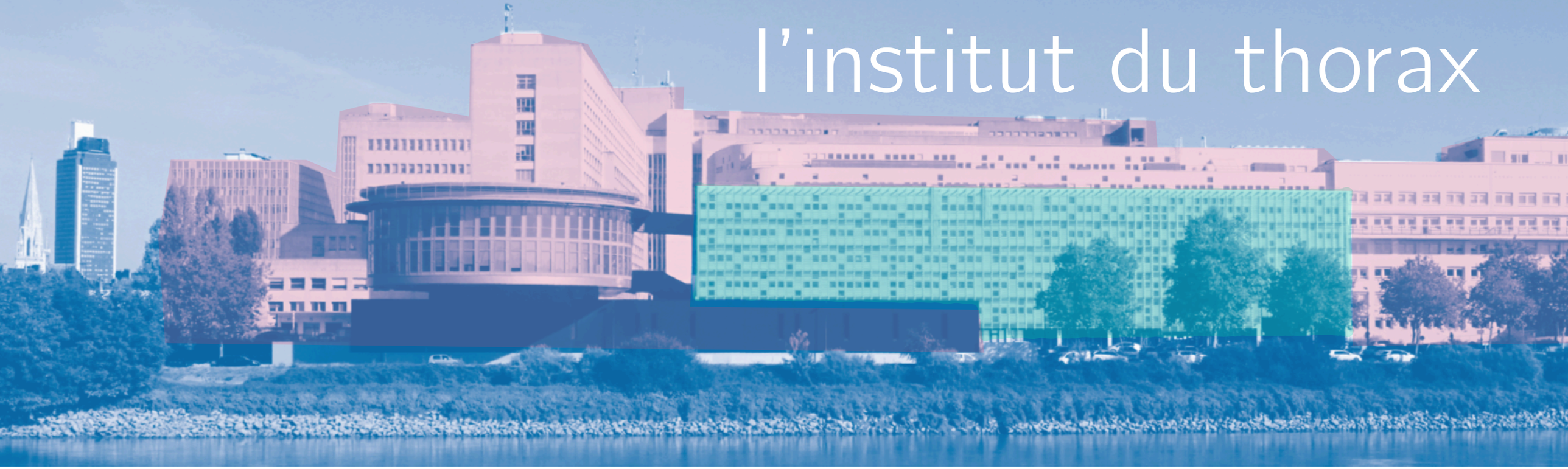
Semantic Beacons: a framework to support federated querying over genomic variants and public Knowledge Graph

Alban Gaignard
CNRS, institut du thorax, Nantes, France

Inserm webinar

March 20, 2026

Introduction



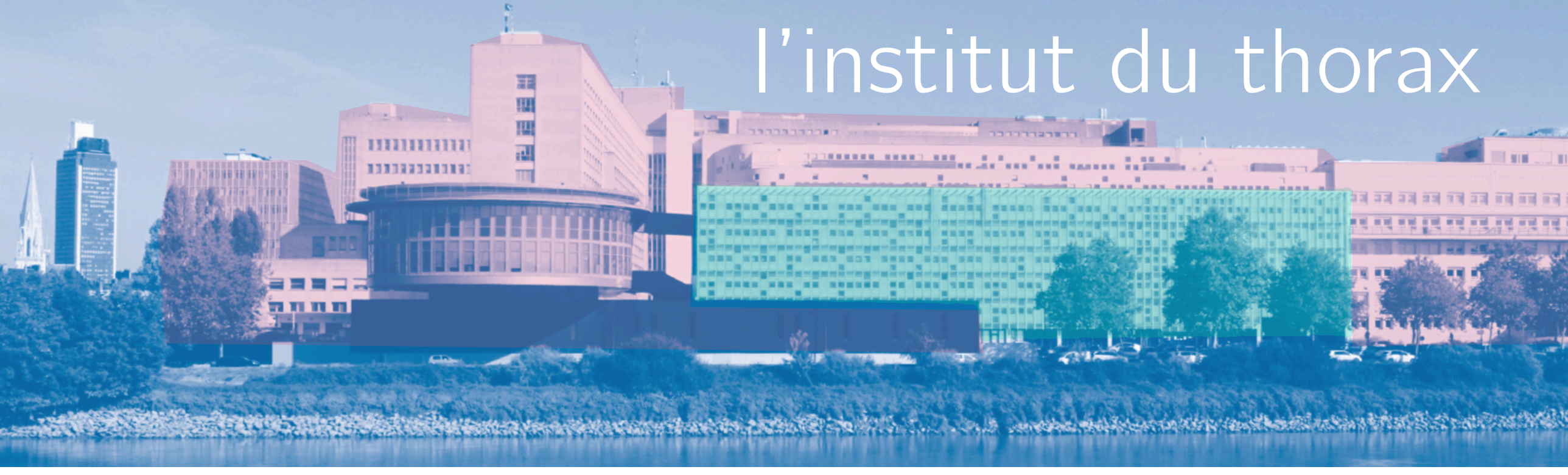
l'institut du thorax

Better understanding of cardio-vascular and metabolic diseases

Gene-function associations

Translational medicine

Real proximity between the university hospital and the research lab



l'institut du thorax

Better understanding of cardio-vascular and metabolic diseases

Gene-function associations

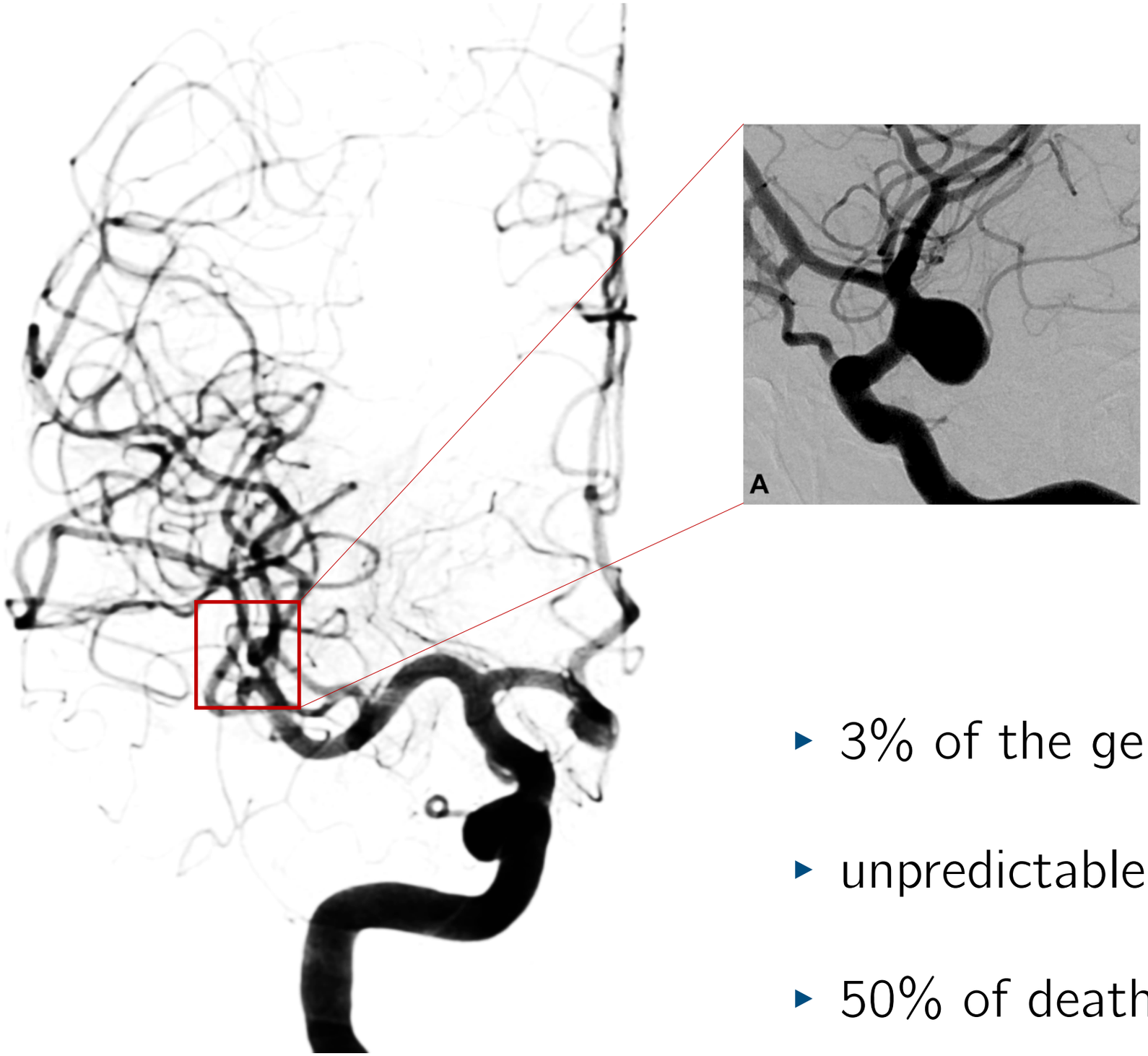
Translational medicine

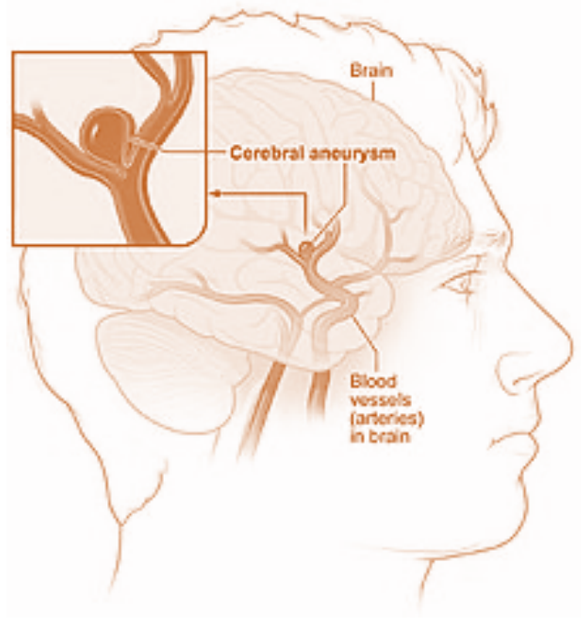
Real proximity between the university hospital and the research lab

Bioinformatics

- Massive production of genomic sequence & health data
  $\longrightarrow$ high performance computing
- Integration of multi-modal and multi-scale data
- Predictive models

Intracranial aneurysms





- 3% of the general population
- unpredictable rupture
- 50% of death in case of rupture

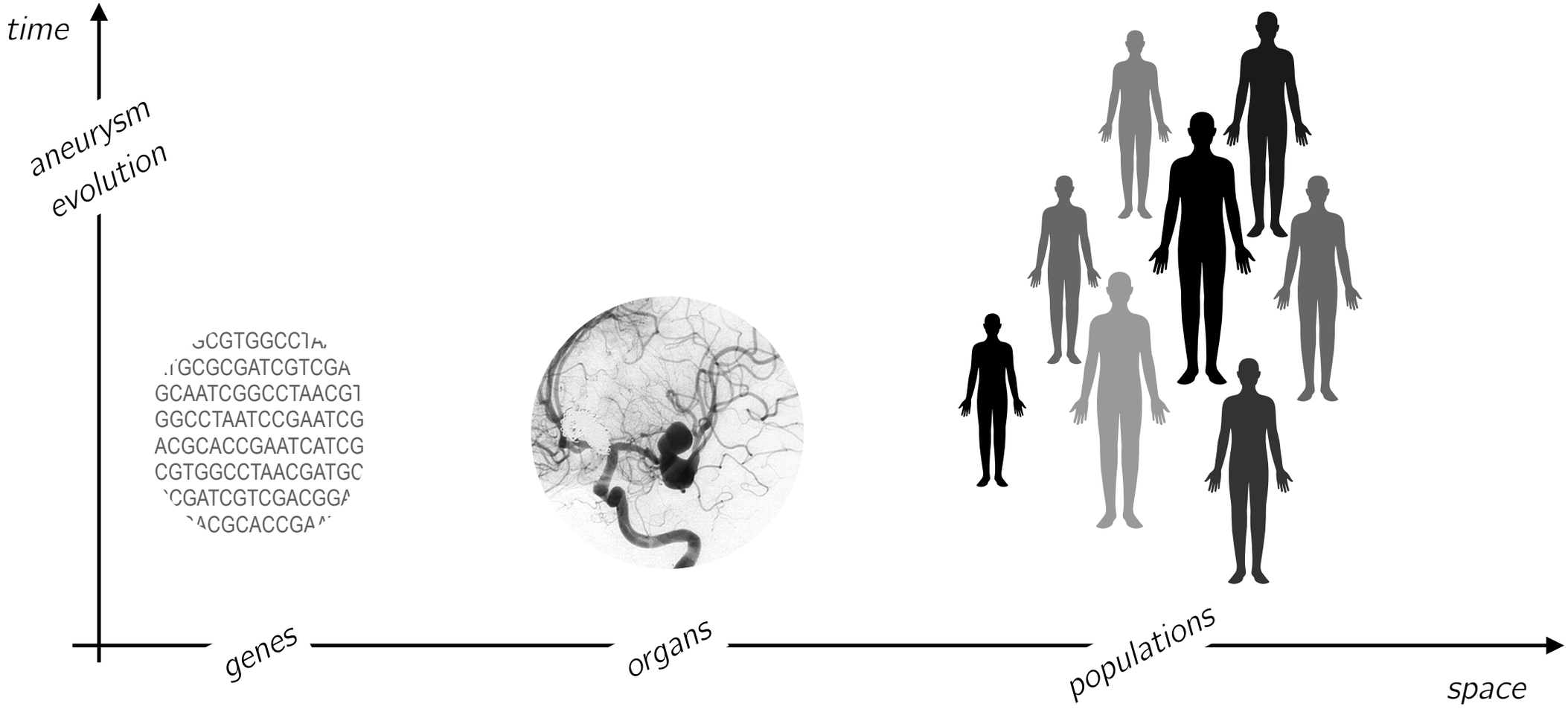
Multi-factorial disease $\longrightarrow$ multi-scale data

Multi-factorial disease $\longrightarrow$ multi-scale data

- Inter-disciplinary efforts needed for a better understanding of the pathology
- Specific data produced at very specific scales

