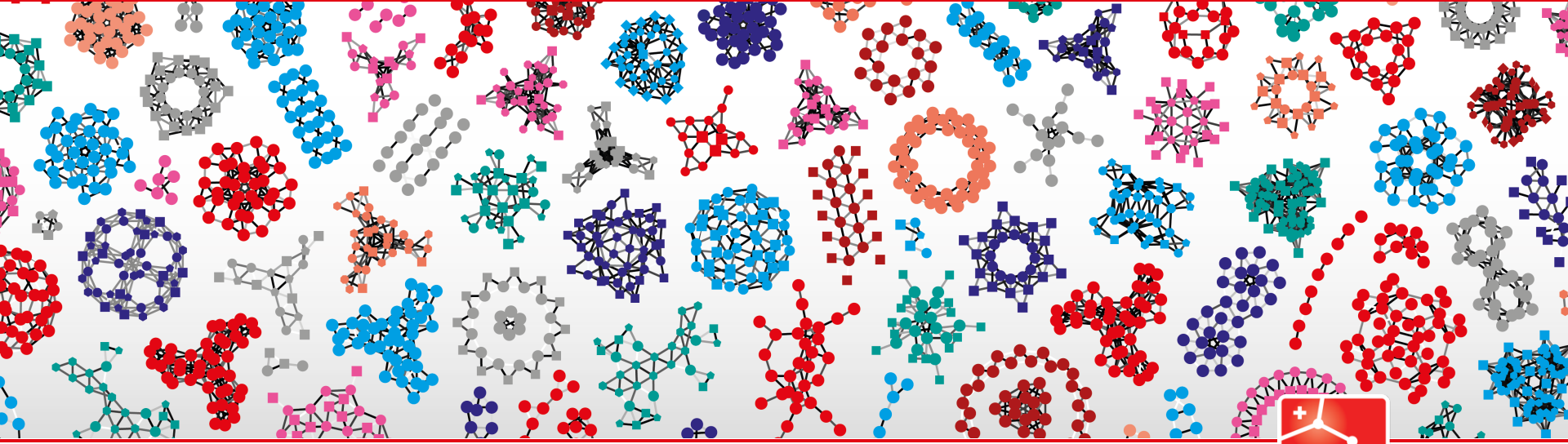


Swiss Institute of Bioinformatics

Querying the Bgee Gene Expression Database

Tarcisio Mendes de Farias, Fréderic Bastian, and Marc Robinson-Rechavi

bgee@sib.swiss / tarcisio.mendes@sib.swiss

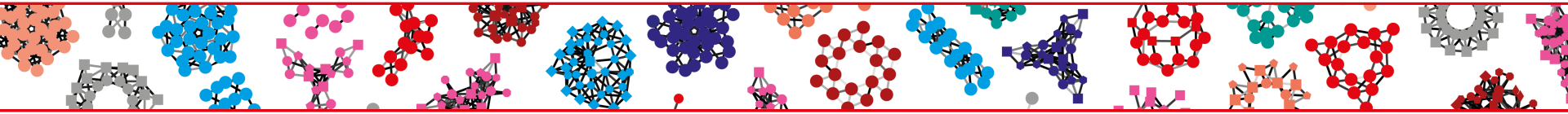
SIB Scientist

elixir SWITZERLAND | UNIL | Université de Lausanne | Bgee

www.bgee.org
@BgeeDB

www.sib.swiss

Overview



Introduction – Bgee Database

- **Reference of healthy gene expression in human**
- **Information of tissue specificity, and functionally-relevant tissues for each gene**
- **Human data from RNA-Seq, Affymetrix, EST data**
- **High-quality curation and consistent re-analyses**
- **Multi-species gene expression database**

Introduction – Bgee Database

H. sapiens human | *M. musculus* mouse | *D. rerio* zebrafish | *D. melanogaster* fruit fly | *C. elegans* nematode | *P. paniscus* bonobo | *P. troglodytes* chimpanzee | *G. gorilla* gorilla | *M. mulatta* macaque

R. norvegicus rat | *B. taurus* cattle | *S. scrofa* pig | *M. domestica* opossum | *O. anatinus* platypus | *G. gallus* chicken | *A. carolinensis* green anole | *X. tropicalis* western clawed frog

