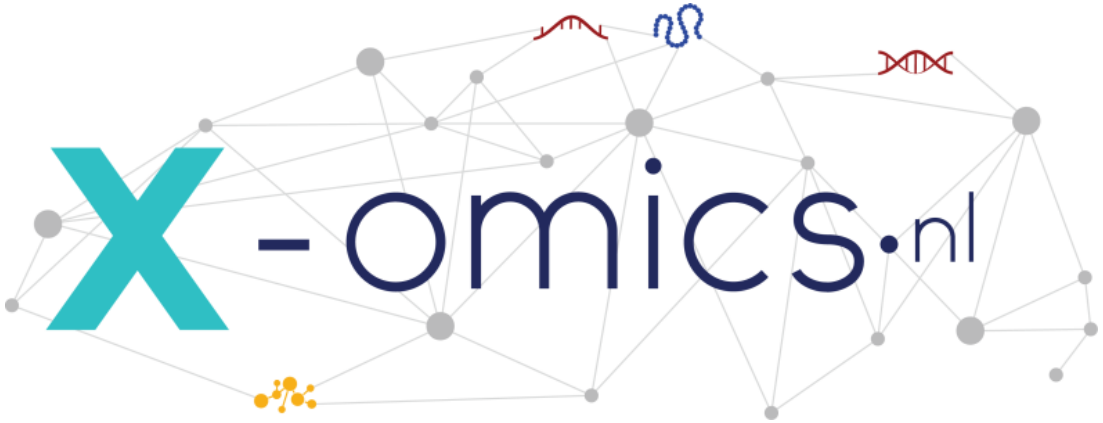
Workshop Series

Strategies to overcome your challenges in multi-omics data integration

Monday 22nd June 2020

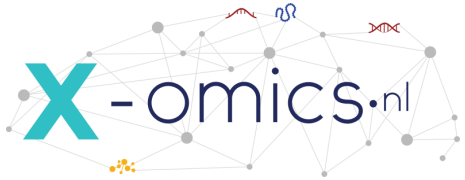
Linked data in practice:

An RDF-based approach with SPARQLing Genomics



Data standards and multi-omics data integration

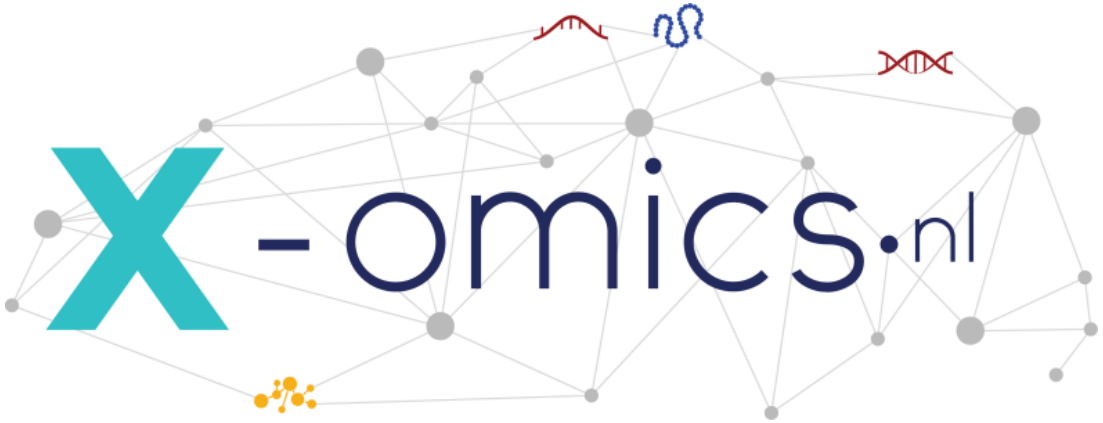
12.10	General intro
12.30	RDF and SPARQL intro and demo
13.10	Interactive Quiz
13.20	Ontologies and FAIR data
13.40	Q&A + Open Discussion


X-omics.nl

Monday 22nd June 2020

Linked data in practice

General Introduction



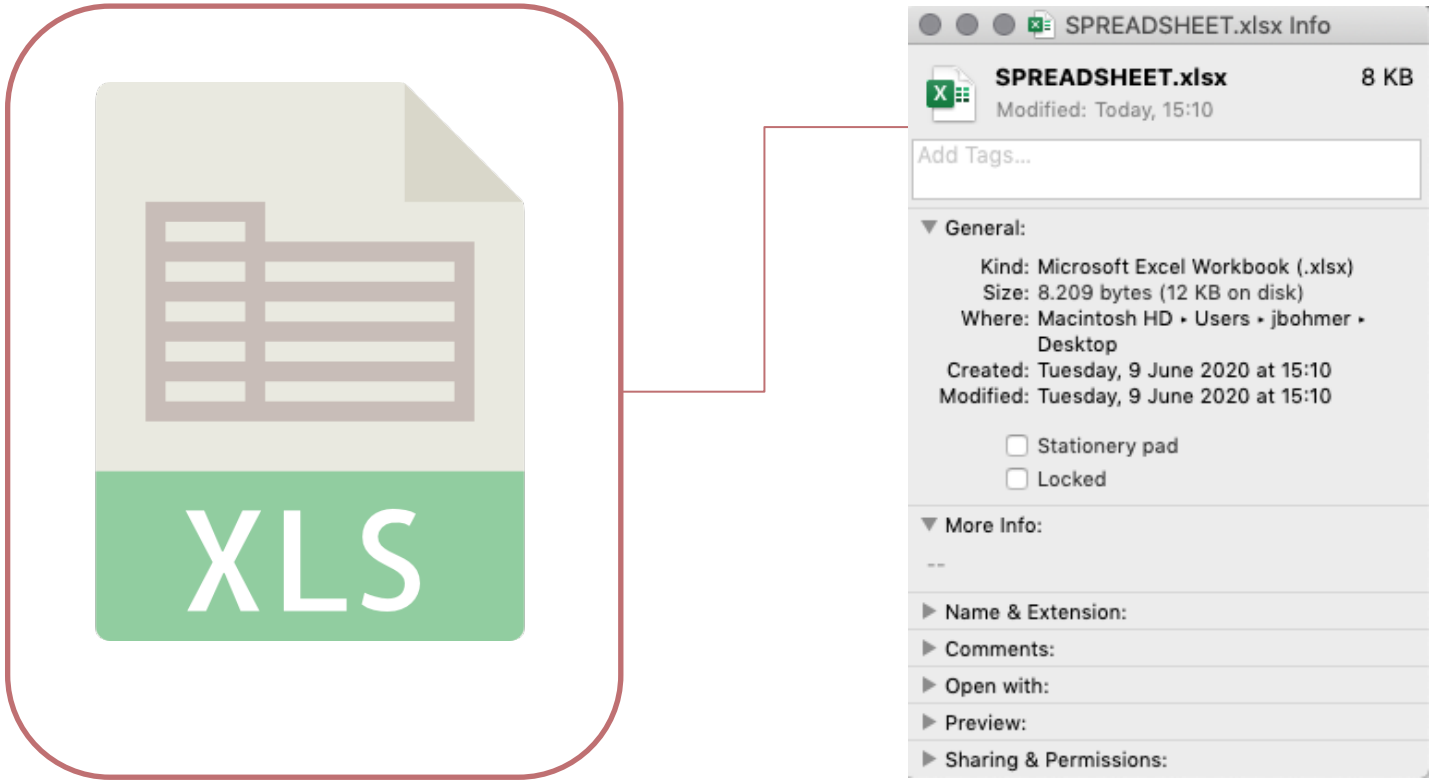
General introduction

1. Reasoning
2. Data and metadata
3. RDF and SPARQL
4. Current omics data landscape
5. SPARQLing Genomics project

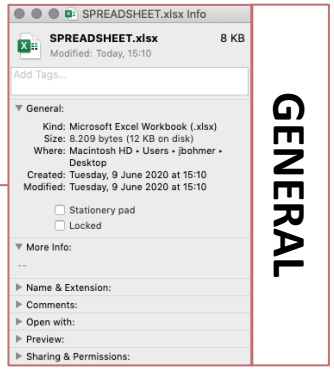
REASON

- Interoperability across data, platforms, research lines, disciplines
- Harmonizing data and metadata
- FAIR data requirements
- Reusable and trustworthy data and information

DATA and METADATA



DATA and METADATA



Collaborator
License
Keywords
Access Rights
…

EXTRA

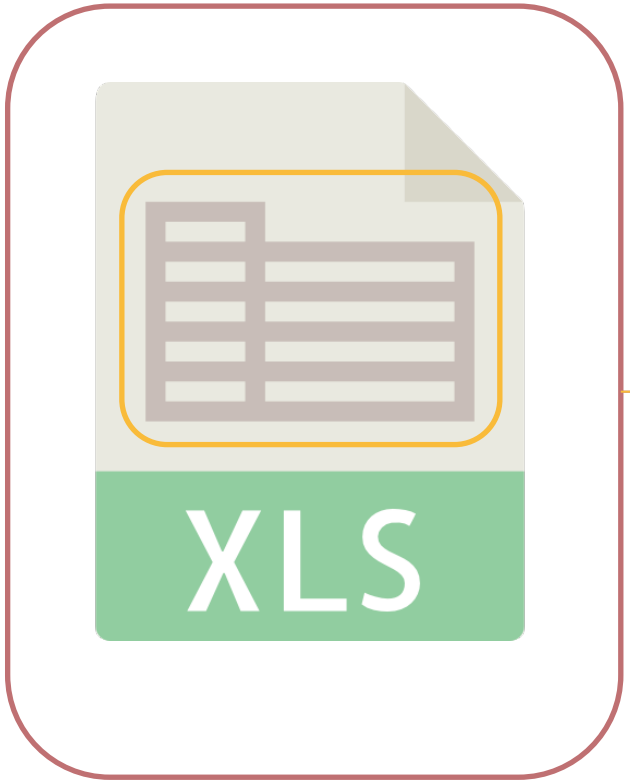
https://www.flaticon.com/authors/smashicons

