Work4Graph – 20201207

CC BY

Utilisation de bases de données et de représentations graphe pour la gestion de connaissances

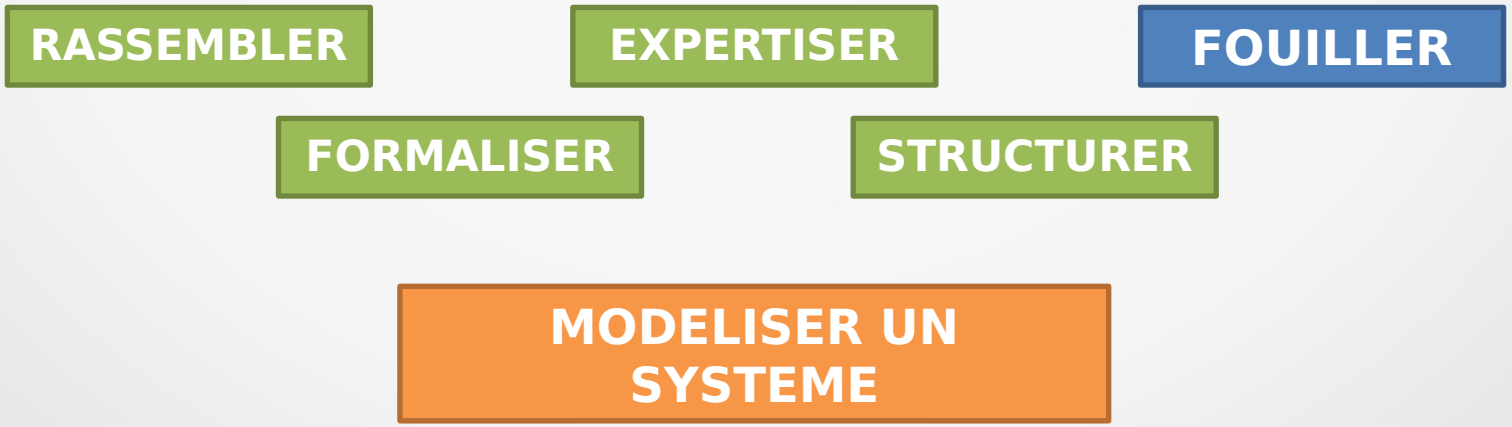
sebastien.carrere@inrae.fr

LIPME, INRAE/CNRS

Bases de connaissances

- Une base de connaissance sert à **rassembler** - de manière **centralisée** - **l'expertise** d'un domaine généralement formalisée de manière déclarative.
- Une base de connaissance regroupe des **connaissances spécifiques** à un domaine spécialisé donné, sous une forme **exploitable** par un ordinateur. Elle peut contenir des **règles** (dans ce cas, on parle de base de règles), des faits ou d'autres représentations. Si elle contient des règles, un moteur d'inférence - simulant les **raisonnements déductifs** logiques - peut être utilisé pour déduire de **nouveaux faits**.

