Making Data Deeply FAIR through Lightweight Standards, the KnetMiner and ELIXIR Cases

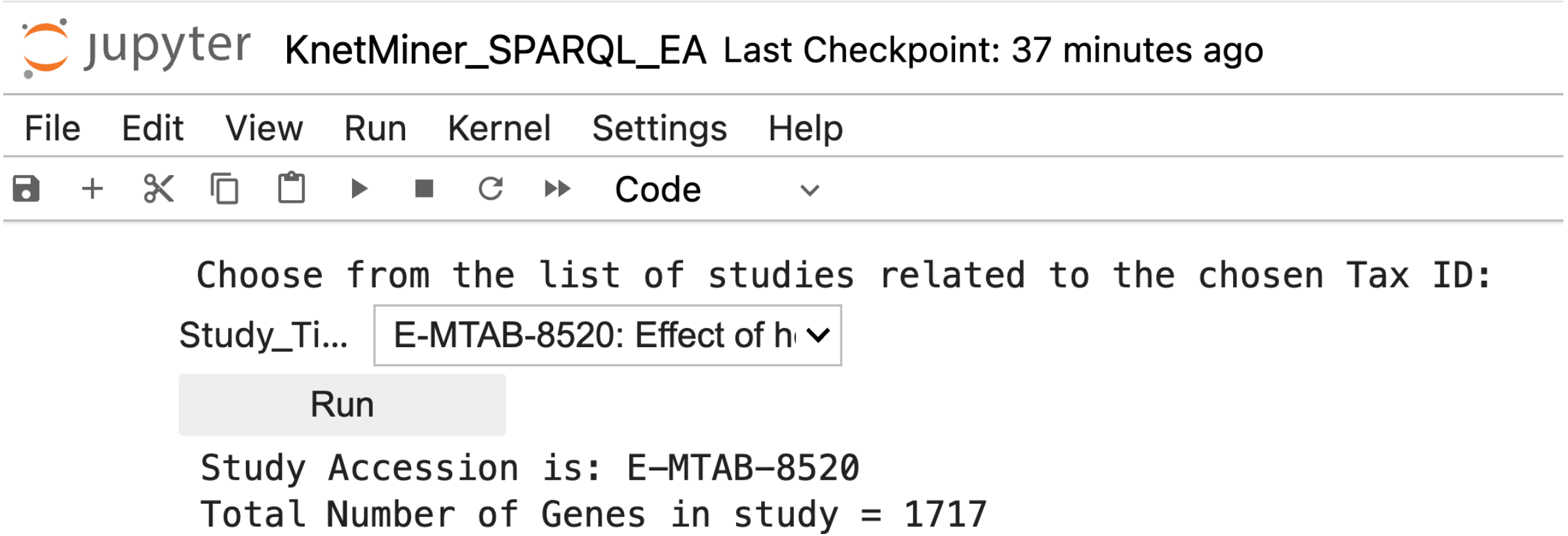
Marco Brandizi, Rothamsted Research

Nov 20th, 2023, ELIXIR Plant Biology Community Meeting

Background credits: Gianni del Bufalo, Marche countryside (my birthplace) https://www.flickr.com/photos/bygdb/14668615409

Let's start from a Use Case...

KnetMiner



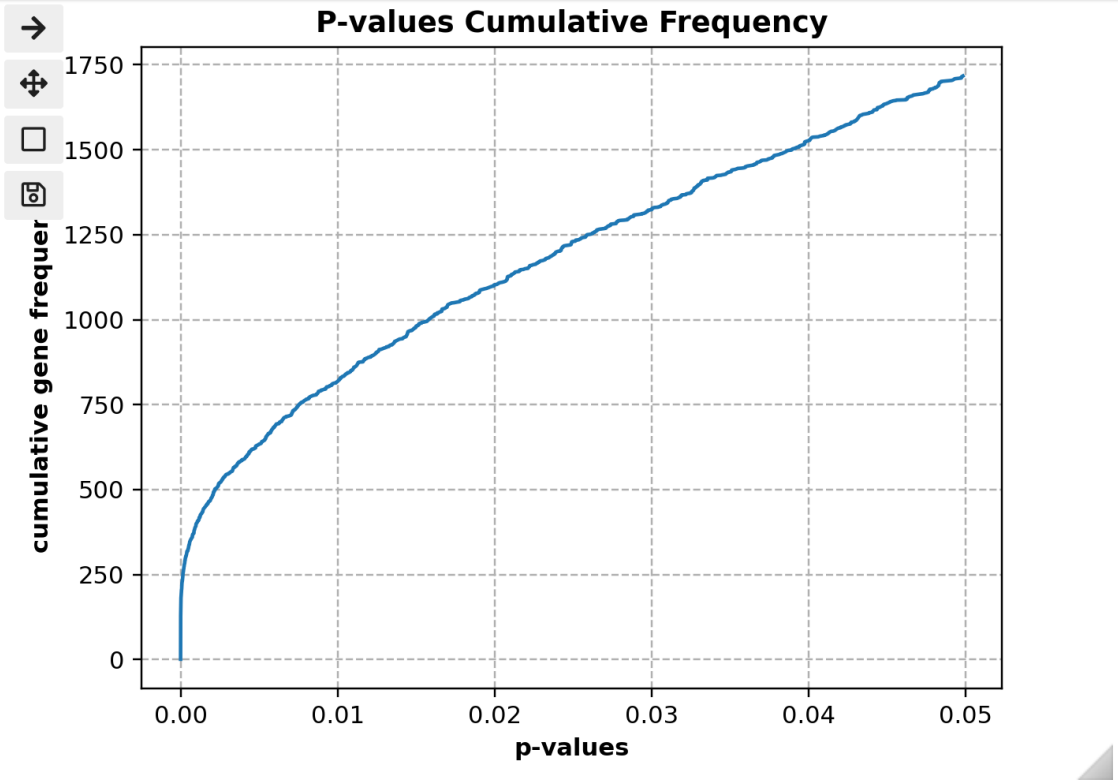
Let's start from a Use Case...