DATA and METADATA

	A	B	C	D	E	F
1	Sample	Population	Has Omni Genotypes	Has Axiom Genotypes	Has Affy 6.0 Genotypes	Has Exome/LOF Genotypes
2	HG00096	GBR	1			1
3	HG00097	GBR	1			1
4	HG00098	GBR	1			1
5	HG00099	GBR	1			1

https://www.flaticon.com/authors/smashicons
ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/20130606_sample_info.xlsx

DATA and METADATA



GENERAL and EXTRA **METADATA**

Author
Modification Date
Access Rights
Keywords
…

DATA CONTENT as **SPECIFIC METADATA**

Sample-ID
Population-ID
Gender
…

https://www.flaticon.com/authors/smashicons

METADATA-Standard

Standardised **METADATA:**

Creator:	…
Creation Date:	…
Gender:	…
Sample-ID:	…
Data-Type:	…
Access Condition:	…
… :	…

METADATA-Standard

Standardised **METADATA:**

Creator:	*Free Text*
Creation Date:	*DD.MM.YYY*
Gender:	*CONTROLLED*
Sample-ID:	*SYNTAX*
Data-Type:	*CONTROLLED*
Access Condition:	*CONTROLLED*

Semantic ambiguity

How many ways can you say ‘female’?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

EMBL-EBI

https://www.slideshare.net/ConnectedDataLondon/ontology-services-for-the-biomedical-sciences; Courtesy of N. Silvester, European Nucleotide Archive, EMBL- EBI

X-omics.nl

Semantic ambiguity and ontology terms

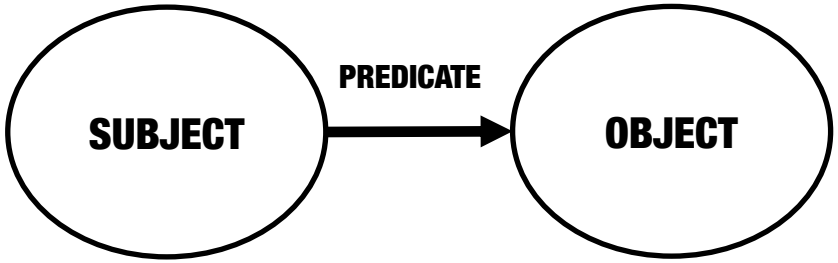
Meta-data element	Preferred ontology term for meta-data	Value types	Issue Number	Issue (open / closed)
Individual ID	NCIT:C164337	ID [string]	#3	closed
Gender	SIO:010029	Male SIO:010048 Female SIO:010052 Unknown / Undetermined	#4	closed
Genotypic sex	PATO:0020000	UNKNOWN_KARYOTYPE, XX, XY, XO, XXY, XXX, XXYY, XXXY, XXXX, XYY, OTHER_KARYOTYPE	#69	closed

Linked Data with RDF and SPARQL

RDF

Resource Description Framework

Expresses natural facts in triplets



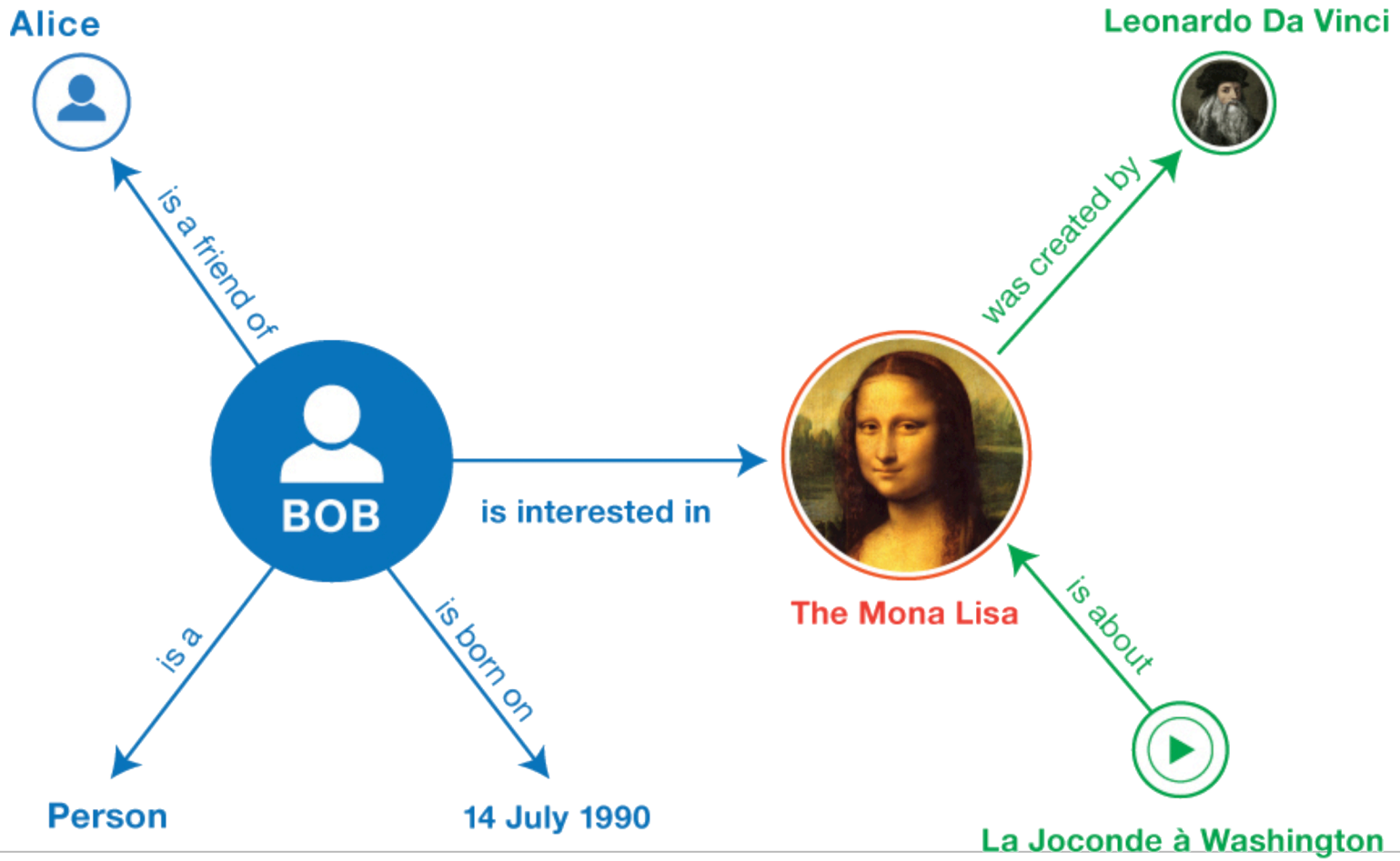
SPARQL

Protocol to query RDF content

- `SELECT` • raw values
- `CONSTRUCT` • Information
- `ASK` • True/False

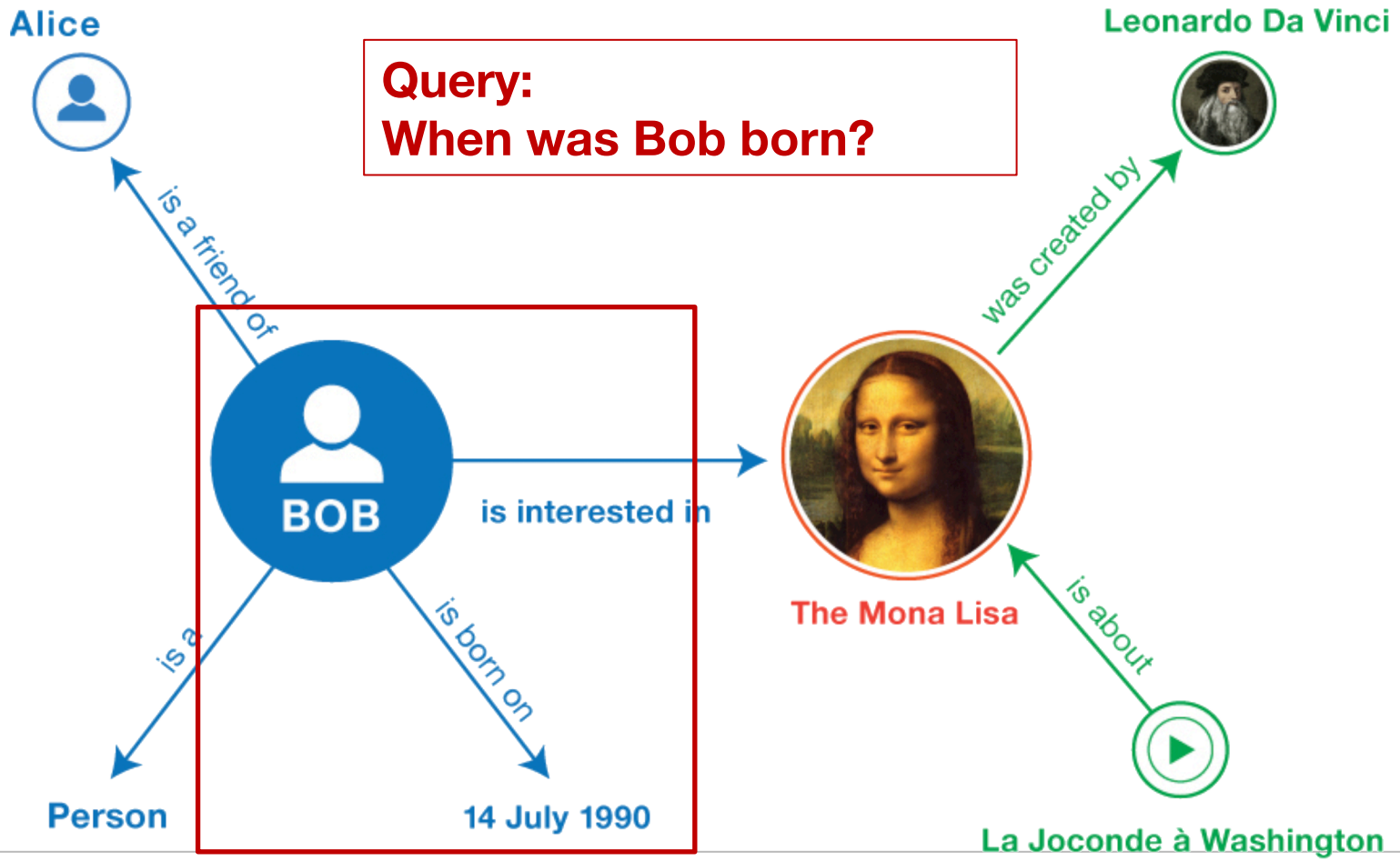
https://en.wikipedia.org/wiki/Resource_Description_Framework
https://en.wikipedia.org/wiki/SPARQL