Source Wikipédia



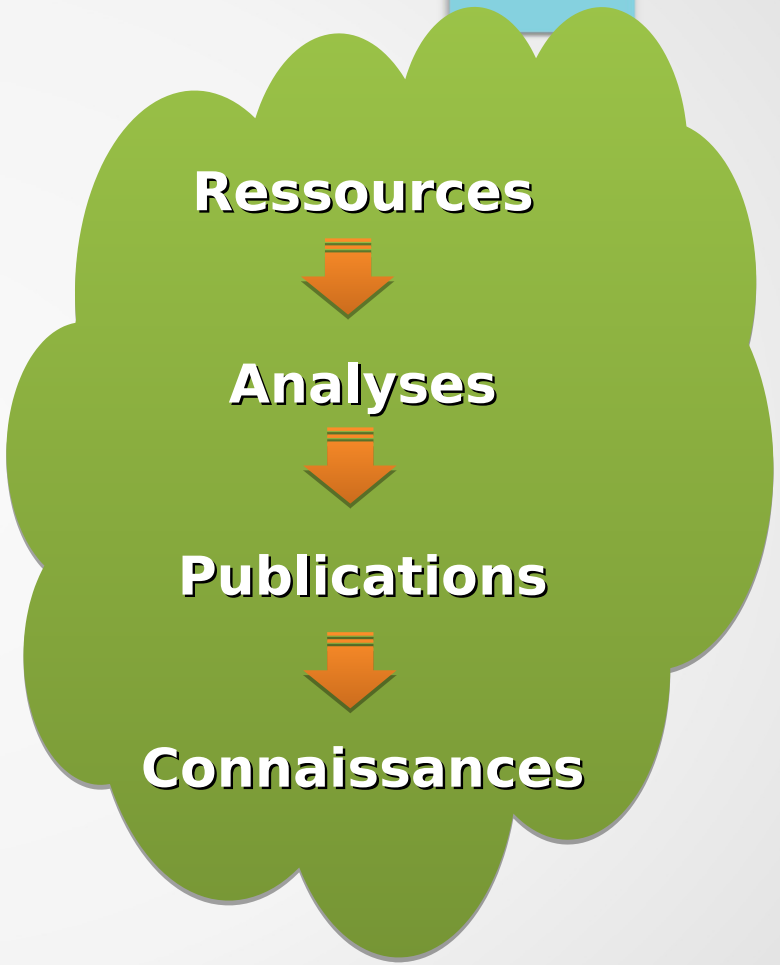
Cas d'école : *M. truncatula*

1997 : Premières EST au NCBI

2003: Début du projet de séquençage

2013/07/31: Mt4.0v1

2017/03: Annotation Séquence PacBio



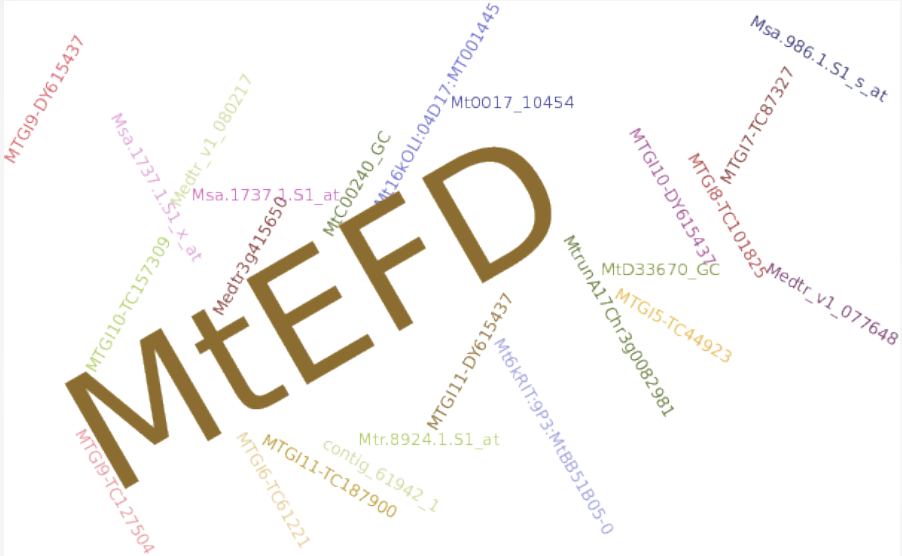
Cas d'école : *M. truncatula*

20 ans de production de ressources

RESSOURCE	DESCRIPTION	TYPE
affx-1	Affymetrix Chip	Oligos
NimbleGen-GPL16373	NimbleGen custom IRHS_Medtr_102K_v1 - design sequences	Oligos
Mt1.0.mRNA	mRNA sequences predicted from the genome assembly release 1	mRNA
IMGA-Mt2-gene	gene sequences predicted from the genome assembly release 2	Genes
IMGA-Mt3-gene	gene sequences predicted from the genome assembly release 3	Genes
IMGA-Mt3.5.1-gene	gene sequences predicted from the genome assembly release 3.5.1	Genes
JCVI-Mt4.0-gene	gene sequences predicted from the genome assembly release 4.0 at JCVI	Genes
MtrunA17r5.0-ANR	**gene sequences predicted from the genome assembly release 5.0 at INRA**	**Genes**
Mt.genomic.GBKDec2009	Genbank mRNA sequences tagged as "genomic" (Dec 2009)	mRNA
Mt.mRNA.GBKJan2009	Genbank mRNA sequences tagged as "complete CDS" (Jan 2009)	mRNA
Mt16kOLIplus-2004	Set of 70mers representing 16.086 tentative consensus sequences of the TIGR Gene Index version 5 plus 384 probes primarily representing transcription factors (H. Kuster et al)	Oligos
Mt20120830	gene, ncRNA and missing sequences predicted from the genome assembly release 3.9 (used for Nimblegen array)	Genes
Mt6kRIT-Jan2003	cDNA clone sequences representing 5648 EST-clusters from M.truncatula root nodules, AM roots, and uninfected roots. (H. Kuster et al)	cDNA
MtCDJan2003	EST Navigation System - release January 2003 (Journet et al. 2002)	EST
MtGI5 → 11	gene Index release 5 – 2002 → 11 2011	EST
MtSCDJan2003	EST Navigation System - release December 2003 (includes SSH)	EST
MtSCDJun2006	EST Navigation System - release June 2006	EST
NCR-2003	311 EST sequences of nodule-specific genes. Mergaert et al, 2003	EST

Cas d'école : *M. truncatula*

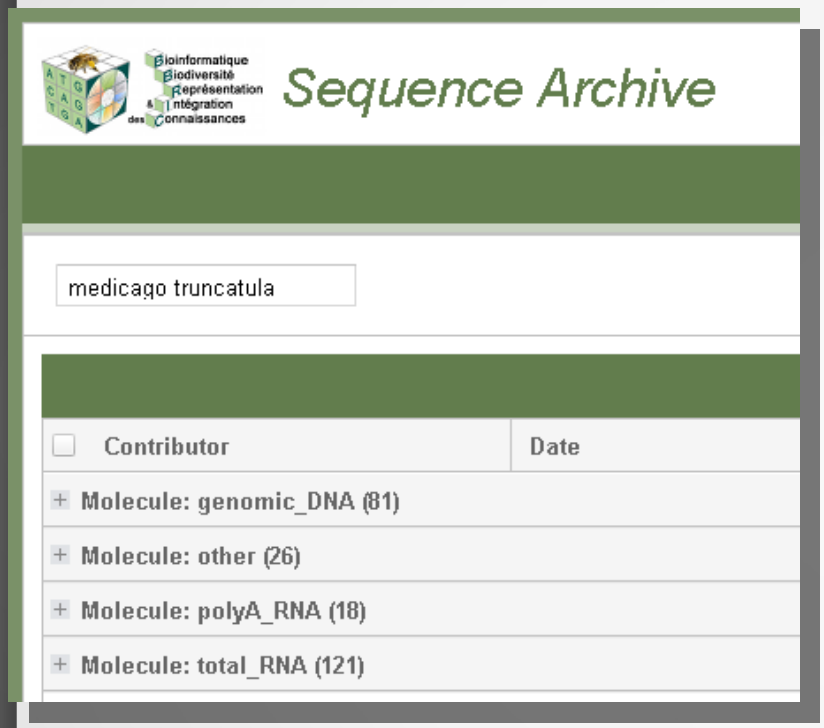
20 ans de production de ressources



Un objet biologique : X façons de le nommer

Cas d'école : *M. truncatula*

20 ans de production de données et connaissances



" *Medicago truncatula* "
- 1499 GEO datasets

" *Medicago truncatula* affymetrix "
→ 21 publications

"*Medicago truncatula* interaction "
→ 822 publications

"*Medicago truncatula* transcriptome "
→ 260 publications

"*Medicago truncatula* regulation "
→ 797 publications

Cas d’école : *M. truncatula*

Comment intégrer/structurer ces résultats ?

