Journées PEPR Santé Numérique
14-15 Octobre 2025

Semantic Beacons

federated querying over genomic pheno-clinical data and public knowledge graphs leveraging international standards and semantic web technologies

Alexandrina Bodrug
l'**institut du thorax**, équipe 1 génétique
INSERM umr1087, CNRS umr6291,
Nantes Université

Projet financé par
ANR-22-PESN-0008

Research context

PEPR Santé Numérique Neurovasc

The ICAN project

ICAN dataset: individuals carrying **intracranial aneurysms** that benefited from **neuroimaging**, **exome sequencing** and **pheno-clinical data** collection

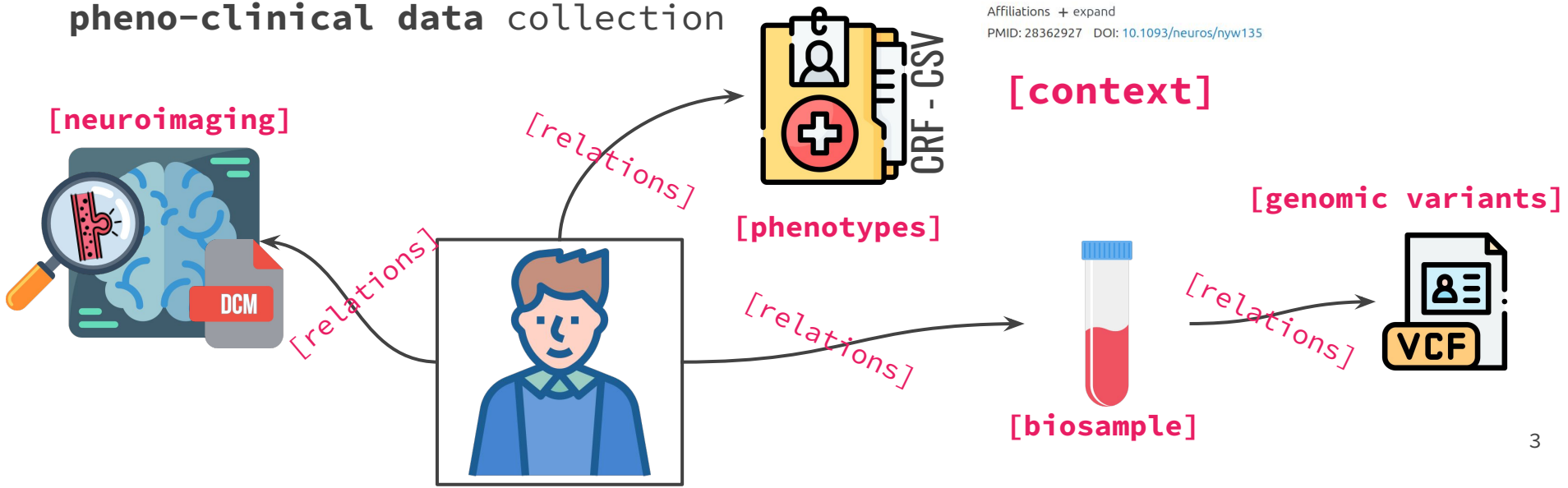
> Neurosurgery. 2017 Apr 1;80(4):621-626. doi: 10.1093/neuros/nyw135.

Understanding the Pathophysiology of Intracranial Aneurysm: The ICAN Project

Romain Bourcier [1], Stéphanie Chatel [2], Emmanuelle Bourcereau [2], Solène Jouan [1], Hervé Le Marec [2], Benjamin Daumas-Duport [1], Mathieu Sevin-Allouet [3], Benoit Guillon [3], Vincent Roualdes [4], Tanguy Riem [4], Bertrand Isidor [5], Pierre Lebranchu [6], Jérôme Connault [7], Thierry Le Tourneau [2], Alban Gaignard [2], Gervaise Loirand [2], Richard Redon [2], Hubert Desal [1]; ICAN Investigators

Affiliations + expand

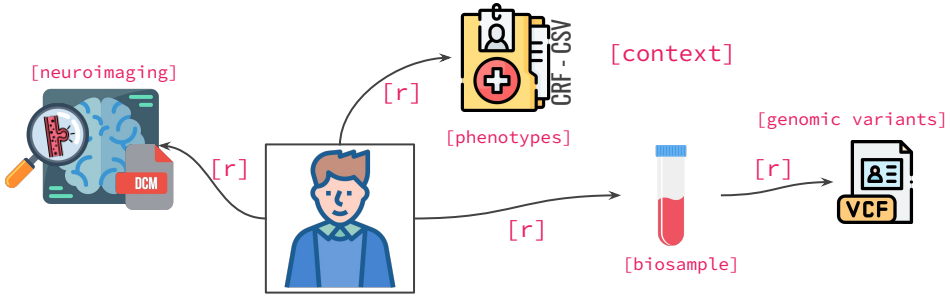
PMID: 28362927 DOI: 10.1093/neuros/nyw135



PEPR Santé Numérique Neurovasc

Multi-centric, multi-domain, multi-modal **heterogeneous** dataset.

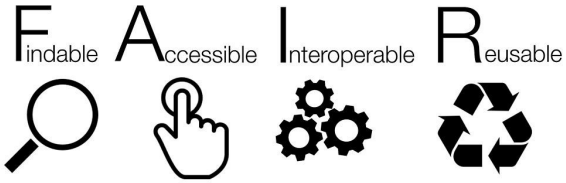
Research **health** dataset → **sensitive** data with privacy constraints



[ICAN dataset]

Challenge: How to increase **FAIRness**?

WP2 Task2.1: FAIR genomic data demonstrator



We identified 3 main challenges

What do we want to achieve?

(1) **Share** sensitive genomic data (in a safe way)

(2) **Semantify, integrate and query** multi-modal health data

(3) Integrate genomic health data with **public knowledge bases**

Why do we want to achieve it?