D. pseudo obscura

D. simulans

D. virilis

D. yakuba

D. mojavensis

D. ananassae

E. europaeus

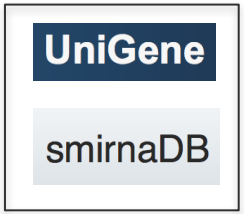
C. porcellus

O. cuniculus

F. catus

C. lupus familiaris

E. caballus

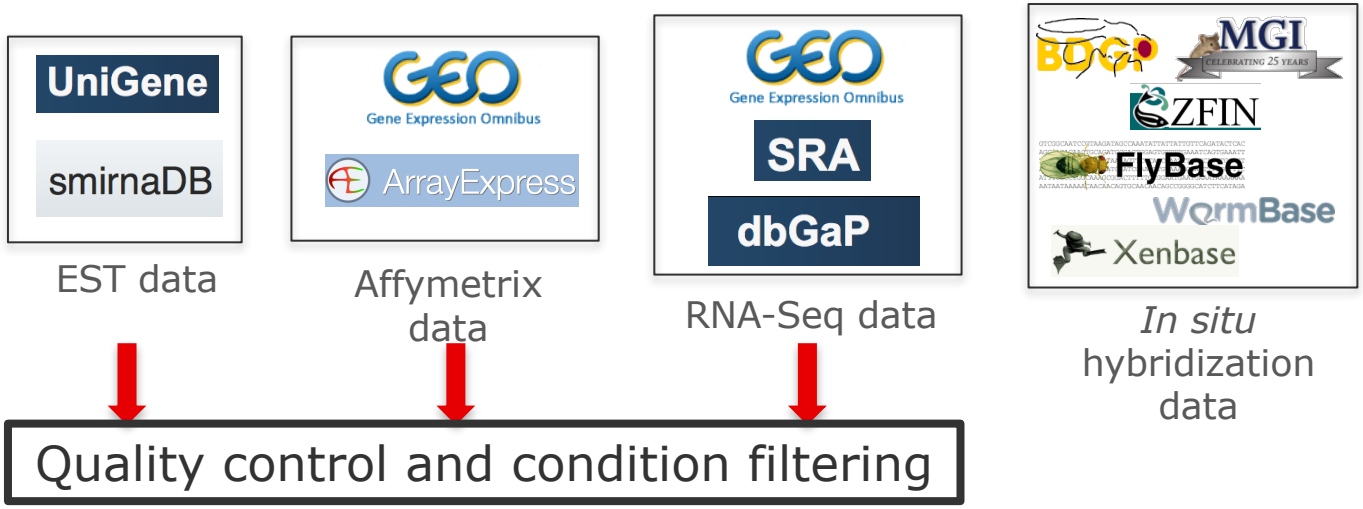


EST data

Affymetrix data

RNA-Seq data

In situ hybridization data



- Select only healthy wild-type expression data

- Specific quality controls for each data type

  E.g.: IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics.
  Rosikiewicz M., Robinson-Rechavi M., 2014, Bioinformatics