KnetMiner

jupyter KnetMiner_SPARQL_EA Last Checkpoint: 39 minutes ago

File Edit View Run Kernel Settings Help

Code JupyterLa



After choosing the p-value and number of genes with the slider, click the 'Run Analysis' s results.

Pvalues 0.0001

p-value = 0.0001, Number of genes = 228

Run Analysis

Let's start from a Use Case...

KnetMiner

The enrichment table below has 153 rows.
To view the whole table, see the 'View whole tables section' or click on the download link below:

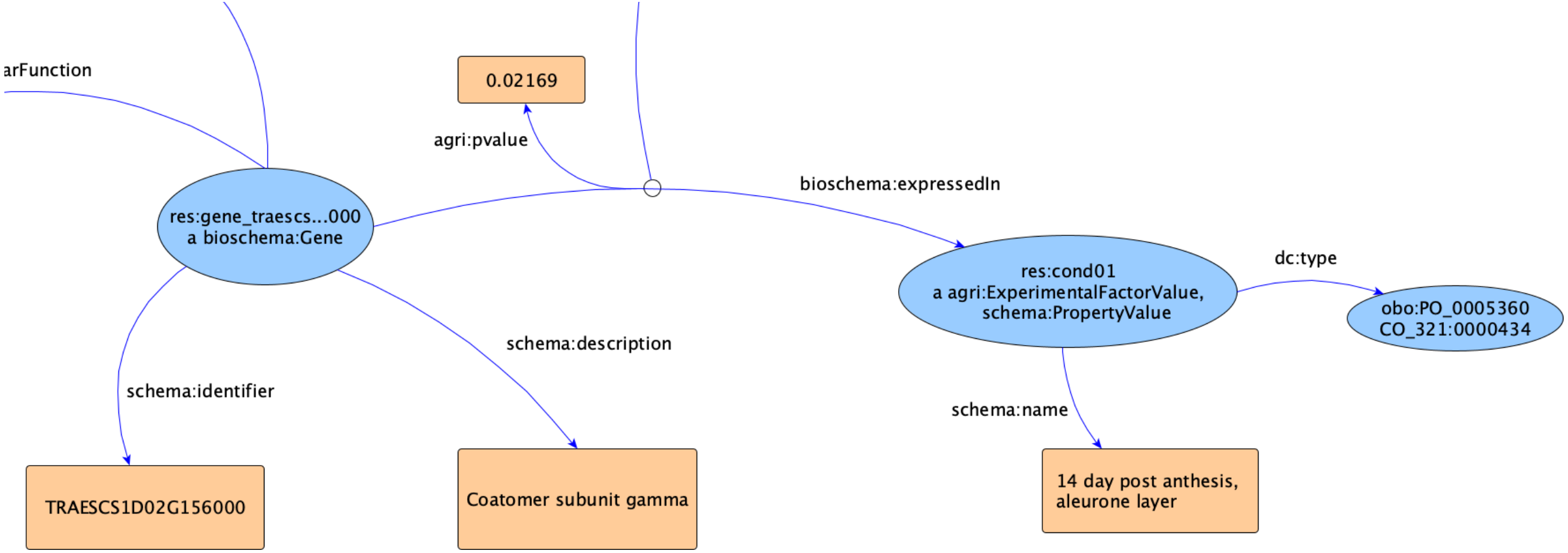
Download enrichment table CSV file

	Ontology Term	Preferred Name	odds ratio	exact p-value	adj p-value	Reference Genes	User/Study Genes
0	TO_0000430	germination rate	17.746196	2.489665e-77	3.809187e-75	5626	107
1	TO_0006002	proline content	10.730036	3.963930e-57	3.032406e-55	8960	107
2	TO_0000276	drought tolerance	5.943980	1.637385e-36	8.350665e-35	16360	112
3	TO_0000190	seed coat color	14.368467	4.736225e-22	1.811606e-20	1150	28
4	TO_0002661	seed maturation	4.699963	4.384546e-16	1.341671e-14	6291	48
5	TO_0000253	seed dormancy	4.632023	5.420875e-14	1.382323e-12	5296	41
6	TO_0000112	disease resistance	2.867433	1.414227e-11	3.091096e-10	15631	70
7	TO_0000043	root morphology trait	2.772096	4.127774e-10	7.894368e-09	13868	62
8	TO_0000495	chlorophyll content	3.268355	6.032035e-10	1.025446e-08	7763	
9	TO_0000259	heat tolerance					22