Facilitate answering 'simple' biological inquiries

→ *Allow **collaborators** to check for their genomic variation of interest in the ICAN dataset.*

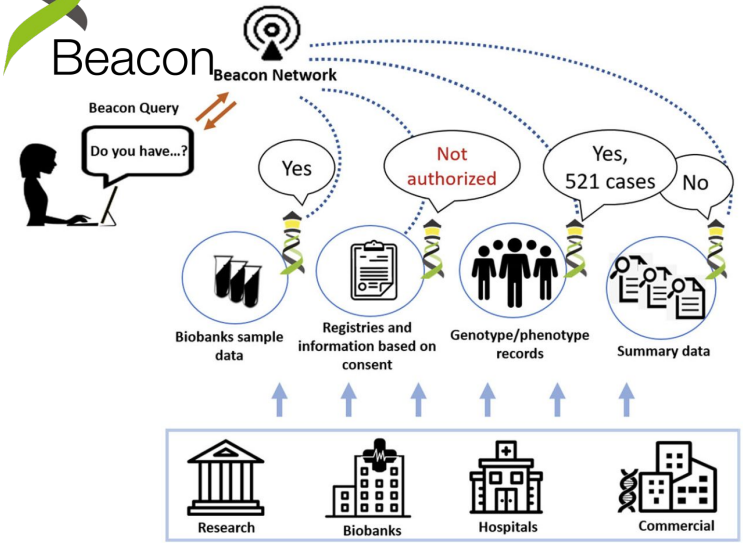
→ *Quickly retrieve information about genomic variation presence within phenotypic subgroups.*

→ *Quickly enrich genomic variation with **fresh knowledge** about gene location and protein function **without duplicating** databases locally.*

(1) Sharing sensitive genomic health data

Technologies from organisations facilitating biomedical research

The **Beacon standard for genomic discovery** is a key approach for decentralized biomedical research (promoted by **Elixir** and **GA4GH**)



Framework → defines the rules for querying genomic health datasets.

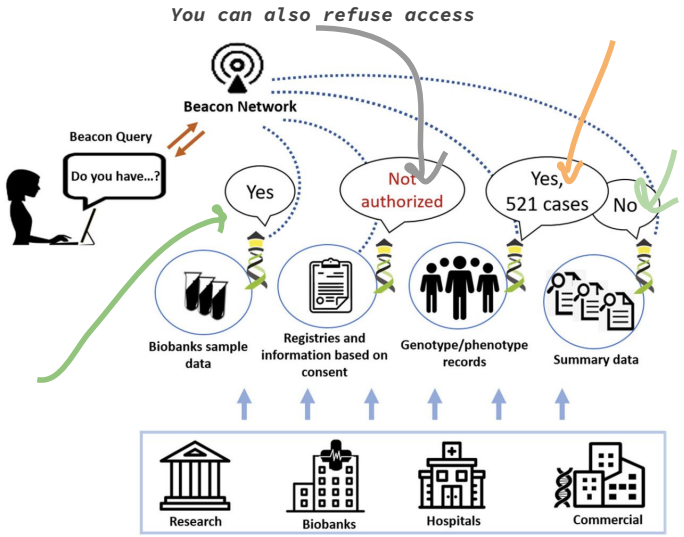
Model → structures the data
Following **standard models** 'Variation', 'Sample', 'Dataset', 'Individual', etc

"Beacon v2 is an **API specification** established by the Global Alliance for Genomics and Health initiative (GA4GH) that defines a **standard for federated discovery** of genomic and phenotypic data." Rueda *et al.* 2022

Deploying the Beacon standard for genomic discovery

→ **We deployed a Beacon API using the ICAN dataset**, locally at ITX
→ We are investigating how to deploy a similar solution in a hospital environment

As a Beacon API deployer, you can grant or revoke access at different ***response granularities.***



Boolean *response*

e.g. you allow anyone to discover the existence of variation in your biocollections.

Aggregated *response*

e.g. you allow broad collaborators to access statistics about the variants in your biocollections.

Record level *response*

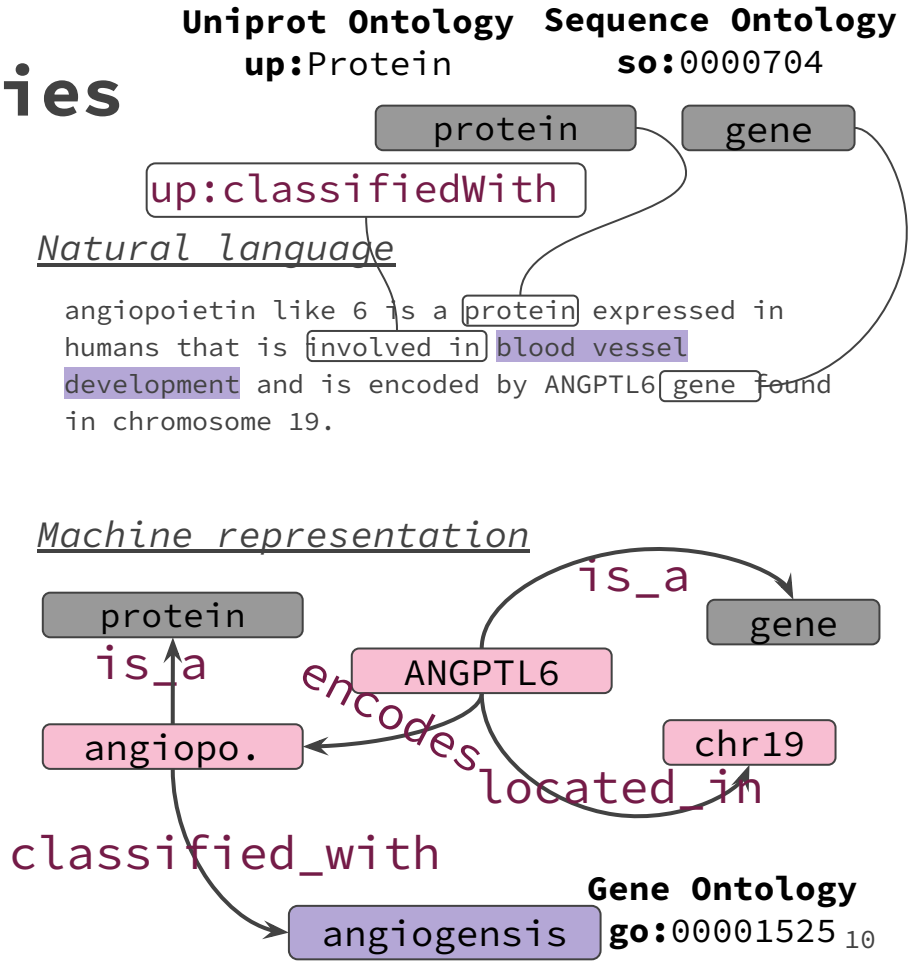
e.g. you allow close collaborators to access detailed information about the variants in your shared biocollection that you manage.

