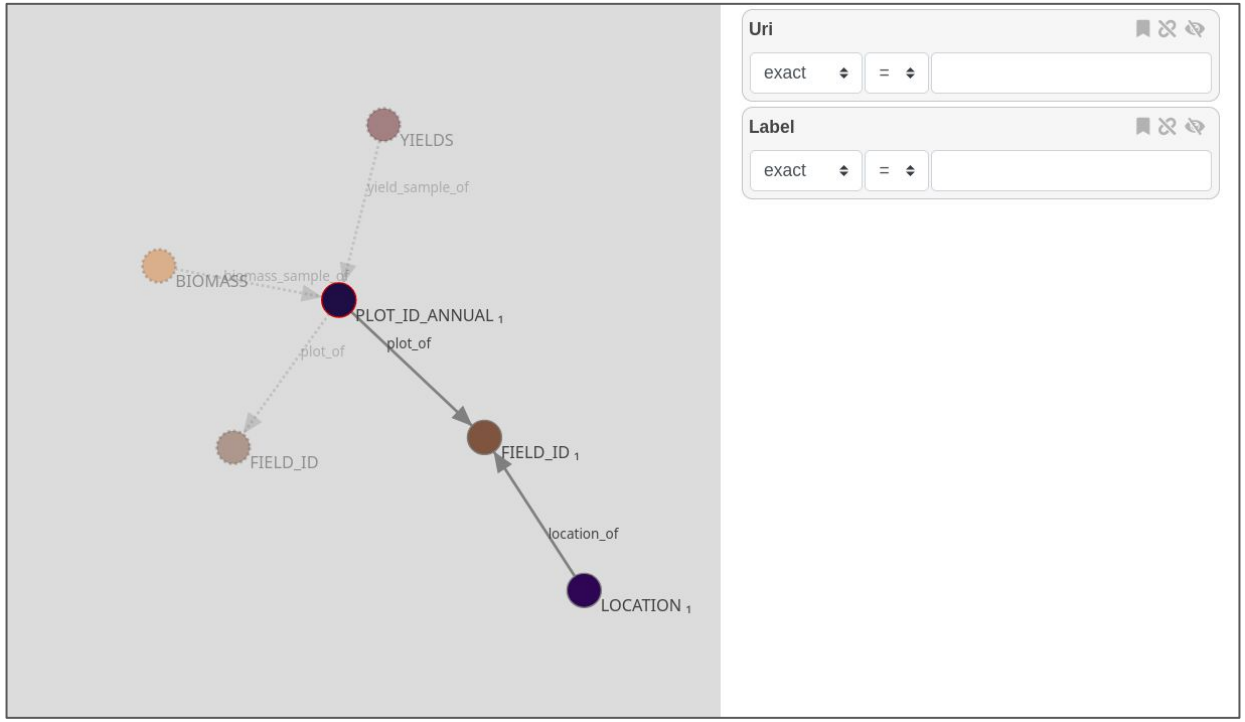
Construction itérative / progressive des requêtes d’entité en entité en spécifiant les attributs

AskOmics: Construction itérative (1)