Try it! https://github.com/Rothamsted/knetgraphs-gene-traits

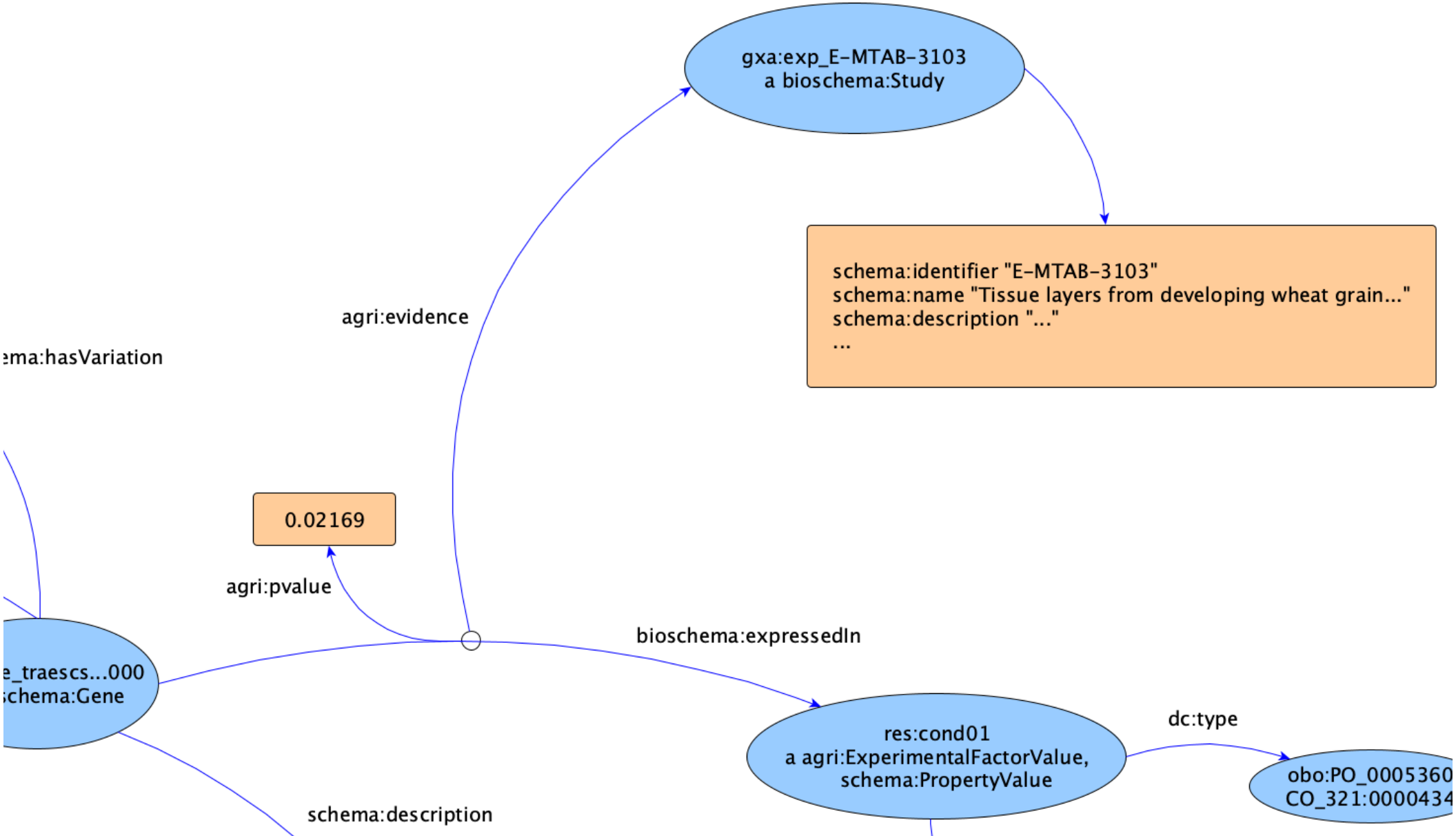
Behind the Scenes

KnetMiner



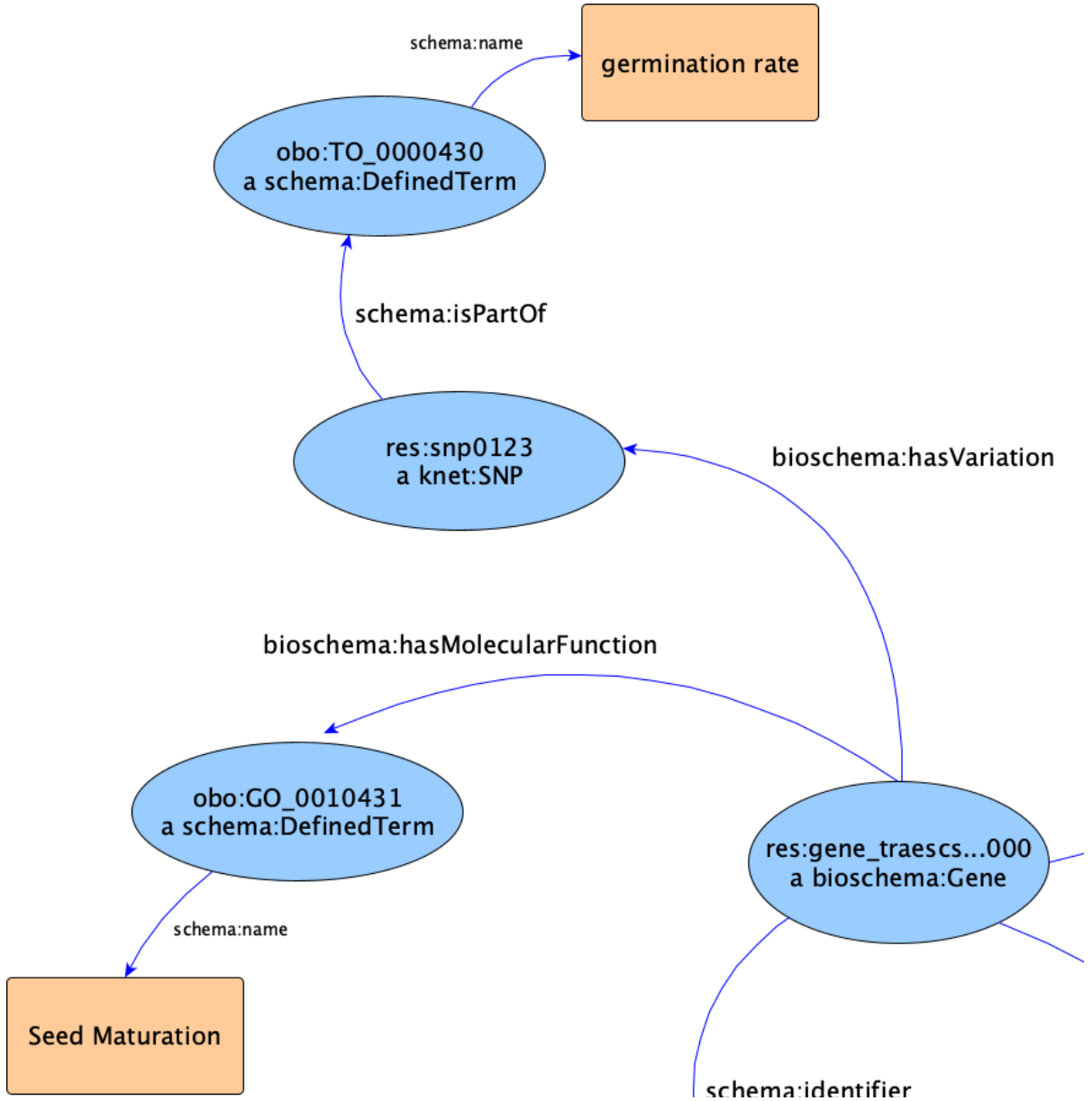
Behind the Scenes

KnetMiner



Behind the Scenes

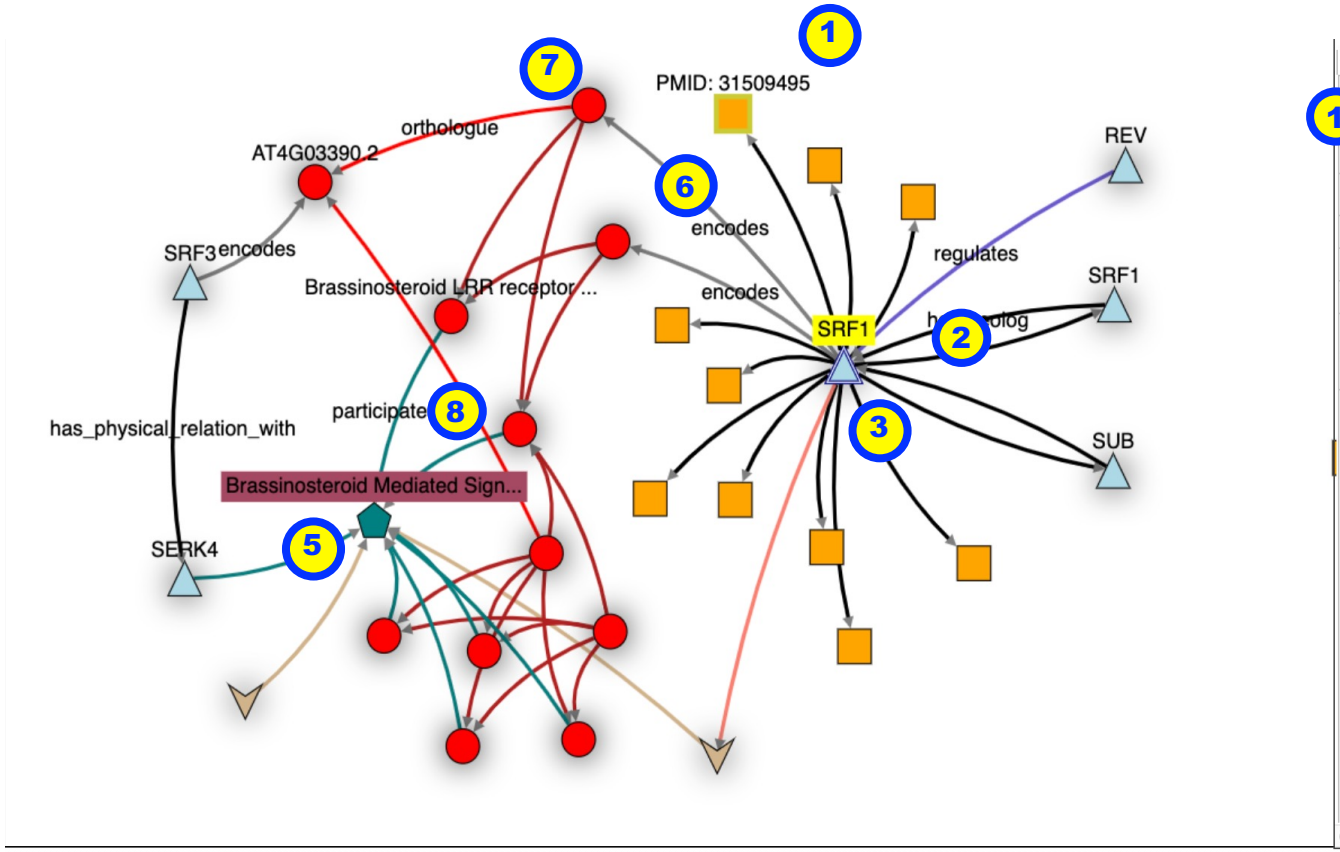
KnetMiner

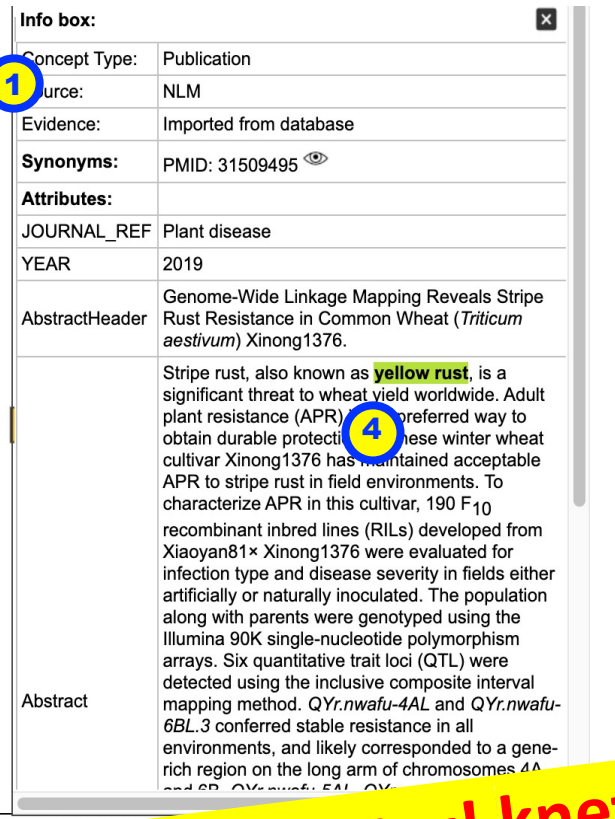


Where it started from…

KnetMiner®

Based on publications, which genes are *related* to the yellow rust disease? In which biological processes are their *encoded* proteins *involved*?





Have a try! knetminer.com

Integrating with EBI GXA

KnetMiner

Based on publications, which genes are *related* to the yellow rust disease?

Knetminer data (include PubMed, ENSEMBL, Text Mining on mentions)

In which conditions *are expressed*?