Comment représenter toutes ces connaissances ?

Comment trouver une information pertinente ?

Etape n°1

Collecter les Connaissances

+

Curation

Collecte des connaissances

Côté utilisateur → K.I.S. !

ObjetA | Relation | ObjetB

EFD Is an ERF transcription factor involved in the control of nodule number and differentiation in **Medicago truncatula**.

Vernié T, Moreau S, de Billy F, Plet J, Combier JP, Rogers C, Oldroyd G, Frugier F, Niebel A, Gamas P.

Plant Cell. 2008 Oct;20(10):2696-713. doi: 10.1105/tpc.108.059857. Epub 2008 Oct 31.

objectA	organismA	genotypeA	Relation	objectB	organismB	genotypeB
EFD	Medicago truncatula	jemalong A17	controls	Nodule number and differentiation	Medicago truncatula	jemalong A17

Collecte des connaissances

Côté utilisateur → K.I.S. !

objecttypeA	objectA	organismA	genotypeA	relationshiptype	objecttypeB	objectB	organismB	genotypeB
gene	MtEFD	Medicago truncatula	jemalong A17	regulates	biological process	root nodule differentiation	Medicago truncatula	jemalong A17

Vocabulaire contrôlé

condition	Relationship organism	reference
efd-1 mutant phenotype (fix-, poorly differentiatied nodules)	Sinorhizobium meliloti 2011	PMID:18978033

