



PROGRAMME
DE RECHERCHE
SANTÉ
NUMÉRIQUE



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

Alban Gaignard

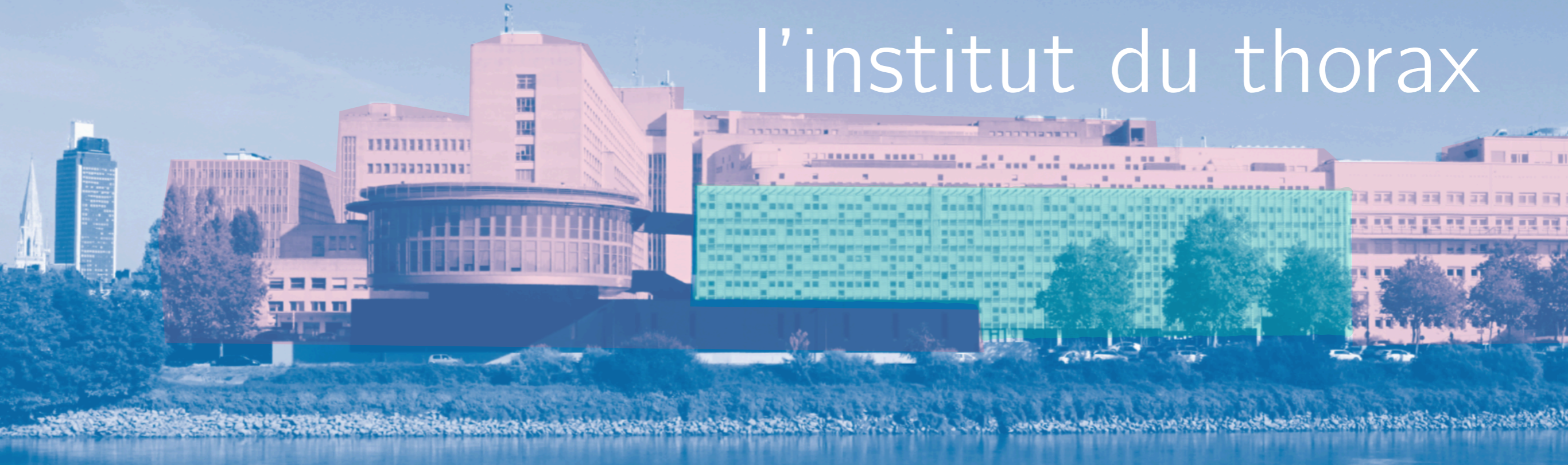
CNRS, institut du thorax, Nantes, France

Inserm webinar

March 20, 2026

Introduction

l'institut du thorax



Better understanding of cardiovascular 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 cardiovascular and metabolic diseases

Gene-function associations

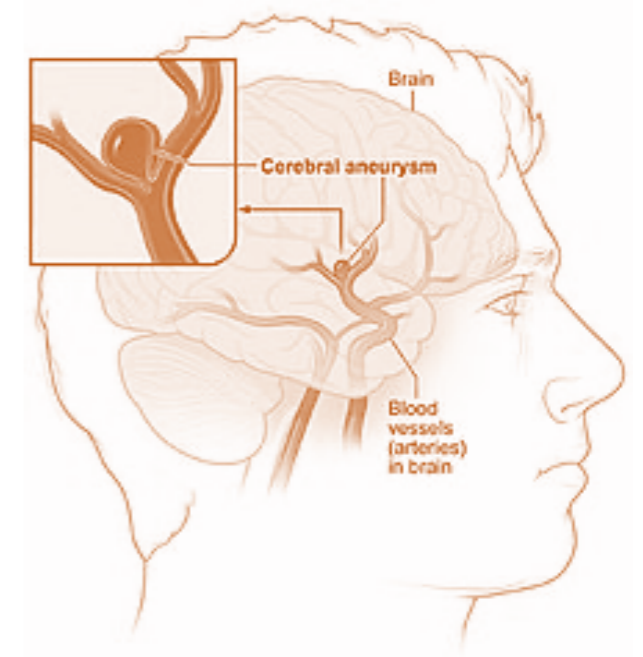
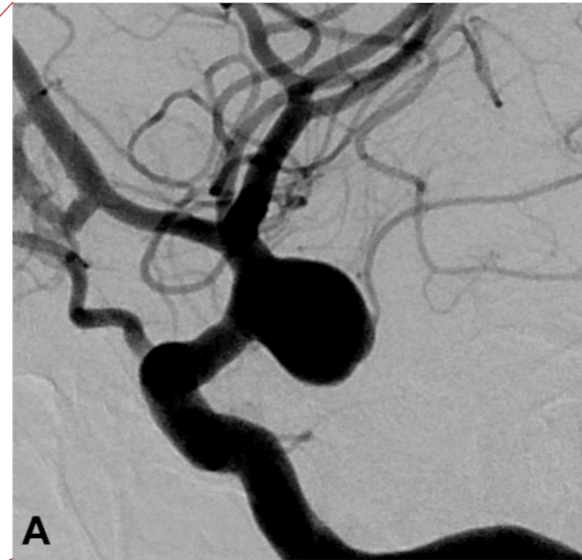
Translational medicine

Real proximity between the university hospital and the research lab

Bioinformatics

- ▶ Massive production of genomic sequence & health data
→ 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

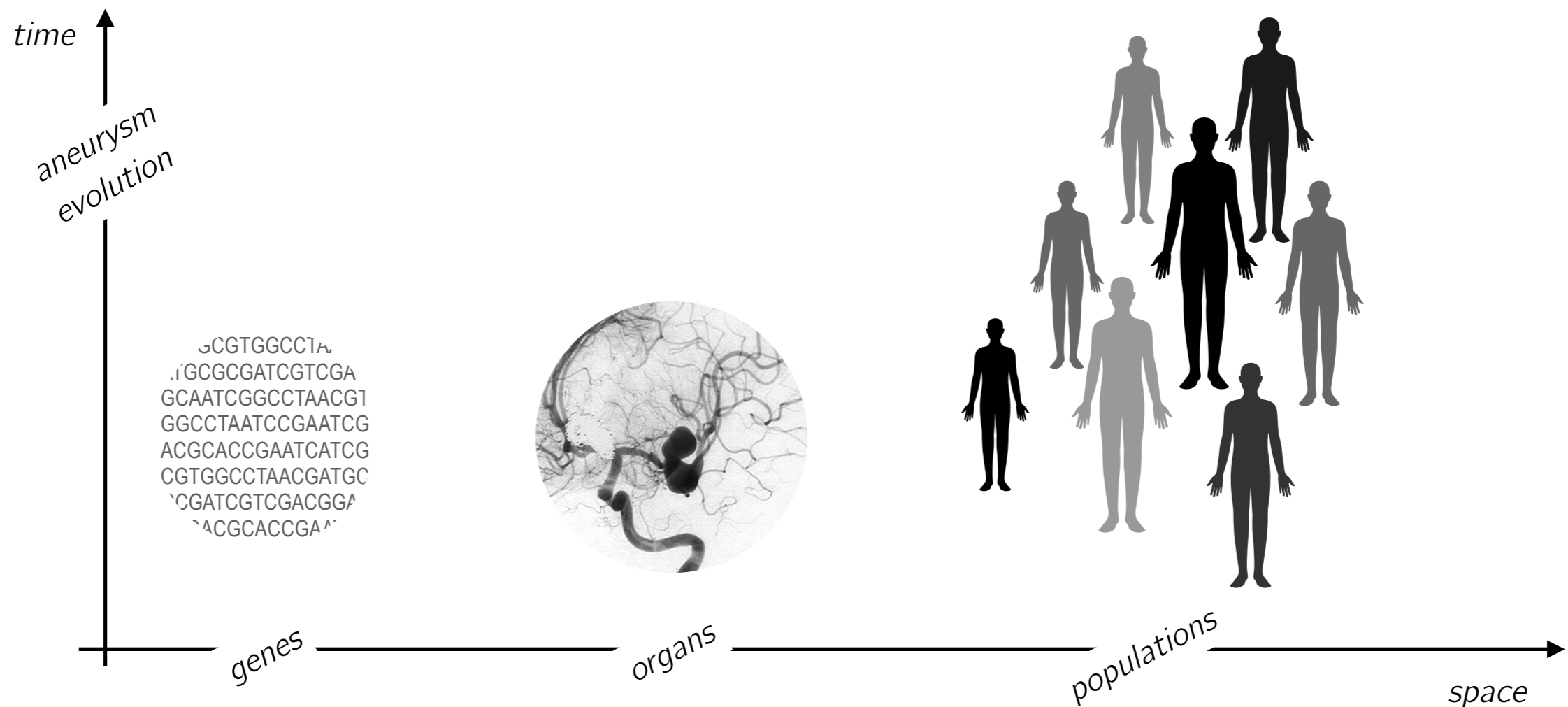
Multi-factorial disease → multi-scale data