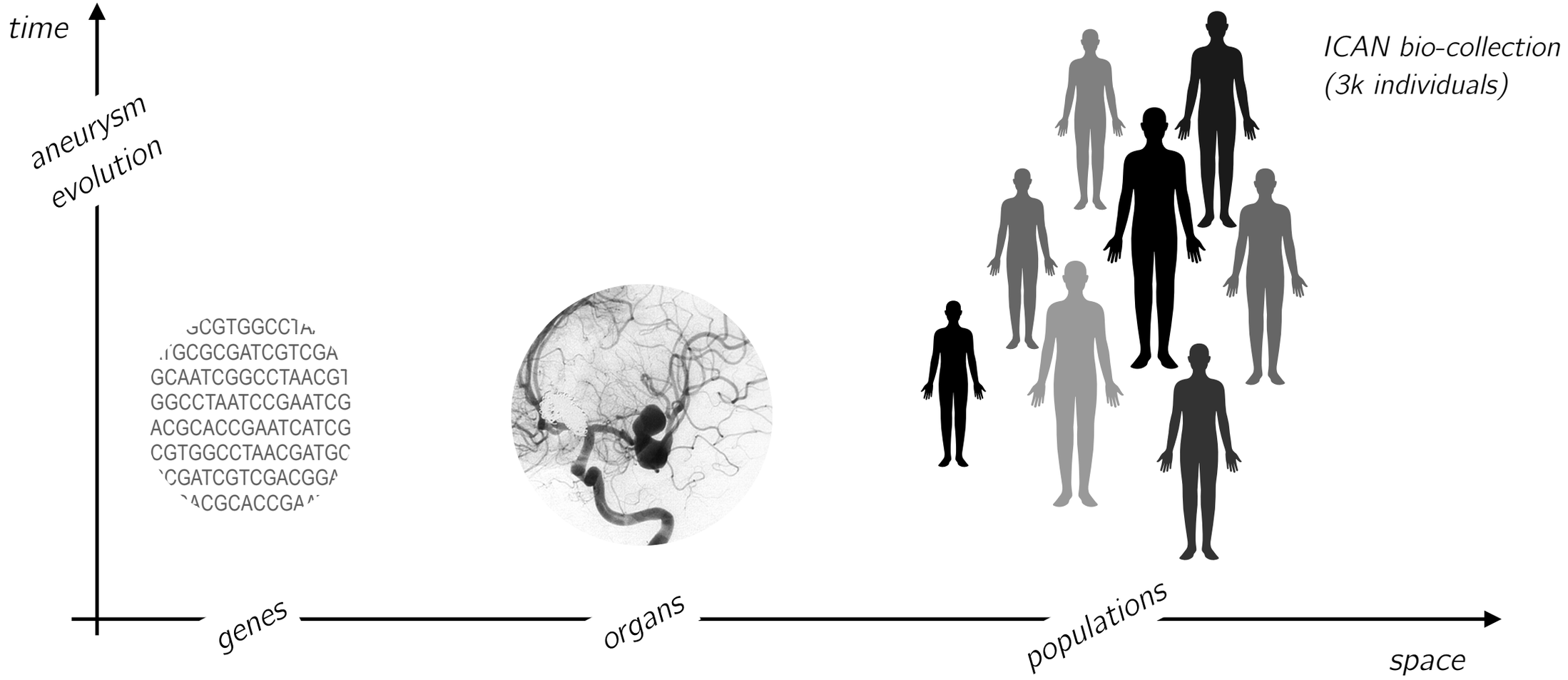
Multi-factorial disease $\longrightarrow$ multi-scale data

- Inter-disciplinary efforts needed for a better understanding of the pathology
- Specific data produced at very specific scales



Multi-factorial disease $\longrightarrow$ multi-scale data

- Inter-disciplinary efforts needed for a better understanding of the pathology
- Specific data produced at very specific scales



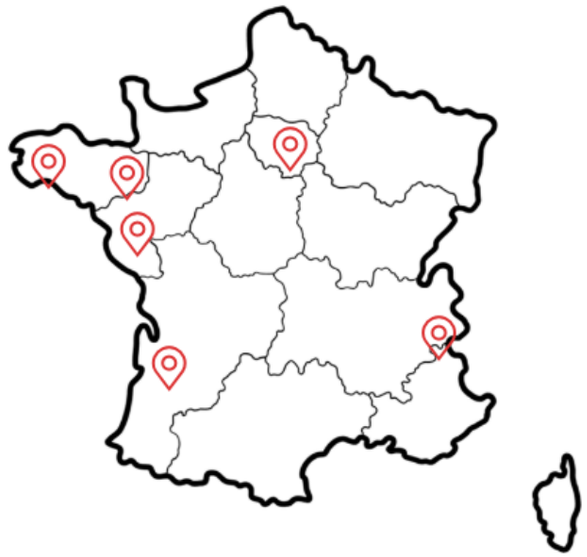
Neurovasc project (PEPR SN)

- Neurovasc: a national programme funded for 4 years by the french research agency to build a **digital infrastructure** to manage and exploit **intracranial aneurysm data**
  - 3 Research Institutes (Inria, Inserm, IMT Atlantique)
  - 2 Clinical Research Teams (Brest & Nantes academic hospitals)
  - 3 Universities (Bordeaux, Paris-Saclay, Nantes)

WP 1: A model of interoperable infrastructure between research and healthcare
- Task 1.1. Managing clinical data
- Task 1.2. Managing imaging data
- Task 1.3. Managing genetic data

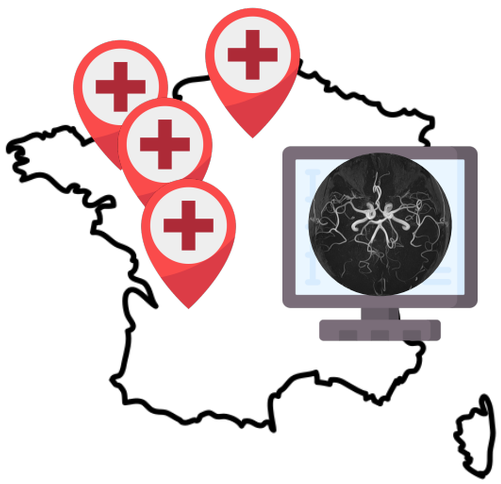
WP 2: Interoperable datasets and predictive models for ICA diagnosis and outcomes
- Task 2.1: FAIR genomic data demonstrator
- Task 2.2: Mining healthcare circuits following ICA diagnosis
- Task 2.3: A non-additive model for global genetic-risk prediction

WP 3: Proof-of-Concept studies with digital companions
- Task 3.1: Accompanying patients with diagnosed unruptured ICA
- Task 3.2: Post-stroke prevention through mobile applications and digital monitoring

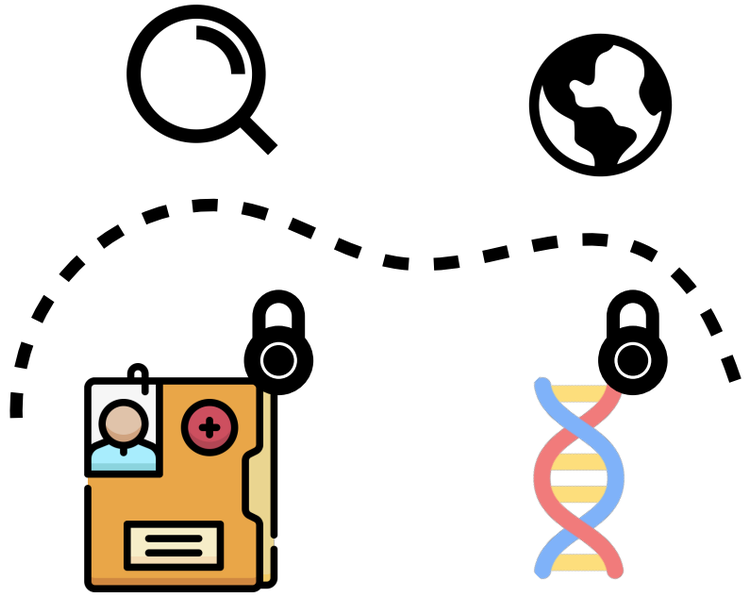


Data integration & sharing challenges

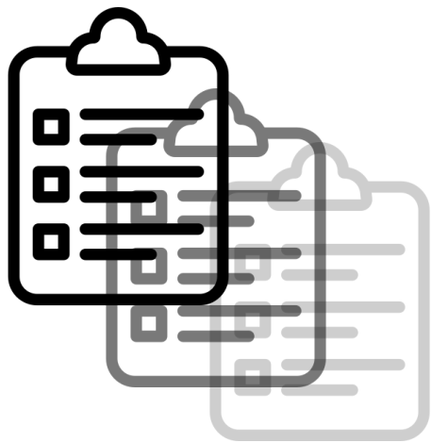
❶ How to **collect high-quality medical images** from multiple hospitals/ MRIs ?

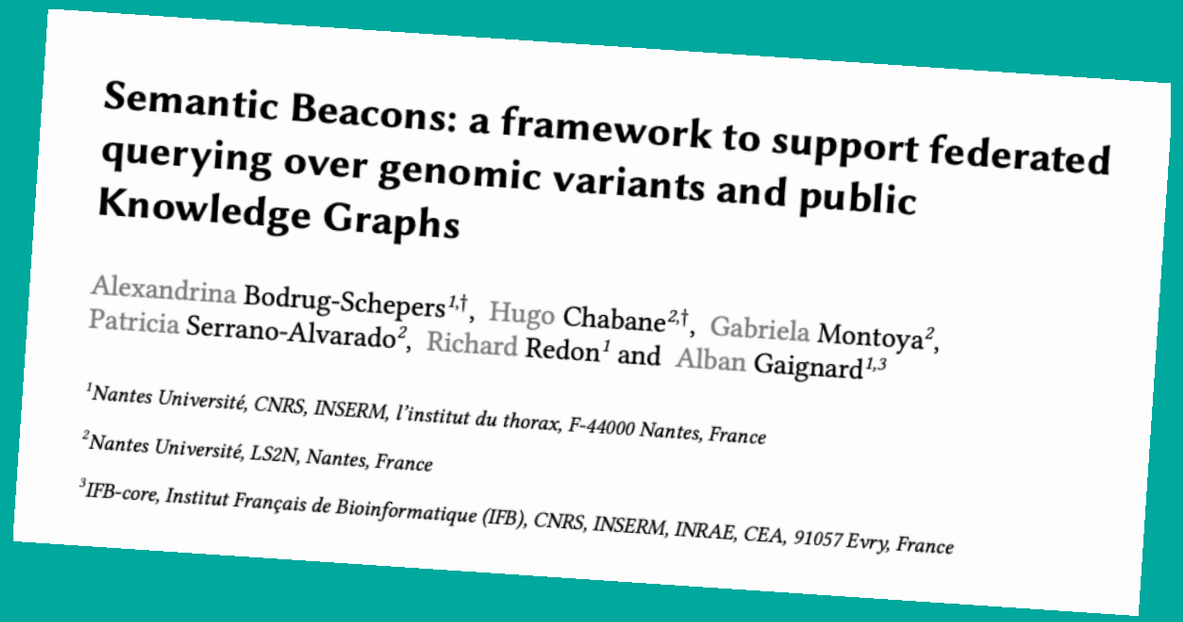


❷ How to **interlink and query multi-modal and multi-scale data** while preserving privacy constraints ?



❸ How to mine and **model patient trajectories** from EHR data ? can we predict clinical outcomes ?





Semantic Beacons: a framework to support federated querying over genomic variants and public Knowledge Graphs

Alexandrina Bodrug-Schepers[1,†], Hugo Chabane[2,†], Gabriela Montoya[2], Patricia Serrano-Alvarado[2], Richard Redon[1] and Alban Gaignard[1,3]

[1]Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France

[2]Nantes Université, LS2N, Nantes, France

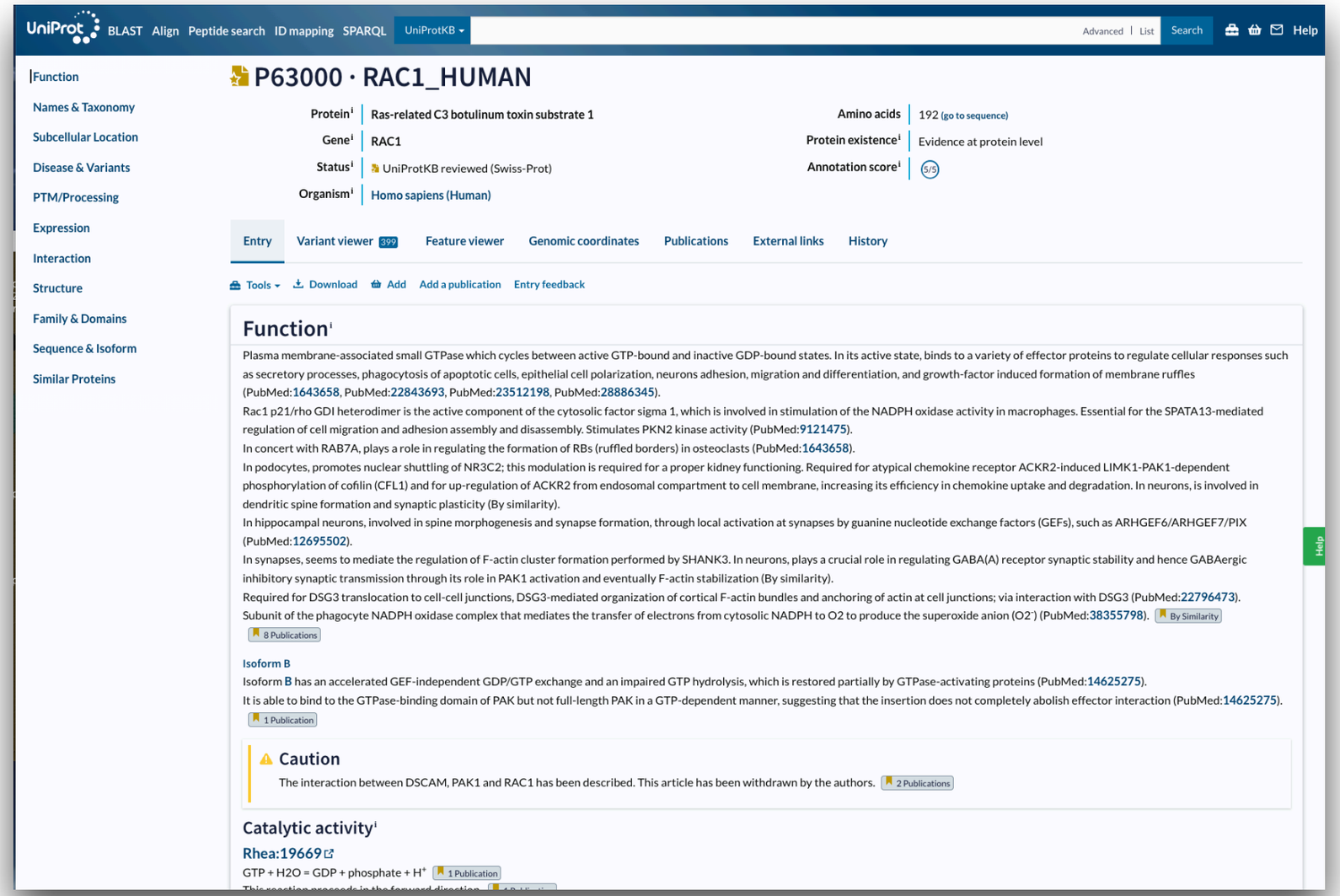
[3]IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