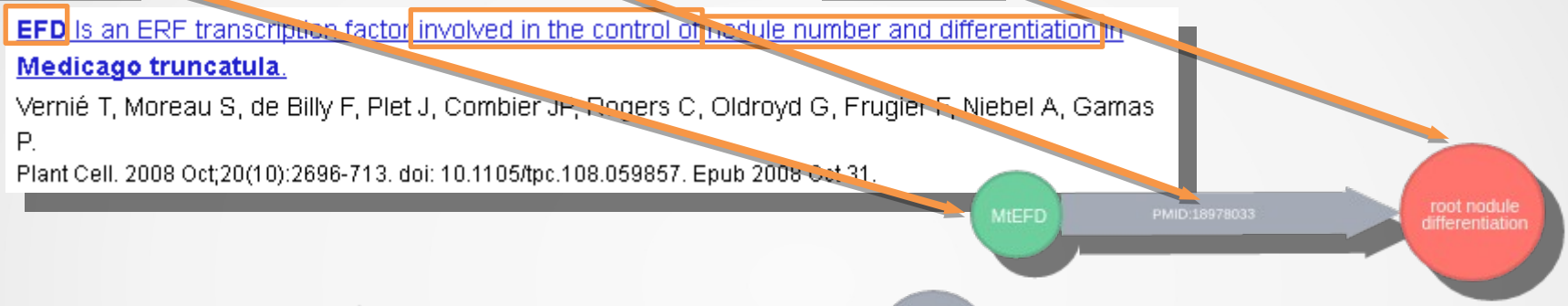
Etape n°2

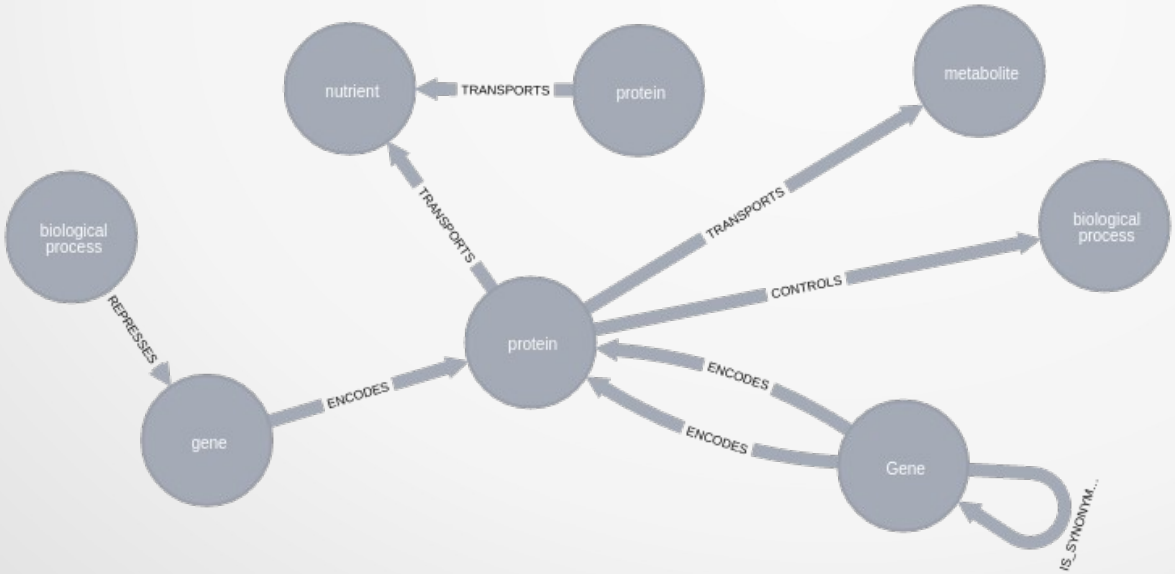
Structurer les données collectées

Structuration de l'information

Formalisme

ObjetA — Relation — ObjetB

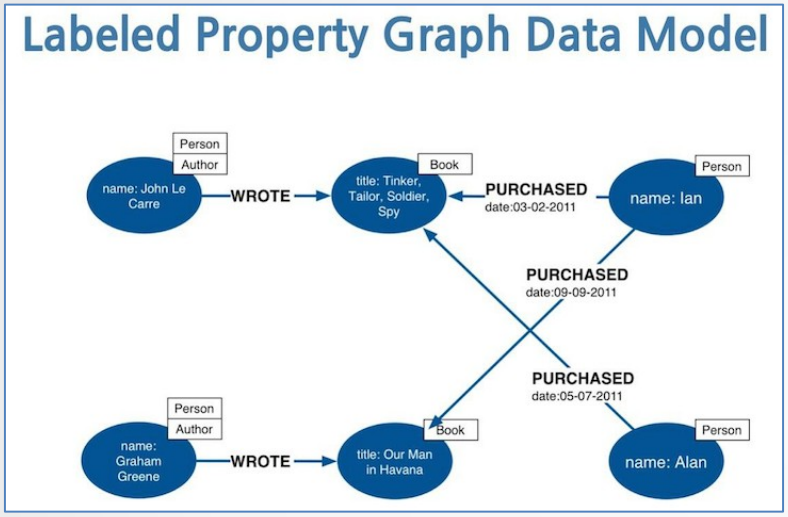




Graphe de relations

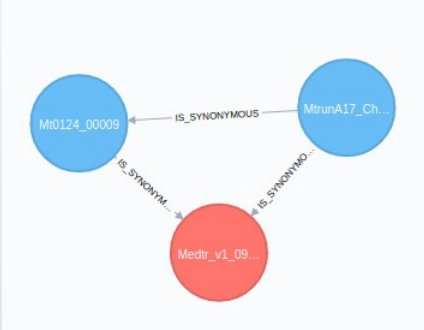
Structuration de l'information

neo4j



- Base de données orientée graphes
- Les nœuds et les arêtes sont typés
- Les nœuds et les arêtes peuvent avoir des propriétés
  → annotations, organisme, condition
- Langage de requête avec des algorithmes implémentés (ex : shortestPath)

```
MATCH p=(n)-[r:IS_SYNONYMOUS]-()
WHERE n.object = "Medtr_v1_090570"
RETURN p
```




```
MATCH (n:BIOLOGICAL_PROCESS)-[r]-(x)
WHERE r.reference = 'PMID:24483147'
RETURN count (distinct r),
       r.reference AS reference,
       n.subtype AS subtype
```

count

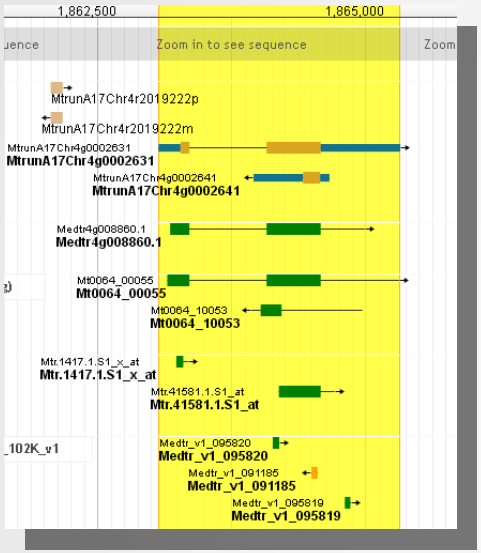
(distinct r)	reference	subtype
16793	PMID:24483147	N-fixing & arbusc. mycorrhizal symbioses

Etape n°3

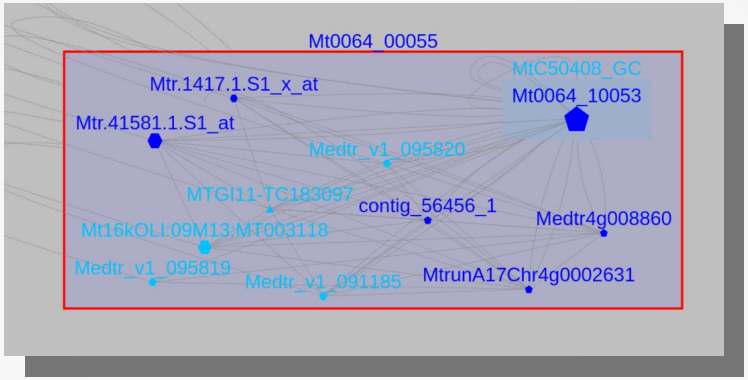
Construire la base de données

Workflow Build

- Calcul des **liens de synonymie** entre différents datasets



Utilisation de l'assemblage V5 comme pivot
Meilleur alignement + Intersection



GMAP

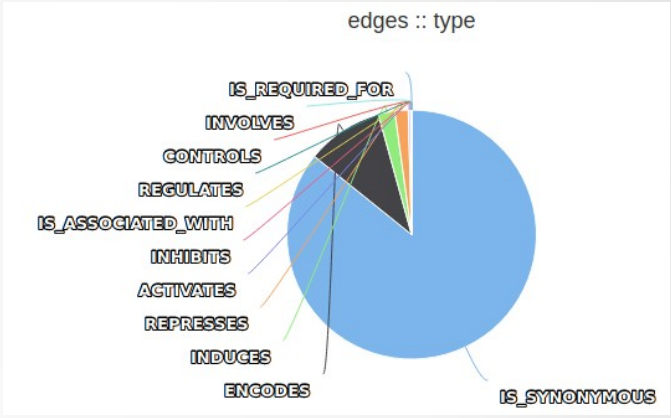
- Import des fichiers Excel des curateurs et **correction des typos**
- Construction de fichiers tabulés avec **identification unique des nœuds** (signature = type +accession + organisme + genotype)