Multi-factorial disease → multi-scale data

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

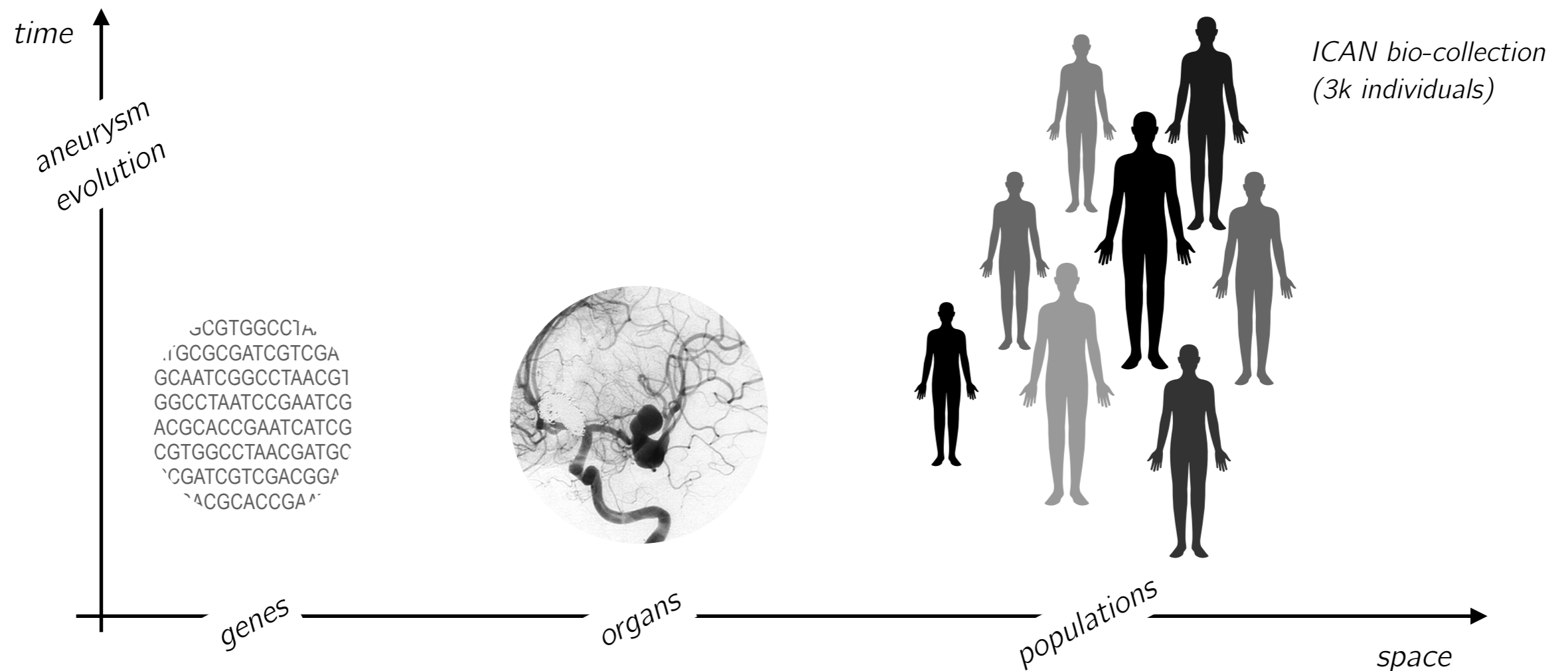
Multi-factorial disease → multi-scale data

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

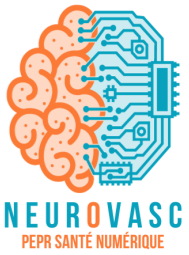


Multi-factorial disease → multi-scale data

- ▶ 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

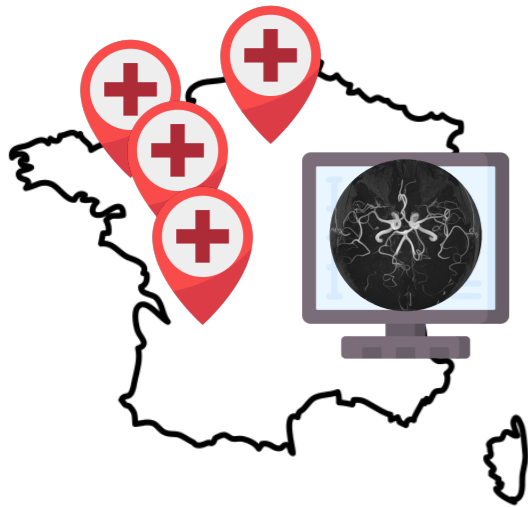
WP 3: Proof-of-Concept studies with digital companions

- Task 3.1: Accompanying patients with diagnosed unruptured ICA
- Task 3.2: Post-stroke prevention through mobile applications and digital monitoring

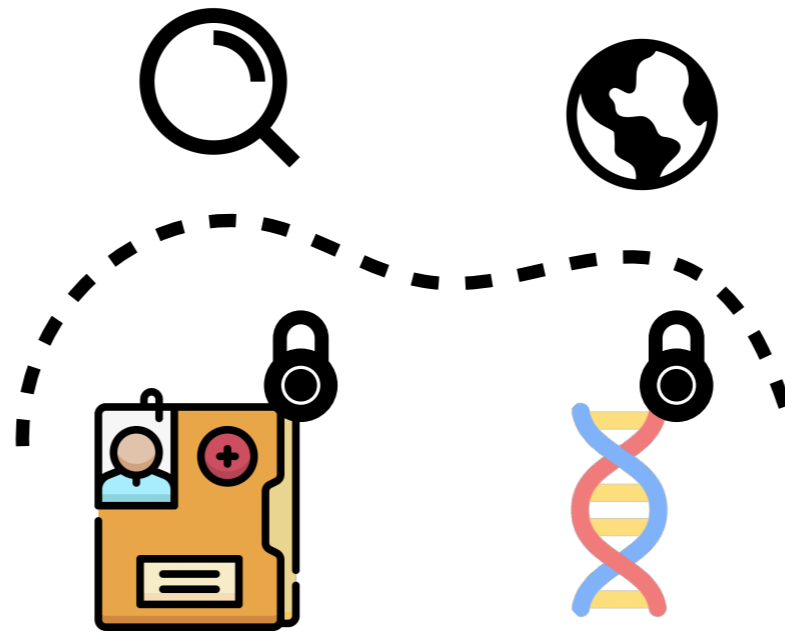


Data integration & sharing challenges

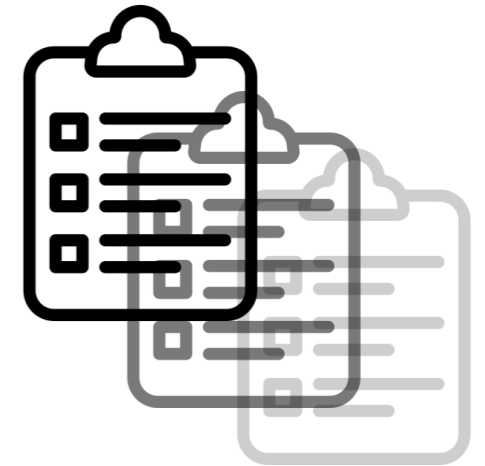
❶ How to **collect high-quality medical images** from multiple hospitals/ MRIs ?



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 ?





Alexandrina
Bodrug

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

Alexandrina Bodrug-Schepers^{1,†}, Hugo Chabane^{2,†}, Gabriela Montoya²,
Patricia Serrano-Alvarado², Richard Redon¹ and Alban Gaignard^{1,3}

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

²Nantes Université, LS2N, Nantes, France

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

Bridging genomic data and public knowledge bases ?

The UniProt public knowledge graph

The screenshot displays the UniProt entry for P63000 · RAC1_HUMAN. The interface includes a top navigation bar with search options (BLAST, Align, Peptide search, ID mapping, SPARQL, UniProtKB) and a search bar. A left sidebar lists various categories: Function, Names & Taxonomy, Subcellular Location, Disease & Variants, PTM/Processing, Expression, Interaction, Structure, Family & Domains, Sequence & Isoform, and Similar Proteins. The main content area features a header with the protein name and a summary table:

Protein ⁱ	Ras-related C3 botulinum toxin substrate 1	Amino acids	192 (go to sequence)
Gene ⁱ	RAC1	Protein existence ⁱ	Evidence at protein level
Status ⁱ	UniProtKB reviewed (Swiss-Prot)	Annotation score ⁱ	5/5
Organism ⁱ	Homo sapiens (Human)		

Below the summary table are tabs for Entry, Variant viewer (399), Feature viewer, Genomic coordinates, Publications, External links, and History. A toolbar offers options for Tools, Download, Add, Add a publication, and Entry feedback.

Functionⁱ

Plasma membrane-associated small GTPase which cycles between active GTP-bound and inactive GDP-bound states. In its active state, binds to a variety of effector proteins to regulate cellular responses such as secretory processes, phagocytosis of apoptotic cells, epithelial cell polarization, neurons adhesion, migration and differentiation, and growth-factor induced formation of membrane ruffles (PubMed:1643658, PubMed:22843693, PubMed:23512198, PubMed:28886345).

Rac1 p21/rho GDI heterodimer is the active component of the cytosolic factor sigma 1, which is involved in stimulation of the NADPH oxidase activity in macrophages. Essential for the SPATA13-mediated regulation of cell migration and adhesion assembly and disassembly. Stimulates PKN2 kinase activity (PubMed:9121475).

In concert with RAB7A, plays a role in regulating the formation of RBs (ruffled borders) in osteoclasts (PubMed:1643658).

In podocytes, promotes nuclear shuttling of NR3C2; this modulation is required for a proper kidney functioning. Required for atypical chemokine receptor ACKR2-induced LIMK1-PAK1-dependent phosphorylation of cofilin (CFL1) and for up-regulation of ACKR2 from endosomal compartment to cell membrane, increasing its efficiency in chemokine uptake and degradation. In neurons, is involved in dendritic spine formation and synaptic plasticity (By similarity).

In hippocampal neurons, involved in spine morphogenesis and synapse formation, through local activation at synapses by guanine nucleotide exchange factors (GEFs), such as ARHGEF6/ARHGEF7/PIX (PubMed:12695502).

In synapses, seems to mediate the regulation of F-actin cluster formation performed by SHANK3. In neurons, plays a crucial role in regulating GABA(A) receptor synaptic stability and hence GABAergic inhibitory synaptic transmission through its role in PAK1 activation and eventually F-actin stabilization (By similarity).