EBI Gene Expression Atlas

Try it with SPARQL: https://tinyurl.com/2qerv5wn

The Idea

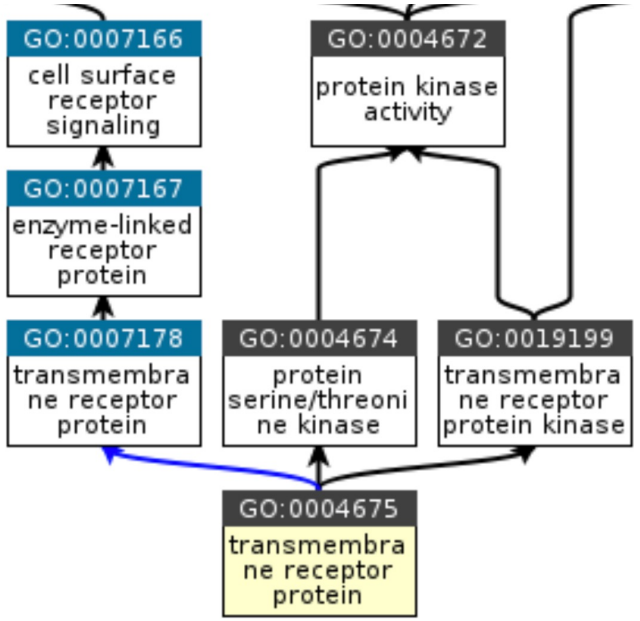
KnetMiner

- Adopt the Knowledge **graph data model**
  - Network of nodes and relationships, each with properties (Ehrlinger, Wöß, 2016)
  - Typical features coming from Knowledge representation, formal logics
  - Made with a variety of techniques, eg, manual curation, imports, machine learning (Gabrilovich, Usunier, 2016)
- Focus on **exploratory research**
  - Yes: find gene candidates, interesting articles
  - No: perform a precise ANOVA analysis of gene significance, compare gene expression across experiments in a precise way
- Coherently, focus on **lightweight schematisation**, integrate with more formal models, ie, OBO ontologies, when useful
  - Simple and complementary approach
  - Suitable for integrating a high number of heterogeneous datasets, web sources, “noisy” data from the world **wild** web
  - Data can be presented directly to the user

Simple, informal

KnetMiner

```
“transmembrane receptor protein serine/threonine kinase activity
(GO:0004675)”
  subClassOf: ”protein serine/threonine kinase activity (GO:0004674)”
    AND “transmembrane receptor protein kinase activity (GO:0019199)”
    AND Restriction:
      onProperty: part of (BFO_0000050)
      someValueFrom: “transmembrane receptor protein
                      serine/threonine kinase signalling pathway
                      (GO:0007178)”
```