- **Import** dans la base de données

Quelques chiffres

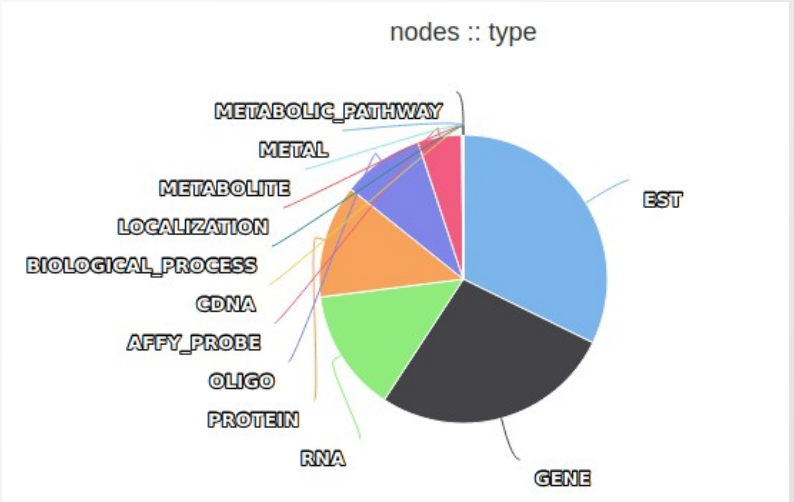
LeGOO

- Release Name: **20200619**
- Number of nodes: **1211527**
- Number of relations: **5681356**
- Number of organisms: **36**
- Number of references (PMID/DOI): **935**



Taille de la base de données : 4.7Go
Taille des input : ~ 2.5Go

Temps build : ~30min (*)
- nettoyage
- + formatage
- + import neo4J
- + indexation elasticsearch

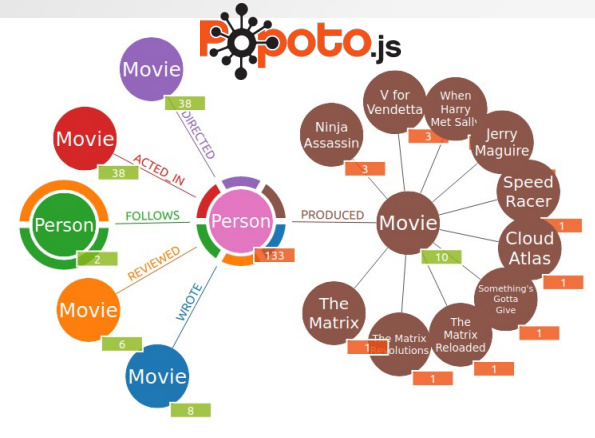


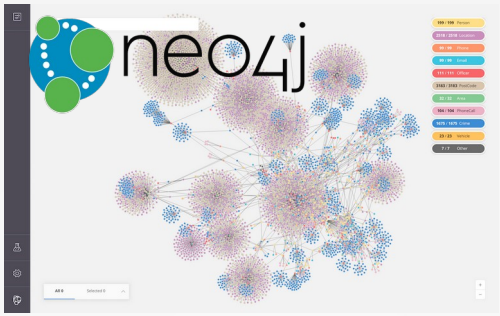
* sur un serveur avec assez de RAM
- nécessité de tout avoir en mémoire pour calculer des scores de pertinence
→ #relations avec edge:type!= « IS_SYNONYMOUS » ou avec un synonyme pertinent

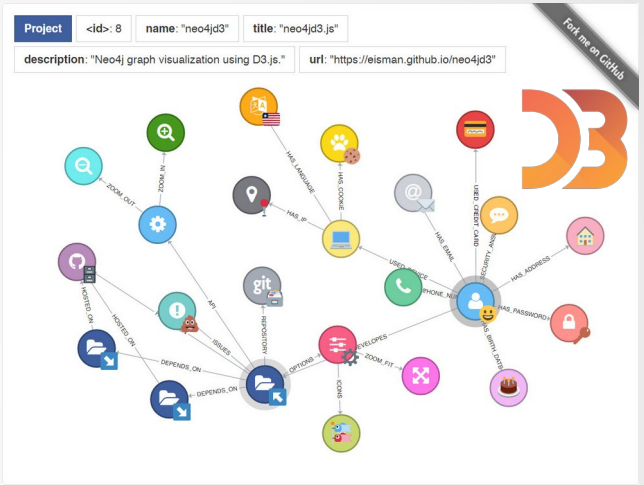
Etape n°4

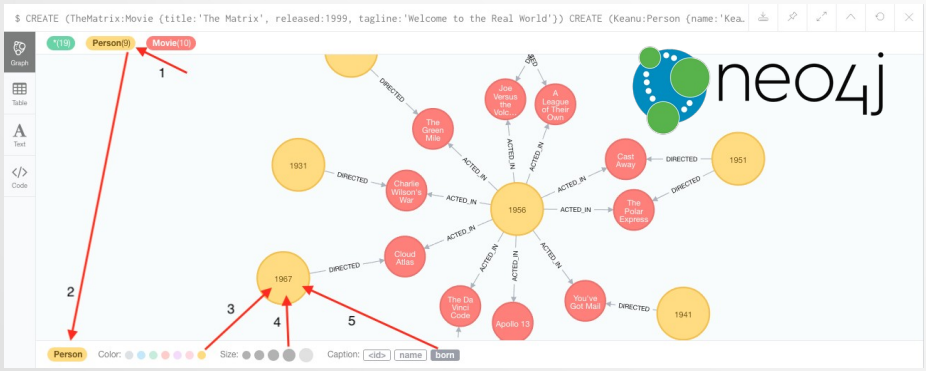
Représenter les connaissances collectées