Required for DSG3 translocation to cell-cell junctions, DSG3-mediated organization of cortical F-actin bundles and anchoring of actin at cell junctions; via interaction with DSG3 (PubMed:22796473).

Subunit of the phagocyte NADPH oxidase complex that mediates the transfer of electrons from cytosolic NADPH to O₂ to produce the superoxide anion (O₂⁻) (PubMed:38355798). [By Similarity](#)

8 Publications

Isoform B

Isoform B has an accelerated GEF-independent GDP/GTP exchange and an impaired GTP hydrolysis, which is restored partially by GTPase-activating proteins (PubMed:14625275).

It is able to bind to the GTPase-binding domain of PAK but not full-length PAK in a GTP-dependent manner, suggesting that the insertion does not completely abolish effector interaction (PubMed:14625275).

1 Publication

Caution

The interaction between DSCAM, PAK1 and RAC1 has been described. This article has been withdrawn by the authors. 2 Publications

Catalytic activityⁱ

Rhea:19669 [↗](#)

GTP + H₂O = GDP + phosphate + H⁺ 1 Publication

This reaction proceeds in the forward direction. [↗](#)

The UniProt public knowledge graph

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB

P63000 · RAC1_HUMAN

Protein¹ Ras-related C3 botulinum toxin domain protein 1
 Gene¹ RAC1
 Status¹ UniProtKB reviewed (Swiss-Prot)
 Organism¹ Homo sapiens (Human)

Entry Variant viewer 399 Feature viewer

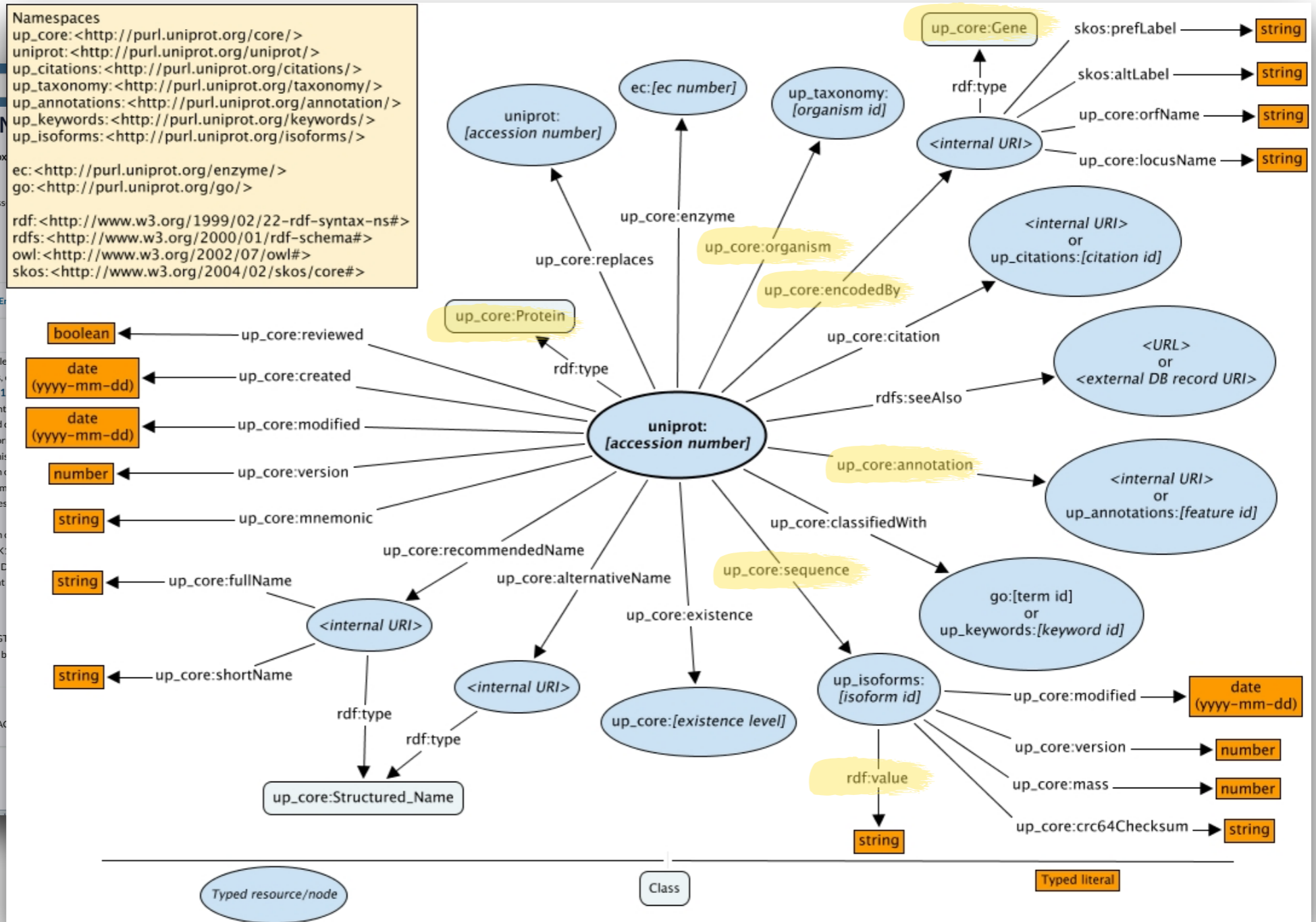
Tools Download Add Add a publication

Function¹

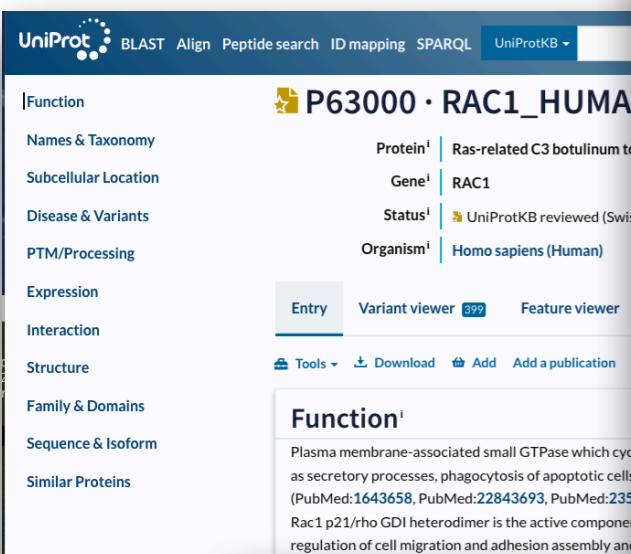
Plasma membrane-associated small GTPase which cycles between GTP-bound and GDP-bound states as secretory processes, phagocytosis of apoptotic cells, and cell migration and adhesion assembly and regulation. Rac1 p21/rho GDI heterodimer is the active component. In concert with RAB7A, plays a role in regulating the formation of podocytes, promotes nuclear shuttling of NR3C2; this phosphorylation of cofilin (CFL1) and for up-regulation of dendritic spine formation and synaptic plasticity (By similarity). In hippocampal neurons, involved in spine morphogenesis (PubMed:12695502). In synapses, seems to mediate the regulation of F-actin and inhibitory synaptic transmission through its role in PAK1. Required for DSG3 translocation to cell-cell junctions. Dimeric subunit of the phagocyte NADPH oxidase complex that is involved in the production of superoxide anion.

Catalytic activity¹

Rhea:19669
 GTP + H₂O = GDP + phosphate + H⁺



The UniProt public knowledge graph



UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB

P63000 · RAC1_HUMAN

Function: Plasma membrane-associated small GTPase which cycles as secretory processes, phagocytosis of apoptotic cells, Rac1 p21/rho GDI heterodimer is the active component regulation of cell migration and adhesion assembly and

Namespaces

- up_core: <http://purl.uniprot.org/core/>
- uniprot: <http://purl.uniprot.org/uniprot/>
- up_citations: <http://purl.uniprot.org/citations/>
- up_taxonomy: <http://purl.uniprot.org/taxonomy/>
- up_annotations: <http://purl.uniprot.org/annotation/>
- up_keywords: <http://purl.uniprot.org/keywords/>
- up_isoforms: <http://purl.uniprot.org/isoforms/>

ec: <http://purl.uniprot.org/enzyme/>
go: <http://purl.uniprot.org/go/>

rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
owl: <http://www.w3.org/2002/07/owl#>
skos: <http://www.w3.org/2004/02/skos/core#>