RDF triplets



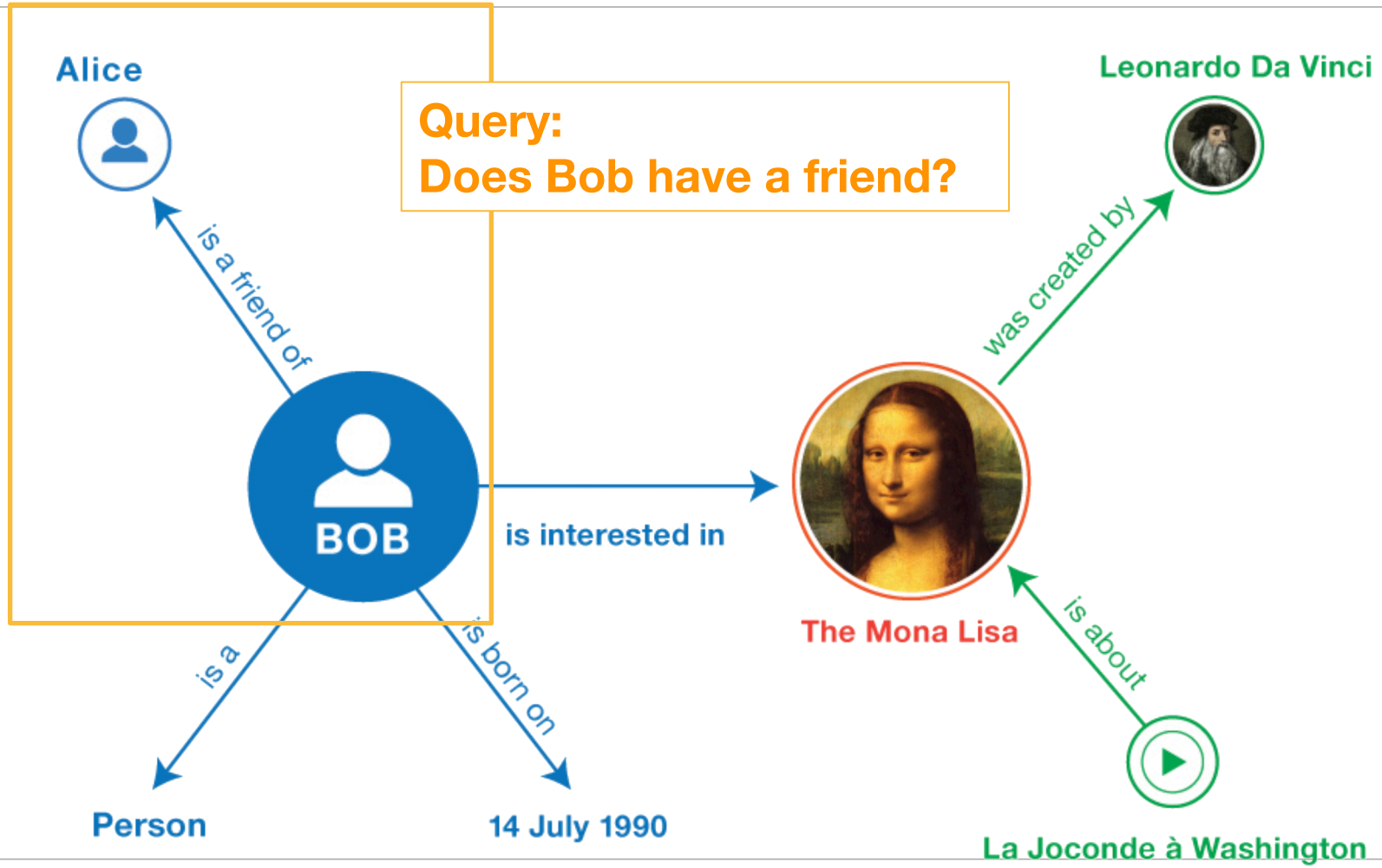
https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-primer/index.html

RDF triplets



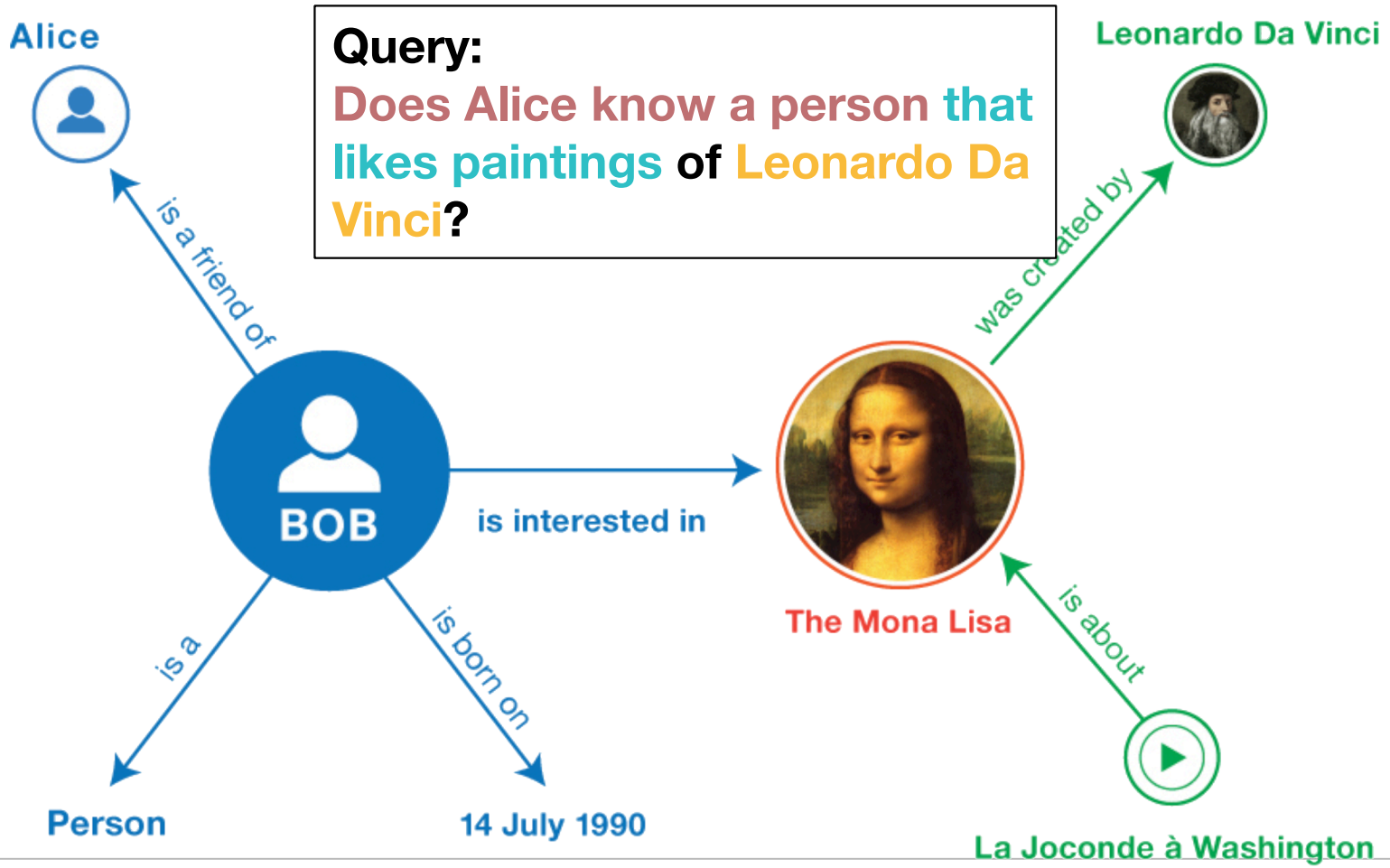
https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-primer/index.html