Alexandrina Bodrug

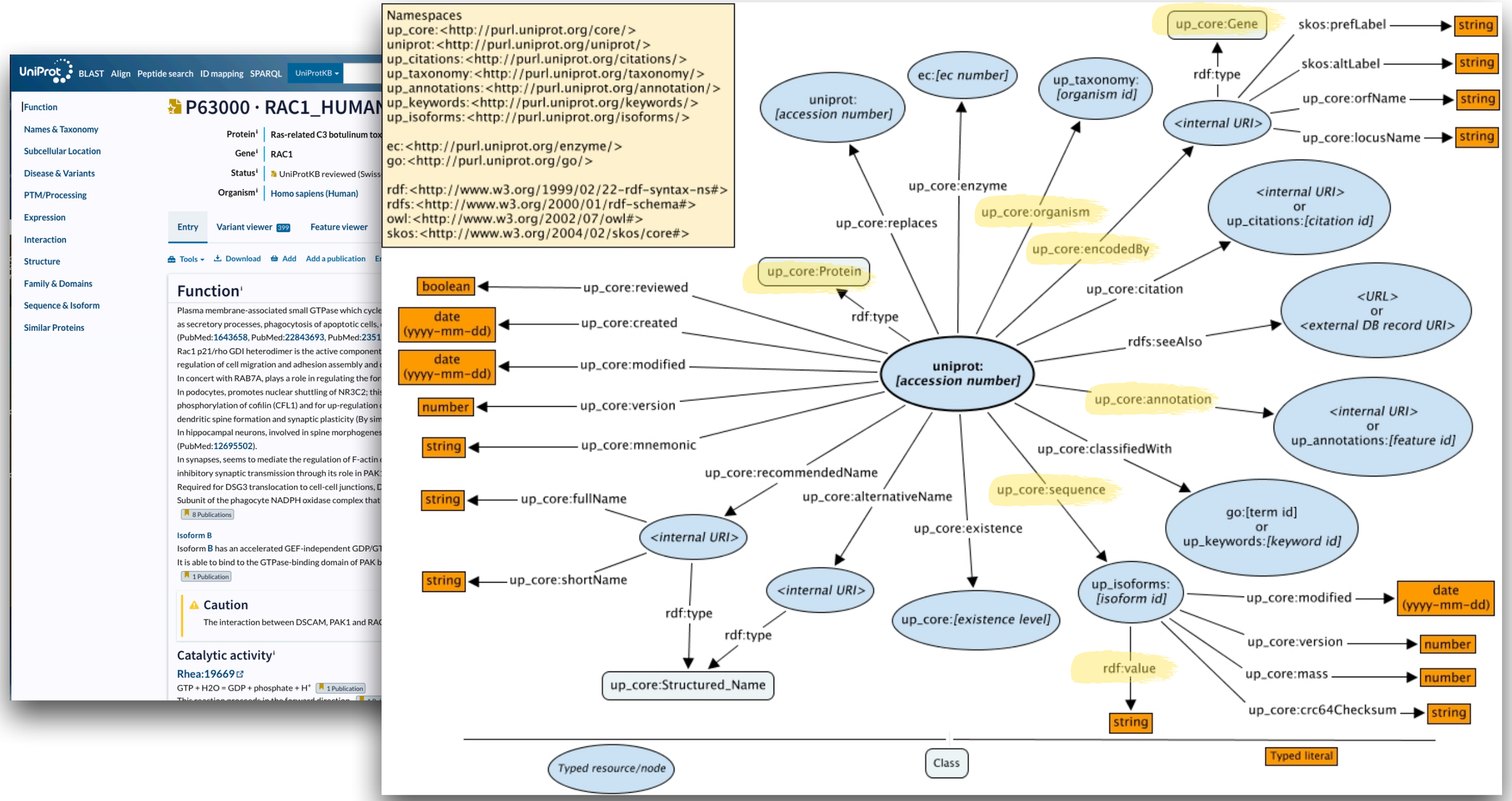
Bridging genomic data and public knowledge bases ?

Alexandrina Bodrug-Schepers, et al.. **Semantic Beacons: a framework to support federated querying over genomic variants and public Knowledge Graphs**. SWAT4HCLS 2025, Feb 2025, Barcelona, Catalonia, Spain. ⟨hal-04908530v2⟩

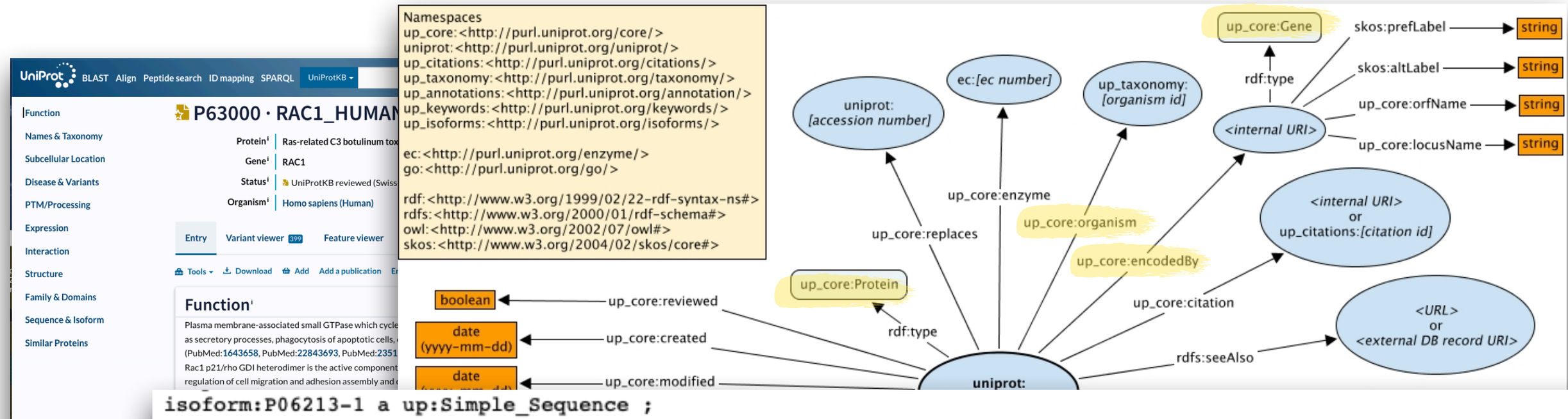
The UniProt public knowledge graph



The UniProt public knowledge graph

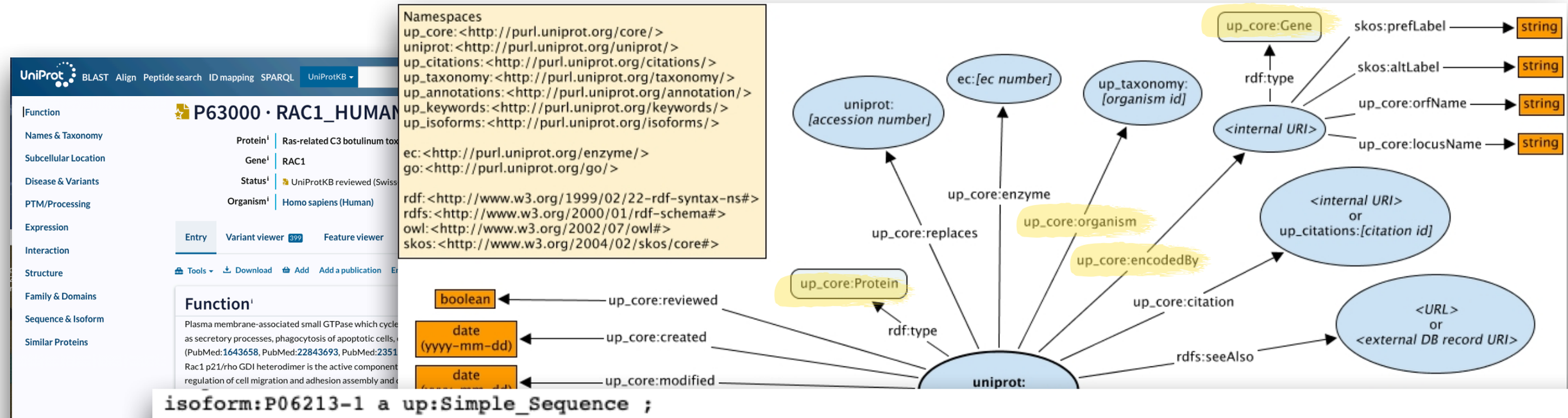


The UniProt public knowledge graph




```
isoform:P06213-1 a up:Simple_Sequence ;
  up:modified "2010-10-05"^^xsd:date ;
  up:version 4 ;
  up:precursor true ;
  up:mass 156333 ;
  up:md5Checksum "8eb04104be33f0a0cb376ed145e684ff" ;
  skos:prefLabel "Long" ;
  skos:altLabel "HIR-B" ;
  rdf:value
"MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHLQILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFPNLTVIRGS
RLFFNYALVIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRILDSVEDNYIVLNKDDNEECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTICKS
HGCTAEGLCCHSECLGNCSQPDDPTKCVACRNFYLDGRCVETCPPPYYHFQDWRCVNFSFCQDLHHKCKNSRRQGCHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKV
CHLLEGEKTIDSVTSAQELRGCTVINGSLIINIRGGNNLAAELEANLGLIEEISGYLKIRRSYALVSLSFFRKLRLIRGETLEIGNYSFYALDNQNLRQLWDWSKHNLTITQG
KLFFHYNPKLCLSEIHKMEEVSGTKGRQERNDIALKTNGDQASCENELLKFSYIRTSFDKILLRWEPYWPPDFRDLLGFMLFYKEAPYQNVTEFDGQDACGSNSWTVVDIDPP
LRSNDPKSQNHPGWLMRGLKPWTQYAIFVKTLVTFSDERRTYGAKSDIIYVQTDATNPSVPLDPISVSNSSSQIILKWKPPSDPNGNITHYLVFWERQAEDSELFELDYCLKG
LKLPSRTWSPPFESEDSQKHNQSEYEDSAGECCSCPKTDSQILKELEESSFRKTFEDYLHNVVFVPRKTSSGTGAEDPRPSRKRRSLGDVGNVTVAVPTVAAFPNTSSTSVPT
SPEEHRPFEKVVNKESLVISGLRHFTGYRIELQACNQDTPEERCSVAAYVSARTMPEAKADDIVGPVTHEIFENNVVHLMWQEPKEPNGLIVLYEVSYRRYGDEELHLCVSRK
HFALERGCRLRGLSPGNYSVRIRATSLAGNGSWTEPTYFYVTDYLDVPSNIAKIIIGPLIFVFLFSVVIGSIYLFLRKRQPDGPLGPLYASSNPEYLSASDVFPCSVYVPDEW
EVSREKITLLRELGQGSFGMVYEGNARDIIKGEAETRVAVKTVNESASLRERIEFLNEASVMKGFTCHHVVRLLGVVSKGQPTLVVMELMAHGDLKSYLRSLRPEAENNPGRP
PPTLQEMIQMAAEIADGMAYLNAKKFVHRDLAARNCMVAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMAPESLKDGVFTTSSDMWSFGVVLWEITSLAEQPYQGLSN
EQVLKFVMDGGYLDQPDNCPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEMEFEDMENVPLDRSSHCQREEAGGRDGGSSLGFKRSY
EEHIPYTHMNGGKKNGRILTLPRSNPS" .
```

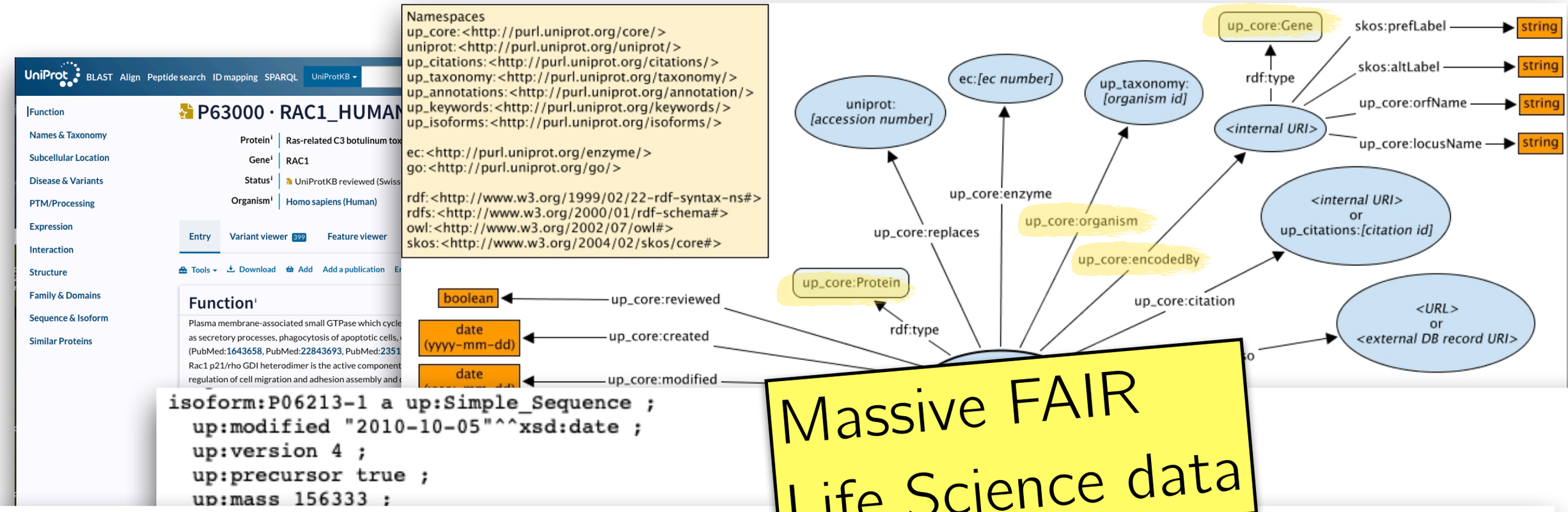
The UniProt public knowledge graph



There are 217,505,202,099 triples is this release. All triples are available in the default graph. There are 22 named graphs corresponding to specific datasets.