UniGene

smirnaDB

EST data

Gene Expression Omnibus

ArrayExpress

Affymetrix data

Gene Expression Omnibus

SRA

dbGaP

RNA-Seq data

MGI

CELEBRATING 25 YEARS

ZFIN

FlyBase

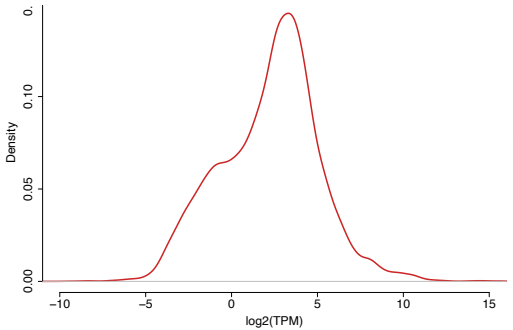
WormBase

Xenbase

In situ hybridization data

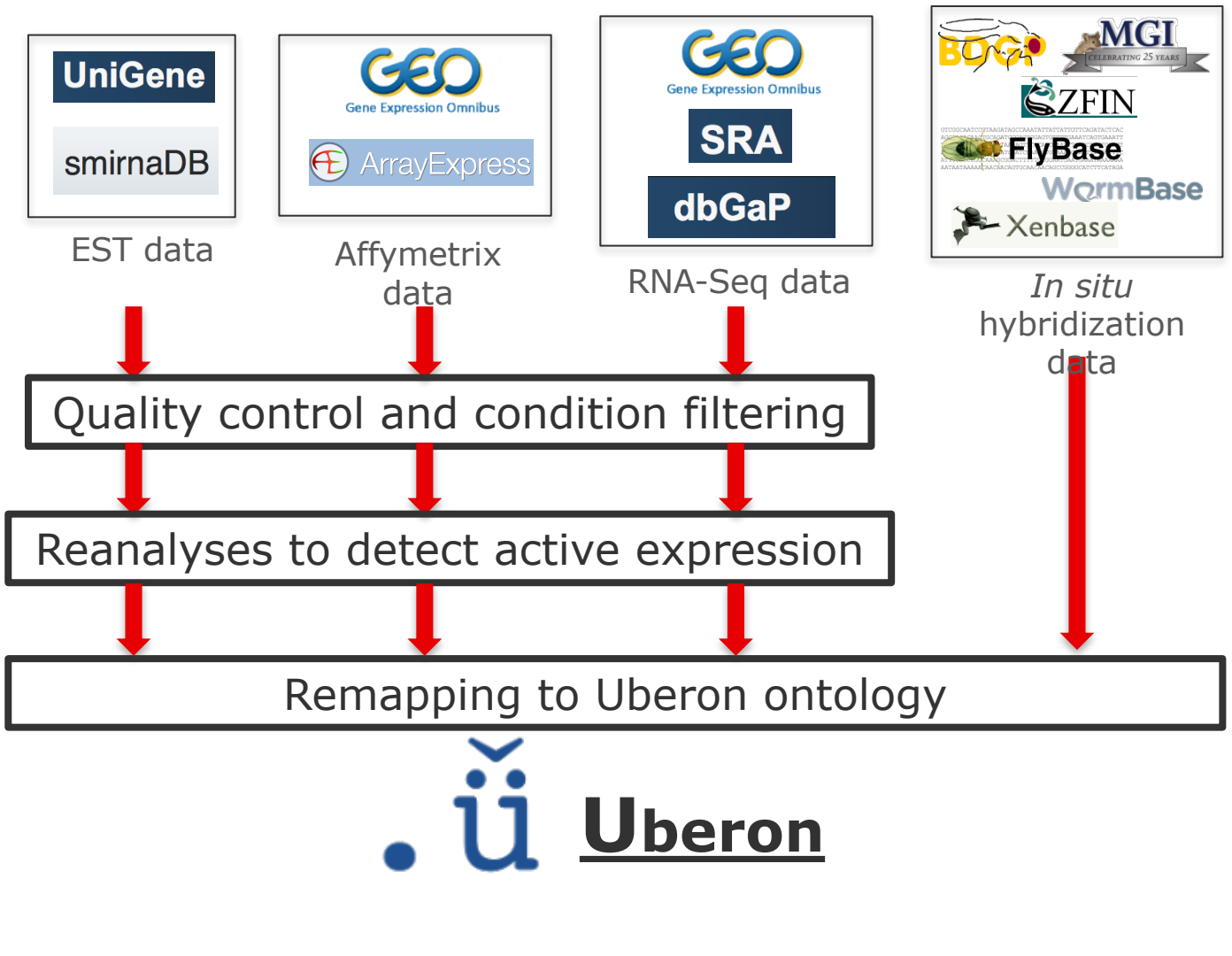
Quality control and condition filtering

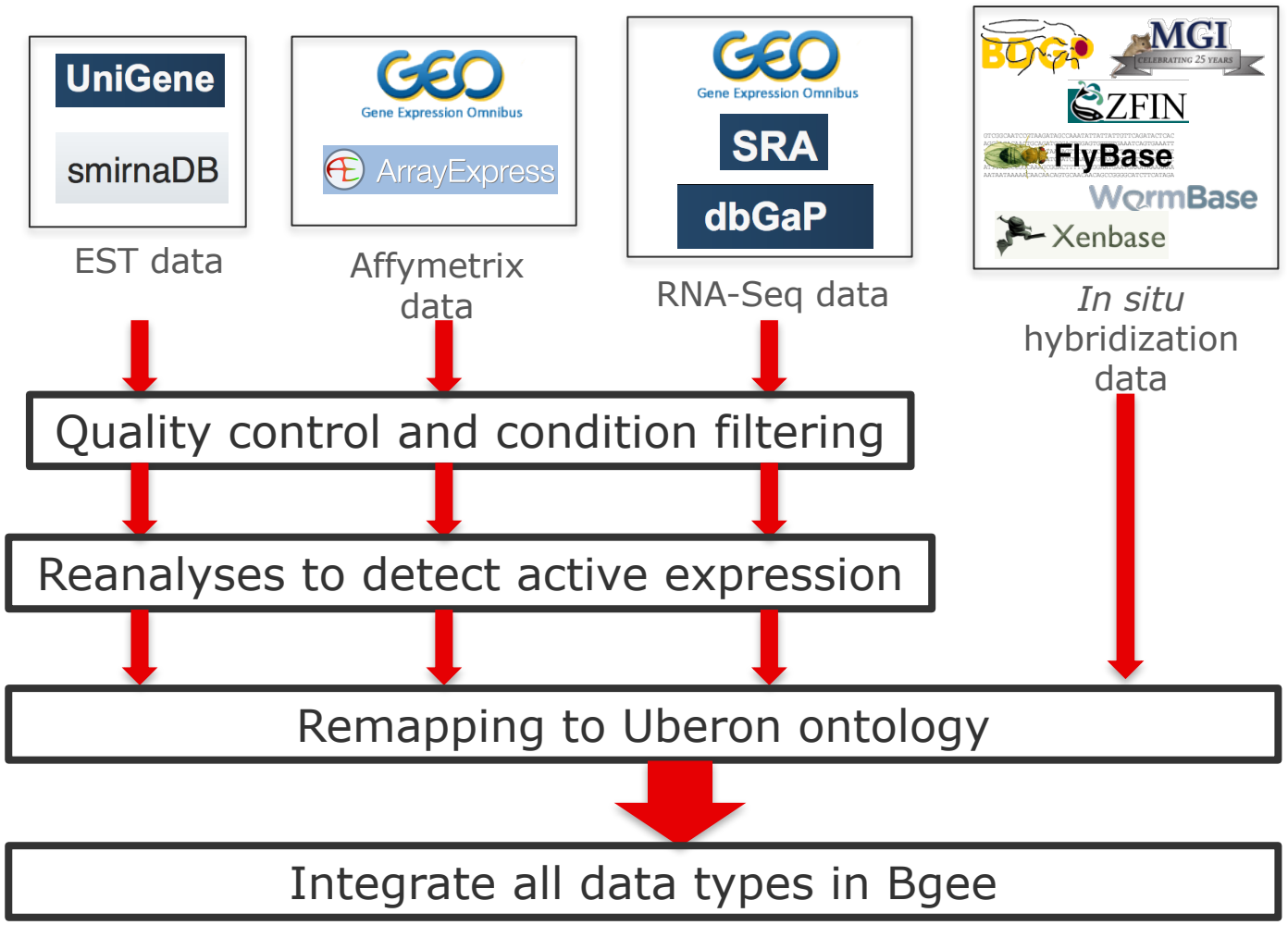
Reanalyses to detect active expression

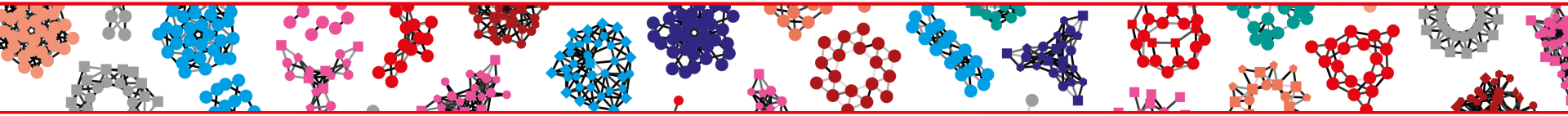


Continuous data (e.g. RNA-Seq)

Discrete data (e.g. In situ)


UniGene
smirnaDB
EST data
Gene Expression Omnibus
ArrayExpress
Affymetrix data
Gene Expression Omnibus
SRA
dbGaP
RNA-Seq data
MGI
CELEBRATING 25 YEARS
ZFIN
FlyBase
WormBase
Xenbase
In situ hybridization data
Quality control and condition filtering
