umr1087.univ-nantes.fr

Federated querying of genomic health data leveraging semantic web technologies and the Beacon standard for genomic discoverability

Alexandrina Bodrug-Schepers [1], Hugo Chabane [2], Gabriela Montoya [2], Patricia Serrano-Alvarado [2], Richard Redon [1], 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

SCIENTIFIC BACKGROUND

Biomedical challenge

Intracranial aneurysms (**ICA**) are neurovascular disorders characterised by saccular dilatations of cerebral vessels. They occur in 3 to 6% of the general population. Rupture is the only complication of ICA and although uncommon *i.e.* in less than 1% of carriers, it leads to death or severe disability in 2/3rds of rupture cases. Understanding ICA formation and rupture is challenging because it is a multifactorial complex disease with few screening solutions.

Project context

The ICAN project [1], a French non-interventional multicenter nationwide study, recruited individuals with ICA and collected neuroimaging, genetic, and phenotypic data. One of the objectives of the **PEPR Santé Numérique Neurovasc** is to make the ICAN dataset more aligned with **FAIR principles** (Findable, Accessible, Interoperable, Reusable) [2].

DATASET : ICAN BIOCOLLECTION AND PHENO-CLINICAL DATA

Individuals carrying ICA were recruited for the ICAN research projet and benefited from **neuroimaging, biosampling**, **genetic screening** and **pheno-clinical data** collection. The ICAN dataset is **heterogenic**, multi-centric and **multimodal**. As a genomic health dataset it has **data privacy** protection and **data governance** constraints.

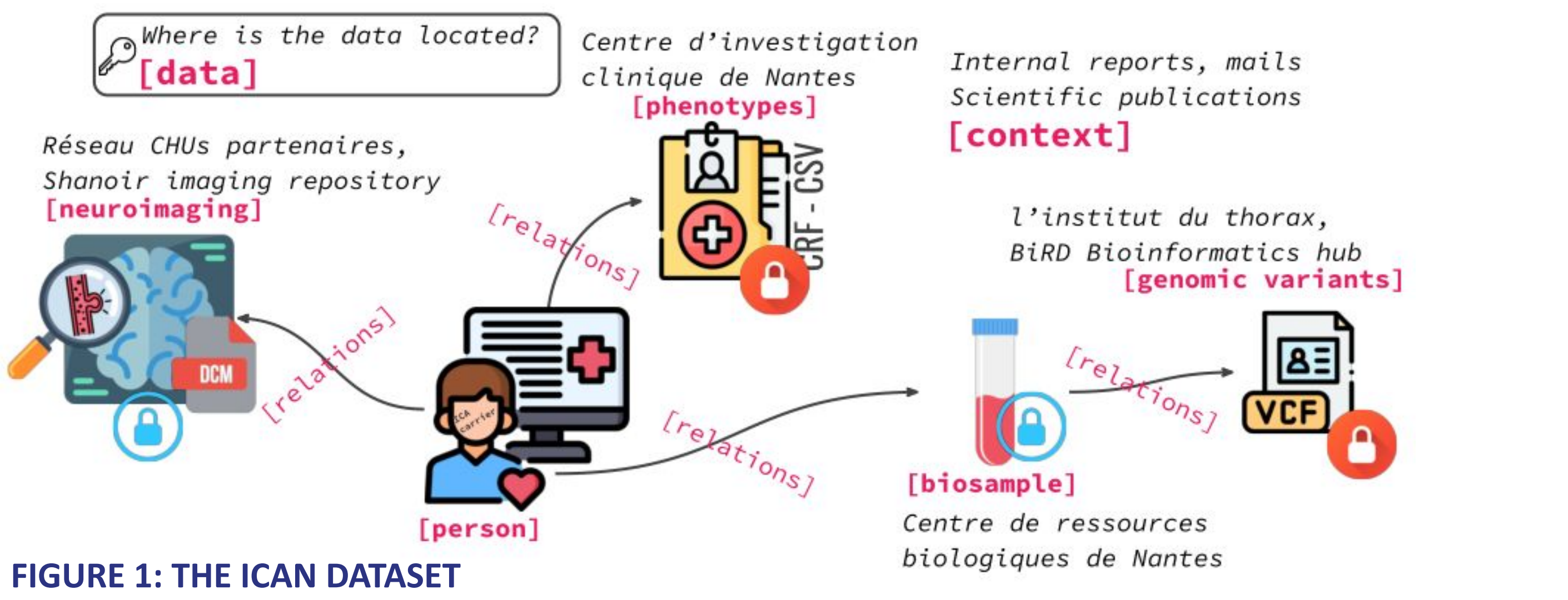


FIGURE 1: THE ICAN DATASET

The ICAN dataset is made of several multimodal domain-specific datasets. These are located in specialized laboratories that have technologies tailored to their specific needs. Privacy risk is higher with pheno-clinical data than with brain imaging, so different sharing constrains apply. Relationships between data chunks are often hard to retrieve and sparsely documented. Context of the study describing for example inclusion and exclusion criteria, can be found in scientific publications, internal reports and mails.

We simulated a synthetic dataset to be used for our publicly available **FAIR technological demonstrators**. We simulated phenotypic data by mimicking per column variable problabilities and eliminating mutually exclusive variables (*i.e.* no gestational diabetes in males). Genetic data was simulated using the methods described in the poster ***« Incorporating complex genetic model into risk stratification »*** (Laporte *et al.*)

INFO POINT : SYNTHETIC DATASET

FAIRification STRATEGIES

Existing solutions

Organisations such as **GA4GH**, **ELIXIR**, and **F-EGA** are collectively developing technical standards, policies, and software to facilitate genomic biomedical research. For the **FAIRification of the ICAN dataset**, we explored two main approaches.