Graph	Documentation	Triples	Distinct subjects	Distinct predicates	Distinct classes	Distinct objects	License
uniparc	Documentation	160,189,731,200	40,455,837,024	29	6	46,916,767,863	http://creativecommons.org/licenses/by/4.0/
uniprot	Documentation	44,256,643,227	9,441,439,078	124	121	8,462,262,751	http://creativecommons.org/licenses/by/4.0/
uniref	Documentation	10,224,623,630	1,393,813,725	14	3	1,409,539,937	http://creativecommons.org/licenses/by/4.0/
obsolete	Documentation	2,102,255,458	277,358,373	10	3	286,609,935	http://creativecommons.org/licenses/by/4.0/
citationmapping	Documentation	625,262,380	123,810,071	12	4	29,448,749	http://creativecommons.org/licenses/by/4.0/
taxonomy	Documentation	60,041,721	26,918	21	4	4,698,602	http://creativecommons.org/licenses/by/4.0/
citations	Documentation	31,212,544	419,769	19	5	8,870,230	http://creativecommons.org/licenses/by/4.0/
proteomes	Documentation	8,984,258	1,999,807	33	11	3,777,324	http://creativecommons.org/licenses/by/4.0/
chebi	Documentation	3,419,539	221,830	24	6	1,828,527	http://creativecommons.org/licenses/by/4.0/
rhea	Documentation	1,962,186	138,720	67	3	540,446	http://creativecommons.org/licenses/by/4.0/

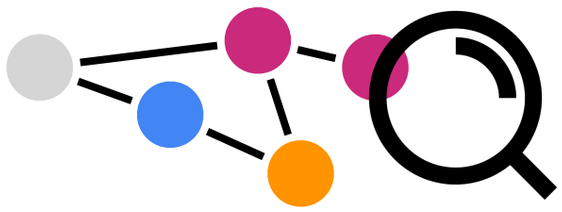
The UniProt public knowledge graph



Graph	Documentation	Triples	Distinct subjects	Distinct predicates	Distinct classes	Distinct objects	License
uniparc	Documentation	160,189,731,200	40,455,837,024	29	6	46,916,767,863	http://creativecommons.org/licenses/by/4.0/
uniprot	Documentation	44,256,643,227	9,441,439,078	124	121	8,462,262,751	http://creativecommons.org/licenses/by/4.0/
uniref	Documentation	10,224,623,630	1,393,813,725	14	3	1,409,539,937	http://creativecommons.org/licenses/by/4.0/
obsolete	Documentation	2,102,255,458	277,358,373	10	3	286,609,935	http://creativecommons.org/licenses/by/4.0/
citationmapping	Documentation	625,262,380	123,810,071	12	4	29,448,749	http://creativecommons.org/licenses/by/4.0/
taxonomy	Documentation	60,041,721	26,918	21	4	4,698,602	http://creativecommons.org/licenses/by/4.0/
citations	Documentation	31,212,544	419,769	19	5	8,870,230	http://creativecommons.org/licenses/by/4.0/
proteomes	Documentation	8,984,258	1,999,807	33	11	3,777,324	http://creativecommons.org/licenses/by/4.0/
chebi	Documentation	3,419,539	221,830	24	6	1,828,527	http://creativecommons.org/licenses/by/4.0/
rhea	Documentation	1,962,186	138,720	67	3	540,446	http://creativecommons.org/licenses/by/4.0/

Knowledge graphs for clinical & genomic data

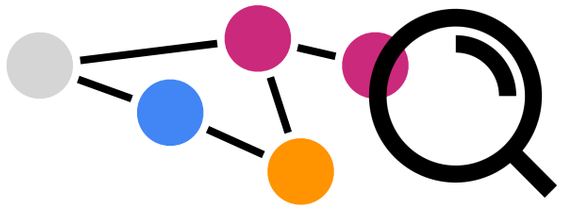
A clinical and genomic intracranial aneurysm knowledge graph



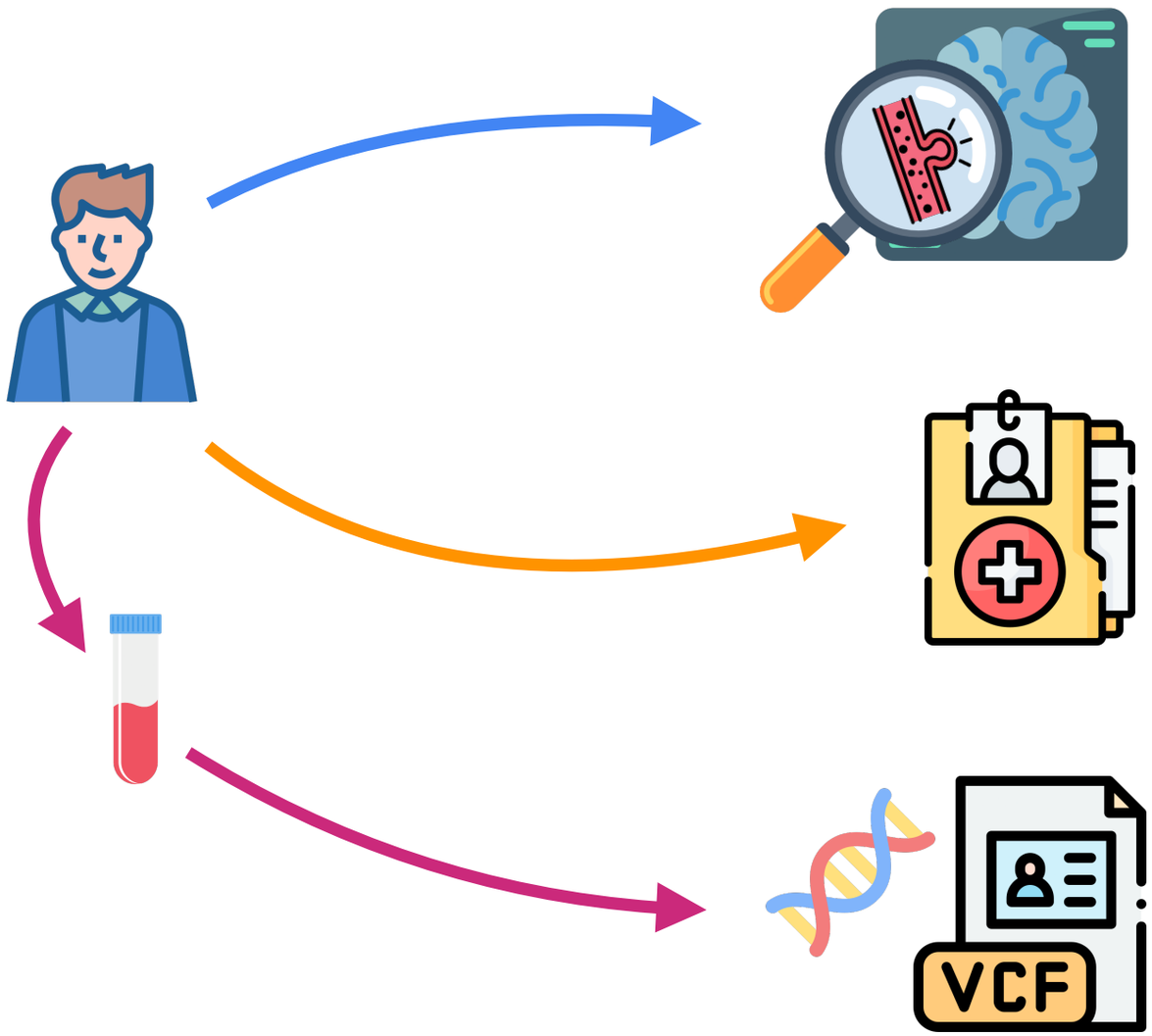
to find & exchange phenotypes/variants with reference terminologies !

Knowledge graphs for clinical & genomic data

A clinical and genomic intracranial aneurysm knowledge graph



to find & exchange phenotypes/variants with reference terminologies !



Anatomical structures ? Neuro-vascular tissues ?

- **UBERON**
- **NCIT**

Clinical data / phenotypes ?

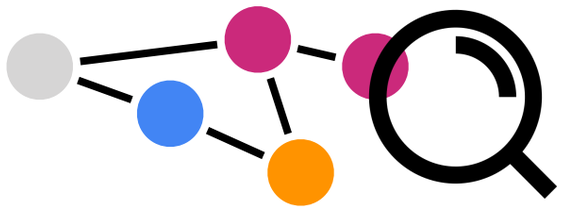
- **SPHN**
- **HPO**
- **DUO**

Genomic data ?

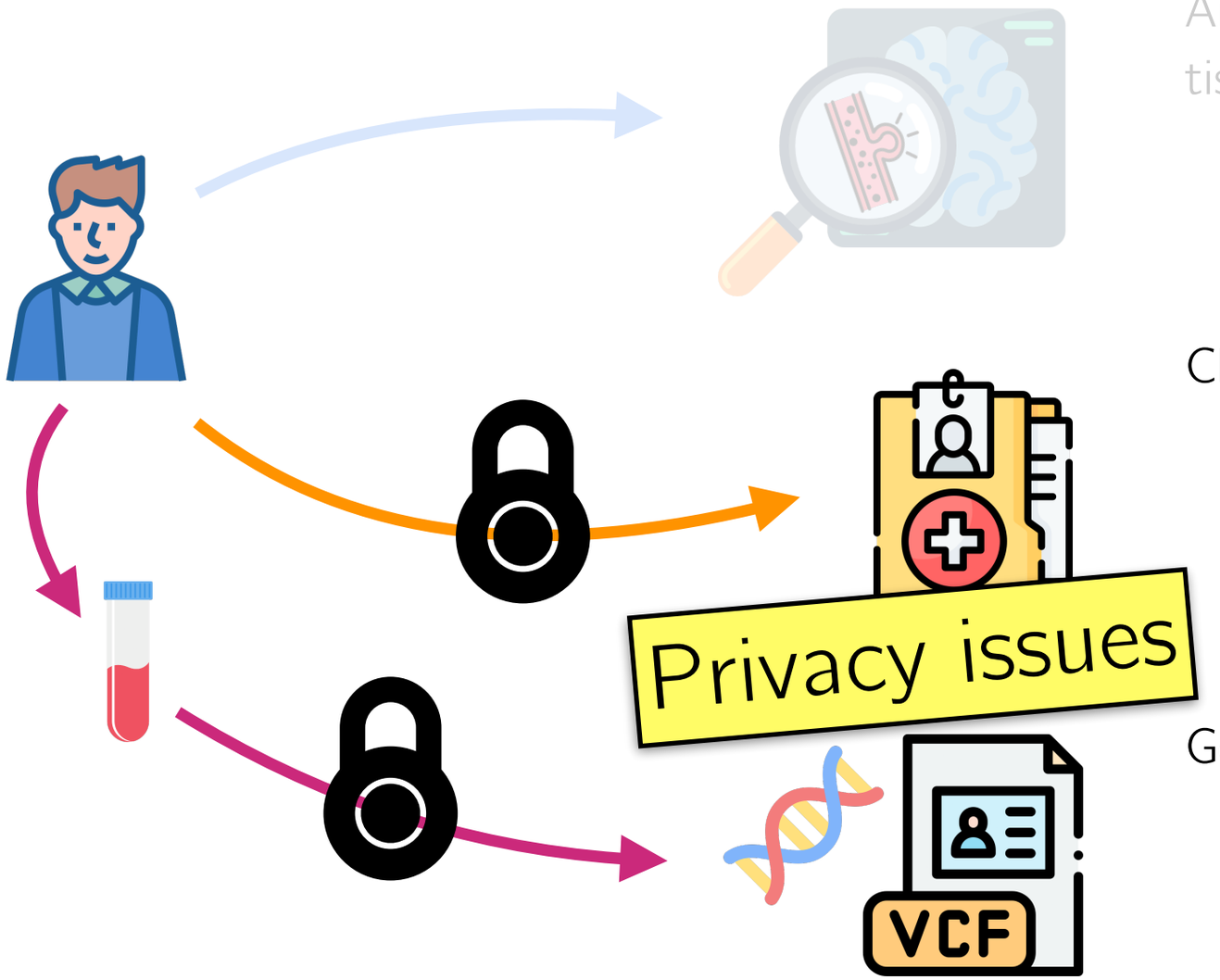
- **FALDO**
- **SO / GENO**
- **SIO**

Knowledge graphs for clinical & genomic data

A clinical and genomic intracranial aneurysm knowledge graph



to find & exchange phenotypes/variants with reference terminologies !



Anatomical structures ? Neuro-vascular tissues ?

- UBERON
- NCIT

Clinical data / phenotypes ?

- **SPHN**
- **HPO**
- **DUO**

Genomic data ?