RDF triplets



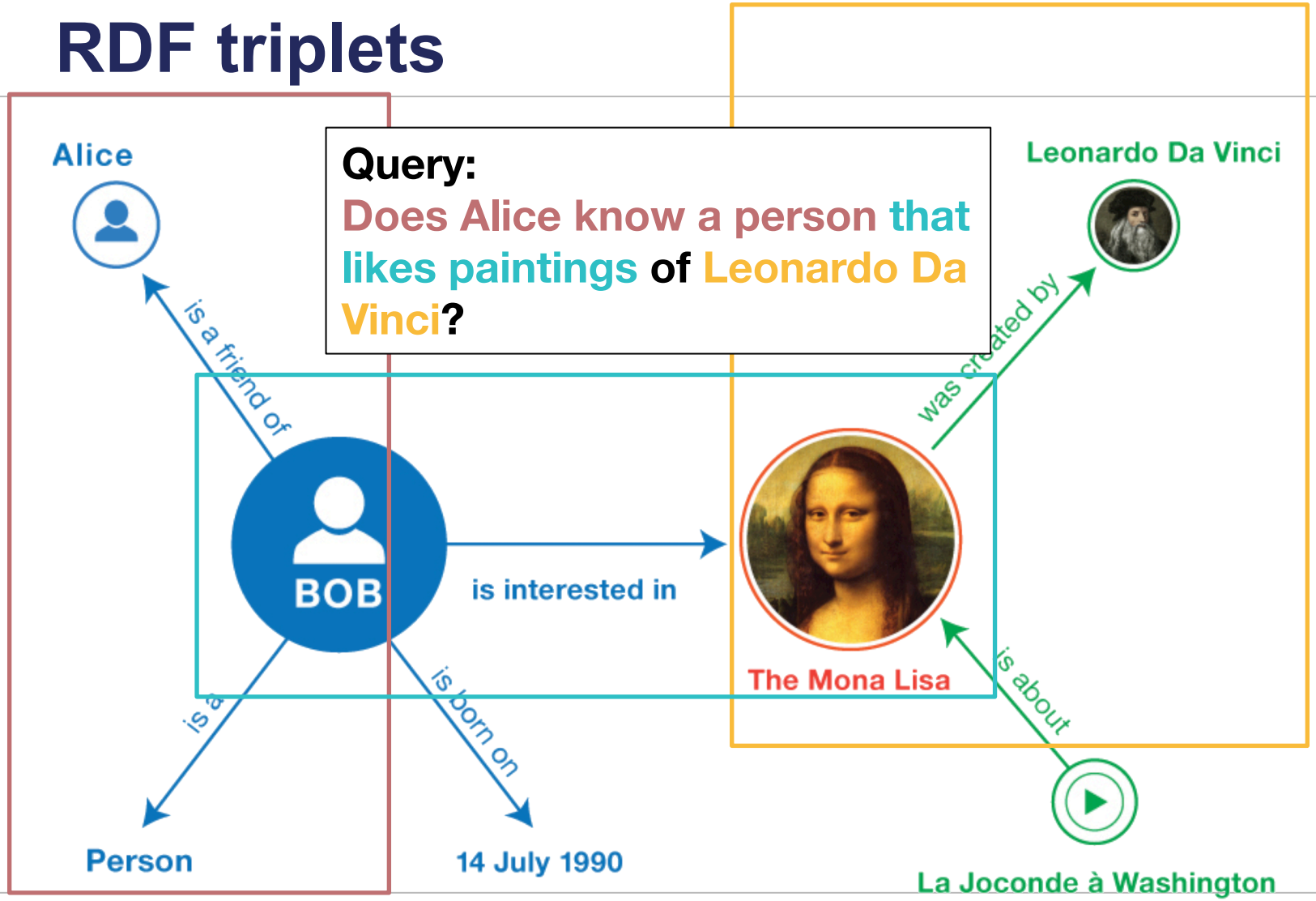
https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-primer/index.html

RDF triplets



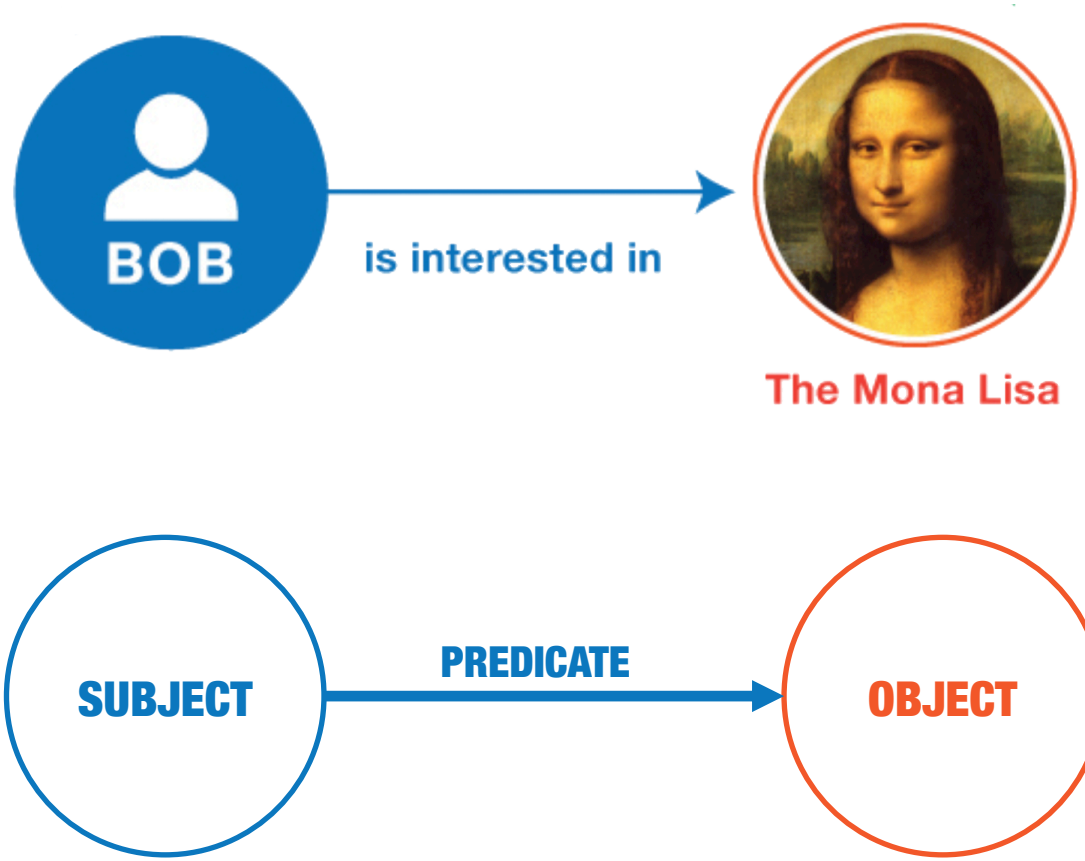
https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-primer/index.html

RDF triplets



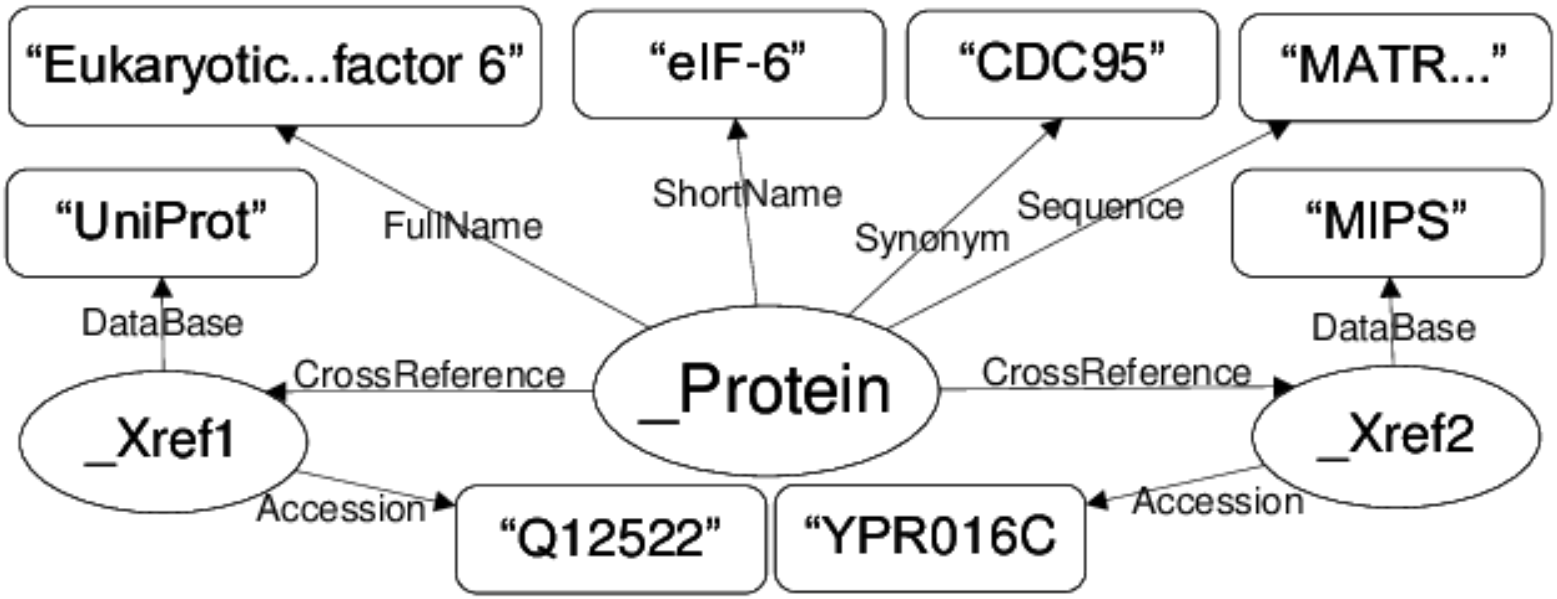
https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-primer/index.html