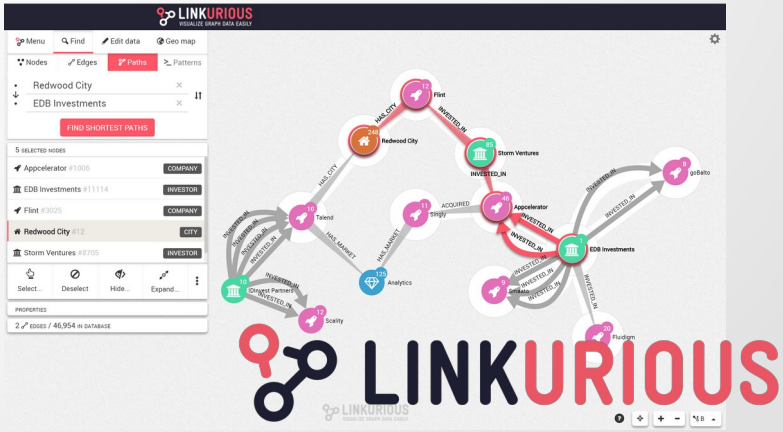
Représentation de l'information







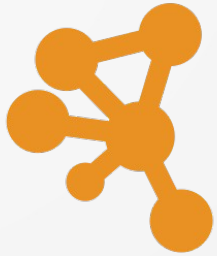




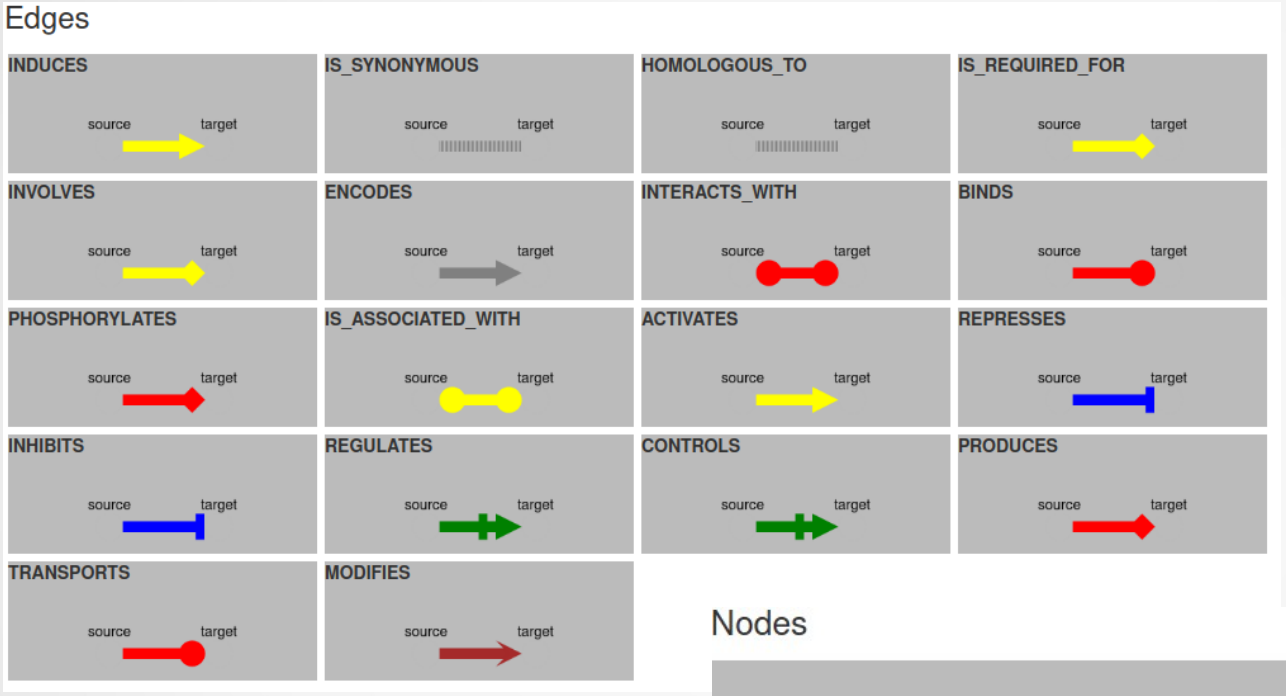
Représentation de l’information

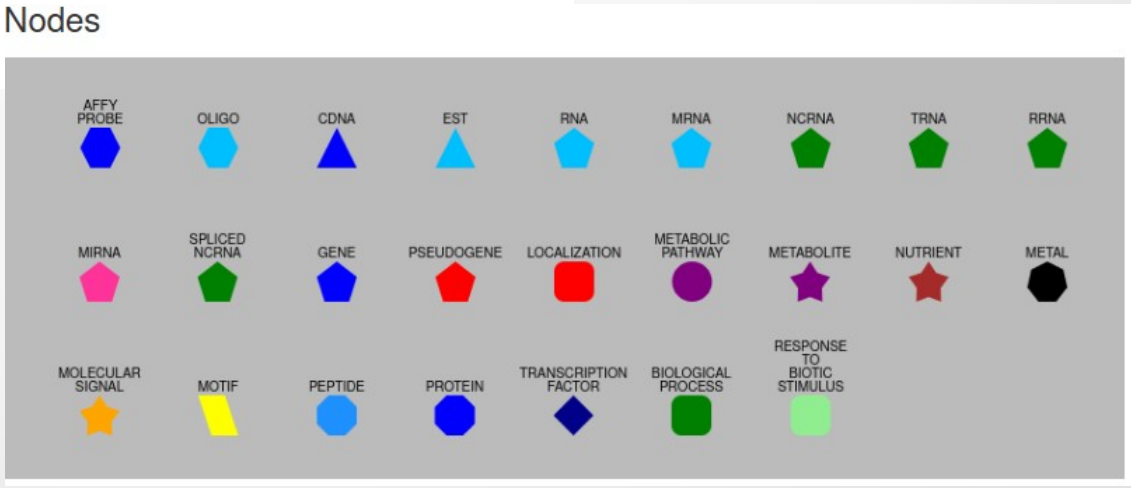
Cytoscape.js

- Prise en main simple
- Customisation simple
- Nombreux plugins
- Grosse communauté (Cytoscape)