(2) Semantify, integrate and query multi-modal data

Semantic web technologies

Ontologies : structured representation of **domain knowledge** (concepts, hierarchy, relationships)

Knowledge bases/graphs (KG): structured representation of entities, facts and their relationships, enabling **reasoning** and **queries**

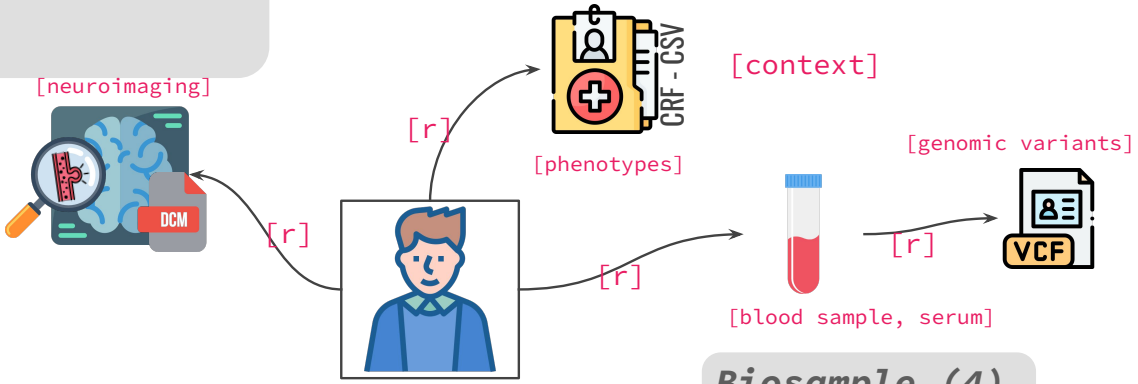


ICAN KG: shared vocabulary between machines and people

Selecting biomedical ontologies to represent the ICAN dataset concepts

Clinical data **(1)**
SPHN
HPO/MONDO
DUO

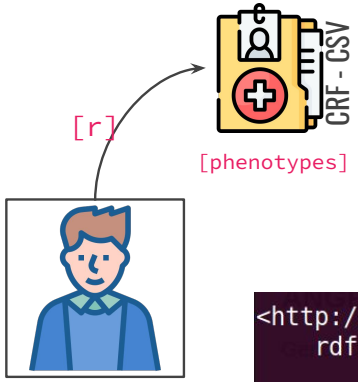
Neuroimaging **(2)**
UBERON
NCIT



Genomics **(3)**
FALDO
SO/GENO
SIO

Biosample **(4)**
UBERON

(1) Data model and ontologies for pheno-clinical data



Clinical data

SPHN - Swiss Personalized Health Network interoperability framework

HPO - Human Phenotype Ontology

```
<http://ican.ressource.org/individual/pid_SIM00108/diagnosis/Cerebral_berry_aneurysm> a sphn:Diagnosis ;
    rdfs:label "Affected family member"^^xsd:string,
        "Cerebral berry aneurysm"^^xsd:string ;
    sphn:hasCode obo:HPO_0007029,
        obo:HPO_0032320 ;
    sphn:hasSubjectAge <http://ican.ressource.org/individual/pid_SIM00108/ageAtDiagnosis/70> ;
    sphn:hasSubjectPseudoIdentifier <http://ican.ressource.org/individual/pid_SIM00108> .

<http://ican.ressource.org/individual/pid_SIM00108/diagnosis/Diabetes_mellitus_type_2> a sphn:Diagnosis ;
    rdfs:label "Diabetes mellitus type 2"^^xsd:string ;
    sphn:hasCode obo:HPO_0005978 ;
    sphn:hasSubjectPseudoIdentifier <http://ican.ressource.org/individual/pid_SIM00108> .
```

Machine readable **R**esource **D**escription **F**ramework (**RDF**) file format

Explore and reason over ICAN dataset KG

In **individuals with multiple ICA,** which **genomic variants reside** in **genes coding for proteins involved in blood vessel formation?**

currently possible in our demonstrator

Virtual machine
(Glicid/IfB)

ICAN dataset Knowledge Graph

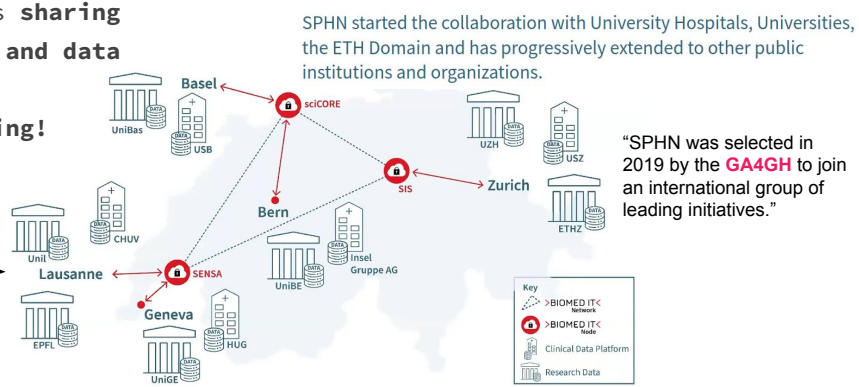
Apache Jena Fuseki

sparql server/ triplestore database

Queries that can be sent to several institutions **sharing common vocabularies and data models.**

→ **Distributed querying!**

Collaborations we're thinking about



"SPHN was selected in 2019 by the **GA4GH** to join an international group of leading initiatives."