RDF triplets

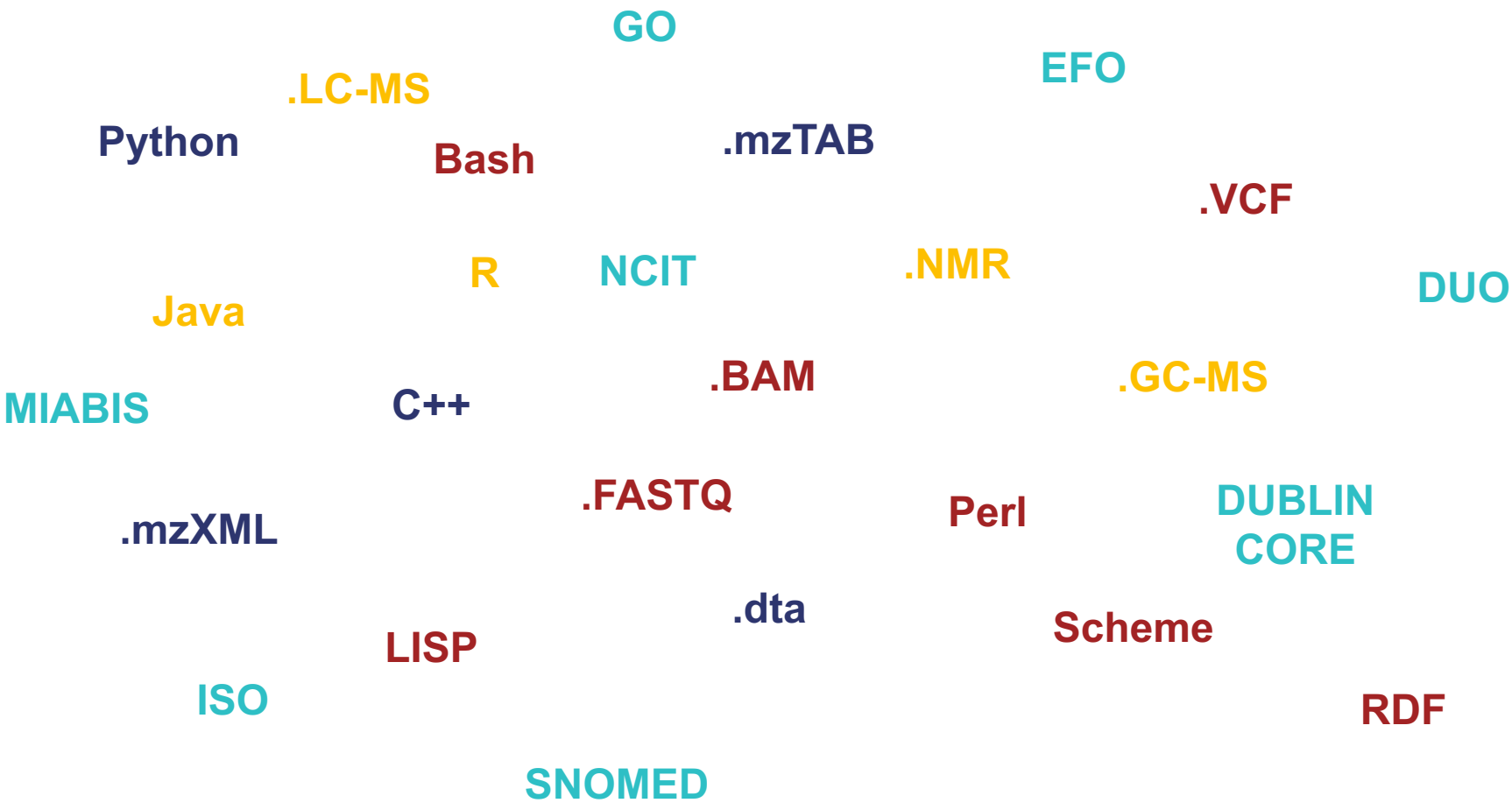


https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-primer/index.html

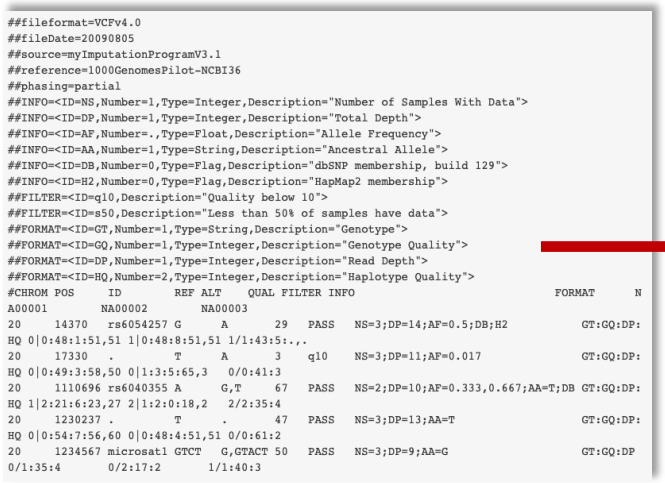
RDF triplets in a knowledge graph



Current Omics data landscape

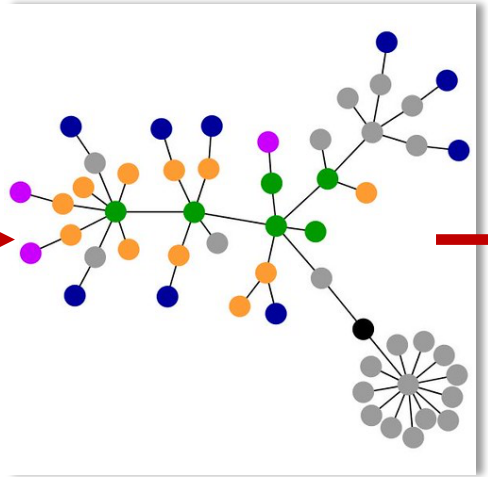


SPARQLing Genomics – Pilot 2019

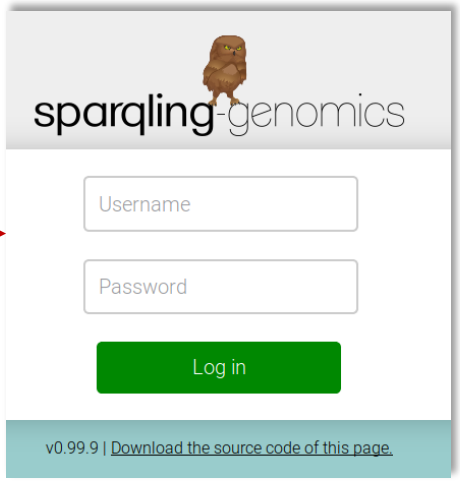


```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      N
A00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:
HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:
HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:
HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:
HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP
0/1:35:4       0/2:17:2       1/1:40:3
```

Variant Call Format - VCF
[Genomics]



RDF Triple Store
[Virtuoso]



Web Interface
[SPARQL]

SPARQLing Genomics – Pilot 2019

```
## Individuals
## ------------------------------------------------------------------------

:vcf2rdf-0.99.9
rdf:type :SoftwareProgram ;
    rdfs:label   "vcf2rdf" ;
    rdfs:comment "Tool to convert VCF files to RDF." .

:table2rdf-0.99.9
    rdf:type :SoftwareProgram ;
    rdfs:label   "table2rdf" ;
    rdfs:comment "Tool to convert tabular data to RDF." .

:folder2rdf-0.99.9
    rdf:type :SoftwareProgram ;
    rdfs:label   "folder2rdf" ;
    rdfs:comment "Tool to extract RDF from files in a directory." .

:sg-web-0.99.9
    rdf:type :SoftwareProgram ;
    rdfs:label   "SPARQLing genomics web interface" ;
    rdfs:comment "Tool for querying multiple SPARQL endpoints using a web interface" .
```

Dublin Core Metadata Initiative

GENEONTOLOGY
Unifying Biology

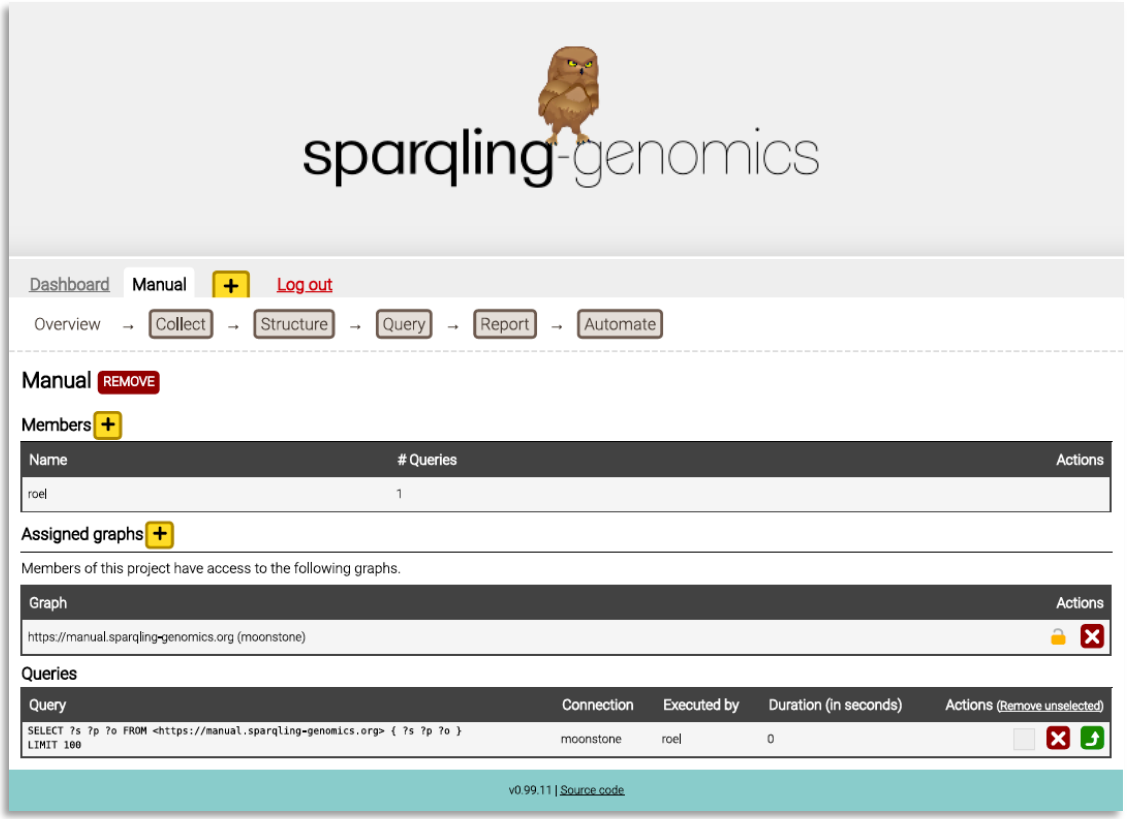
VCF2RDF
TABLE2RDF
XML2RDF
JSON2RDF
…

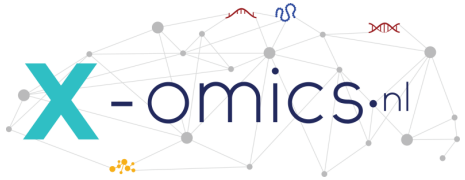
Data Use Ontology

https://www.sparqling-genomics.org/
https://dublincore.org/
http://geneontology.org/
https://github.com/EBISPOT/DUO