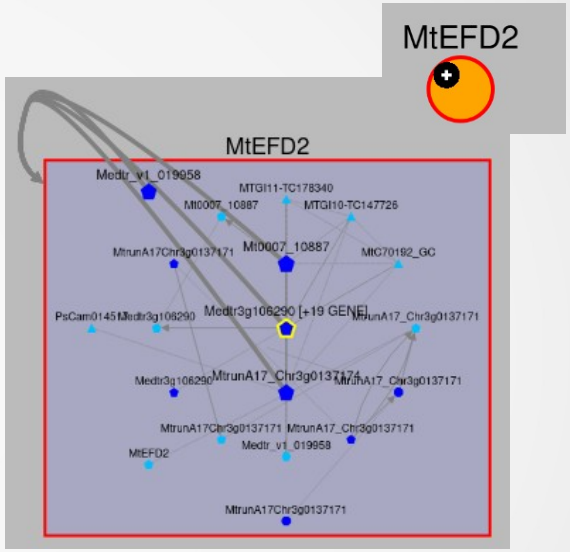
Représentation de l'information

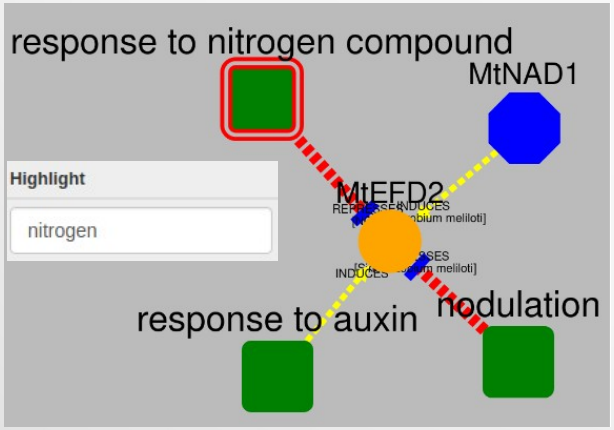


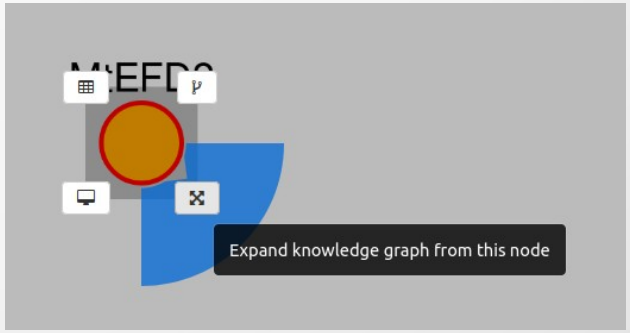


Représentation de l’information

- cytoscape.js-panzoom
- cytoscape.js-expand-collapse
- cytoscape.js-cxtmenu
- cytoscape.js-undo-redo
- cytoscape.js-view-utilities







Résultats

LeGOO

https://www.legoo.org

An Expertized Knowledge Database For The Model Legume *Medicago truncatula*.