Reanalyses to detect active expression
Remapping to Uberon ontology
Uberon


UniGene
smirnaDB
EST data
Gene Expression Omnibus
ArrayExpress
Affymetrix data
Gene Expression Omnibus
SRA
dbGaP
RNA-Seq data
MGI
CELEBRATING 25 YEARS
ZFIN
FlyBase
WormBase
Xenbase
In situ hybridization data
Quality control and condition filtering
Reanalyses to detect active expression
Remapping to Uberon ontology
Integrate all data types in Bgee



Data schema – GenEx semantic model

Bgee (Gene webpage)

Gene: APOC1 - ENSG00000130208 - *Homo sapiens* (human)

General information

Ensembl ID	ENSG00000130208
Name	APOC1
Description	apolipoprotein C-I [Source:HGNC Symbol;Acc:HGNC:607]
Organism	*Homo sapiens* (human)
Synonym(s)	apo-ci, apoc-i, b2r526, q6ib97

Expression

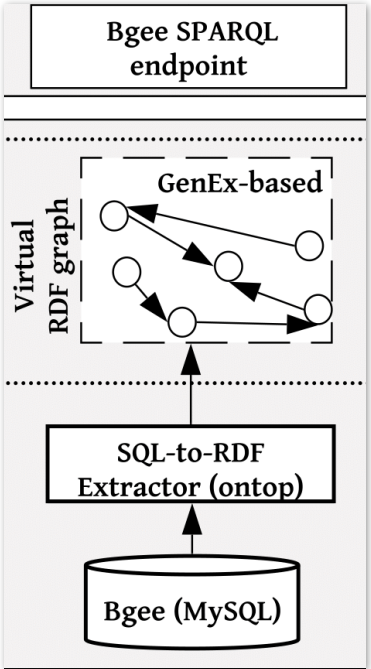
Show 10 entries

Search:

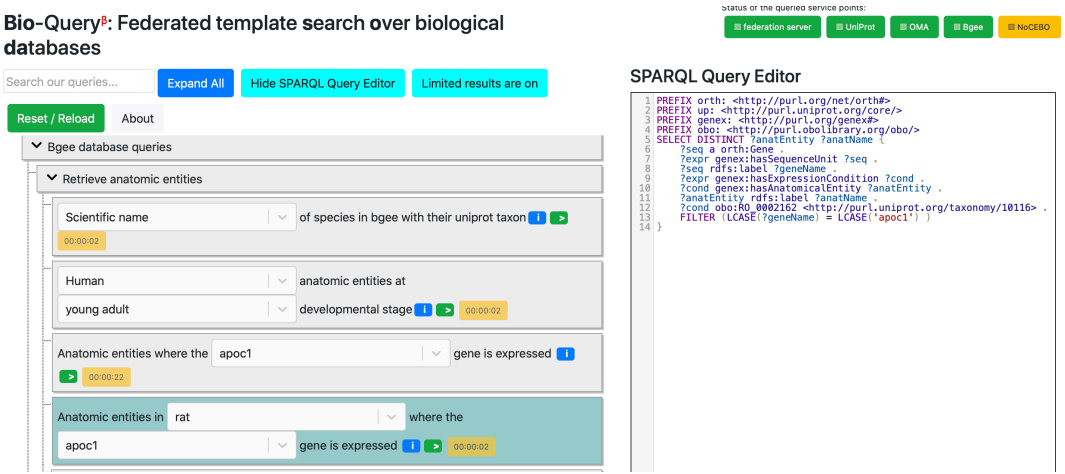
Anat. entity ID	Anatomical entity	Developmental stage(s)	Rank score	Expression score	Sources
UBERON:0002107	liver	[+] 13 stages			
UBERON:0001114	right lobe of liver	[+] 5 stages			
UBERON:0035825	left adrenal gland cortex	[+] 4 stages			
UBERON:0001233	right adrenal gland	[+] 4 stages			
UBERON:0035827	right adrenal gland cortex	[+] 2 stages			
UBERON:0001234	left adrenal gland	[+] 5 stages			
UBERON:0001235	adrenal cortex	[+] 1 stage			
UBERON:0002038	substantia nigra	[+] 9 stages			

Soon accessible through SPARQL

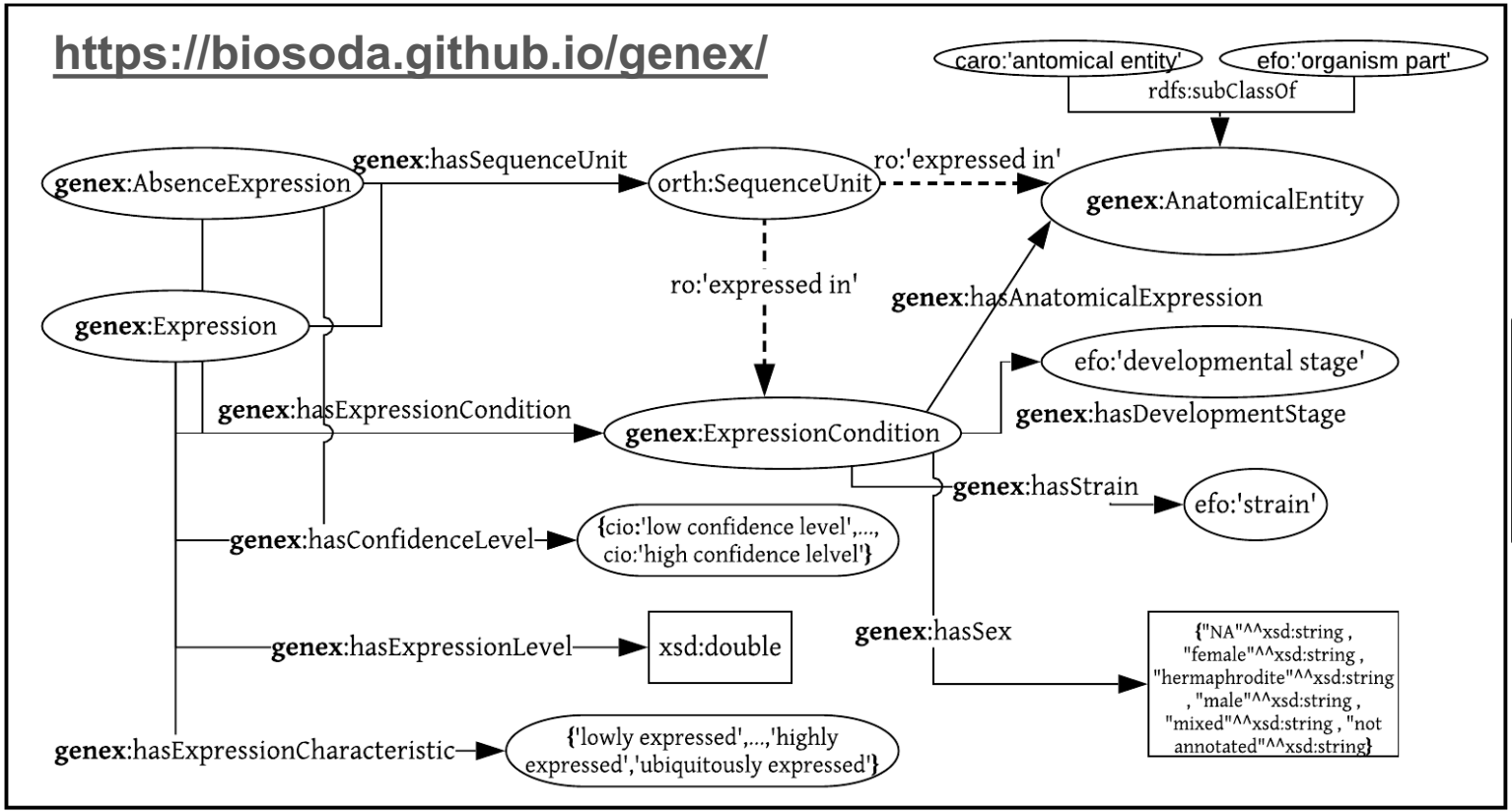
- Bgee SPARQL endpoint: http://biosoda.expasy.org/rdf4j-server/repositories/bgeelight