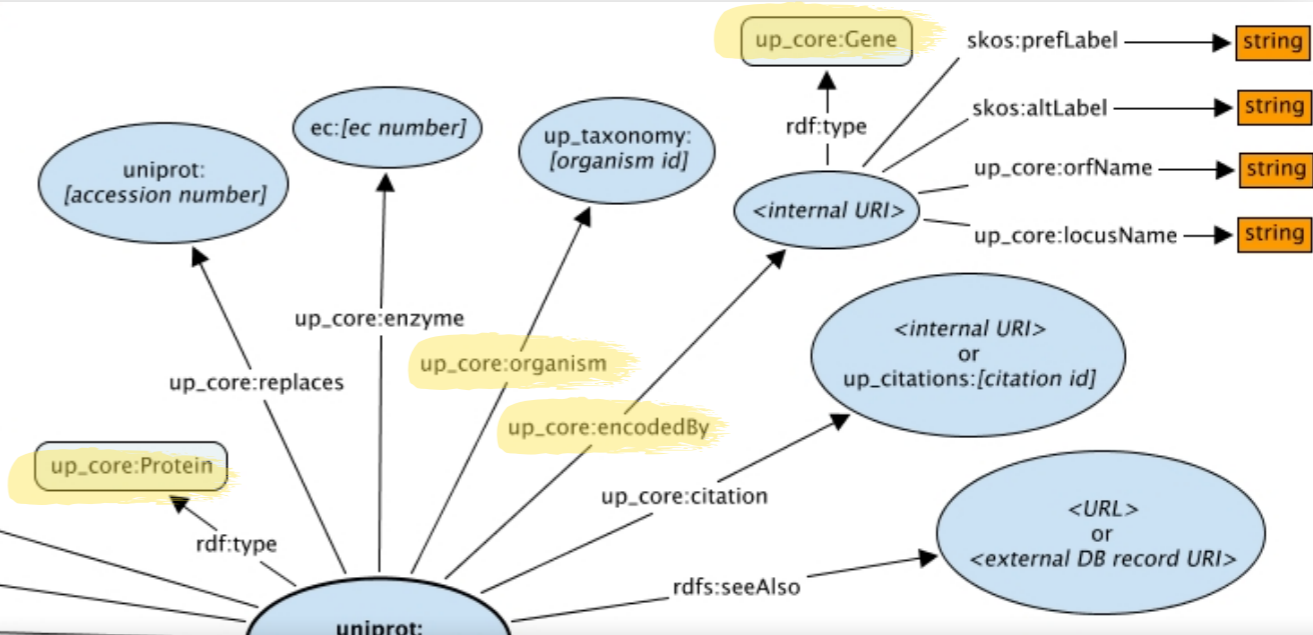


Diagram illustrating the UniProt public knowledge graph structure. The central node is **uniprot:**. It is connected to various entities and properties:

- uniprot: [accession number]** (Property)
- ec: [ec number]** (Property)
- up_taxonomy: [organism id]** (Property)
- up_core:Gene** (Class)
- up_core:Protein** (Class)
- up_core:enzyme** (Class)
- up_core:organism** (Class)
- up_core:encodedBy** (Property)
- up_core:citation** (Property)
- up_core:replaces** (Property)
- up_core:reviewed** (Property, type: boolean)
- up_core:created** (Property, type: date (yyyy-mm-dd))
- up_core:modified** (Property, type: date (yyyy-mm-dd))
- <internal URI>** (Property)
- <internal URI> or up_citations: [citation id]** (Property)
- <URL> or <external DB record URI>** (Property)

```

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
"MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCVIEGHLQILLMFKTRPEDFRDLSFPKLIIMITDYLLLFRVYGLESKDLFPNLTVIRGS
RLFFNYALVIFEMVHLKELGLYNLMNITRGSVRIEKNELCYLATIDWSRILDSVEDNYIVLNKDDNEECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTICKS
HGCTAEGLCCHSECLGNC SQPDDPTKCVACRNFYLDGRCVETCPPPYHFQDWRVCVNF SFCQDLHHKCKNSRRQGCHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKV
CHLLEGEKTI DSVTSAQELRGCTVINGSLI INIRGGNNLAAELEANLGLIEEISGYLKIRRSYALVSLSFPRKRLRIRGETLEIGNYSFYALDNQNLRLQLDWWSKHNLITIQG
KLFPHYNPKLCLSEIHKMEEVSGTKGRQERNDIALKTNGDQASCENELLKFSYIRTSFDKILLRWEPYWPPDFRDLLGFMLFYKEAPYQNVTEFDGQDACGSNSWTVVDIDPP
LRSNDPKSQNHGWL MRGLKPWTQYAI FVKTLVTFSDERRTYGAKSDI IYVQTDATNPSVPLDPI SVSNSSSQI ILLKWKPPSDPNGNITHYLVFVERQAEDSELFELDYLKLG
LKLPSRTWSPFFESEDSQKHNSQSEYEDSAGECCSCPKTD S QILKELEESSFRKTFEDYLHNVVVFPKRTSSGTGAEDPRPSRKRRSLGDVGNVTVAVPTVAAFPNTSSTSVPT
SPEEHRPF EKVVNKE SLVISGLRHFTGYRIELQACNQDTP EERC SVAAVVSARTMPEAKADDIVGPVTHEIFENNVVHLMWQEPKEPNGLIVLYEVS YRRYGDEELHLCVSRK
HFALERGCRLRGLSPGNYSVRIRATSLAGNGSWTEPTYFYVTDYLDVPSNIAKIIIGPLIFVFLFSVVIGSIYLF LRKRQPDGPLGPLYASSNPEYLSASDVFP CSVYVPDEW
EVSREKITLLRELGQGSFGMVYEGNARDI IKGEAETRVAVKTVNESASLRERIEFLNEASVMKGF TCHHVVRLLGVVSKGQPTLVVMELMAHGD LKSYLRSLRPEAENNPGRP
PPTLQEMI QMAAE IADGMAYLNAKKFVHRDLAARNCMVAHDFTVKIGDFGMTRDI YETDYRKGKGLLPVRWMAPE SLKDG VFTTSSDMWSFGVVLWEITSLAEQPYQGLSN
EQVLK FVMDGGYLDQPDNCPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEV SFFHSEENKAP ESELEMEFEDMENVPLDRSSH CQREEAGGRDGGSSLGFKRSY
EEHIPYTHMNGGKKNGRILTLPRSNPS" .
    
```

The UniProt public knowledge graph

The image shows a UniProtKB entry for P63000 (RAC1_HUMAN) on the left and a diagram of the UniProt knowledge graph on the right. The diagram illustrates the central 'uniprot:' node and its relationships to various entities and properties.

Namespaces:

- up_core: <http://purl.uniprot.org/core/>
- uniprot: <http://purl.uniprot.org/uniprot/>
- up_citations: <http://purl.uniprot.org/citations/>
- up_taxonomy: <http://purl.uniprot.org/taxonomy/>
- up_annotations: <http://purl.uniprot.org/annotation/>
- up_keywords: <http://purl.uniprot.org/keywords/>
- up_isoforms: <http://purl.uniprot.org/isoforms/>
- ec: <http://purl.uniprot.org/enzyme/>
- go: <http://purl.uniprot.org/go/>
- rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- owl: <http://www.w3.org/2002/07/owl#>
- skos: <http://www.w3.org/2004/02/skos/core#>

Graph Structure:

- uniprot:** (Central node)
 - up_core:replaces → uniprot:[accession number]
 - up_core:enzyme → ec:[ec number]
 - up_core:organism → up_taxonomy:[organism id]
 - up_core:encodedBy → up_core:Gene
 - up_core:citation → <internal URI> or up_citations:[citation id]
 - rdfs:seeAlso → <URL> or <external DB record URI>
 - up_core:reviewed → boolean
 - up_core:created → date (yyyy-mm-dd)
 - up_core:modified → date (yyyy-mm-dd)
- up_core:Gene**
 - rdf:type → <internal URI>
 - skos:prefLabel → string
 - skos:altLabel → string
 - up_core:orfName → string
 - up_core:locusName → string

UniProtKB Entry (P63000 - RAC1_HUMAN):

Function¹
 Plasma membrane-associated small GTPase which cycle as secretory processes, phagocytosis of apoptotic cells, Rac1 p21/rho GDI heterodimer is the active component regulation of cell migration and adhesion assembly and

```

isoform:P06213-1 a up:Simple_Sequence ;
up:modified "2010-10-05"^^xsd:date ;
up:version 4 ;
up:precursor true ;
up:mass 156333 ;
    
```

There are 217,505,202,099 triples in 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	160,189,731,20040,455,837,024	29	6	46,916,767,863	http://creativecommons.org/licenses/by/4.0/	
uniprot	Documentation	44,256,643,227	9,441,439,078	124	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

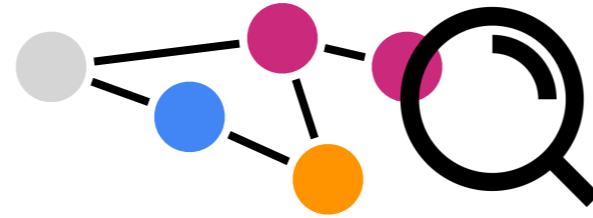
The image illustrates the UniProt public knowledge graph. On the left, a screenshot of the UniProt website shows the entry for P63000 (RAC1_HUMAN), including its function, taxonomy, and various annotations. In the center, a diagram shows the graph structure with nodes representing entities like 'uniprot:[accession number]', 'up_core:Protein', 'up_core:Gene', and 'up_core:enzyme', connected by relationships such as 'rdf:type', 'up_core:replaces', 'up_core:organism', 'up_core:encodedBy', and 'up_core:citation'. A large yellow callout box in the foreground reads 'Massive FAIR Life Science data'.