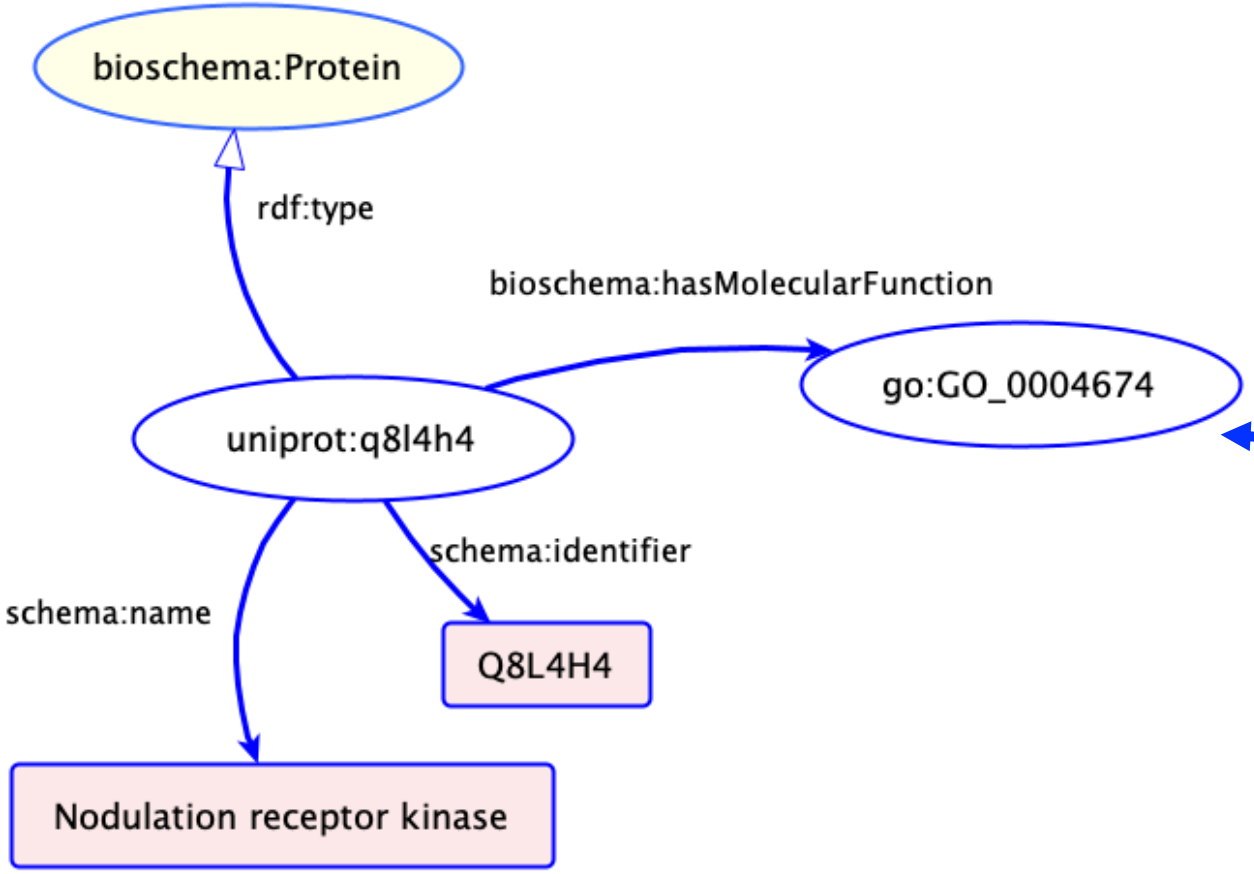


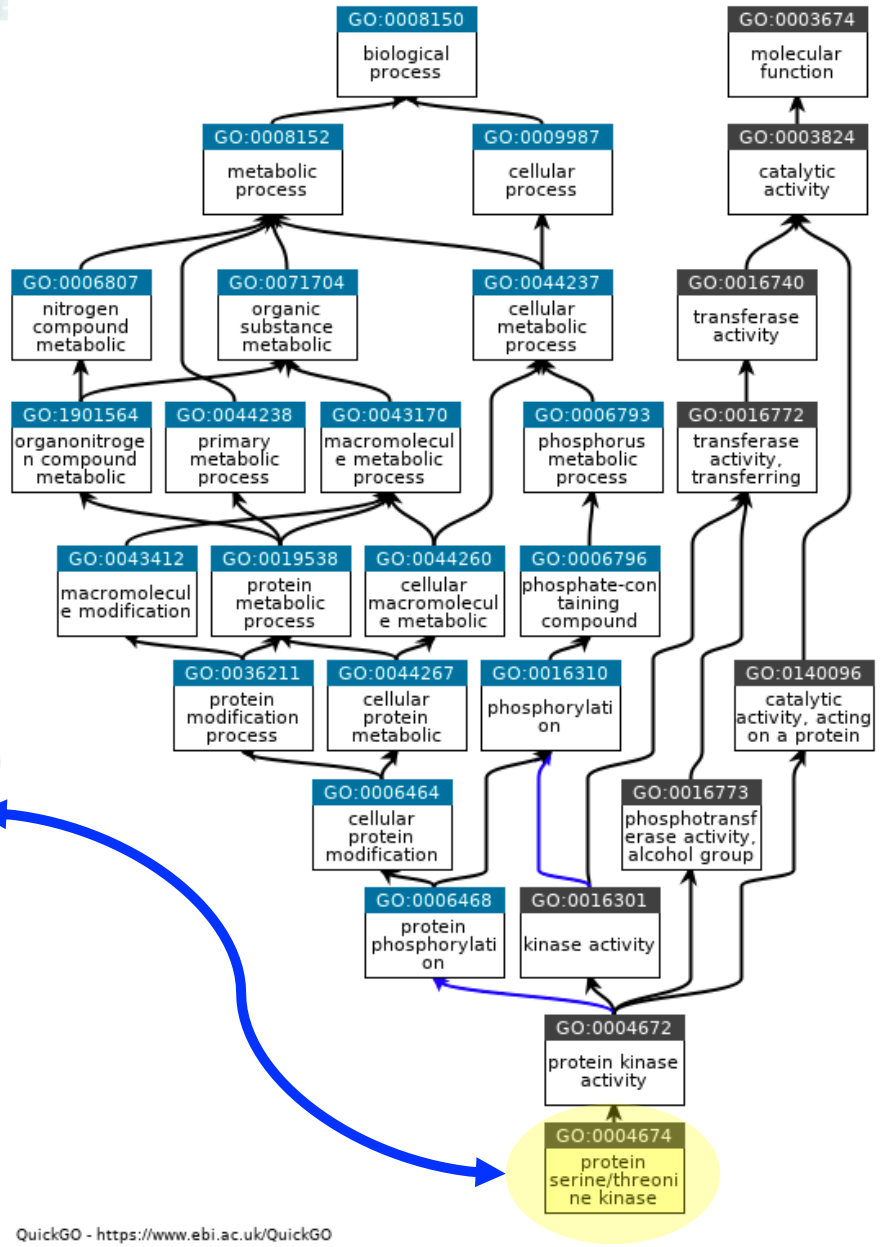
versus:

```
“transmembrane receptor protein serine/threonine kinase activity
(GO:0004675)”
  schema:partOf: “transmembrane receptor protein serine/threonine
                  kinase signalling pathway (GO:0007178)”
  subClassOf: ”protein serine/threonine kinase activity
                (GO:0004674)”,
                “transmembrane receptor protein kinase activity
                 (GO:0019199)”
```