- A user-friendly webpage to query the Virtual RDF graph of Bgee http://biosoda.expasy.org



A portion of the GenEx semantic model



*Note:
genex:isExpressedIn
⊑
ro:'expressed in'

Querying organs where a gene is expressed for a given species

#**Anatomic entities** in **rat** where the **apoc1** gene is expressed

```
PREFIX orth: <http://purl.org/net/orth#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX genex: <http://purl.org/genex#>
PREFIX obo: <http://purl.obolibrary.org/obo/>

SELECT DISTINCT ?anatEntity ?anatName {
  ?seq a orth:Gene .
  ?seq rdfs:label ?geneName .     # ?seq lscr:xrefEnsemblGene ?geneEns
  ?seq genex:isExpressedIn ?cond.
  ?cond genex:hasAnatomicalEntity ?anatEntity .
  ?anatEntity rdfs:label ?anatName .
  ?cond obo:RO_0002162 <http://purl.uniprot.org/taxonomy/10116> . #rat
  FILTER (LCASE(?geneName) = LCASE('Apoc1') )  }
```

Querying organs where a gene is expressed for a given species

#Anatomic entities in rat where the apoc1 gene is expressed

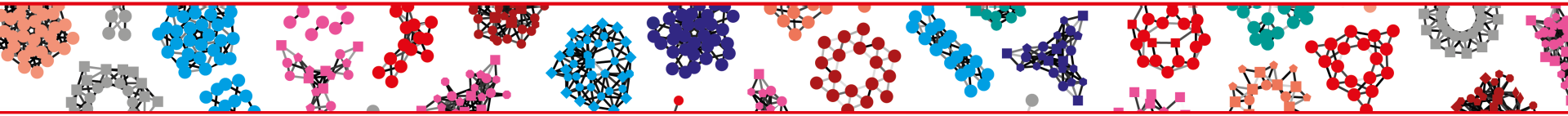
```
PREFIX orth: <http://purl.org/net/orth#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX genex: <http://purl.org/genex#>
PREFIX obo: <http://purl.obolibrary.org/obo/>

SELECT DISTINCT ?anatEntity ?anatName {
   ?seq a orth:Gene .
   ?seq rdfs:label ?geneName .
   ?seq genex:isExpressedIn ?cond.
   ?cond genex:hasAnatomicalEntity ?anatEntity .
   ?anatEntity rdfs:label ?anatName .
   ?cond obo:RO_0002162 <http://purl.uniprot.org/taxonomy/10116> . #rat
   FILTER (LCASE(?geneName) = LCASE('Apoc1') )  }
```

Querying organs where a gene is expressed for a given species and developmental stage

```
PREFIX orth: <http://purl.org/net/orth#>
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX genex: <http://purl.org/genex#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
```