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

Graph	Documentation	Triples	Distinct subjects	Distinct predicates	Distinct classes	Distinct objects	License
uniparc	Documentation	160,189,731,20040,455,837,024	29	6	46,916,767,863	http://creativecommons.org/licenses/by/4.0/	
uniprot	Documentation	44,256,643,227	9,441,439,078	124	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

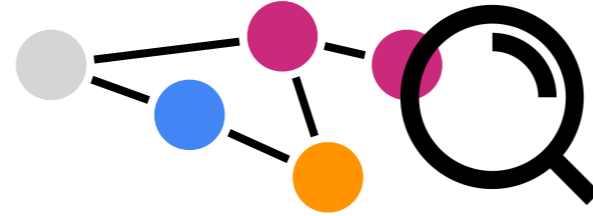
A clinical and genomic
intracranial aneurysm
knowledge graph



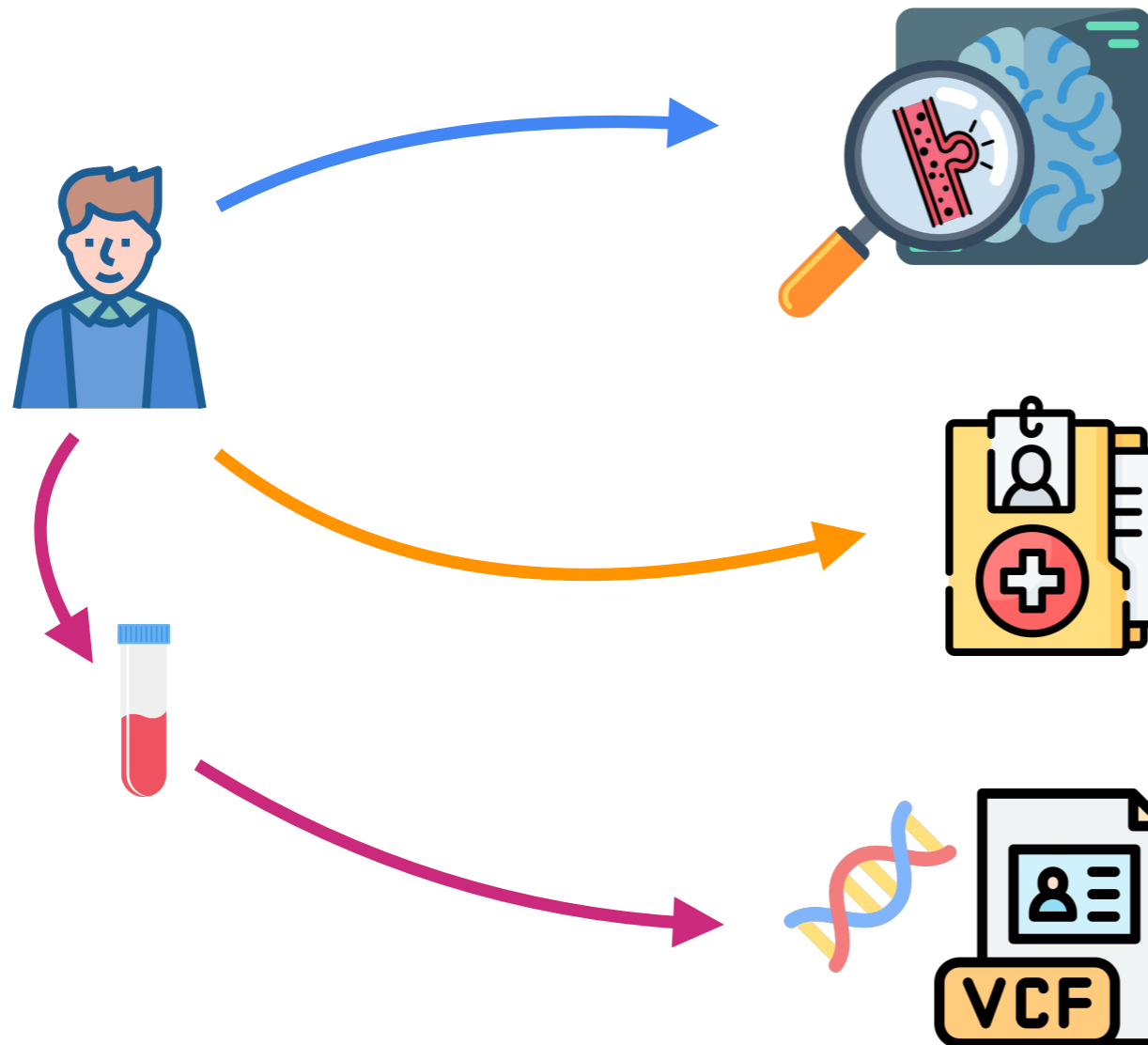
to find & exchange
phenotypes/variants with
reference terminologies !

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 ?

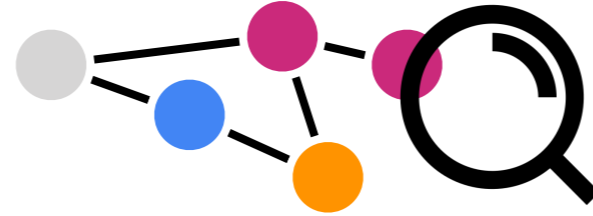
- ▶ **SPHN**
- ▶ **HPO**
- ▶ **DUO**

Genomic data ?

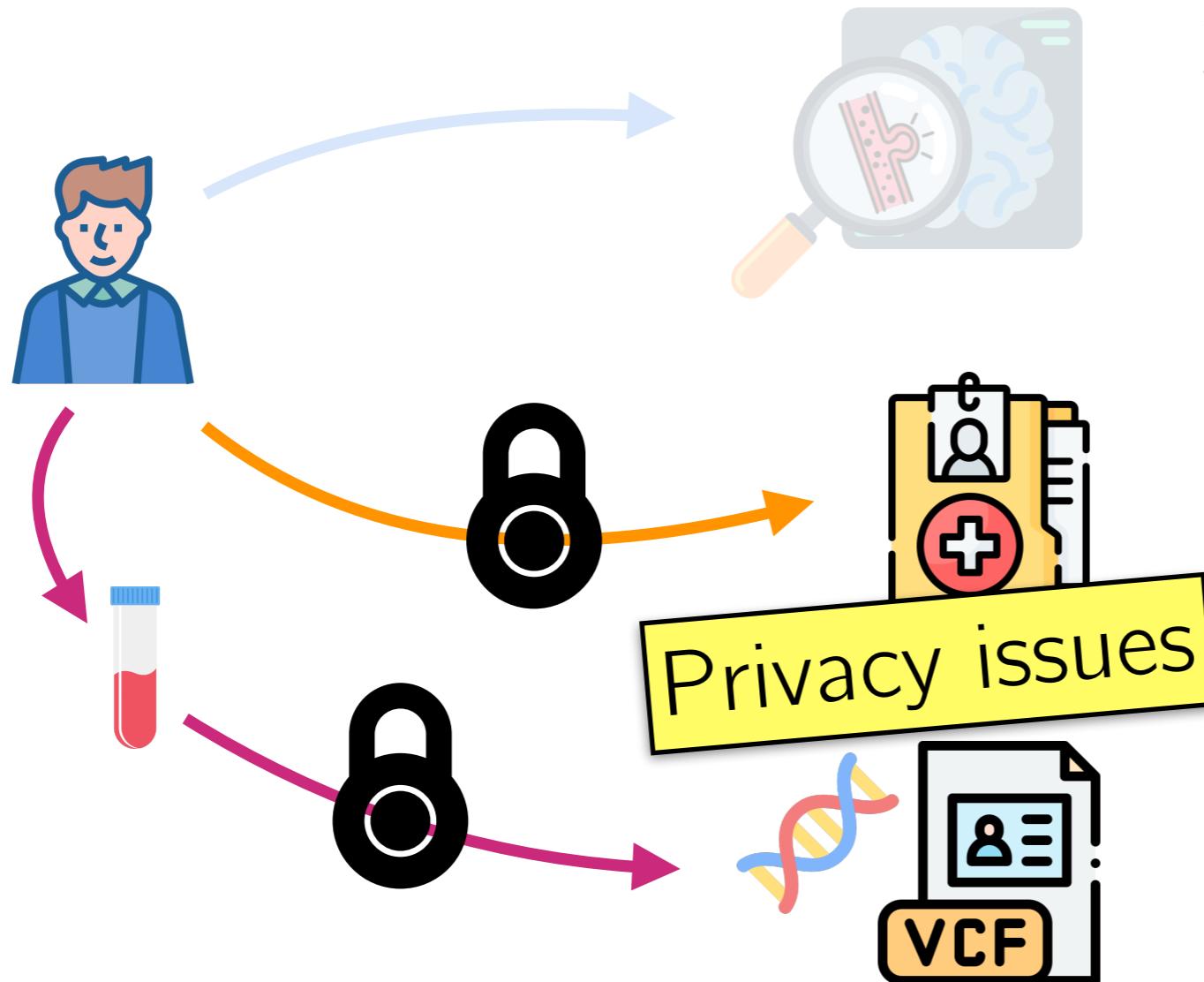
- ▶ **FALDO**
- ▶ **SO / GENO**
- ▶ **SIO**

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)