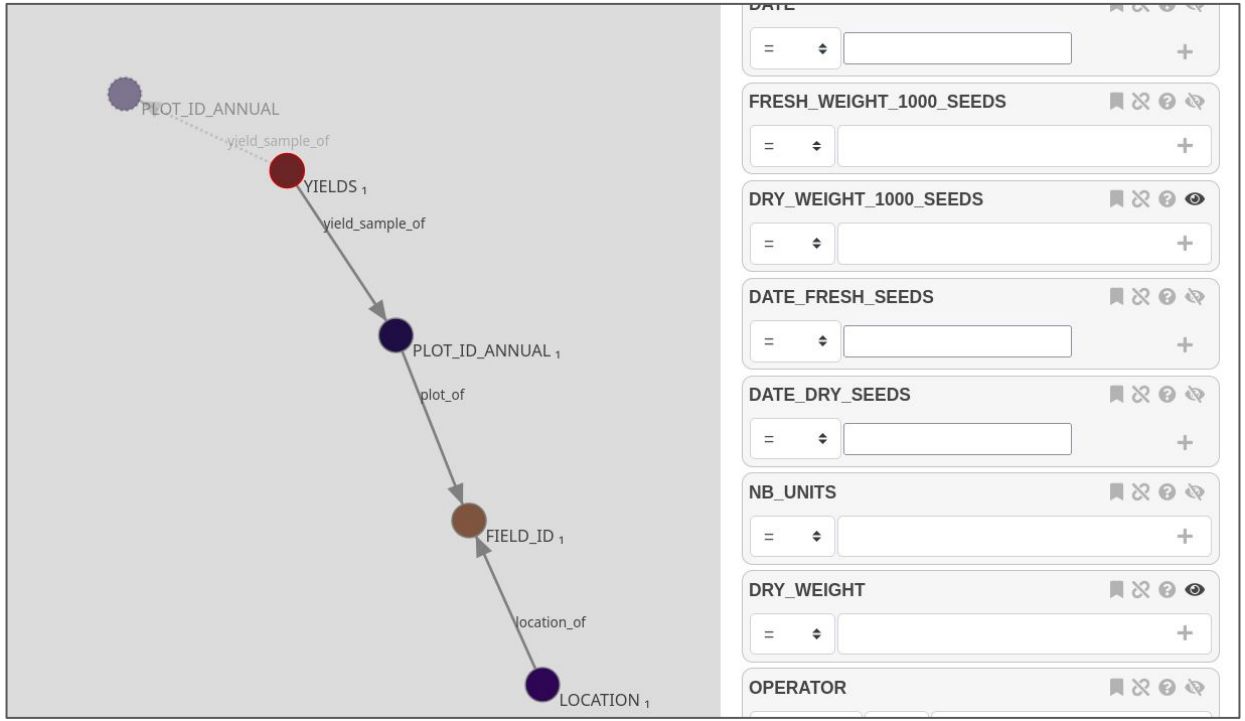
“Tous les champs Y1 dont la localisation est ‘*région Est*”

AskOmics: Construction itérative (2)



“Toutes les parcelles de tous les champs Y1 dont la localisation est ‘*région Est*”

AskOmics: Construction itérative (3)



“Les rendements (poids sec & poids sec 1000 graines) de toutes les parcelles de tous les champs Y1 dont la localisation est ‘*région Est*”””