EffectorK

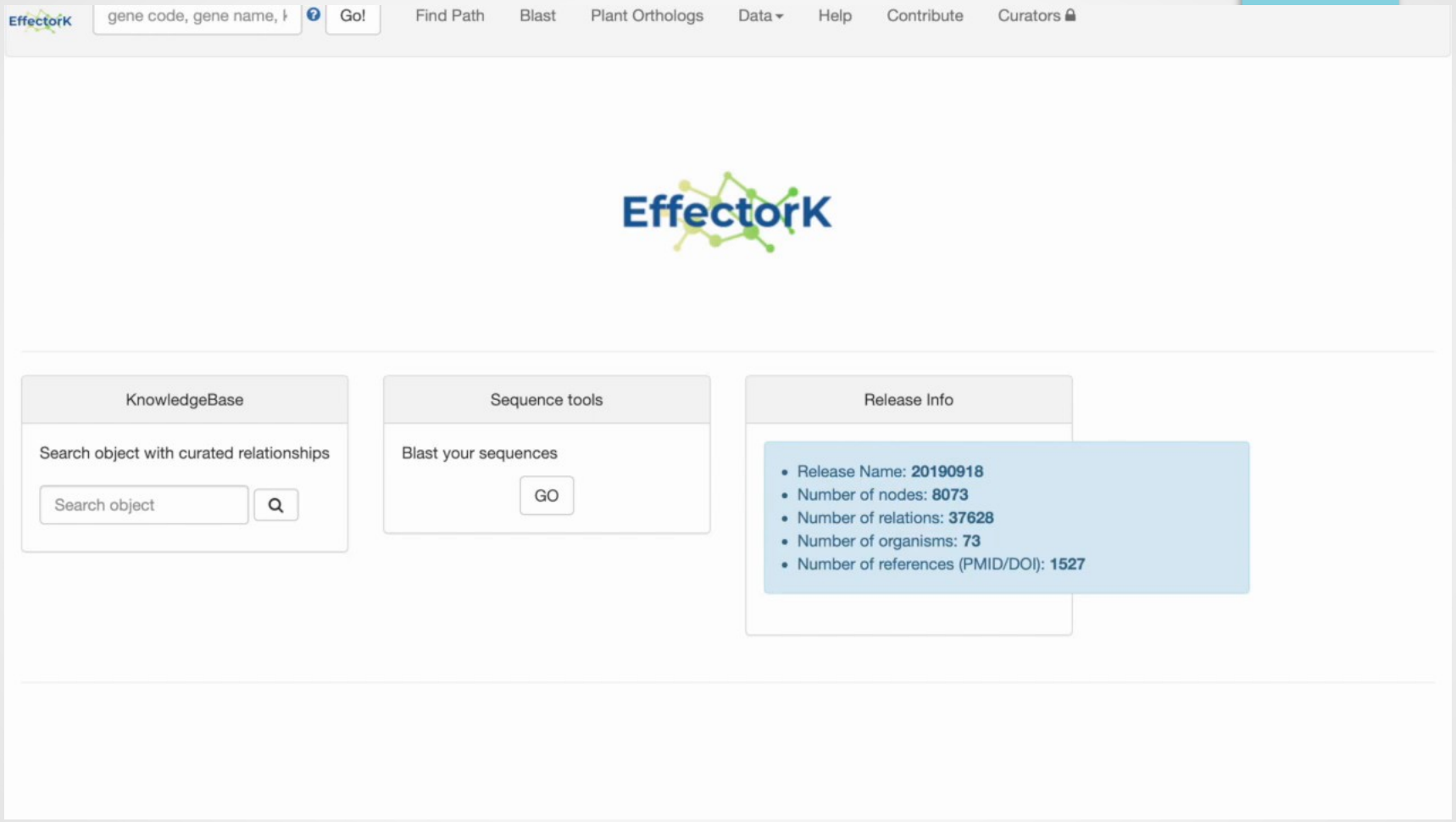
http://www.effectork.org

A comprehensive resource to mine for pathogen effector targets in the *Arabidopsis* proteome

XanthoK

An Expertized Knowledge Database For The bacteria *Xanthomonas campestris pv. campestris*.

Demo (sans l’effet démo)



Source: https://youtu.be/B1QZudpRCKk
Auteur : Nemo Peeters (LIPM)

Mais la visu graph ...

XANTHOK | Xop | "Xop" results

Locus XC_1553

Type	Name	Genotype	3rd party organism
PROTEIN	XC_1553	8004	Xanthomonas campestris pv. campestris

Synonymous 2

Type	Name	Genotype	3rd party organism
GENE	XCC8004_a19341	8004	Xanthomonas campestris pv. campestris
GENE	XC_RS07755	8004	Xanthomonas campestris pv. campestris

Litterature 4

Title	Authors	Journal	Year	Relations
Systematic Functional Analysis of Sigma (σ) Factors in the Phytopathogen Xanthomonas campestris Reveals Novel Roles in Regulation of Virulence and Viability.[PudMed] [DOI]	Yang LY, Yang LC, Gan YL, Wang L, Zhao WZ, He YQ, Jiang W, Jiang BL, Tang JL.	Front Microbiol	2018	Not currated
RpfC regulates the expression of the key regulator hrpX of the hrp/T3SS system in Xanthomonas campestris pv. campestris.[PudMed] [DOI]	Jiang BL, Jiang GF, Liu W, Yang LC, Yang LY, Wang L, Hang XH, Tang JL.	BMC Microbiol	2018	Not currated
Identification of a novel type III secretion-associated outer membrane-bound protein from Xanthomonas campestris pv.[PudMed] [DOI]	Li L, Li RF, Ming ZH, Lu GT, Tang JL.	Sci Rep	2017	Not currated
The type III effectors of Xanthomonas.[PudMed] [DOI]	White FF, Potnis N, Jones JB, Koebnik R.	Mol Plant Pathol	2009	Not currated

Knowlege

INDUCED_BY (2) | INTERACTS_WITH (36)

Type	Name	Condition	Xref	Source Description	Target Description	Genotype	3rd party organism
PROTEIN	AT5G63790.1	Y2H (XopAC-H469A screened)		XopAC, XopAC, AvrAC	ANAC102,NAC102, NAC domain containing protein 102, IPR003441:NAC domain	Col-0	Arabidopsis thaliana
PROTEIN	AT5G51440.1	Y2H		XopAC, XopAC, AvrAC	HSP20-like chaperones superfamily protein, IPR008978:HSP20-like chaperone, IPR031107:Small heat shock protein HSP20, IPR002068	Col-0	Arabidopsis thaliana
PROTEIN	AT5G42270.1	Y2H (XopAC-H469A		XopAC, XopAC, AvrAC	VAR1,FTSH5, FtsH extracellular protease family, IPR003959:ATPase, AAA-type, core, IPR005936:Peptidase, FtsH, IPR027417:P-loop containing nucleoside triphosphate hydrolase, IPR000642,	Col-0	Arabidopsis thaliana

Merci

PLANT & CELL PHYSIOLOGY

LeGOO: An Expertized Knowledge Database for the Model Legume *Medicago truncatula* FREE

Sébastien Carrère ✉, Marion Verdenaud, Clare Gough, Jérôme Gouzy, Pascal Gamas

Author Notes

Plant and Cell Physiology, Volume 61, Issue 1, January 2020, Pages 203–211, https://doi.org/10.1093/pcp/pcz177

Published: 17 September 2019 **Article history** ▾

Molecular Plant Pathology Open Access

ORIGINAL ARTICLE | Open Access

EffectorK, a comprehensive resource to mine for *Ralstonia*, *Xanthomonas*, and other published effector interactors in the *Arabidopsis* proteome

Manuel González-Fuente, Sébastien Carrère, Dario Monachello, Benjamin G. Marsella, Anne-Claire Cazalé, Claudine Zischek, Raka M. Mitra, Nathalie Rezé, Ludovic Cottret ... **See all authors**

First published: 15 August 2020 | https://doi.org/10.1111/mpp.12965 | Citations: 2

Manuel González-Fuente and Sébastien Carrère contributed equally to this work.