- **FALDO**
- **SO / GENO**
- **SIO**

Beacon protocol

Beacon: a standard and exchange protocol for more decentralized biomedical research (promoted by Elixir and GA4GH)

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). **Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond**. *Human mutation*, *43*(6), 791–799. https://doi.org/10.1002/humu.24369

Beacon protocol

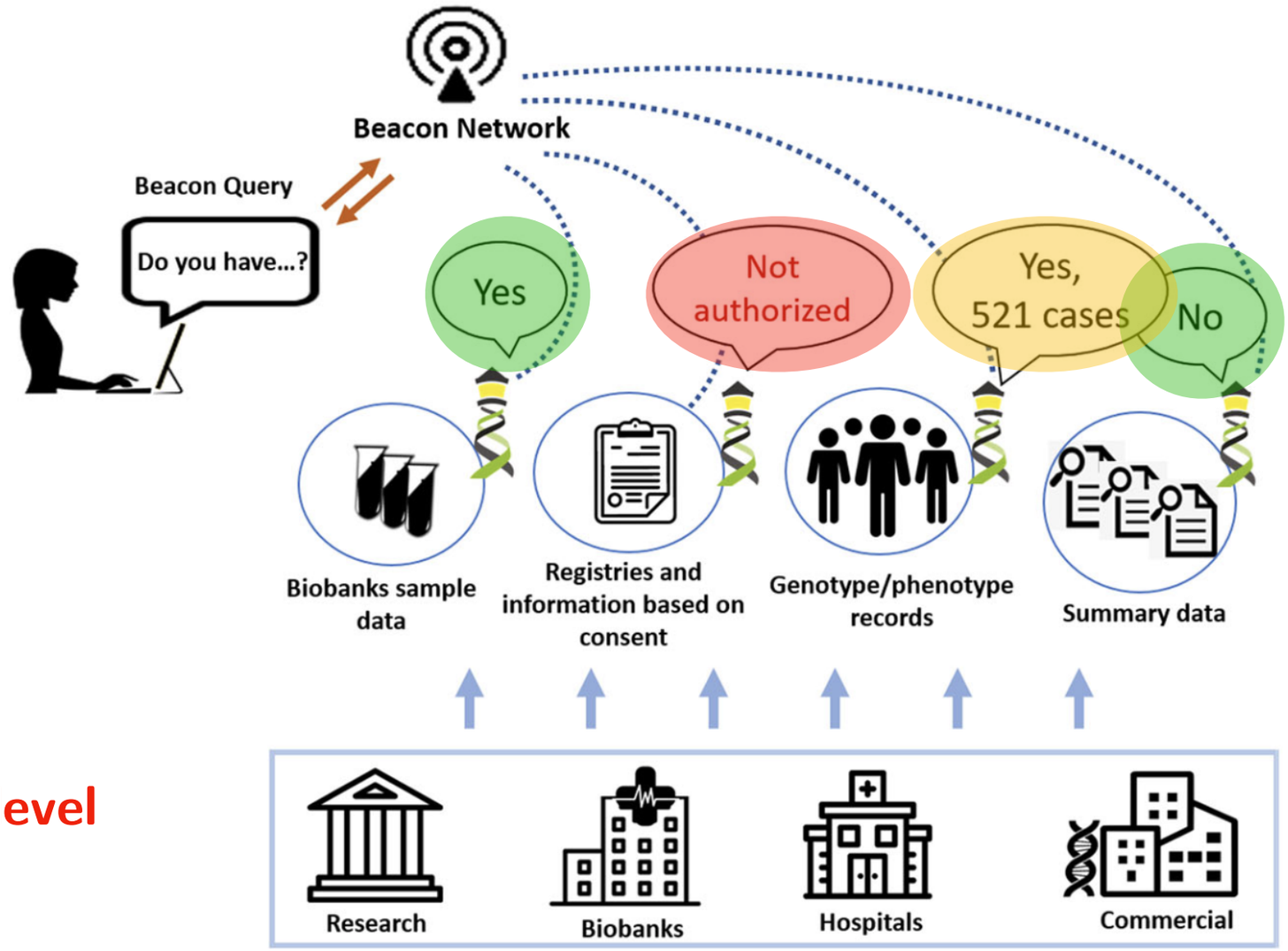
Beacon: a standard and exchange protocol for more decentralized biomedical research (promoted by Elixir and GA4GH)

- **Metadata model** for ‘Variation’, ‘Sample’, ‘Dataset’, ‘Individual’, etc
- Different **access models**: **boolean**, **aggregated data**, **record level**
- A framework with a **reference implementation**

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). **Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond**. *Human mutation*, *43*(6), 791–799. https://doi.org/10.1002/humu.24369

Beacon protocol

Beacon: a standard and exchange protocol for more decentralized biomedical research (promoted by Elixir and GA4GH)

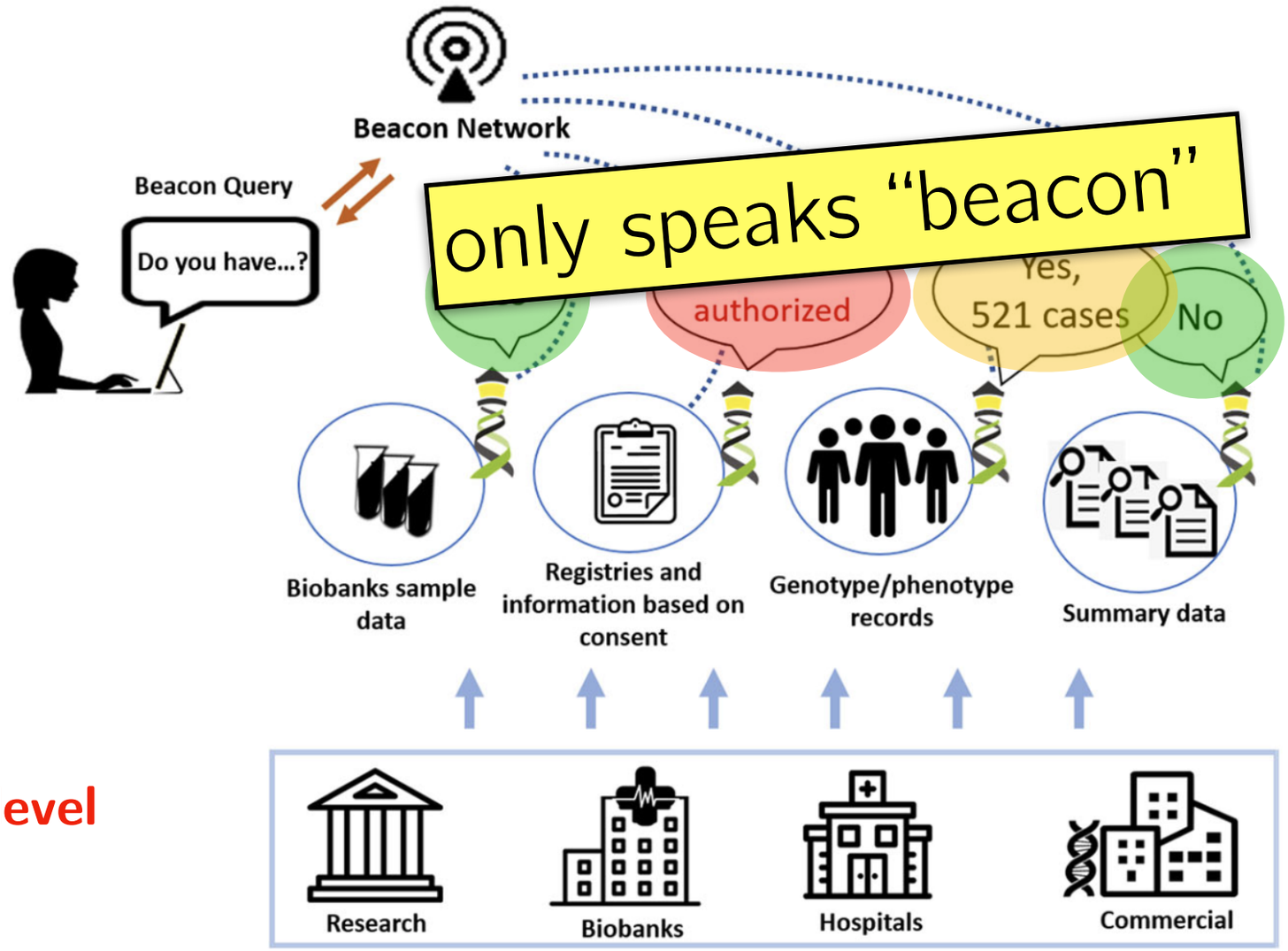


- **Metadata model** for 'Variation', 'Sample', 'Dataset', 'Individual', etc
- Different **access models**: **boolean**, **aggregated data**, **record level**
- A framework with a **reference implementation**

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). **Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond**. *Human mutation*, *43*(6), 791–799. https://doi.org/10.1002/humu.24369

Beacon protocol

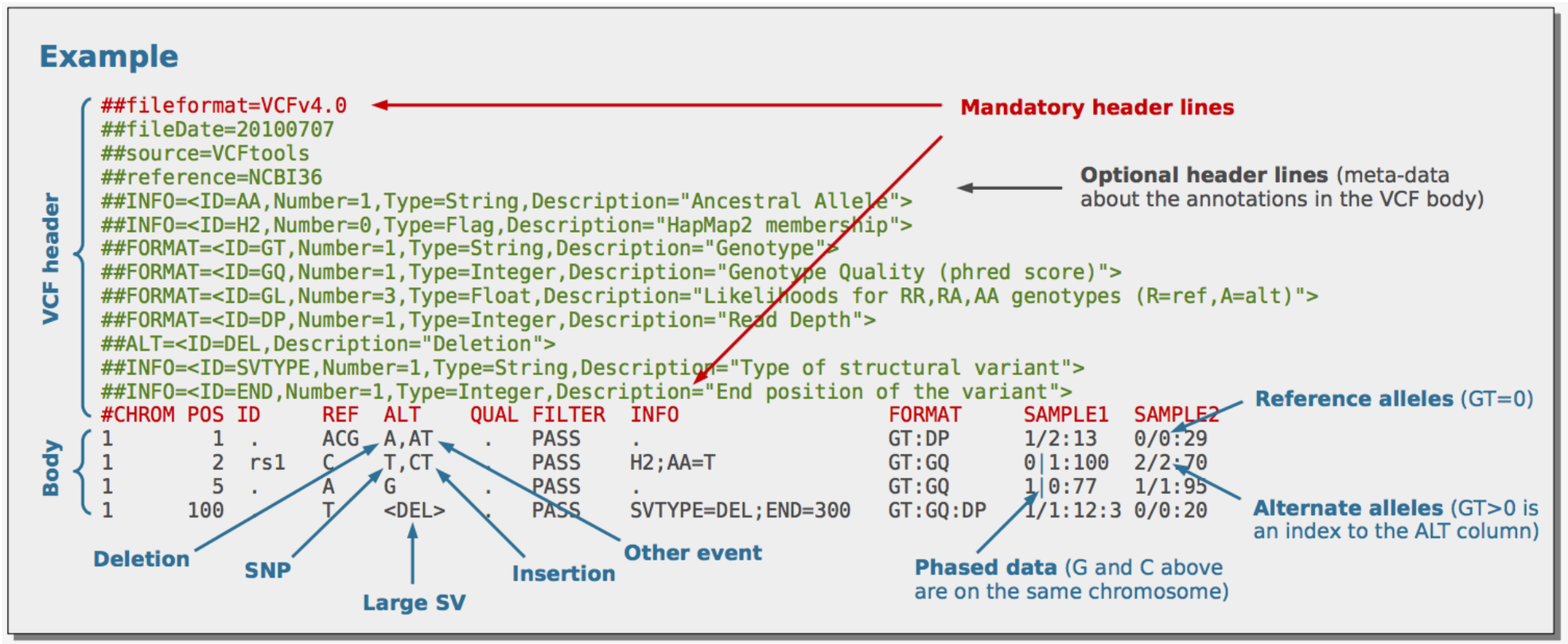
Beacon: a standard and exchange protocol for more decentralized biomedical research (promoted by Elixir and GA4GH)



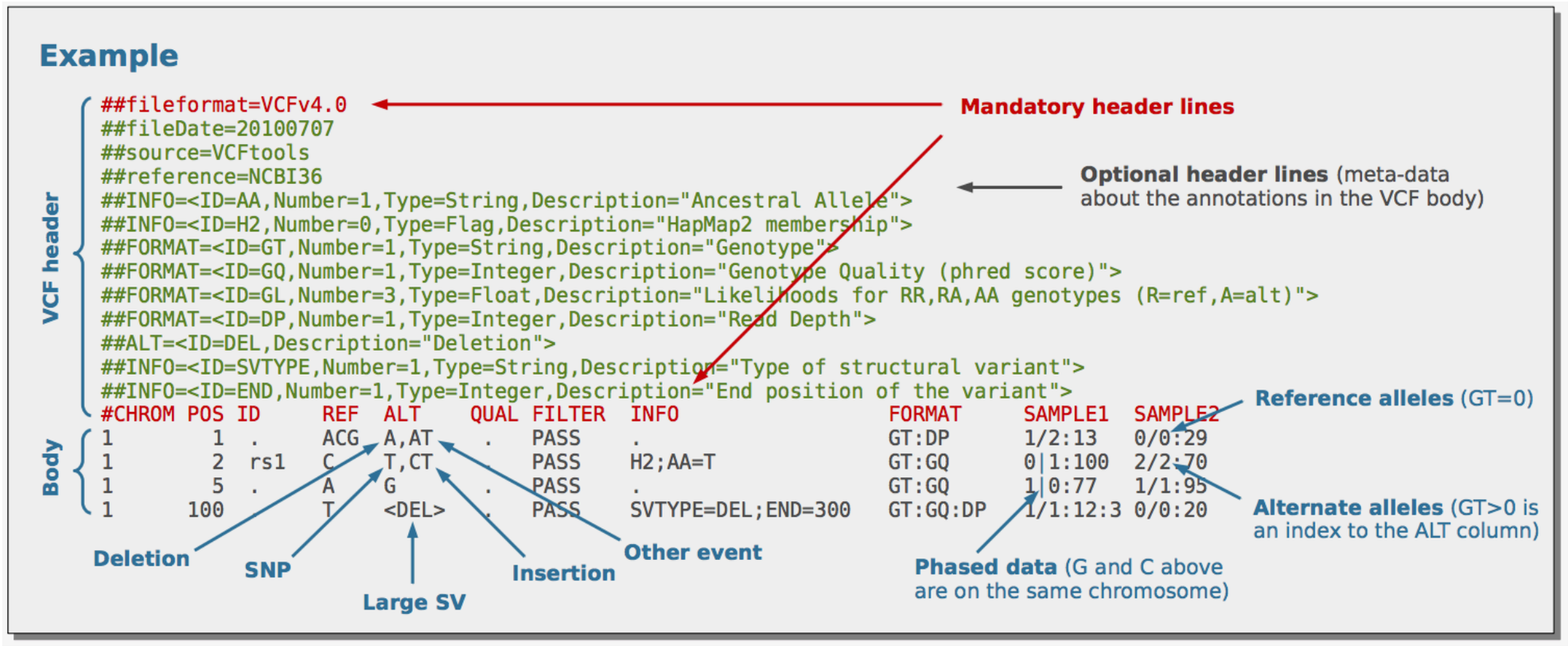
- **Metadata model** for ‘Variation’, ‘Sample’, ‘Dataset’, ‘Individual’, etc
- Different **access models**: **boolean**, **aggregated data**, **record level**
- A framework with a **reference implementation**