#Anatomical entities at the Mouse's adult stage where the apoc1 gene is expressed.

```
SELECT DISTINCT ?anatEntity ?anatName {
   ?seq a orth:Gene .
   ?seq rdfs:label ?geneName .
   ?seq  genex:isExpressedIn ?cond .
   ?cond genex:hasAnatomicalEntity ?anatEntity .
   ?anatEntity rdfs:label ?anatName .
   ?cond genex:hasDevelopmentalStage ?stage .
   ?stage rdfs:label ?stageName .
   ?cond obo:RO_0002162 ?taxon . #in taxon property .
   ?taxon up:commonName ?commonName .
FILTER ( LCASE(STR(?commonName)) = LCASE("Mouse") ) .
FILTER (CONTAINS(?stageName,'adult') && LCASE(?geneName) = LCASE('apoc1') ) }
```



Conclusion

Conclusion

- **In this tutorial, we learned how to query gene expression patterns from the Bgee database with SPARQL**
- **We described the main part of the GenEx semantic model used to represent the core data provided by Bgee.**
  - **GenEx documentation: https://biosoda.github.io/genex**
- **SPARQL endpoint**

http://biosoda.expasy.org/rdf4j-server/repositories/bgeelight

- **Soon**: http://sparql.bgee.org/sparql

VoIDext* vocabulary to describe interlinks among distributed and indepenedt datasets on the Web

Documentation: https://biosoda.github.io/voidext

VoIDext Vocabulary Specification Draft

Extended Vocabulary of Interlinked Datasets (VoIDext)

Release 2019-06-30

This version:
http://purl.org/query/voidext

Latest version:
http://purl.org/query/voidext

Previous version (DEPRECATED):
~~Deprecated 2019-03-30 VoIDext release and documentation~~

Author and editor:
Tarcisio Mendes de Farias, SIB Swiss Institute of Bioinformatics

Contributor:
Christophe Dessimoz, University College London
Kurt Stockinger, ZHAW Zurich University of Applied Sciences
To be defined

Download serialization:
Format RDF/XML | Format N Triples | Format TTL

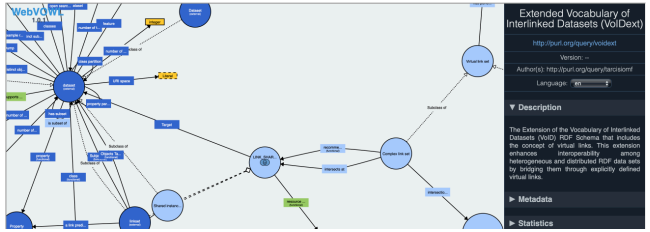
License:
License https://creativecommons.org/licenses/by/3.0/