Beacon protocol



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

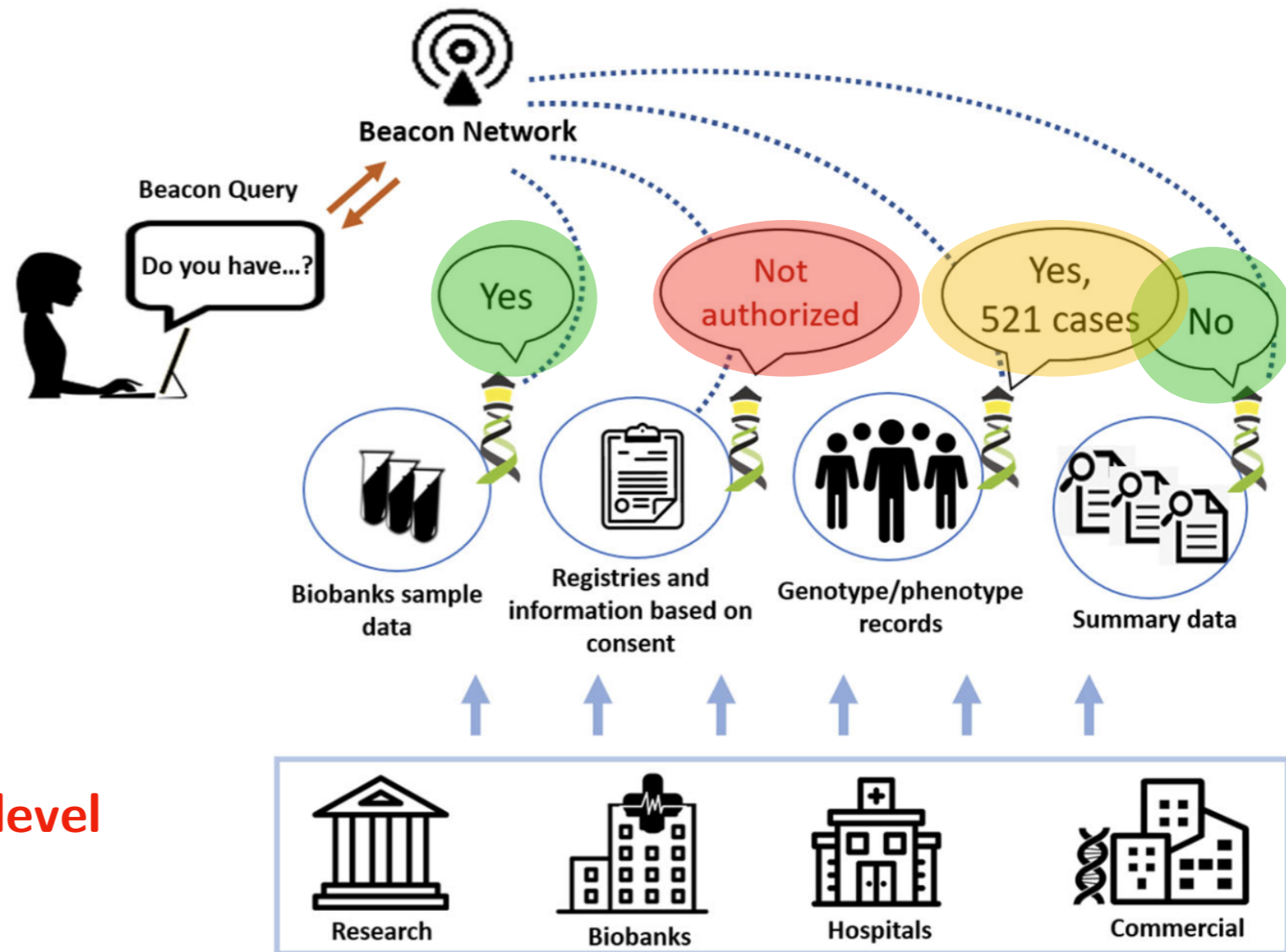


- ▶ **Metadata model** for ‘Variation’, ‘Sample’, ‘Dataset’, ‘Individual’, etc
- ▶ Different **access models**:
boolean, **aggregated data**, **record level**
- ▶ A framework with a **reference implementation**

Beacon protocol



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

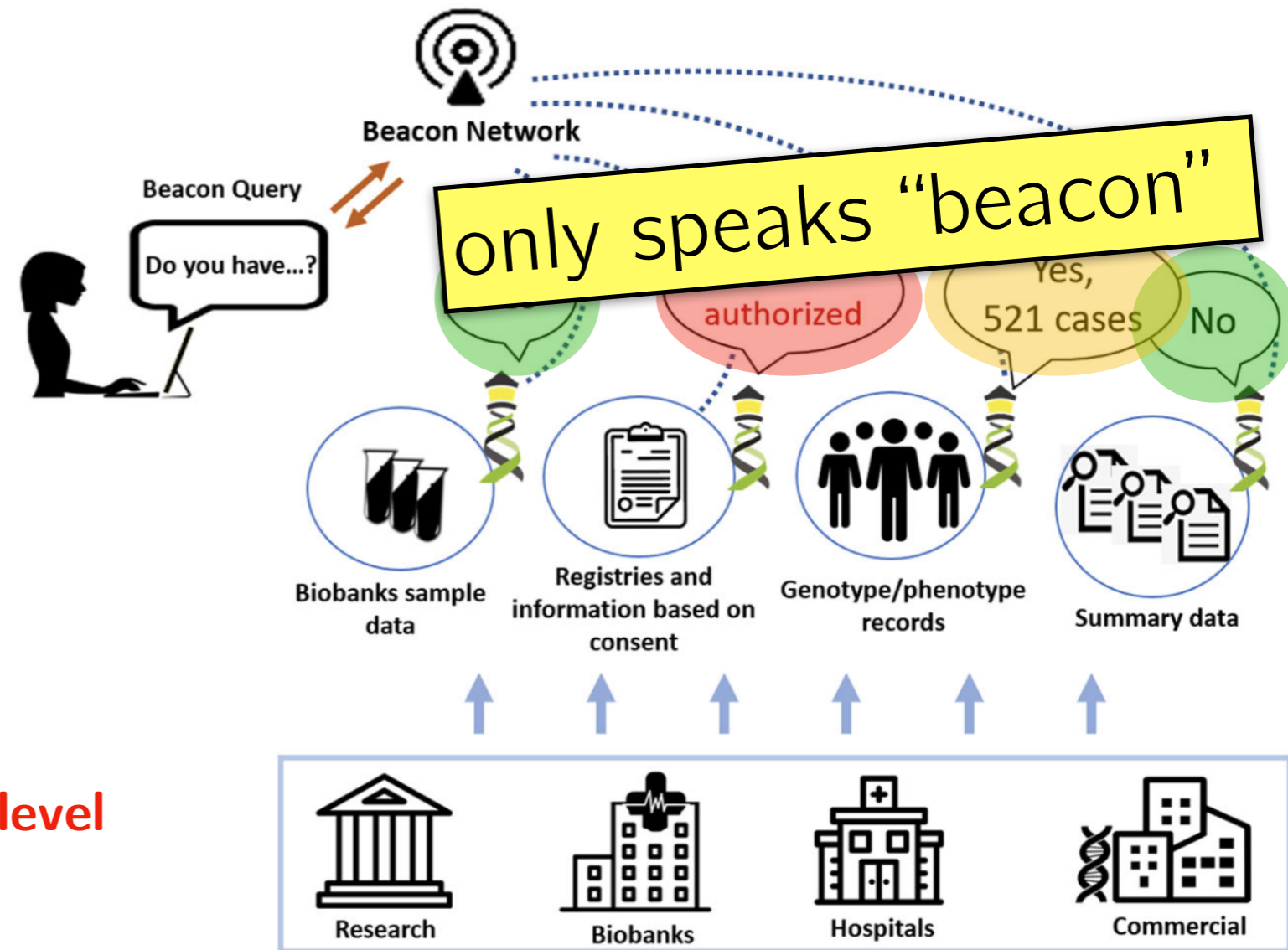


- ▶ **Metadata model** for 'Variation', 'Sample', 'Dataset', 'Individual', etc
- ▶ Different **access models**:
boolean, **aggregated data**, **record level**
- ▶ A framework with a **reference implementation**

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**

Genomic variation data & "annotation"



Example

```

##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
    
```

VCF header

Mandatory header lines (lines starting with ##)

Optional header lines (meta-data about the annotations in the VCF body)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	T			.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion (points to in row 4)

SNP (points to A,AT in row 1)

Large SV (points to SVTYPE=DEL;END=300 in row 4)

Insertion (points to T,CT in row 2)

Other event (points to G in row 3)

Phased data (G and C above are on the same chromosome) (points to 0|1:100 in row 2)

Genomic variation data & “annotation”



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

VCF header

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Body

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

- ▶ 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)

Genomic variation data & “annotation”