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). **Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond**. *Human mutation*, *43*(6), 791–799. https://doi.org/10.1002/humu.24369

Genomic variation data & “annotation”



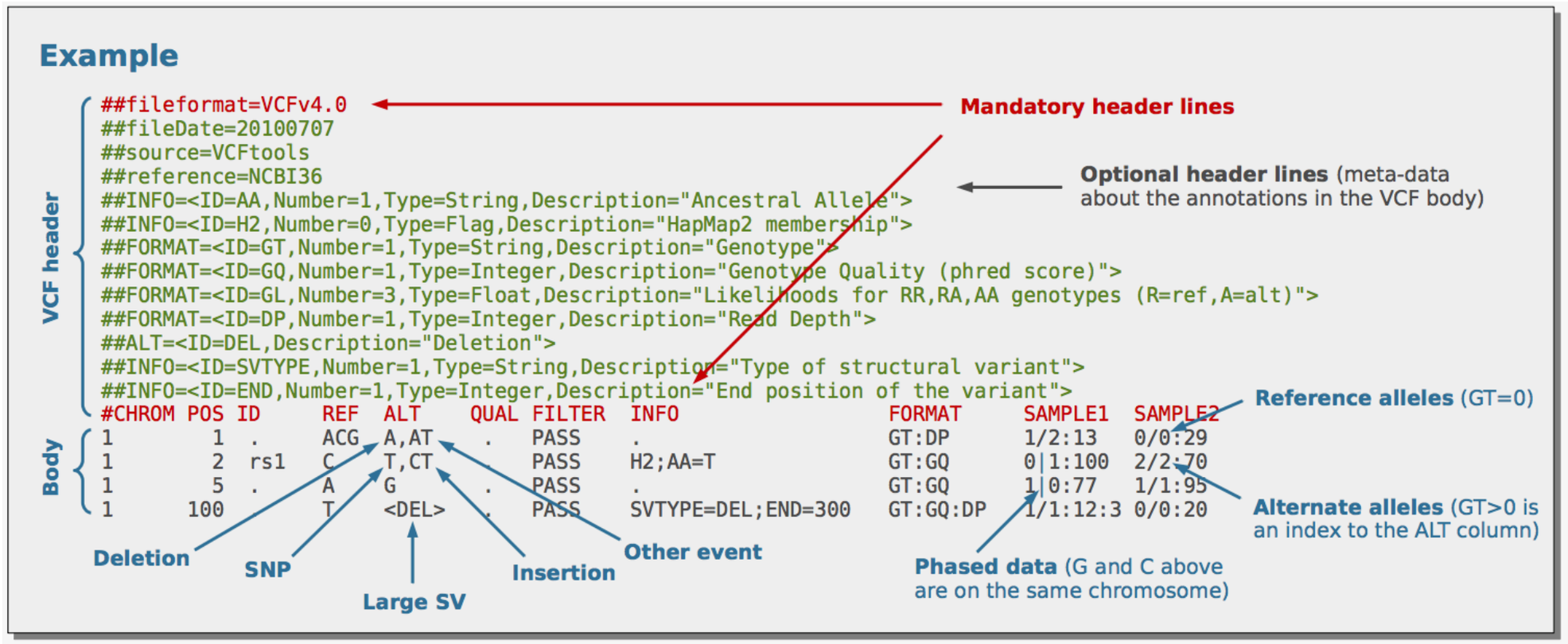
Genomic variation data & “annotation”



- Large tabular file: 1 line per genomic variation, 1 column per individual
- Specific columns for **locating** the variation in the **genome**
- INFO column for **annotations coming from external databases:** *e.g.* pathogenicity scores (CADD v1.7: whole genome annotation database, 625G)

https://cadd.gs.washington.edu/download

Genomic variation data & “annotation”



- Large tabular file: 1 line per genomic variation, 1 column per individual
- Specific columns for **locating** the variation in the **genome**
- INFO column for **annotations coming from external databases:** *e.g.* pathogenicity scores (CADD v1.7: whole genome annotation database, 625G)

Compute and storage intensive variant annotation

https://cadd.gs.washington.edu/download

Issues & Objectives

⚠️ **Genomic variants** must be safely kept **on-site**.

⚠️ **Annotating genomic variants** for biological interpretation is **costly** (data transfer + CPU).

🚀 Massive and diverse reference data already available in the form of **interoperable public knowledge graphs**.

❓ How to enable **on-the-fly annotation** of genomic beacon data with public knowledge graphs ?

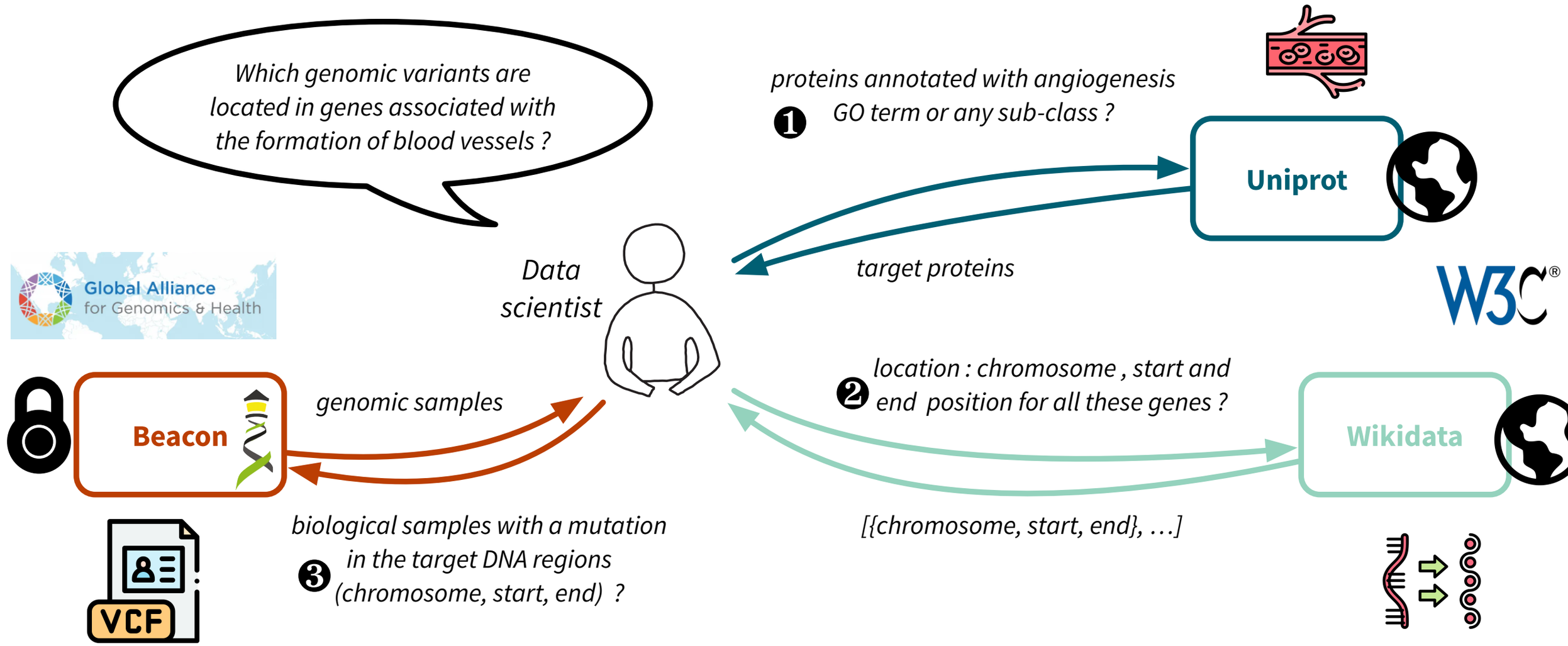
Contributions

(i) a **semantic mapping** for aligning Beacon genomic variations to state-of-the-art ontologies

(ii) an architecture for **on-the-fly FAIRification** of genomic variation data

(iii) a concrete **federated query** showcasing the integration of local genomic variation data and public knowledge graphs

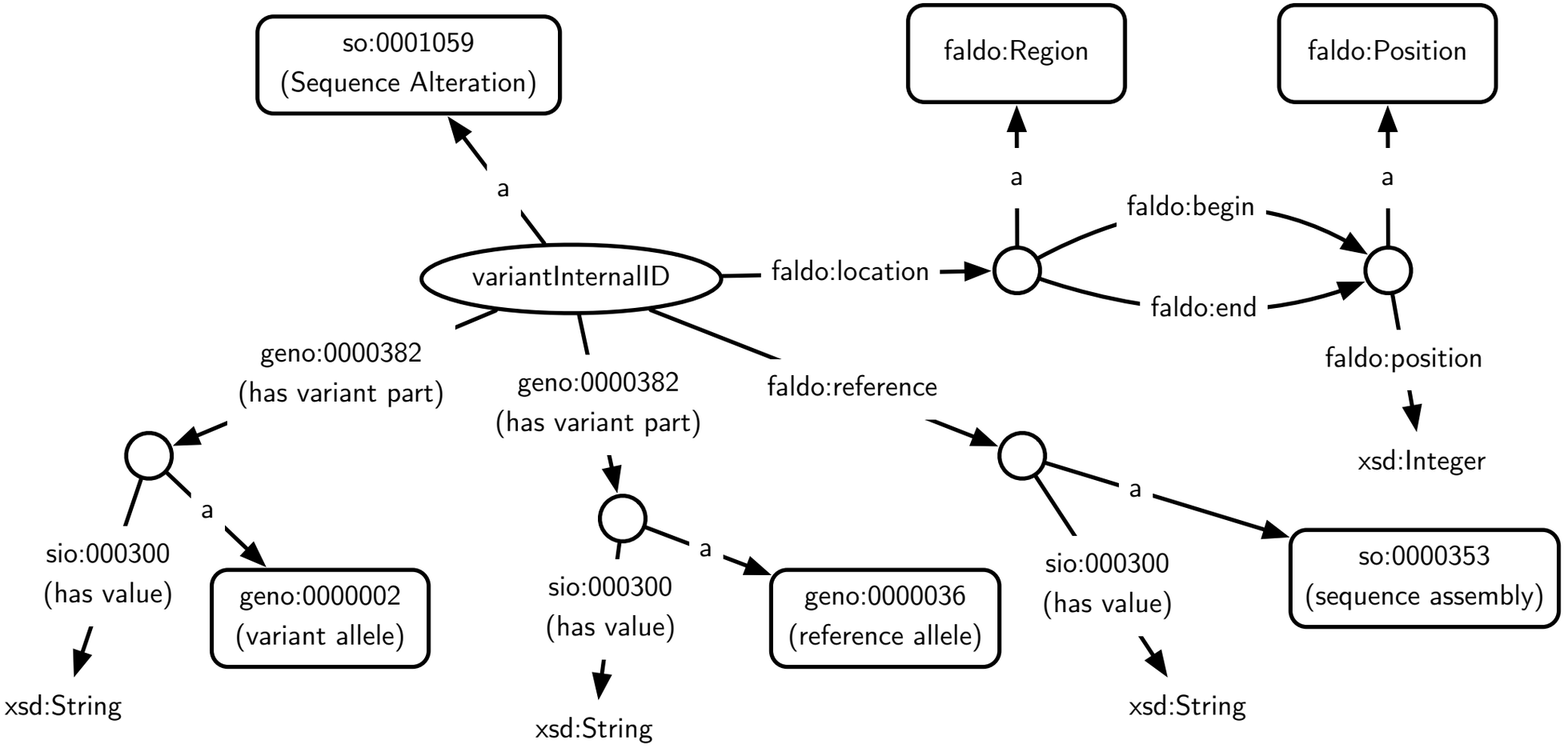
Intracranial aneurysm motivating use case



❶ Semantic representation of genomic variants

No need to create a new ontology ...

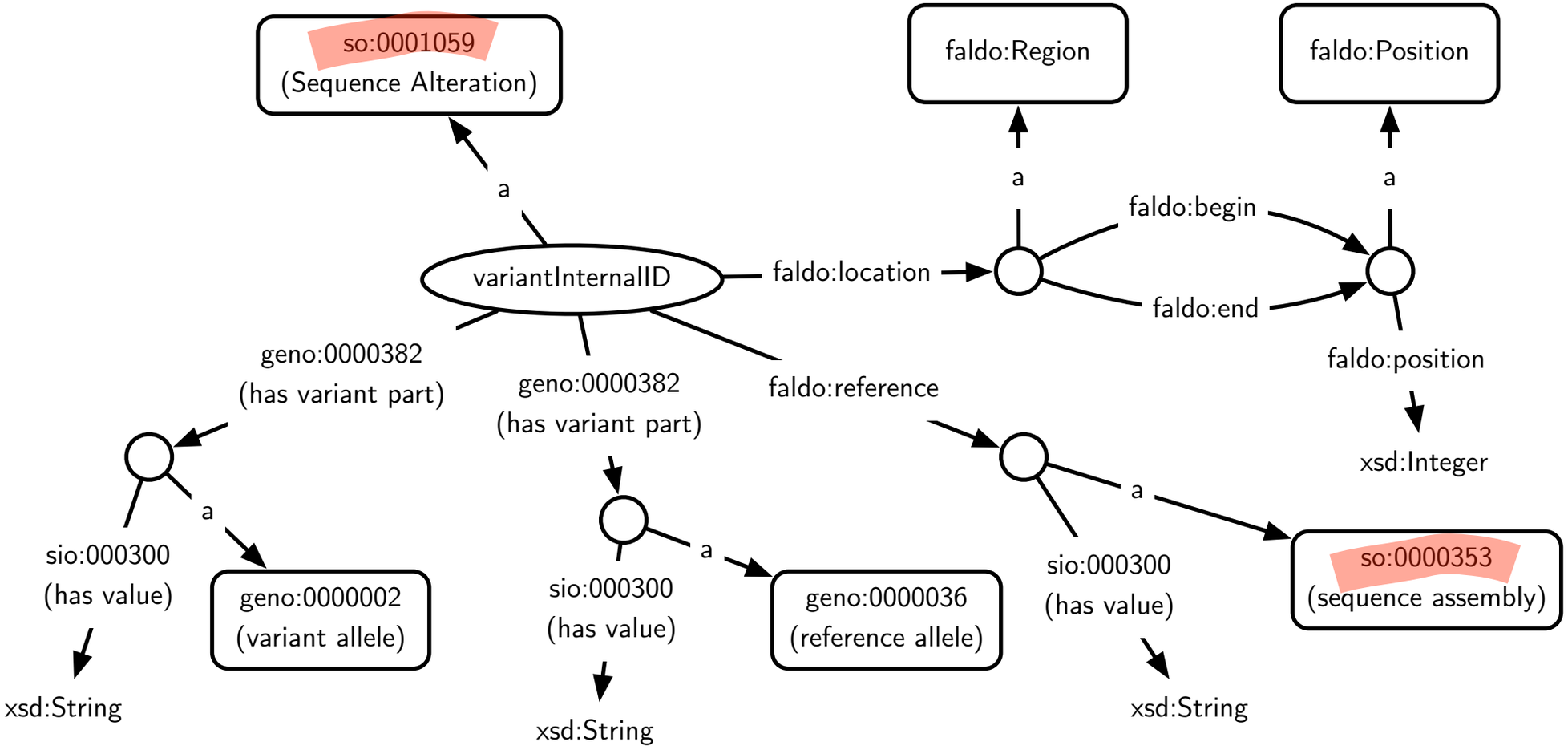
- SO: Sequence Ontology
- GENO: Genotype Ontology
- FALDO: Feature Annotation Location Description Ontology
- SIO: Semantic science Integrated Ontology



❶ Semantic representation of genomic variants

No need to create a new ontology ...