X-omics.nl

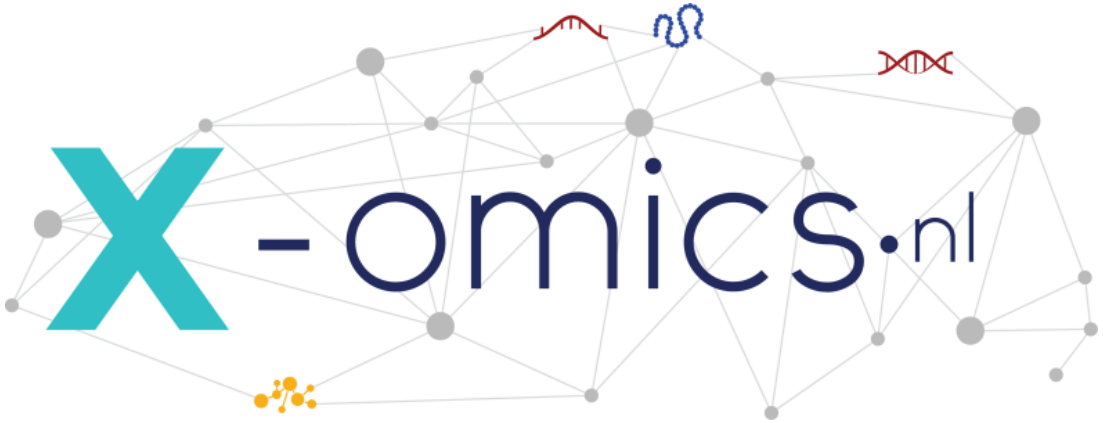
SPARQLing Genomics – Portal 2020




X-omics.nl

Monday 22nd June 2020
Linked data in practice

RDF intro and SPARQLing Genomics demo



Insights from the first XO workshop

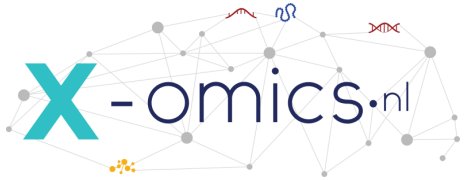
Go to **www.menti.com** and use the code **45 38 35**

Mentimeter

Which skills are needed for Big data and multi-omics data integration?

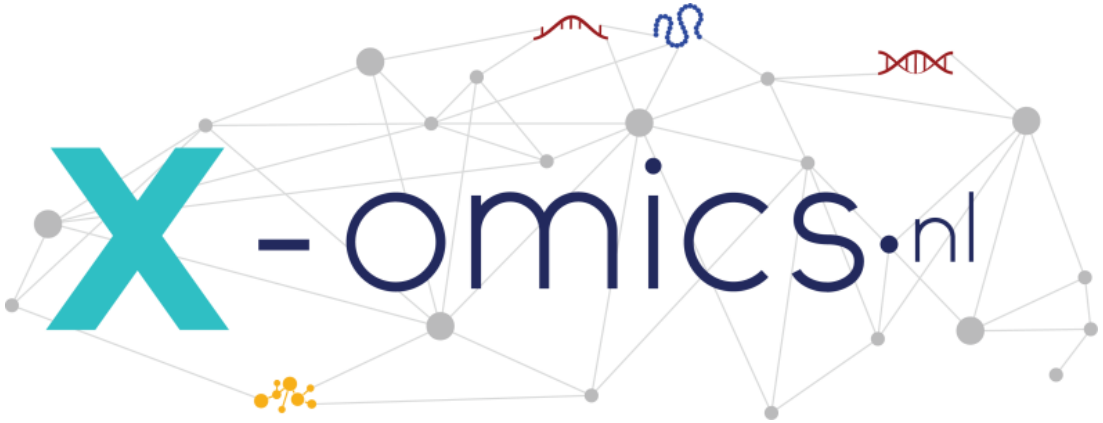
25

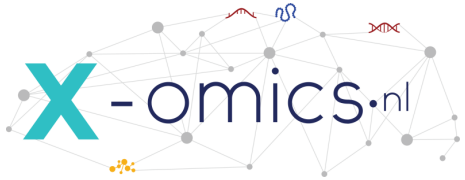
From Mentimeter Quiz, X-Omics Workshop Part 1 "Data standards and multi-omics data integration", 18th June 2020


X-omics.nl

Monday 22nd June 2020
Linked data in practice

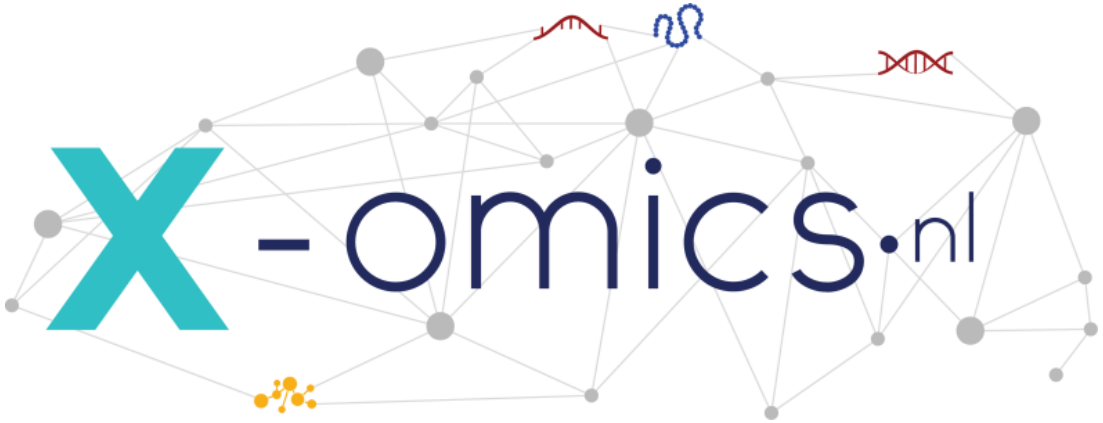
QUIZ TIME!




X-omics.nl

Monday 22nd June 2020
Linked data in practice

We need to talk about ~~data standards~~ ~~vocabularies~~ ontologies



Back to the semantic ambiguity

How many ways can you say ‘female’?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

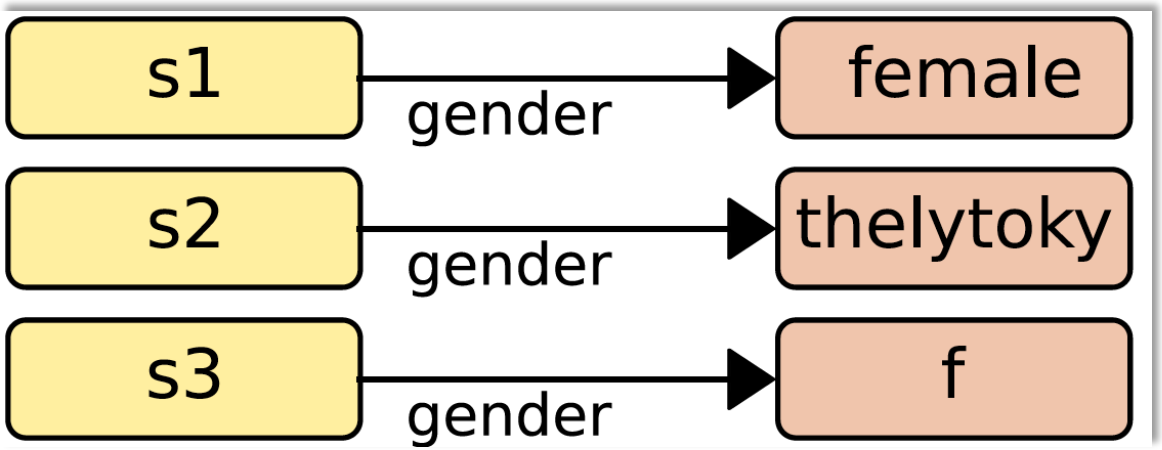
female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

EMBL-EBI

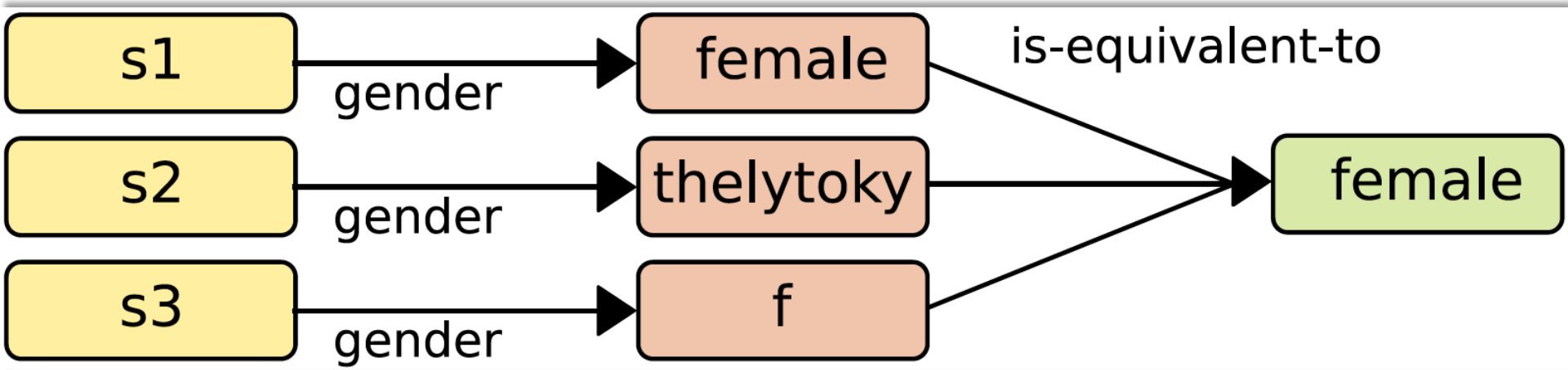
https://www.slideshare.net/ConnectedDataLondon/ontology-services-for-the-biomedical-sciences; Courtesy of N. Silvester, European Nucleotide Archive, EMBL- EBI