Complementary to ”real” ontologies

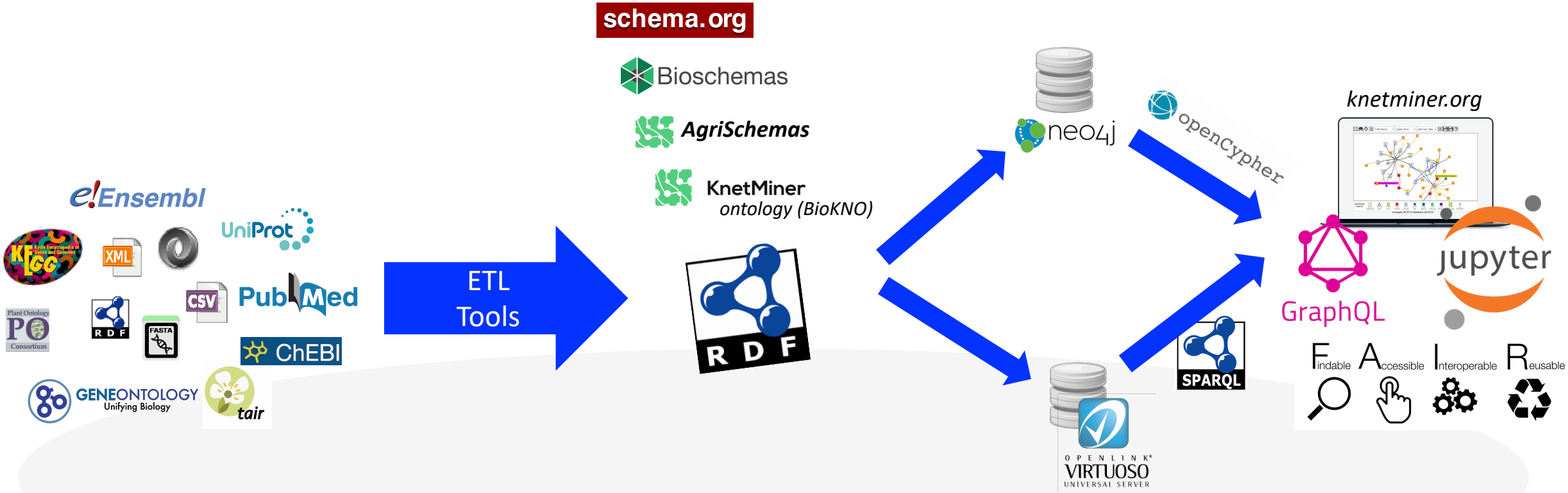
KnetMiner®





And it can be FAIRer

KnetMiner



Based on publications, which genes are *related* to the yellow rust disease? In which biological processes are their *encoded* proteins *involved*?