?

Querying Knowledge Graphs is possible thanks to the SPARQL query language.

@ https://sphn.ch/network/data-coordination-center/the-sphn-semantic-interoperability-framework/

Explore and reason over ICAN dataset KG

In **individuals with multiple ICA**, which **genomic variants reside** in **genes coding for proteins involved in blood vessel formation**?

currently possible in our demonstrator

Virtual machine
(Glicid/IfB)

ICAN dataset Knowledge Graph

Apache Jena Fuseki

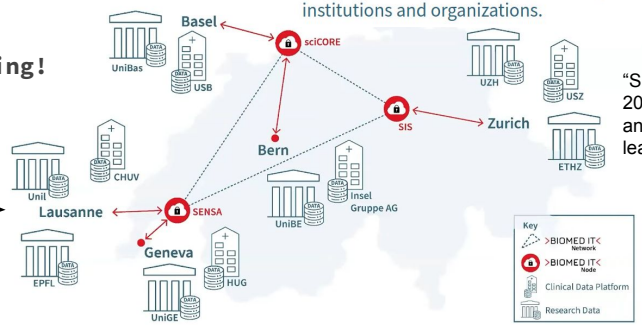
sparql server/ triplestore database

Queries that can be sent to several institutions **sharing common vocabularies and data models.**

→ **Distributed querying!**

Collaborations we're thinking about

SPHN started the collaboration with University Hospitals, Universities, the ETH Domain and has progressively extended to other public institutions and organizations.



"SPHN was selected in 2019 by the **GA4GH** to join an international group of leading initiatives."

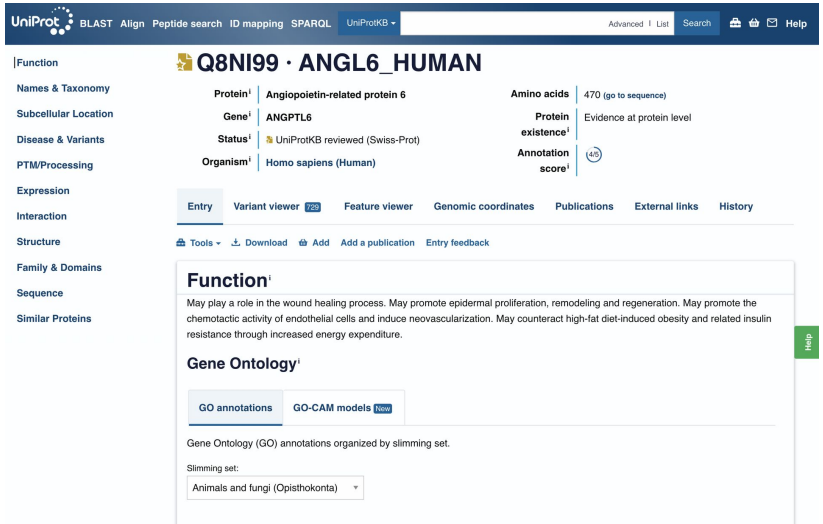
?

knowledge not present in ICAN KG
Knowledge present in public biomedical KG

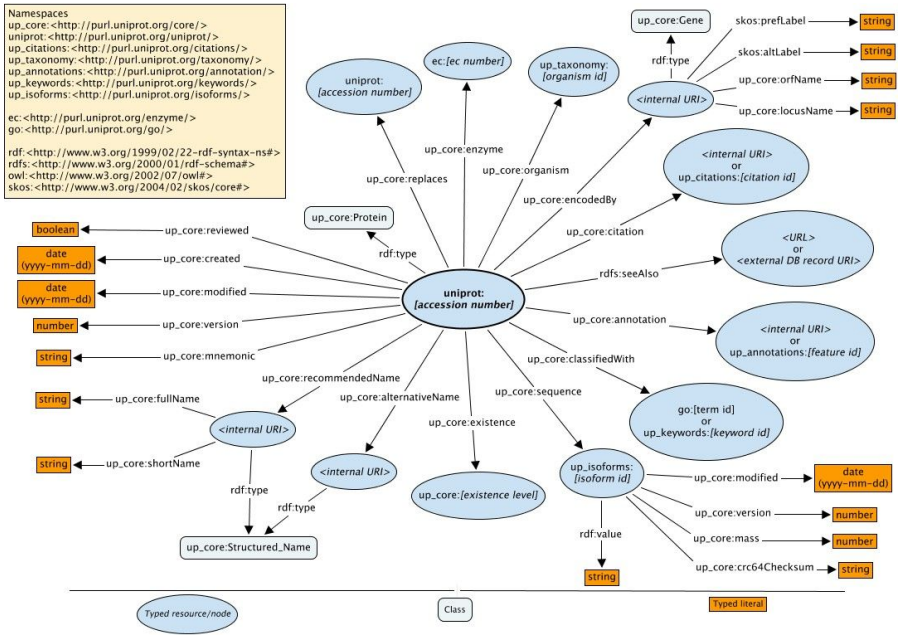
@ https://sphn.ch/network/data-coordination-center/the-sphn-semantic-interoperability-framework/

(3) Integrating with public biomedical KG

Uniprot Knowledge Graph: Human User Interface and sparql endpoint



28 million proteins



58 billion triples (statements)

Integrating genomic health data with public KG

Use Case

Within the ICAN dataset, which genomic variants reside in genes coding for proteins involved in blood vessel formation ?