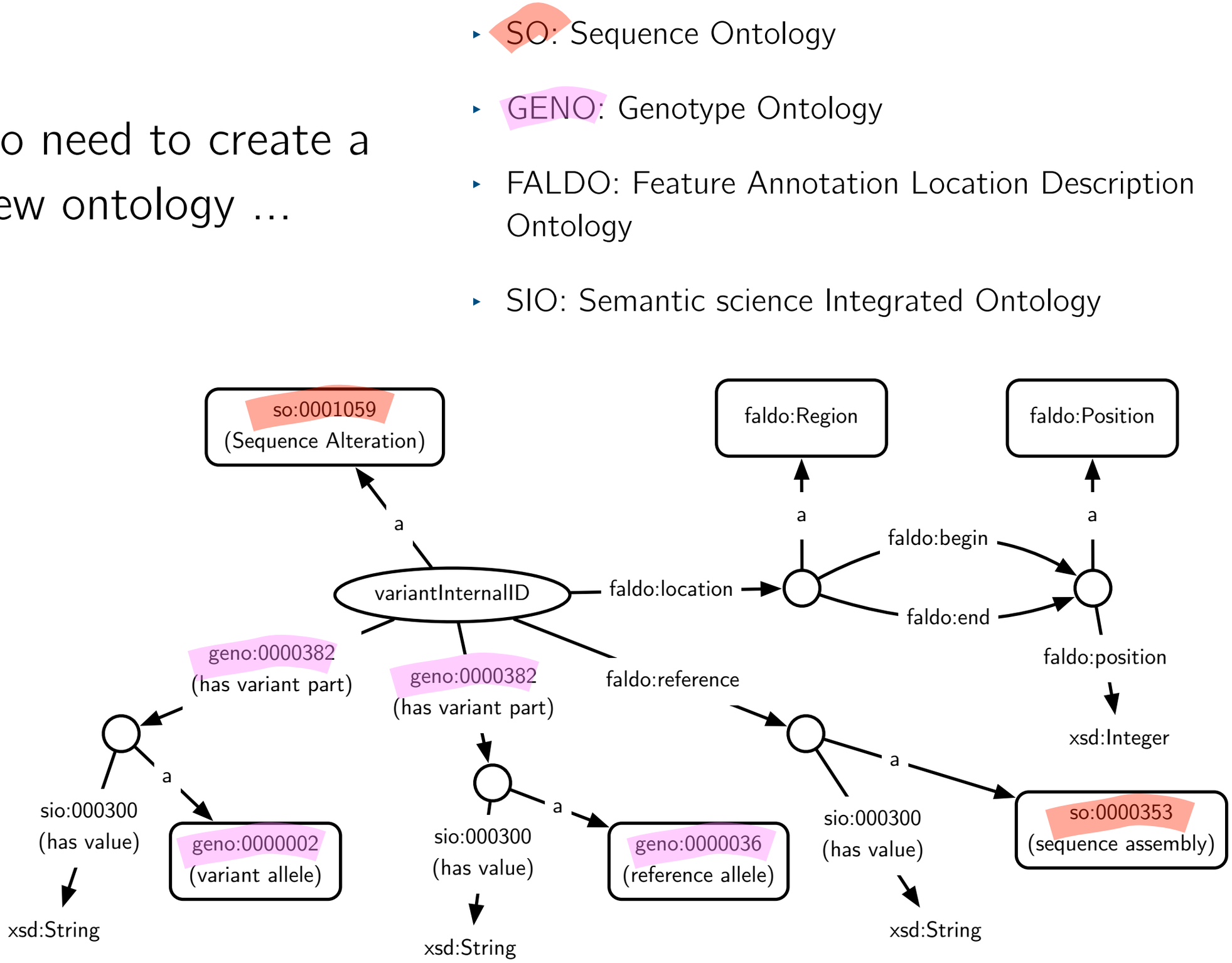
- SO: Sequence Ontology
- GENO: Genotype Ontology
- FALDO: Feature Annotation Location Description Ontology
- SIO: Semantic science Integrated Ontology



❶ Semantic representation of genomic variants

No need to create a new ontology ...

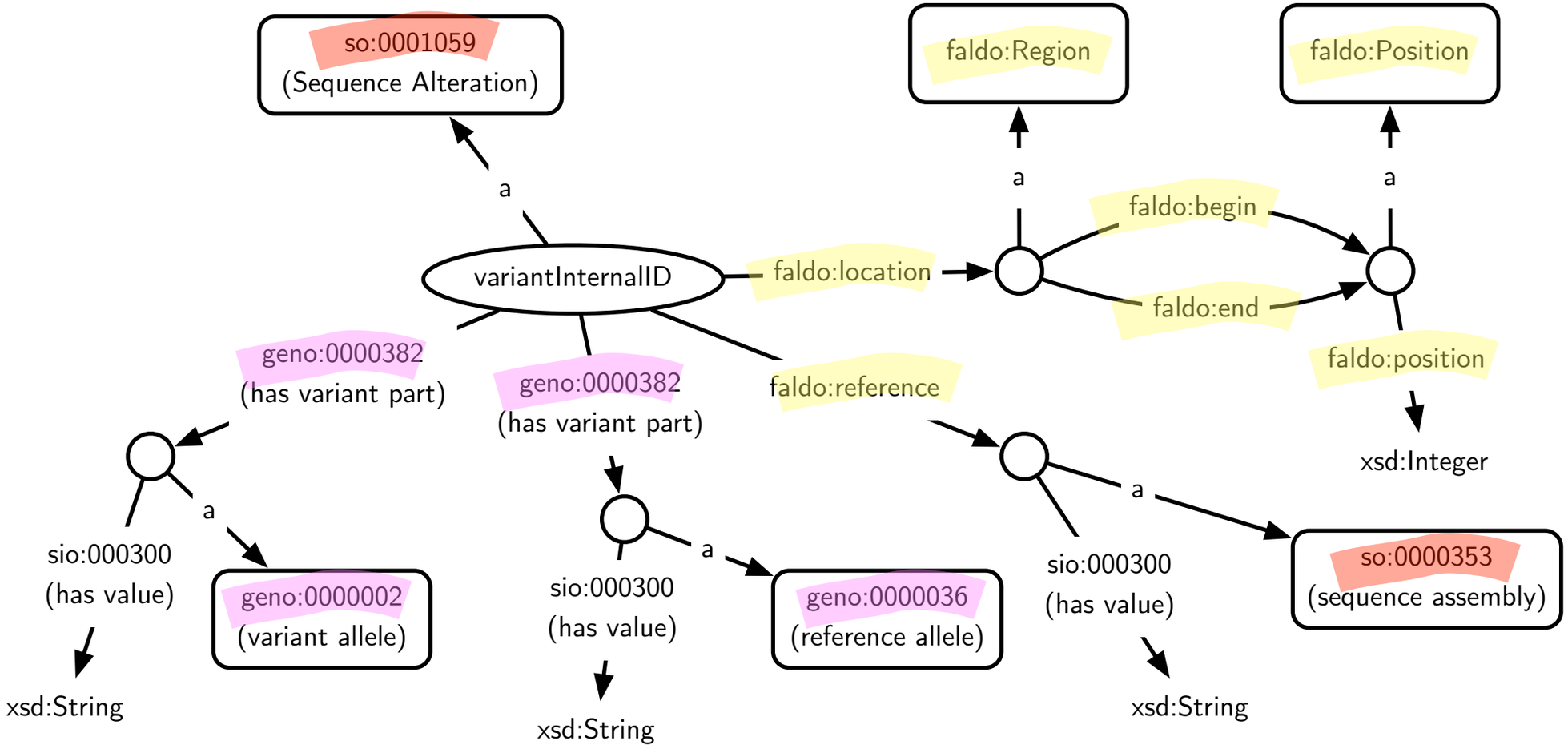
- SO: Sequence Ontology
- GENO: Genotype Ontology
- FALDO: Feature Annotation Location Description Ontology
- SIO: Semantic science Integrated Ontology



❶ Semantic representation of genomic variants

No need to create a new ontology ...

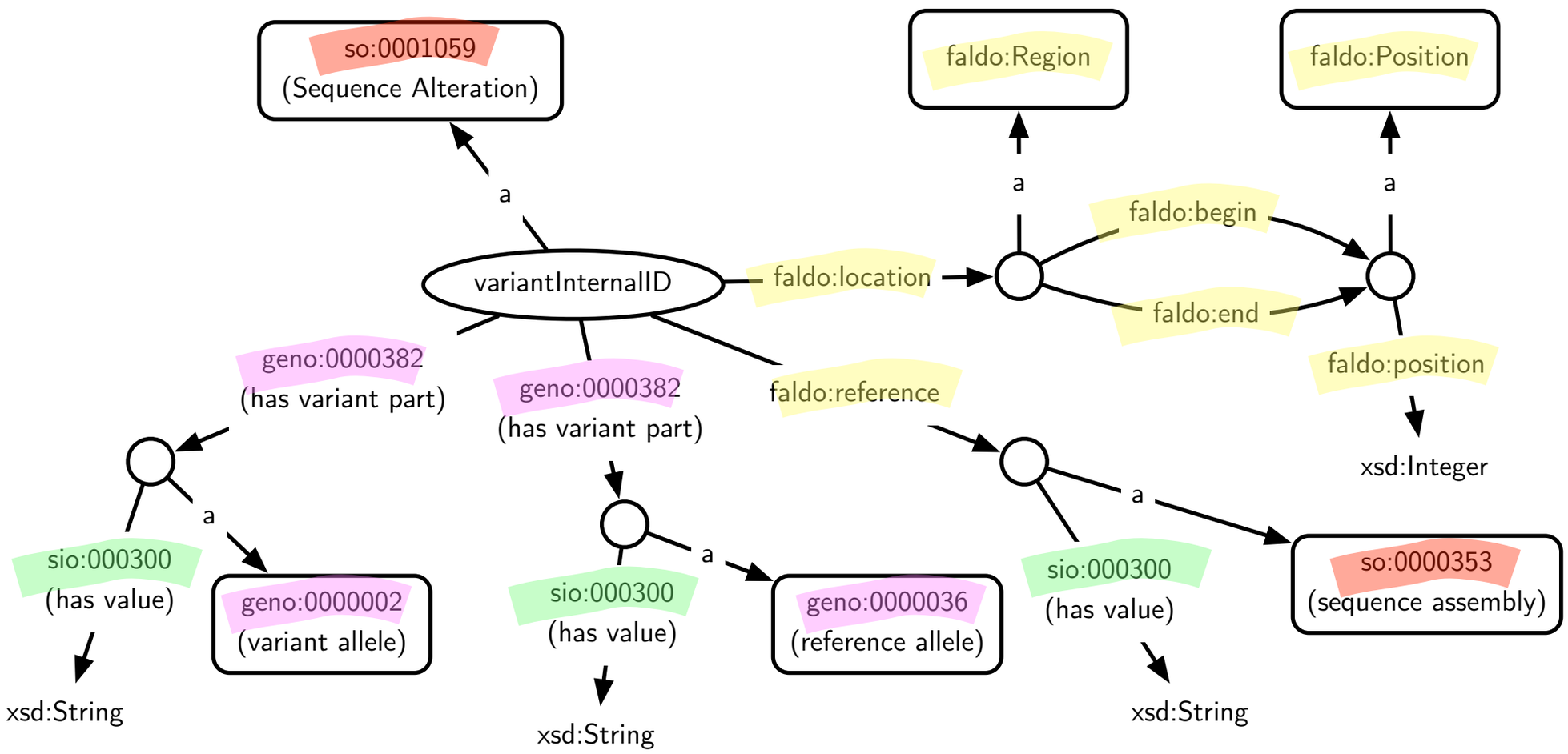
- SO: Sequence Ontology
- GENO: Genotype Ontology
- FALDO: Feature Annotation Location Description Ontology
- SIO: Semantic science Integrated Ontology



❶ Semantic representation of genomic variants

No need to create a new ontology ...