AskOmics: Output

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

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

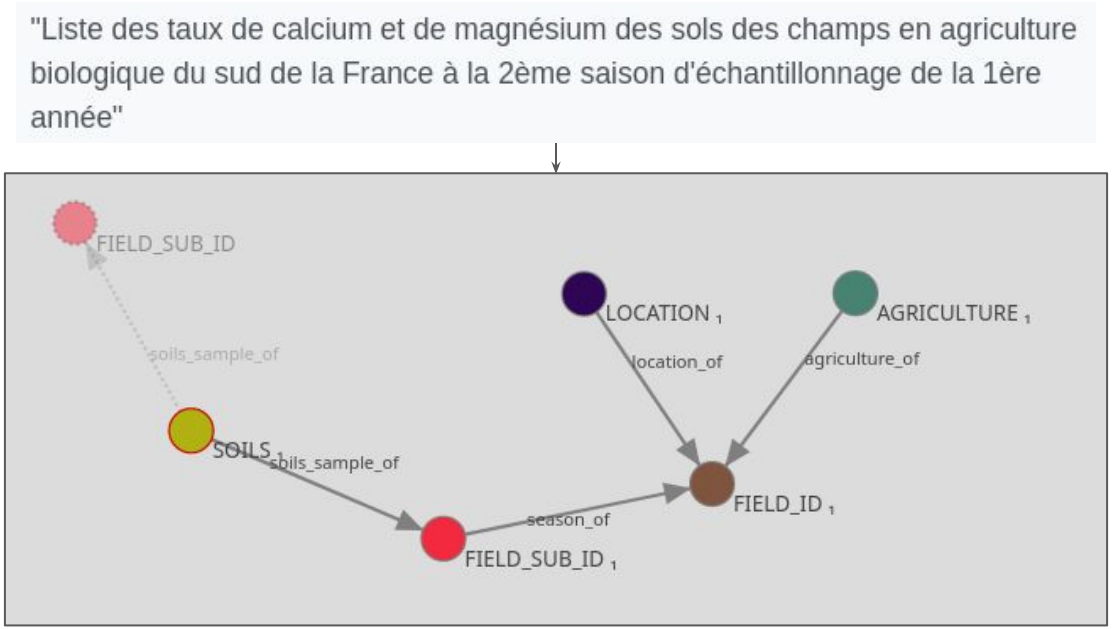
Requêtes : plus généralement

Plusieurs possibilités pour interroger AskOmics (et les données)

- Construction itérative de la requête via l'UI
- Réutilisation & extension d'une requête publique
- Utilisation de formulaires: simple interface de sélection des valeurs
  - "Les rendements de toutes les parcelles de tous les champs XX de la région YY""
- Envoi d'une requête SPARQL 'pure' via l'interface
  - Nécessaire pour les requêtes complexes (ex: opérations sur les résultats)
- Construction itérative puis modification du code SPARQL généré

Partage de requêtes: direct

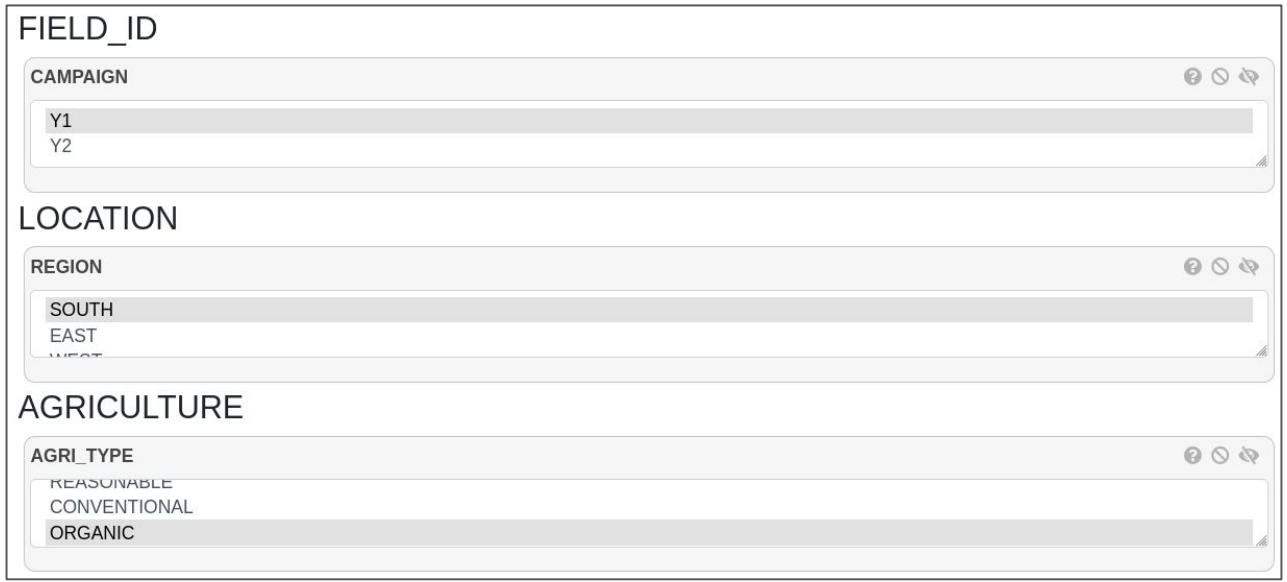
- Possibilité de ‘sauvegarder’ sa requête, et de la rendre publique
- Les autres utilisateurs peuvent la ré-utiliser et itérer dessus



Partage de requêtes: les formulaires

- Création d'une requête 'normale'
- Sélection d'attributs 'modifiables'
- Création du formulaire

Mise à disposition d'une requête complexe mais customisable simplement



AskOmics & DeepImpact : exemples

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?