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
```

VCF header

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Body

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

- ▶ 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

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

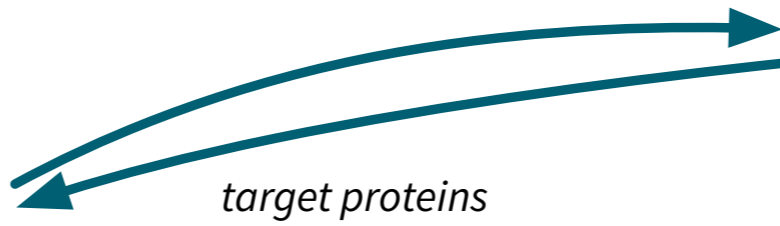
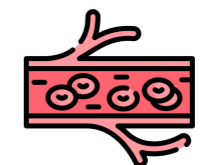
- (i) a **semantic mapping** for aligning Beacon genomic variations to state-of-the-art ontologies
- (ii) an architecture for **on-the-fly FAIRification** of genomic variation data
- (iii) a concrete **federated query** showcasing the integration of local genomic variation data and public knowledge graphs

Intracranial aneurysm motivating use case

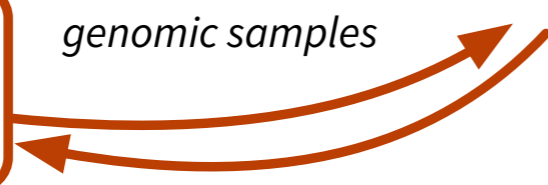
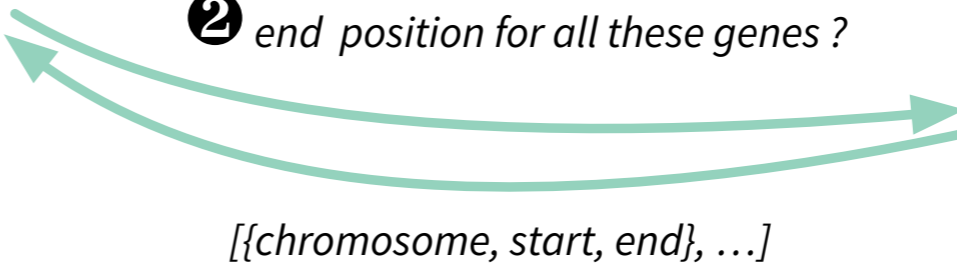
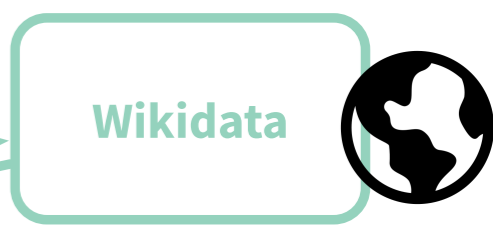
Which genomic variants are located in genes associated with the formation of blood vessels?



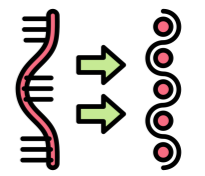
proteins annotated with angiogenesis
① GO term or any sub-class?



② location : chromosome , start and end position for all these genes?



③ biological samples with a mutation in the target DNA regions (chromosome, start, end)?

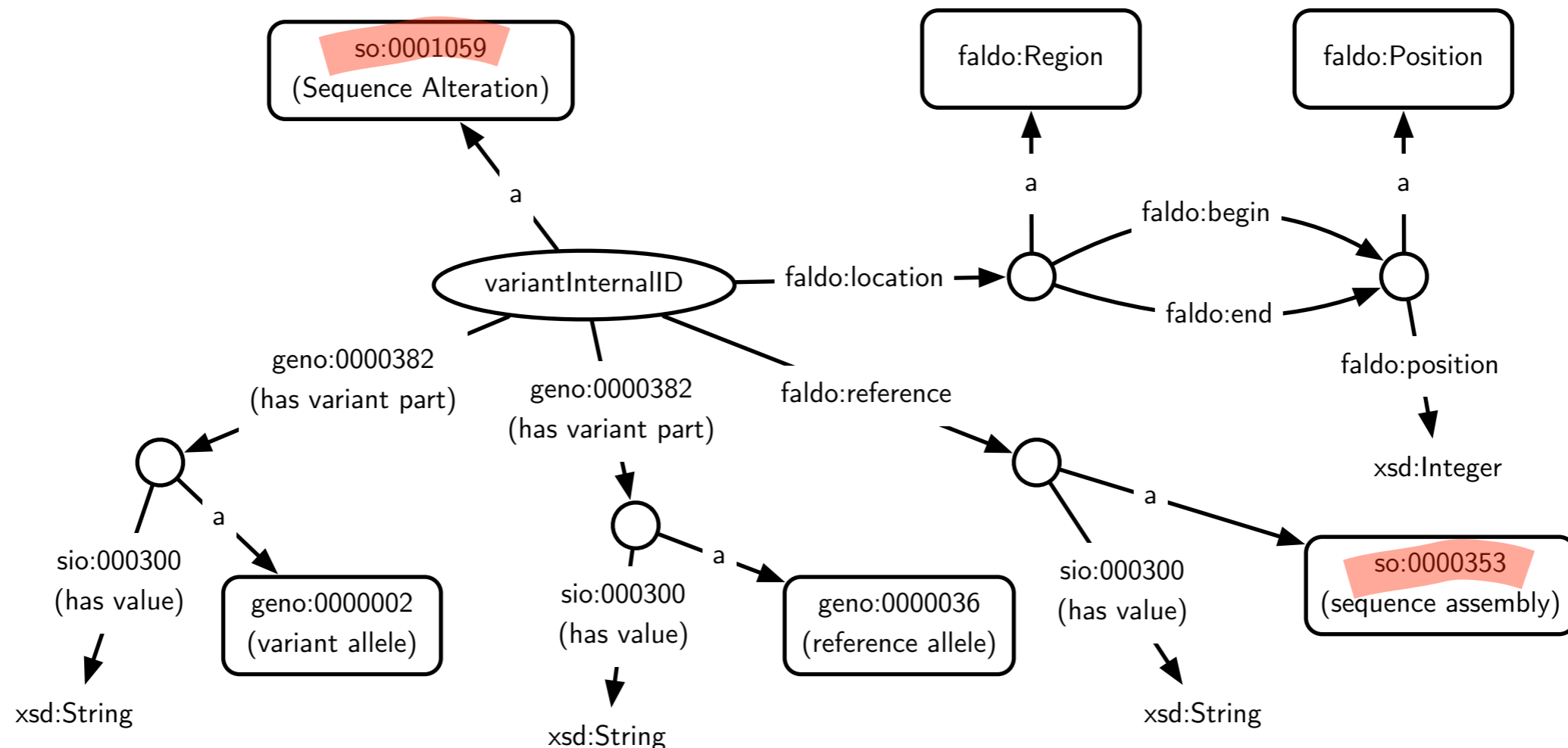


1 Semantic representation of genomic variants



No need to create a new ontology ...

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

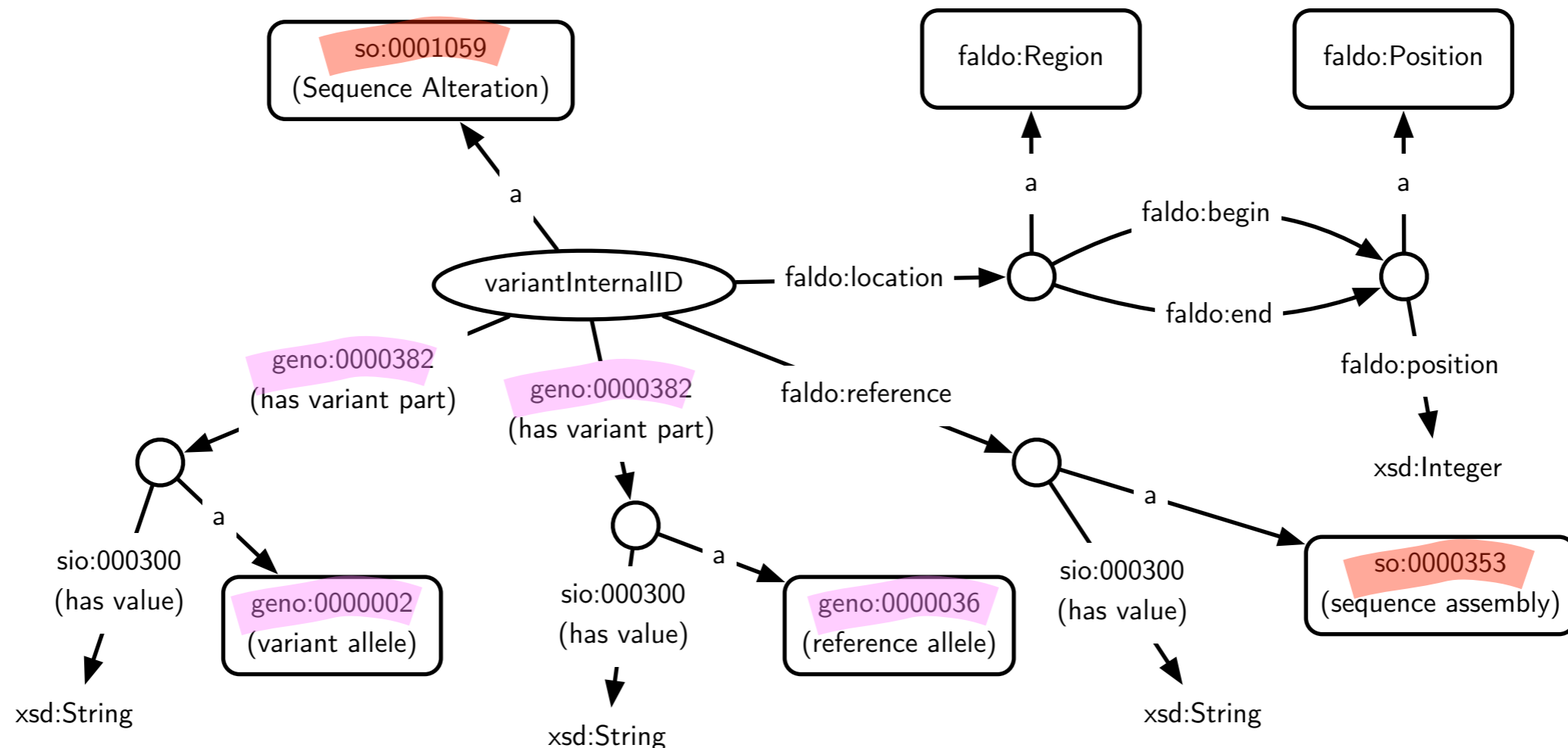


1 Semantic representation of genomic variants



No need to create a new ontology ...

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

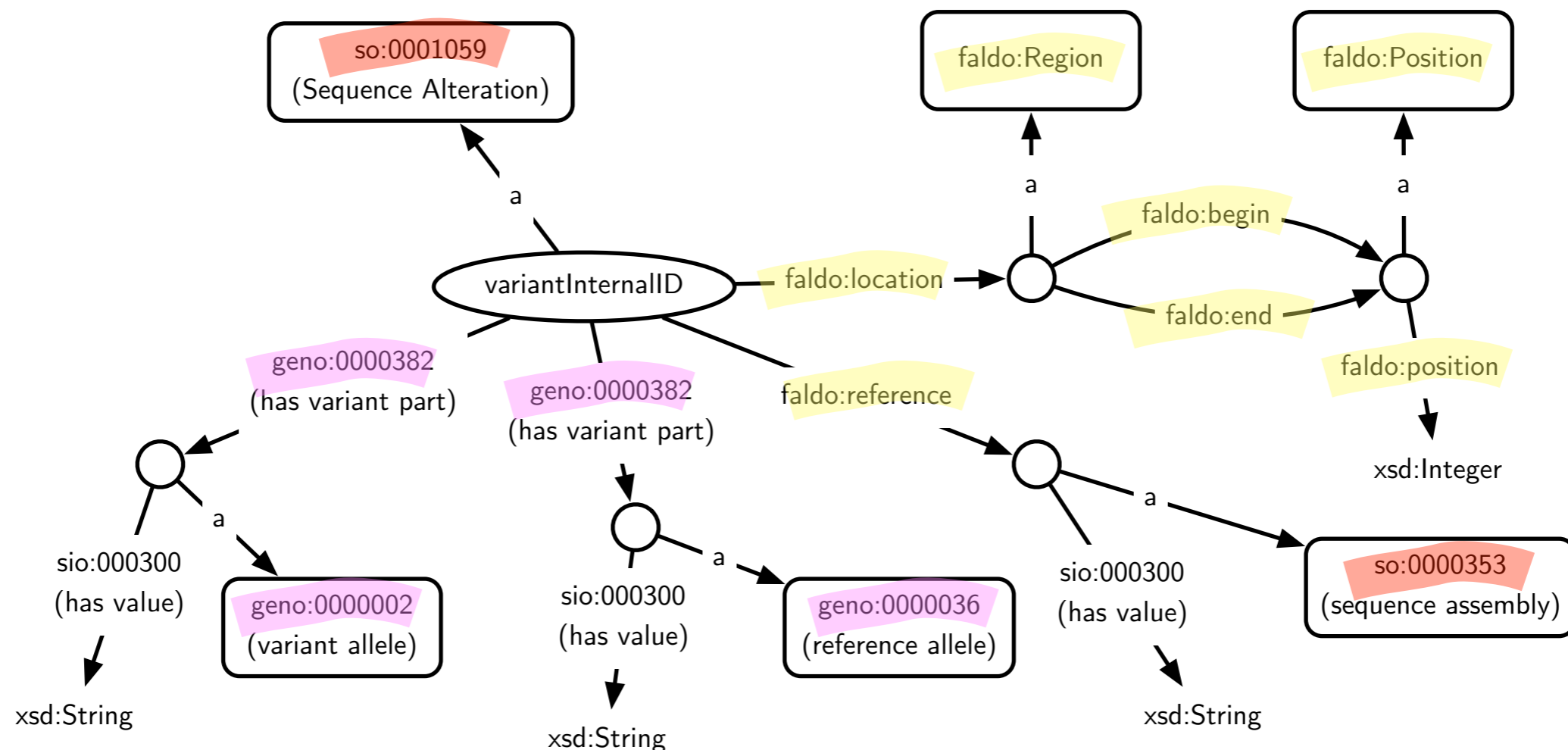


1 Semantic representation of genomic variants



No need to create a new ontology ...

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



② 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 .
```

```
_: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 ;
```