This question can be answered with knowledge from three different sources.
*→ **Federated querying!***

*We want to transform this question into a **single federated query**.*
***Problem** → How do we make two KG and a REST API talk to each other? → Not the same standards!*

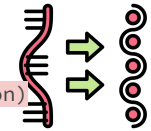
blood vessel formation

UniprotKB
Biomedical Knowledge Graph
Protein organisms
Protein function

protein (function)

Wikidata
Generalistic Knowledge Graph
Gene location
Genome assembly versions
Protein coding genes

gene (location)



protein

aneurysm

Variation (location)

Beacon

API {REST}

ICAN dataset
genomic variants
pheno-clinical descriptions
biosample metadata
study context

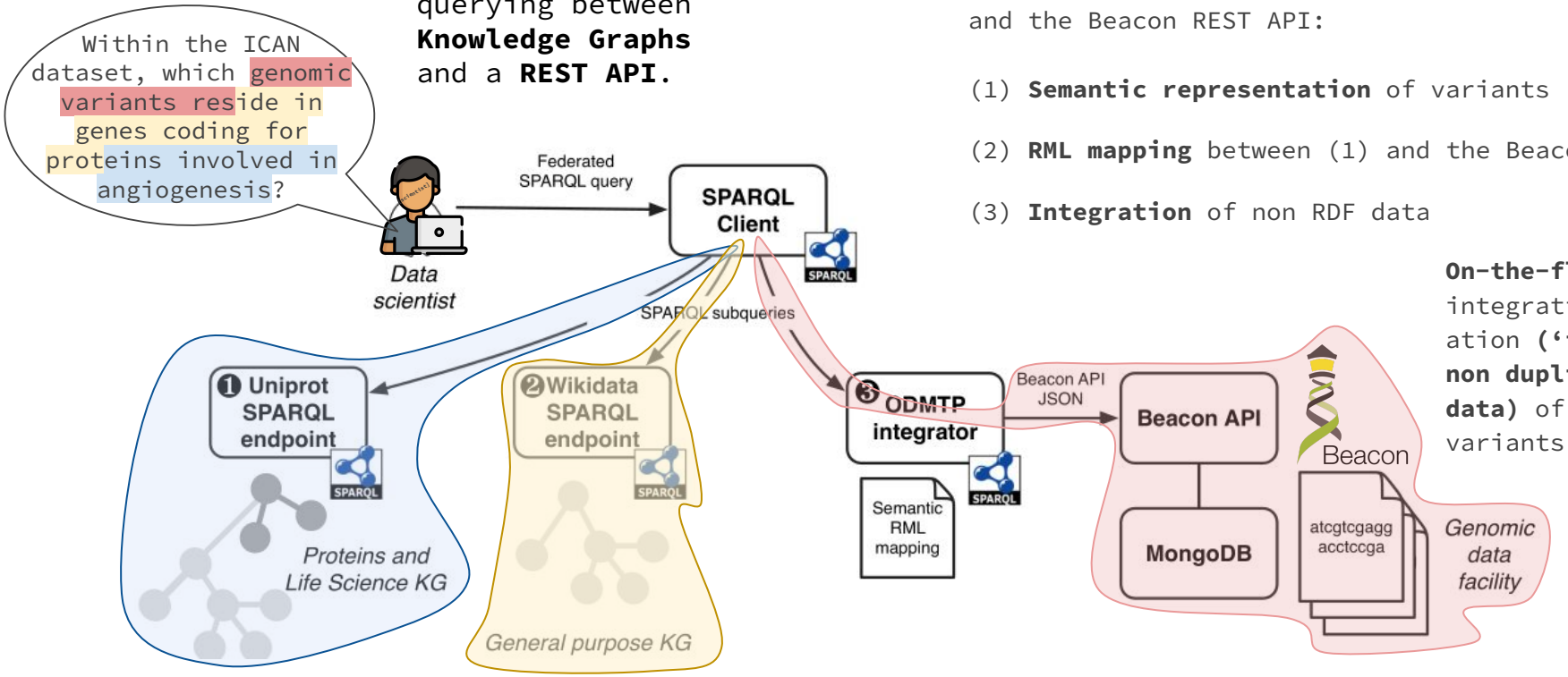
swat4hcls 2025 conference

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]

Semantic Beacons framework

Enabling federated querying between **Knowledge Graphs** and a **REST API.**



To enable communication between the SPARQL client and the Beacon REST API:

(1) **Semantic representation** of variants

(2) **RML mapping** between (1) and the Beacon API

(3) **Integration** of non RDF data

On-the-fly integration/annotation **('fresh' non duplication data)** of genomic variants

Lessons learned

Lessons learned

Beacon standard

(+) Sharing of **sensitive** genomic data
(+) International **standard** (used by F-EGA, promoted by Elixir & GH4GH)
(+) Several existing **implementations**
(+) **Data models** and framework

(-) No semantics
(-) Genomics focused
(-) Limited interop.

Knowledge Graphs

(+) **Federation**
(+) **Interoperability** with domain knowledge
(+) Semantics/Ontologies
(+) Flexible to new data integration
(+) W3C **standard**

(-) Technologies unfamiliar to geneticists and biologists
(-) Steep learning curve
(-) **Data transformation** intensive

Semantic Beacons

(+) **Innovative** approach
(+) **Joint perspective: biomedical + semantic**
(+) 2 standards
(+) Semantics
(+) Federation

(-) Experimental
(-) Slow query execution
(-) Genomics focused

Thank you for your attention

Ressources

Semantic Beacons **conference paper** & talk
HAL Id: hal-04908530

Code repositories
Genomics data transformation **pipeline**
https://gitlab.univ-nantes.fr/bodrug-a/etl4fairdata_AIC
Phenoclinical data transformation **pipeline**
https://gitlab.univ-nantes.fr/bodrug-a/etl4sphn_AIC
Technological demonstrator **ansible project**
https://gitlab.univ-nantes.fr/bodrug-a/demo-aggrvarkg

Virtual machines hosting demonstrators
Semantic Beacons (IfB) - https://134.158.249.80/
Full KG approach (Glicid) - https://cgen-kg-ica.bird.glicid.fr/