X-omics.nl

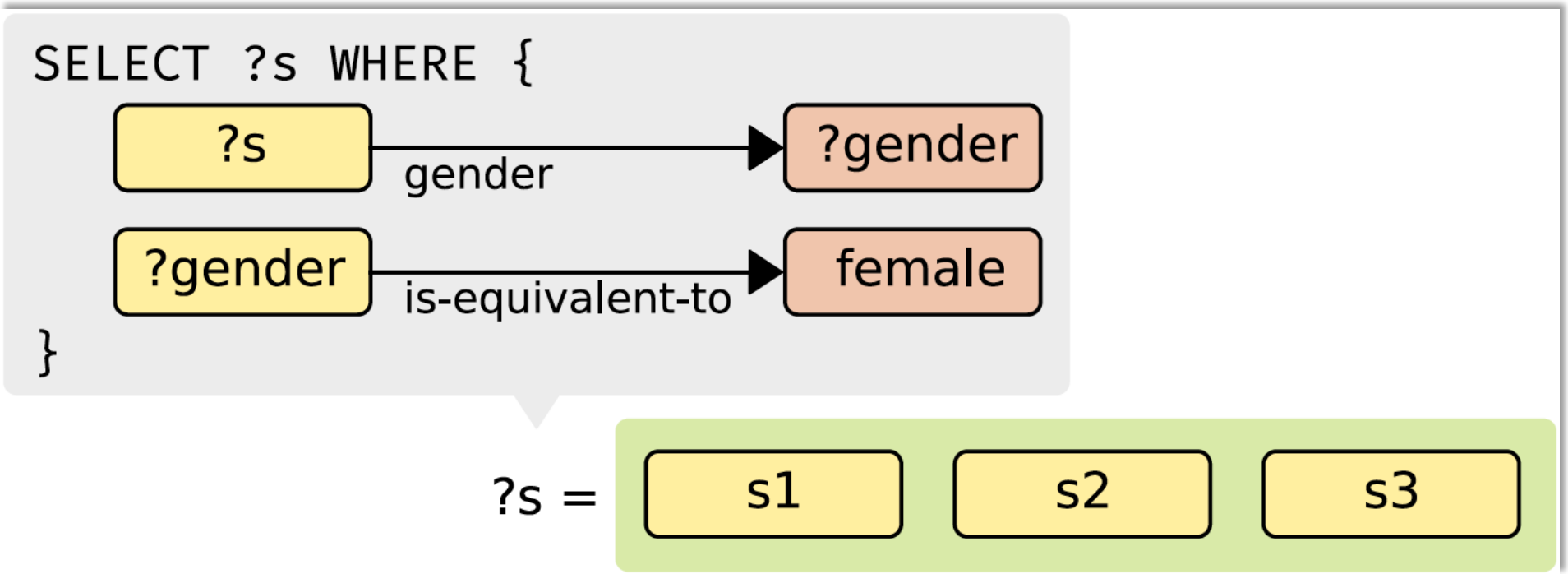
RDF version



RDF version


s1
gender
female
is-equivalent-to
s2
gender
thelytoky
female
s3
gender
f

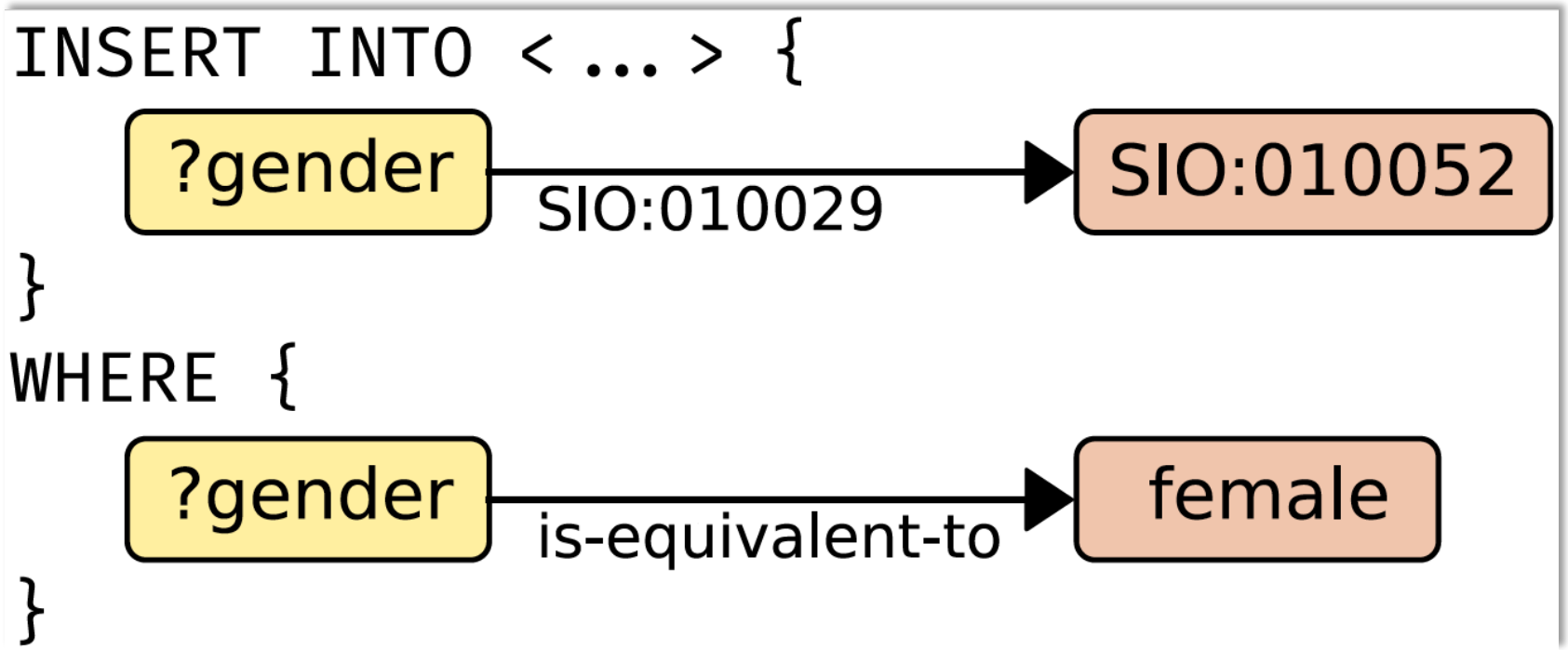
SPARQL query to select “female”



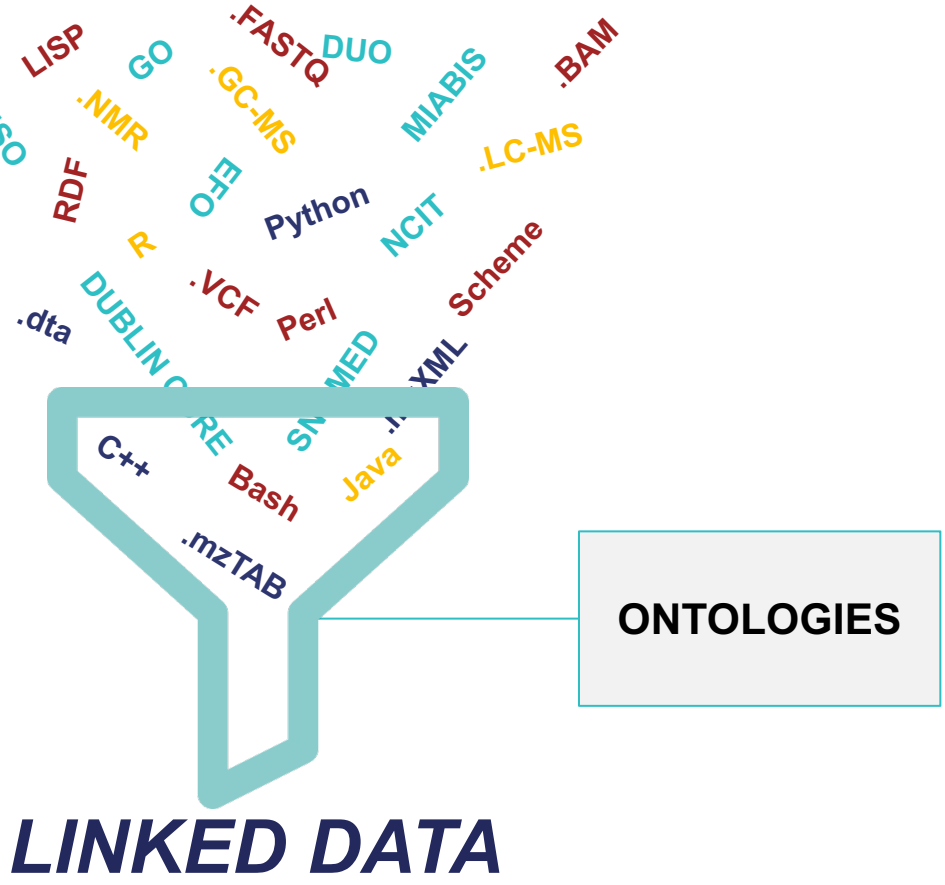
FAIR genome definition of gender

Meta-data element	Preferred ontology term for meta-data	Value types	Issue Number	Issue (open / closed)
Individual ID	NCIT:C164337	ID [string]	#3	closed
Gender	SIO:010029	Male SIO:010048 Female SIO:010052 Unknown / Undetermined	#4	closed
Genotypic sex	PATO:0020000	UNKNOWN_KARYOTYPE, XX, XY, XO, XXY, XXX, XXYY, XXXY, XXXX, XYY, OTHER_KARYOTYPE	#69	closed