Visualization:
Visualize with WebVowl

Compatible with:
http://rdfs.org/ns/void

GitHub

https://github.com/biosoda/voidext

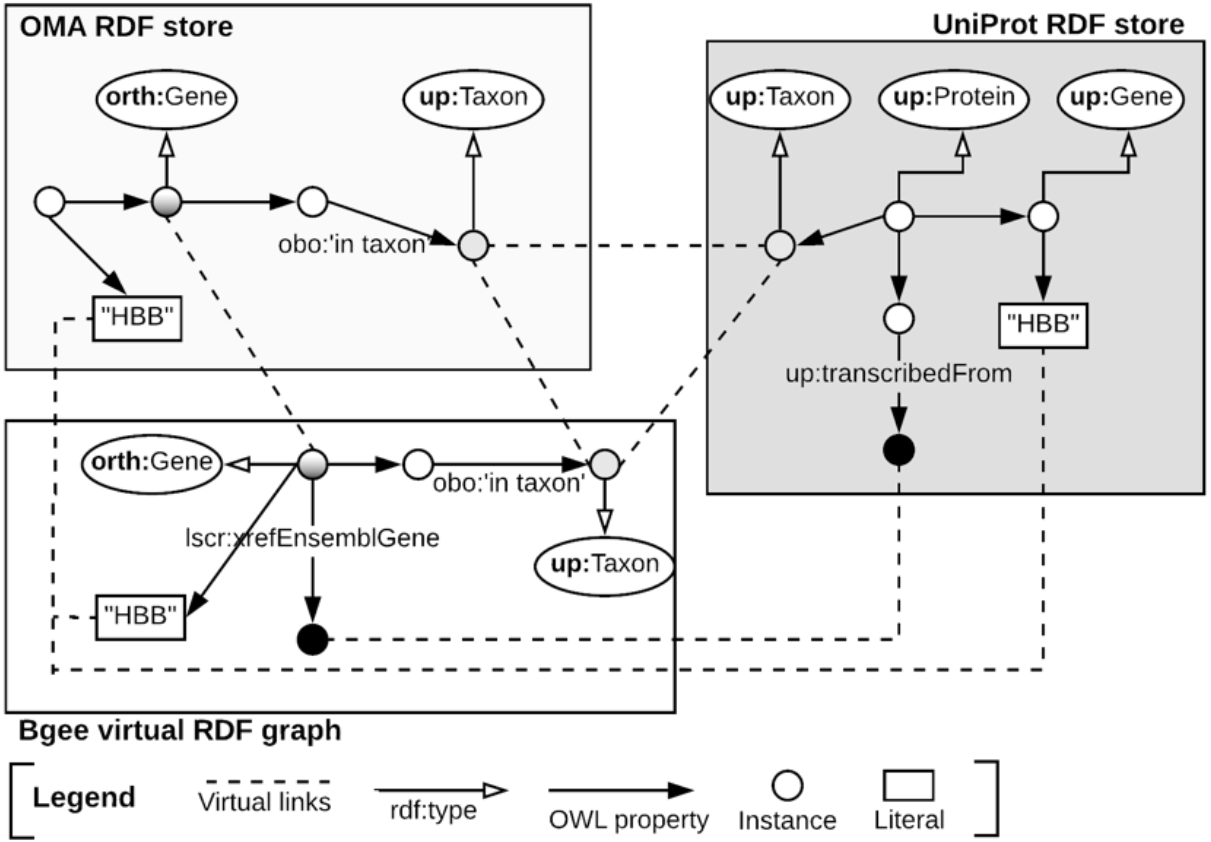


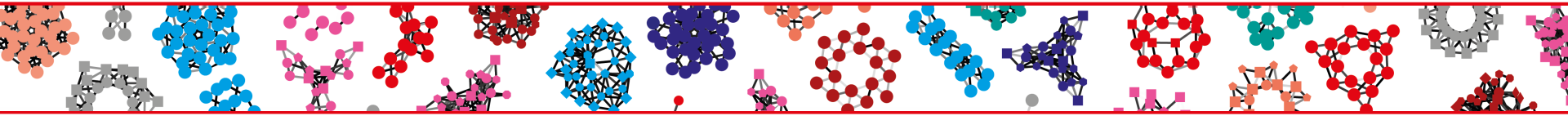
VoIDext RDF schema vocabulary

http://purl.org/query/voidext

*Mendes de Farias T., Stockinger K., Dessimoz C. (2019) VoIDext: Vocabulary and Patterns for Enhancing Interoperable Datasets with Virtual Links. In: On the Move to Meaningful Internet Systems: OTM 2019 Conferences Lecture Notes in Computer Science, vol 11877. Springer, Cham https://doi.org/10.1007/978-3-030-33246-4_38

VoIDext Metadata of interlinks among Bgee, OMA and UniProt





A Conjunctive Federated Query: Bgee and UniProt databases

Bgee-UniProt Federated Query Example – Part 1

#Genes expressed in the human's brain during the infant stage and their UniProt disease descriptions?

```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX genex: <http://purl.org/genex#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX orth: <http://purl.org/net/orth#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX sio: <http://semanticscience.org/resource/>
SELECT DISTINCT ?geneEns ?uniprot ?description {
  SERVICE <http://biosoda.expasy.org/rdf4j-server/repositories/bgeelight> {
        SELECT ?geneEns {
                ?geneB genex:isExpressedIn ?cond ;
                        lscr:xrefEnsemblGene ?geneEns .
                ?cond    genex:hasDevelopmentalStage ?st .
                ?cond    genex:hasAnatomicalEntity ?anat .
                ?st        rdfs:label 'infant stage'^^xsd:string .
                ?anat     rdfs:label 'brain'^^xsd:string .
                ?geneB orth:organism ?o .
                ?o obo:RO_0002162 ?taxon2 .
                 obo:RO_0002162  rdfs:label “in taxon”.
                ?taxon2 up:commonName 'human' .}        LIMIT 10        }
```

Bgee-UniProt Federated Query Example – Part 2

```
SERVICE <http://sparql.uniprot.org/sparql> {
        ?uniprot       rdfs:seeAlso ?gene_xref.
        ?gene_xref  up:transcribedFrom ?geneEns .
        ?uniprot       up:annotation ?annotation .
        ?annotation a up:Disease_Annotation .
        ?annotation rdfs:comment ?description . }}
```

SIB Swiss Institute of Bioinformatics

Bgee team members:

Frederic Bastian, Sara Fonseca Costa, **Tarcisio Mendes De Farias**, Sébastien Moretti, Anne Niknejad, Valentine Rech de Laval, Marc Robinson-Rechavi, Julien Wollbrett