Data model documentation
linkML project -
https://gitlab.univ-nantes.fr/bodrug-a/neugenfair
mkdocs - https://neugenfair-caffb5.univ-nantes.io/mkdocs/

NEUROVASC PEPR SANTÉ NUMÉRIQUE
FRANCE 2030 PROGRAMME DE RECHERCHE SANTÉ NUMÉRIQUE
Projet financé par
ANR-22-PESN-0008

fin

below are question slides

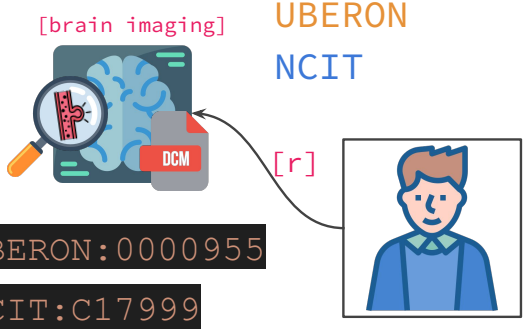
(2) Small bridge to the medical imaging repository Shanoir

Acquisition identifier enabling linkage to the specialized medical imaging repository **Shanoir** (WP1- Neurovasc)

Brain imaging

UBERON

NCIT



Brain → UBERON:0000955

Scan → NCIT:C17999

```
<http://ican.ressource.org/individual/pid_SIM00105/imagingProcedure/AIC_05_7/bodySite/brain> a sphn:BodySite ;
    rdfs:label "Intracranial vasculature"^^xsd:string ;
    sphn:hasCode obo:UBERON_0000955 .
```

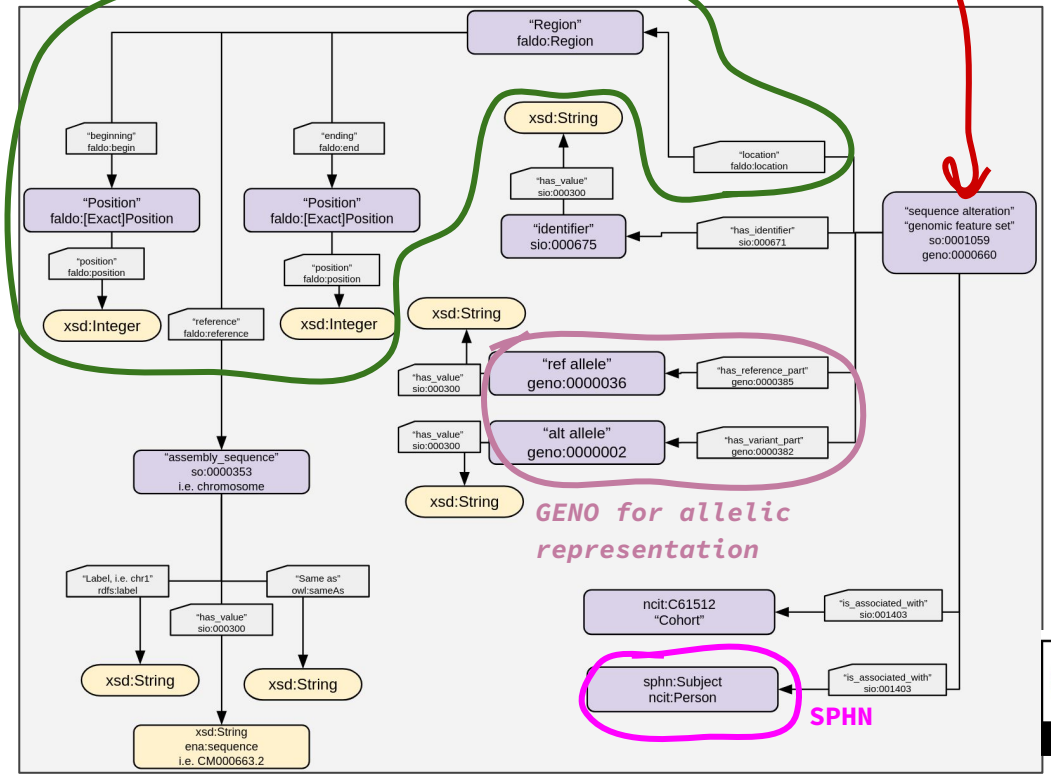
```
<http://ican.ressource.org/individual/pid_SIM00105/imagingProcedure/AIC_05_7> a sphn:ImagingProcedure ;
    rdfs:label "Imaging procedure for the simulated individual from the ICAN Biocollection, the identifier being the [I|U]CAN inclu
sion number"^^xsd:string ;
    sphn:hasBodySite <http://ican.ressource.org/individual/pid_SIM00105/imagingProcedure/AIC_05_7/bodySite/brain> ;
    sphn:hasCode obo:NCIT_C17999 ;
    sphn:hasIdentifier "AIC_05_7"^^xsd:string ;
    sphn:hasSubjectPseudoIdentifier <http://ican.ressource.org/individual/pid_SIM00105> .
```

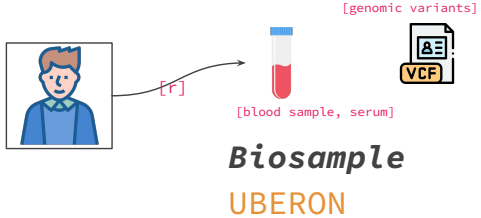
Linked Data -> **R**esource **D**escription **F**ramework (**RDF**) file

(3) Data model and ontologies for genomic variants

FALDO for precise extensive representation of sequence loci

SO and GENO terms for variant concept

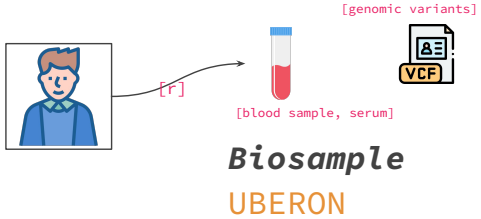




We developed **our own genome variation model**, inspired from DisgeNET KG model.