Agrischemas

KnetMiner

- Kwoledge Graph Patterns and use cases about plant biology, agronomy, food, forestry, weather
- Based on standards, mainly schema.org, bioschemas
- Used to integrate knetminer data and other relevant data (more later)
- Allows for publishing datasets and data endpoints (knetminer.com/data)
- Together with ETL and consuming tools

The story so far

KnetMiner

KnetMiner endpoints (Brandizi et al, 2018)

Agrischemas started (2019)

Use cases (2019)

GXA mapped (2020)

MIAPPE mapping started (Biohackathon 2021)

MIAPPE mapping finalized (Biohachathon 2022)

Real ISA-Tab mapped Text mining/GBCL (2023)

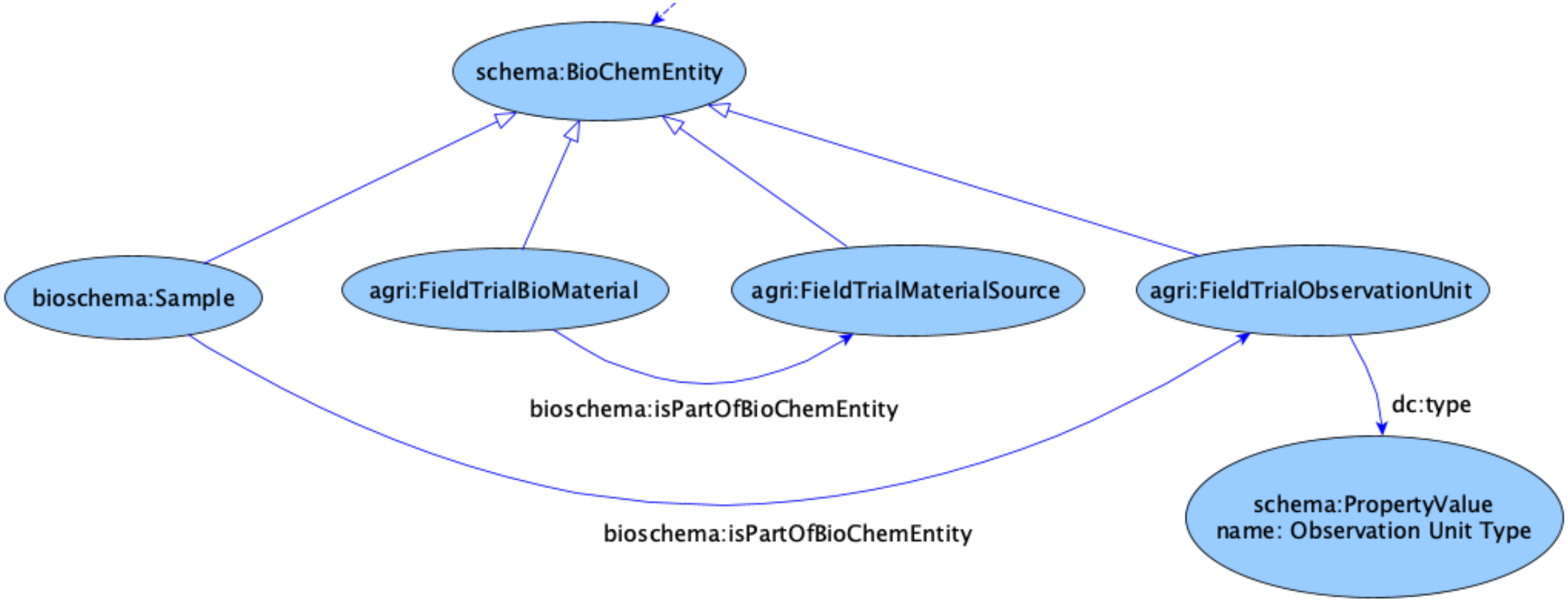
More MIAPPE use cases (German Biohack 2023)

KnetMiner New Generation, focused onNeo4j (2024)

KnetMiner NG, AgriSchemas-Native (2024)

Integrating with Field Trials (MIAPPE/ISA-Tab/BrAPI)

KnetMiner



Based on field trials, which germplasms have least yield loss under drought? Which genes (variants, markers) are associated to them? Which bioprocs, traits, mol functions are most significant?

Use case	Data Types	Data Sources	Status
Molecular Biology	Gene, Protein, Pathway encodes, participates	Via Knetminer: ENSEMBL, UniProt, TILLING, wheat-expression.com, KEGG	Done.
Ontology Annotations	Ontology Term (schema:DefinedTerm) dc:type, schema:additionalType	Via Knetminer: GO, PO, CROP-Onto	Done.
Experiments	Study, agri:StudyFactor, PropertyValue	EBI/GXA, GLTen, MIAPPE/BrAPI sources, ?	GXA Done MIAPPE, mapping done, ISA-Tab converter to be written, real data to be imported GLTen use case drafted
Literature	agri:ScholarlyPublication mentions	Via Knetminer: PubMed	Done
Gene Expression	bioschema:expressedIn, reified statements, agri:evidence, agri:pvalue, agri:baseCondition	EBI/GXA, Via Knetminer: wheat-expression.com	GXA
Host-pathogen interaction	Gene, Phenotype, agri:ScholarlyPublication agri:HostPathogenInteraction agri:evidence	PHI-Base	Use case drafted
Weather	?	?	TO DO
Dataset metadata	Dataset, DataCatalog license, distribution	knetminer.org/data	ongoing

Acknowledgements

- KnetMiner and Rothamsted
  - Mennatullah Shehata, Jupyter project lead
  - Jeremy Parsons, bioinformatics engineer, ETL author
  - Arne De Klerk, product owner
  - Keywan Hassani-Pak, team leader and CEO
  - Lawal Olaotan, UI developer
  - Alumni and past collaborators
  - Chris Baker, IDE department director
  - Chris Rawling, dept director, consultant
  - Brett Drury, GCBL text mining project
- Empats, contractor developers
- ELIXIR
  - Cyril Pommier
  - Sebastian Beier
  - Daniel Arend
- Bioschemas
  - Alasdair Gray
- Germany Hackathon
  - Gabriel Schneider (and FAIRAgro)