- SO: Sequence Ontology
- GENO: Genotype Ontology
- FALDO: Feature Annotation Location Description Ontology
- SIO: Semantic science Integrated Ontology



❷ RML Mapping for the Beacon API

RML is a declarative mapping language to transform any XML, CSV, JSON structured document into RDF

JSONPath expressions to locate specific data fields.

```
{ "response": { "resultSets":
[ { "results":
  [{ "variation": {
    "location": {
         "interval": {
        "start": { "value": 10093466 }
       } } } }
       ]
}]}}
```

→

```
itx:exact_pos_12345 rdf:type faldo:ExactPosition ;
                            faldo:position 10093466 .
```

RDF Mapping Language (RML)

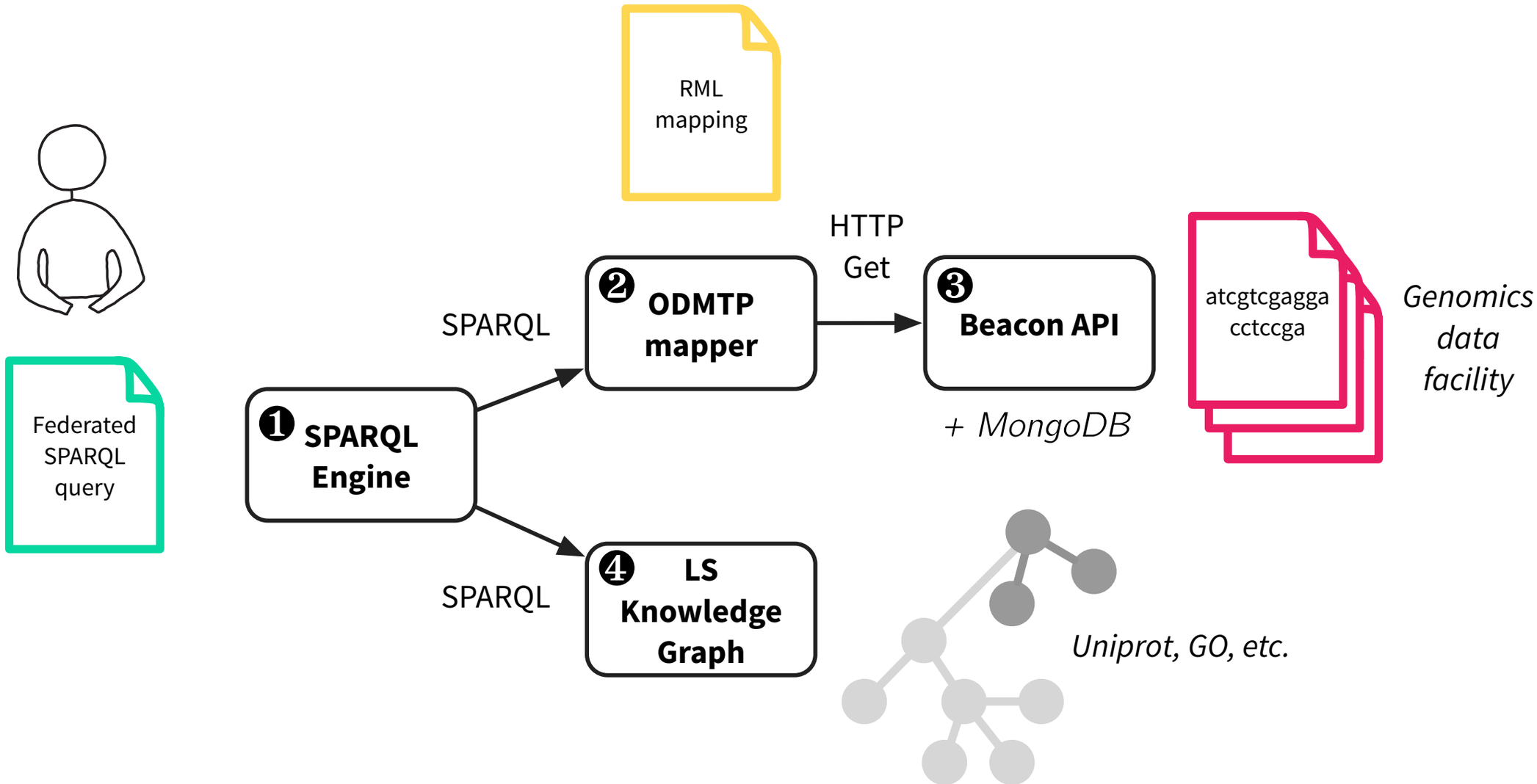
Unofficial Draft 20 June 2024

W3C®

https://rml.io/specs/rml/

```
_:BeginPositionMap a rr:TriplesMap ;
   rml:logicalSource [
      rml:source "reponse_beacon.json" ;
      rml:referenceFormulation ql:JSONPath ;
      rml:iterator "$.response.resultSets[*].results[*].variation.location.interval" ]
   rr:subjectMap [
      rr:template “http://ourlab.org/ressources/exact_pos_{start.value}” ;
      rr:class faldo:ExactPosition ] ;
   rr:predicateObjectMap [
      rr:predicate faldo:position ;
      rr:objectMap [ rml:reference "start.value" ;
         rr:termType rr:Literal ;
[...]
```

❸ Architecture for “Semantic” Beacons



Moreau, B., Serrano-Alvarado, P., Desmontils, E., & Thoumas, D. (2017, October). Querying non-RDF Datasets using Triple Patterns. In *16th International Semantic Web Conference (ISWC2017)*.

Federated SPARQL query

```
SELECT * WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
        up:organism taxon:9606 ;
        up:classifiedWith ?goTerm .
    ?goTerm rdfs:subClassOf* GO:0001525 .
  }
  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinID2)
.

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinID2 ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?startcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    ?wgse wdps:P645 ?endcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    FILTER(lang(?assembly) = "en")
    FILTER(STR(?assembly) = "genome assembly GRCh38")
  }

  ?variant a so:0001059 ;
    faldo:reference/sio:SIO_000300 ?chromosome ;
    faldo:location/faldo:begin/faldo:position ?v_start ;
    faldo:location/faldo:end/faldo:position ?v_end .

  FILTER( (((?v_start >= xsd:integer(?startcoordinate)) &&
            (?v_start <= xsd:integer(?endcoordinate)) ))
        || ((?v_end >= xsd:integer(?startcoordinate)) &&
            (?v_end <= xsd:integer(?endcoordinate))) )
}
LIMIT 10
```

Beacon

Federated SPARQL query

```
SELECT * WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
        up:organism taxon:9606 ;
        up:classifiedWith ?goTerm .
    ?goTerm rdfs:subClassOf* GO:0001525 .
  }
  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinID2)
.

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinID2 ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?startcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    ?wgse wdps:P645 ?endcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    FILTER(lang(?assembly) = "en")
    FILTER(STR(?assembly) = "genome assembly GRCh38")
  }

  ?variant a so:0001059 ;
    faldo:reference/sio:SIO_000300 ?chromosome ;
    faldo:location/faldo:begin/faldo:position ?v_start ;
    faldo:location/faldo:end/faldo:position ?v_end .

  FILTER( (((?v_start >= xsd:integer(?startcoordinate)) &&
            (?v_start <= xsd:integer(?endcoordinate)) ))
        || ((?v_end >= xsd:integer(?startcoordinate)) &&
            (?v_end <= xsd:integer(?endcoordinate))) )
}
LIMIT 10
```

SERVICE clauses for each remote data source

Federated SPARQL query

```
SELECT * WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
        up:organism taxon:9606 ;
        up:classifiedWith ?goTerm .
    ?goTerm rdfs:subClassOf* GO:0001525 .
  }
  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinID2)
.

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinID2 ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?startcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    ?wgse wdps:P645 ?endcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    FILTER(lang(?assembly) = "en")
    FILTER(STR(?assembly) = "genome assembly GRCh38")
  }

  ?variant a so:0001059 ;
    faldo:reference/sio:SIO_000300 ?chromosome ;
    faldo:location/faldo:begin/faldo:position ?v_start ;
    faldo:location/faldo:end/faldo:position ?v_end .

  FILTER( (((?v_start >= xsd:integer(?startcoordinate)) &&
            (?v_start <= xsd:integer(?endcoordinate)) ))
        || ((?v_end >= xsd:integer(?startcoordinate)) &&
            (?v_end <= xsd:integer(?endcoordinate))) )
}
LIMIT 10
```

SERVICE clauses for each remote data source

All human proteins (`taxon:9606`) classified with all sub-classes of “angiogenesis” (`GO:0001525`)

Federated SPARQL query

```
SELECT * WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
        up:organism taxon:9606 ;
        up:classifiedWith ?goTerm .
    ?goTerm rdfs:subClassOf* GO:0001525 .
  }
  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinID2)
.

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinID2 ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?startcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    ?wgse wdps:P645 ?endcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    FILTER(lang(?assembly) = "en")
    FILTER(STR(?assembly) = "genome assembly GRCh38")
  }

  ?variant a so:0001059 ;
    faldo:reference/sio:SIO_000300 ?chromosome ;
    faldo:location/faldo:begin/faldo:position ?v_start ;
    faldo:location/faldo:end/faldo:position ?v_end .

  FILTER( (((?v_start >= xsd:integer(?startcoordinate)) &&
            (?v_start <= xsd:integer(?endcoordinate)) ))
        || ((?v_end >= xsd:integer(?startcoordinate)) &&
            (?v_end <= xsd:integer(?endcoordinate))) )
}
LIMIT 10
```

SERVICE clauses for each remote data source

All human proteins (`taxon:9606`) classified with all sub-classes of "angiogenesis" (`GO:0001525`)

For the matching proteins, get the location of the encoding genes for the GRCh38 reference human genome

Federated SPARQL query

```
SELECT * WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
        up:organism taxon:9606 ;
        up:classifiedWith ?goTerm .
    ?goTerm rdfs:subClassOf* GO:0001525 .
  }
  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinID2)
.

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinID2 ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?startcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    ?wgse wdps:P645 ?endcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    FILTER(lang(?assembly) = "en")
    FILTER(STR(?assembly) = "genome assembly GRCh38")
  }

  ?variant a so:0001059 ;
    faldo:reference/sio:SIO_000300 ?chromosome ;
    faldo:location/faldo:begin/faldo:position ?v_start ;
    faldo:location/faldo:end/faldo:position ?v_end .

  FILTER( (((?v_start >= xsd:integer(?startcoordinate)) &&
            (?v_start <= xsd:integer(?endcoordinate)) ))
        || ((?v_end >= xsd:integer(?startcoordinate)) &&
            (?v_end <= xsd:integer(?endcoordinate))) )
}
LIMIT 10
```

SERVICE clauses for each remote data source

All human proteins (`taxon:9606`) classified with all sub-classes of "angiogenesis" (`GO:0001525`)

For the matching proteins, get the location of the encoding genes for the GRCh38 reference human genome

All local genomic variants matching the localization constraints

Conclusion and future works

Take-home message

- Beacon is great for **privacy**-preserving **genomic data discovery**
- However, it has a **limited interoperability** with public knowledge graphs such as Uniprot
- Many ontologies are available to represent **genomic data as knowledge graphs**
- This approach preserves **decentralization and data source autonomy** through federated SPARQL queries.

- **Future works** include
  - addressing **scalability issues** (costly aggregate queries, non-selective queries on remote sources, distributed joins ...)
  - addressing **security issues in knowledge graph federations** → SAFE-KG ANR project
    - safe federated query formulation (LLM)
    - safe and efficient federated query execution
    - decentralized access and usage policies, traceability and explainability

Acknowledgments

anr®

FRANCE 2030 PROGRAMME DE RECHERCHE SANTÉ NUMÉRIQUE

NEUROVASC PEPR SANTÉ NUMÉRIQUE

Gabriella Montoya

Alexandrina Bodrug-Shepers

Patricia Serrano-Alvarado

Richard Redon

Hugo Chabane

Semantic Beacons: a framework to support federated querying over genomic variants and public Knowledge Graphs

Alexandrina Bodrug-Schepers[1,†], Hugo Chabane[2,†], Gabriela Montoya[2], Patricia Serrano-Alvarado[2], Richard Redon[1] and Alban Gaignard[1,3]

[1]Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France

[2]Nantes Université, LS2N, Nantes, France

[3]IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

https://hal.science/hal-04908530v2

contact: alban.gaignard@univ-nantes.fr

Inserm cnrs Nantes Université CHU NANTES

LS2N l'institut du thorax

Back-up slides

Performance issues

- Search and aggregate query over Beacon are costly
  - the reference implementation could be optimized (better indexing, etc.)
- Annotating all variants of a VCF file could overwhelm the remote KG (e.g. Uniprot)
- Non-selective enough queries on Uniprot can be overwhelming for the beacon endpoint

Security challenges

- Genomic data are **inherently identifying**
  - need AuthN and AuthZ for record-level access
  - can we detect re-identifying patterns at query design time ?
  - can we detect re-identifying risks at query run time ?
- Working with **aggregates is a trade-off**
  but far from perfect:
  - "Beacon reconstruction attack"
    https://doi.org/10.1093/bioinformatics/btaf273


JOURNAL ARTICLE

Beacon Reconstruction Attack: Reconstruction of genomes in genomic data-sharing beacons using summary statistics

Kousar Saleem, A Ercument Cicek, Sinem Sav

Bioinformatics, Volume 41, Issue 6, June 2025, btaf273,
https://doi.org/10.1093/bioinformatics/btaf273
Published: 19 May 2025 **Article history**

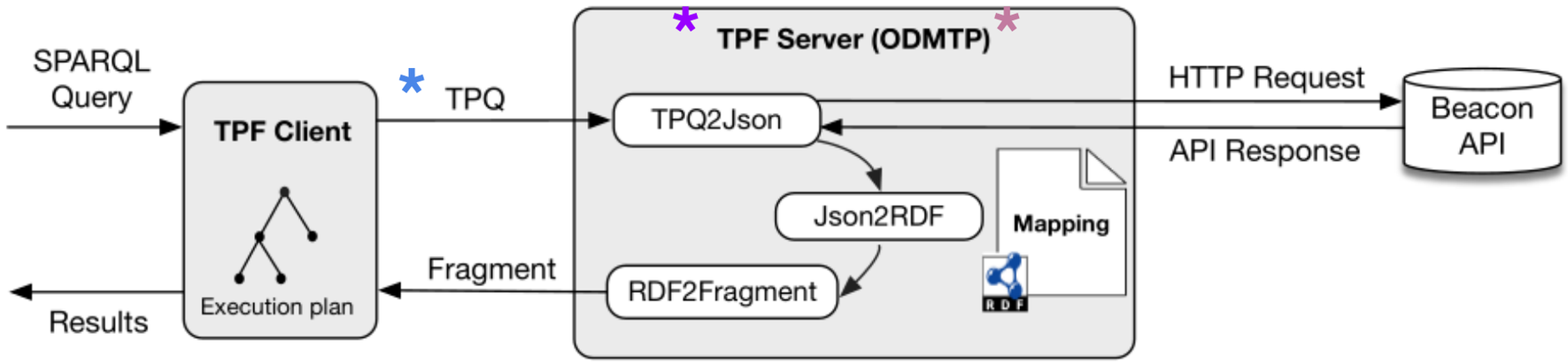
PDF Split View Cite Permissions Share

- Potential **data leaks in distributed joins**
  - can we prevent “data leaks” at query design time ?
  - can we verify it at federated query run time ?

On the fly conversion Beacon API ←→ Linked Data

TPF Server → **converts TPQ into HTTP requests**
(compatible with Beacon API)

API response → mapped to RDF triples (**following our Mapping**)



* Triple Pattern Query

* Triple Pattern Fragment

* On-Demand Mapping using Triple Patterns