Several co-existing solutions → none fitted our needs.

(3) Data model and ontologies for genomic variants



Genomics
FALDO
SO/GENO
SIO

We developed **our own genome variation model**, inspired from DisgeNET KG.

Several co-existing solutions → none fitted our needs.

1) Models modelling VCF → we do not want, we want to model biological variants
2) Models describing a variant as part of a diagnosis, for diagnosis support → we do not want that, we model diagnosis separately with SPHN
3) Non semantic models (VRS?)
4) Modeling variants as part of gene-disease association (DisgeNET)

THE federated query leveraging ICAN KG, Wikidata & UniprotKB

pato:002118 hpo:0007029

In the ICAN dataset subgroup of individuals with multiple ICA, which genomic variants

so:0001059

reside in genes coding proteins involved in angiogenesis?

faldo location model
wikidata location model

wdt:P702

up:Protein , wdt:P352

up:classifiedWith

go:0001525

```
SELECT ?populationGroup ?variantId ?variantStart ?variantEnd ?variantChromosome
      ?proteinName ?proteinID2 ?goTerm ?geneStart ?geneEnd ?geneChromosome ?geneAssembly


WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
    up:organism taxon:9606 ;
    up:classifiedWith go:0001525 .
    ?protein up:mnemonic ?proteinName .
  }

  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinWD) .

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinWD ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?geneStart ;
        wdpq:P1057/wdt:P1813 ?geneChromosome ;
        wdpq:P659/wdt:P2576 ?geneAssembly .
      ?wgse wdps:P645 ?geneEnd ;
        wdpq:P1057/wdt:P1813 ?geneChromosome ;
        wdpq:P659/wdt:P2576 ?geneAssembly .
    FILTER(STR(?geneAssembly) = "hg38")
  }

  ?variant a so:0001059 .
  ?variant sio:000671/sio:000300 ?variantId .
  ?variant faldo:location/faldo:begin/faldo:position ?variantStart .
  ?variant faldo:location/faldo:end/faldo:position ?variantEnd .
  ?variant faldo:location/faldo:reference/rdf:label ?variantChromosome .
  ?variant sio:001403 ?populationGroup .
  ?populationGroup sphn:hasCode  hpo:0007029, pato:002118 .


  FILTER(STR(?geneChromosome) = STR(?variantChromosome))
  FILTER((xsd:integer(?variantStart) >= xsd:integer(?geneStart) &&
      xsd:integer(?variantStart) <= xsd:integer(?geneEnd)) ||
     (xsd:integer(?variantEnd) >= xsd:integer(?geneStart) &&
            xsd:integer(?variantEnd) <= xsd:integer(?geneEnd)))

}
```

coded by

same reference assembly as the ICAN variants

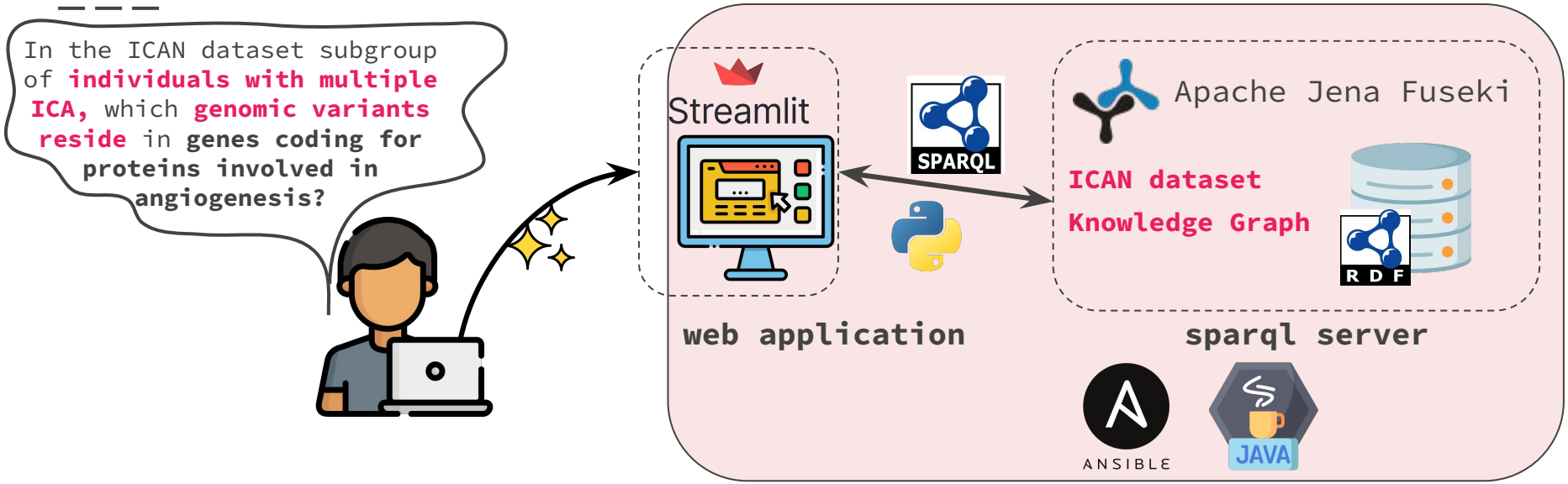
UniProt

WIKIDATA

ICAN Linked Data

https://cgen-kg-ica.bird.glicid.fr/

Full KG technological demonstrator: system architecture



Sparql queries → Exploration and reasoning over integrated ICAN dataset

Virtual machine (Glicid/IfB)

8CPU, 8G RAM - 120G Disk Space

https://cgen-kg-ica.bird.glicid.fr/

Biomedical and life science knowledge graphs

Wikidata – general purpose reliable KG, Gene location, Proteins, Genome assembly versions

Monarch – integrating phenotypes, genes and diseases across species

UniProtKB – protein sequence, spatial and functional information, associated diseases

KGs can be:
Generalistic
Domain specific

Have different goals:
Knowledge representation
Automated reasoning

Beacon API: ontology use

Modified hierarchical ontology query

A Beacon will query for entities associated with the submitted bio-ontology term(s), and by default, all descendant terms. The optional `includeDescendantTerms` parameter can be set to either `true` or `false`. The default and assumed value of `includeDescendantTerms` is `true`, thus if the parameter is not set, then the use of bio-ontology terms in a Beacon request implies that a hierarchical ontology search is requested.