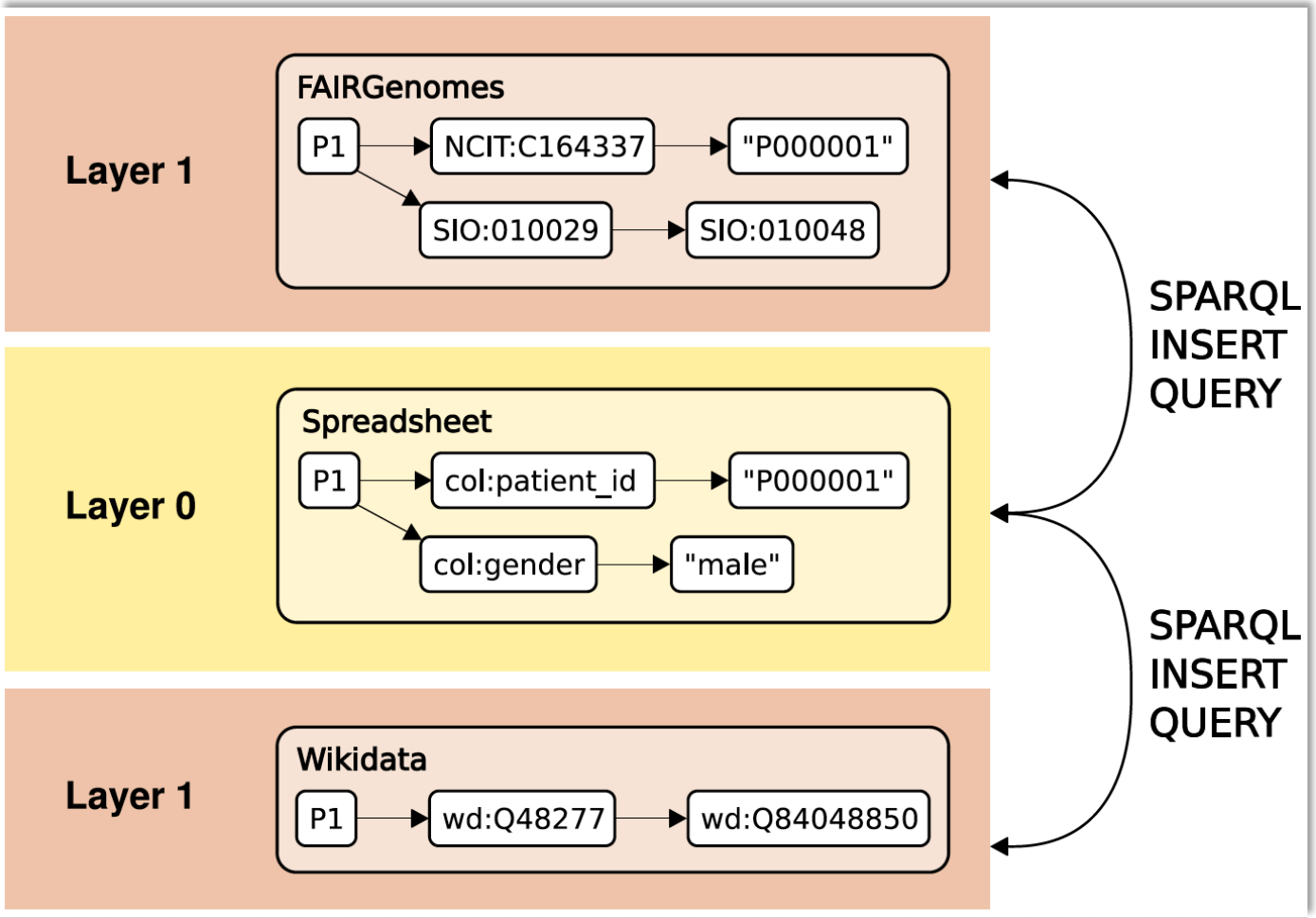
Using SPARQL to ontologise



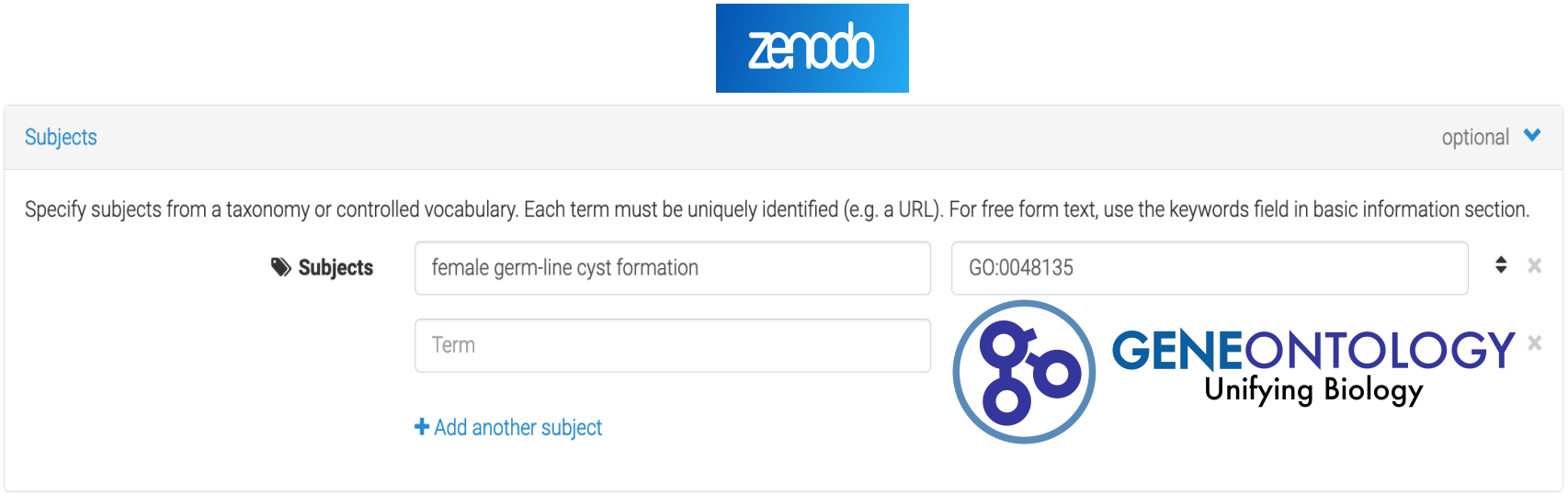
Unifying omics data one term at a time



FAIR data with linked data



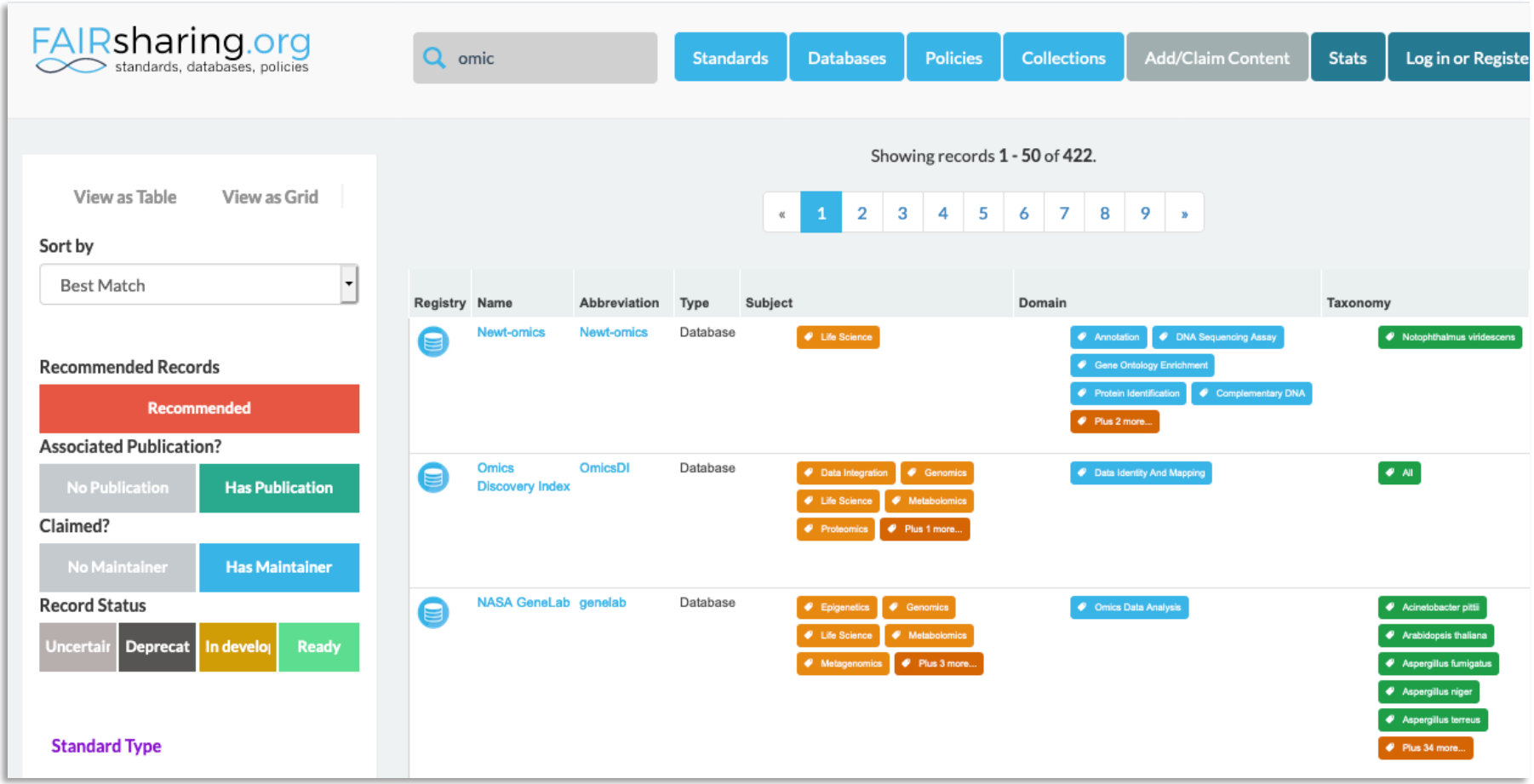
FAIR data work around



https://zenodo.org/
http://geneontology.org/
https://www.ebi.ac.uk/QuickGO/

X-omics.nl

Find standards on FAIRsharing.org



https://fairsharing.org/

SUMMARY

- We can create a knowledge graph of diverse omics data using RDF
- Ontologies help unifying terminology across research domains

- We can transform a spreadsheet into ontologically defined data
- Flexible data management requires easy data transformations, which can be done with SPARQL

- SPARQLing Genomics is a starting point for interoperable omics data

Monday 22nd June 2020
Linked data in practice

Q&A
Open Discussion