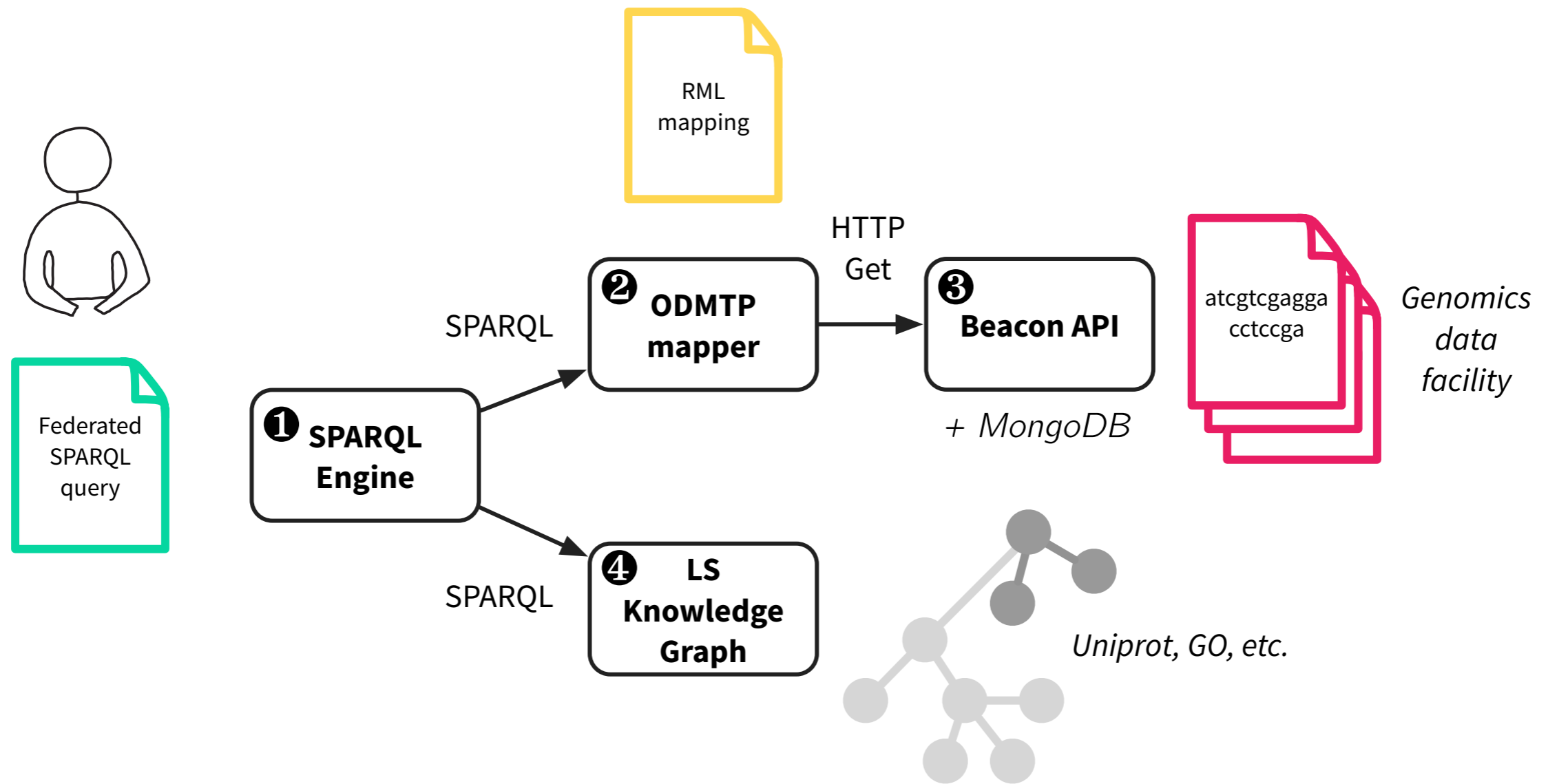
[...]

RDF Mapping Language (RML)

Unofficial Draft 20 June 2024



3 Architecture for “Semantic” Beacons



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
```



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 wdp:P644 ?startcoordinate ;  
      wdpq:P1057/wdt:P1813 ?chromosome ;  
      wdpq:P659/rdfs:label ?assembly .  
    ?wgse wdp: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 wdp:P644 ?startcoordinate ;  
      wdpq:P1057/wdt:P1813 ?chromosome ;  
      wdpq:P659/rdfs:label ?assembly .  
    ?wgse wdp: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

Inserm



Inria
INVENTEURS DU MONDE NUMÉRIQUE

anr[®]

Acknowledgments



PROGRAMME
DE RECHERCHE
SANTÉ
NUMÉRIQUE



Alexandrina
Bodrug-Shepers



Gabriella
Montoya



Patricia
Serrano-Alvarado



Richard
Redon



Hugo
Chabane

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

Alexandrina Bodrug-Shepers^{1,†}, Hugo Chabane^{2,†}, Gabriela Montoya², Patricia Serrano-Alvarado², Richard Redon¹ and Alban Gaignard^{1,3}

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

²Nantes Université, LS2N, Nantes, France

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

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

contact: alban.gaignard@univ-nantes.fr



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>

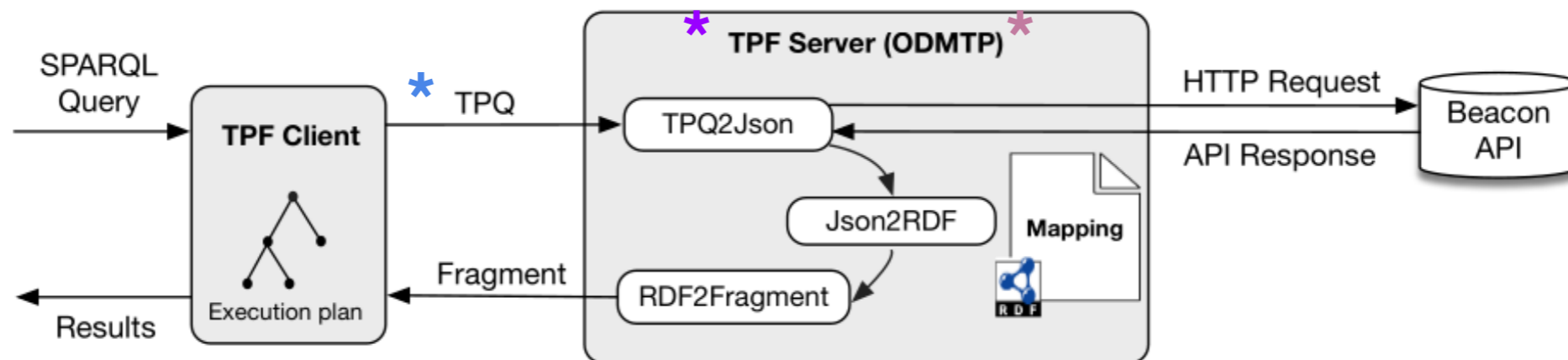
- ▶ Potential **data leaks in distributed joins**
 - can we prevent "data leaks" at query design time ?
 - can we verify it at federated query run time ?



On the fly conversion Beacon API \leftrightarrow Linked Data

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

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



* Triple Pattern Query

* Triple Pattern Fragment

*

On-Demand Mapping using Triple Patterns