Request example of two filters, where one filter excludes matches with descendant terms:

POST

```
"filters": [
    {
        "id": "HP:0100526",
        "includeDescendantTerms": false
    },
    {
        "id": "HP:0005978"
    }
]
```

Semantic similarity query

A Beacon will query for entities that are associated with bio-ontology terms that are similar to the submitted terms. The Beacon API is agnostic to the semantic similarity model implemented by a Beacon and how a Beacon applies the relative thresholds of similarity. A semantic similarity query request contains the required `similarity` parameter with a value set to define the relative threshold level of `high`, `medium` or `low`.

POST request example of two Filters using differing relative similarity thresholds:

```
"filters": [
    {
        "id": "HP:0100526",
        "similarity": "high"
    },
    {
        "id": "HP:0005978",
        "similarity": "medium"
    }
]
```

(Pseudo-)numerical value queries

EXAMPLE OF A FILTER FOR INDIVIDUALS OVER 70 YEARS OF AGE

- age = PATO:0000011, age syntax as ISO 8601

GET POST

- `filters=age:>P70Y`
  - intuitive use but w/o clear scoping (age... when?)
- `filters=PATO_0000011:>P70Y` ("age")
  - using a term for expressing the age quality of the ISO8601 duration
  - computationally more robust but w/o additional quality (age... when?)
- `filters=EFO_0004847:>P70Y` ("age at onset")
  - specific for an "onset" scope of the age value

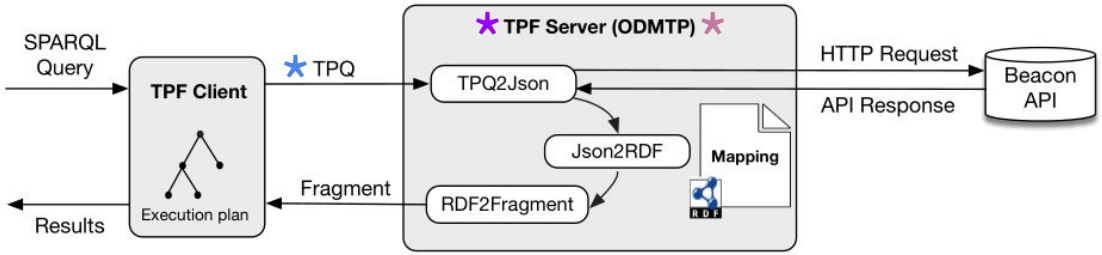
We recommend that implementers provide **term expansions** for equivalent terms, depending on the context. Also, it is up to the implementers to provide the correct tooling for e.g. transformation of input values (e.g. numerical age in years and comparator) to the standardized wire format (e.g. ages/durations are **always** transmitted as ISO8601 periods) as well as the correct deparsing and use (e.g. the ISO values probably will be converted to some numerical format for database matches).

@docs.genomebeacons.org

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

(2) RML Mapping for the Beacon API

Beacon API Response: json

→ Sending **HTTP requests** to the Beacon API "genomicVariants" endpoint

→ *"Do you have variants in chromosome 19 between coordinates 10,093,460 and 10,093,470 ?"*

→ Response in **json format**

RML Mapping

Set of rules that define how to convert the API response in a turtle format that **follows our Semantic representation**

RDF Representation: ttl

→ Follows our **Semantic representation** of genomic variants

→ **Linked Data**

→ Can be queried by SPARQL

(2) RML Mapping for the Beacon API

Beacon API Response: json

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

RDF Representation: ttl

```
* ol:iep10093466 a faldo:ExactPosition ;
        faldo:position 10093466 .
```

RML Mapping

```
_: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:termType rr:BlankNode ;
        rr:template "{start.value}" ;
        rr:class faldo:ExactPosition
  ] ;
  rr:predicateObjectMap [
        rr:predicate faldo:position ;
        rr:objectMap [ rml:reference"start.value" ;
                rr:termType rr:Literal ;
                rr:datatype xsd:integer ]
  ] .
```

* blank node

Increasing FAIRness of pheno-clinical genomic research dataset

---

Data heterogeneity & cie aspects

Semantic web technologies
- (+) Semantic definitions
- (+) Explicit relationships
- (+) High interoperability
- (-) Intense data transformation

Beacon data models:
- (+) International standard
- (-) Lack of semantics
- (-) Limited interoperability

Semantic Beacons: (+)(+)(+)(+)
- (-) Slow query resolution

Data privacy and governance aspects

Semantic web technologies:
- (+) Data Use/Access (DUO/ODRL)
- (-) No built-in access control

Beacon data models:
- (+) Granularity responses: boolean, aggregated, record level
- (-) No built-in access control

Semantic Beacons: (+)(+)
- (-) No built-in access control

Projet financé par
FRANCE 2030 PROGRAMME DE RECHERCHE SANTÉ NUMÉRIQUE
NEUROVASC
l'institut du thorax

Summary of Semantic Beacons

Conclusion

~ **Fresh on-the-fly biological annotations** of genomic data

~ **Minimizing** server side **costs** by using **externally maintained** biological annotations

~ Increasing FAIRness of health data silos using **community agreed ontologies** and **standards**

Next steps

~ Speed of query execution

~ Integrate other Knowledge bases

~ Additional semantic mapping of Beacon specifications

~ Provenance metadata

Beacon API: federations

Can you provide information about missense variants in ANGPTL6 in blood samples from patients with familial forms of intracranial aneurysms?

Beacon *v2* API

The Beacon API v2 represents a simple but powerful **genomics API** for ***federated*** data discovery and retrieval

below are unused slides