The **Beacon standard** [3] focuses on genomic data discoverability and offers a query framework that incorporate **privacy** considerations. The Beacon standard is also a **REST API** specification.

Semantic web technologies make biomedical concepts and their relationships explicit through specialized ontologies, semantic data models, and **knowledge graphs (KGs)**. Many specialised high quality life science KGs are open access.

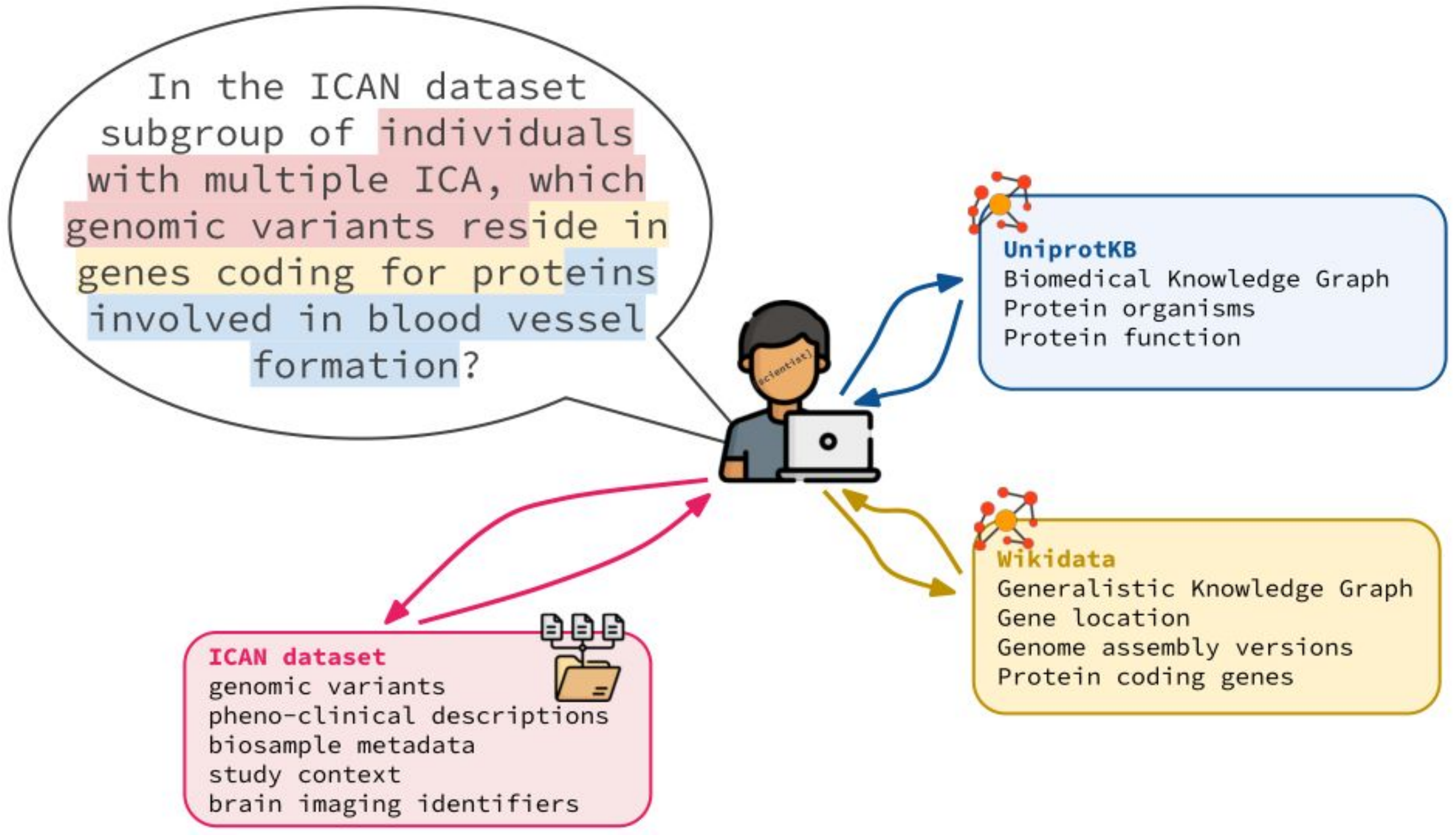
FAIR technological demonstrators

We developped **FAIR genomic technological demonstrators** to showcase different stategies, their advantages and their shortcomings.

CHALLENGES AND MOTIVATION USE CASE SCENARIO

We identified three main challenges to address : (1) **Share sensitive data** in a safe way , (2) **Semantify**, **integrate** and **query multimodal** health data and (3) **Integrate** health data with **public knowledge bases**.

We constructed a motivating question designed to showcase the technological solutions addressing our challenges and enabling the **FAIRification of the ICAN dataset**.


In the ICAN dataset subgroup of individuals with multiple ICA, which genomic variants reside in genes coding for proteins involved in blood vessel formation?
scientist
UniprotKB
Biomedical Knowledge Graph
Protein organisms
Protein function
Wikidata
Generalistic Knowledge Graph
Gene location
Genome assembly versions
Protein coding genes
ICAN dataset
genomic variants
pheno-clinical descriptions
biosample metadata
study context
brain imaging identifiers

KGs structure complex concepts and their relationships. They can be generalistic or more specialized, as is the case with **biomedical KGs**. Some example include the biomedical **UniprotKB** focusing on protein knowledge, and the generalistic cross-domain open KG **Wikidata**. All KGs are not open access. Their goals can include : acting as knowledge integration hubs, FAIR sharing, automated reasoning and inference aid, or as a domain specific decision support.

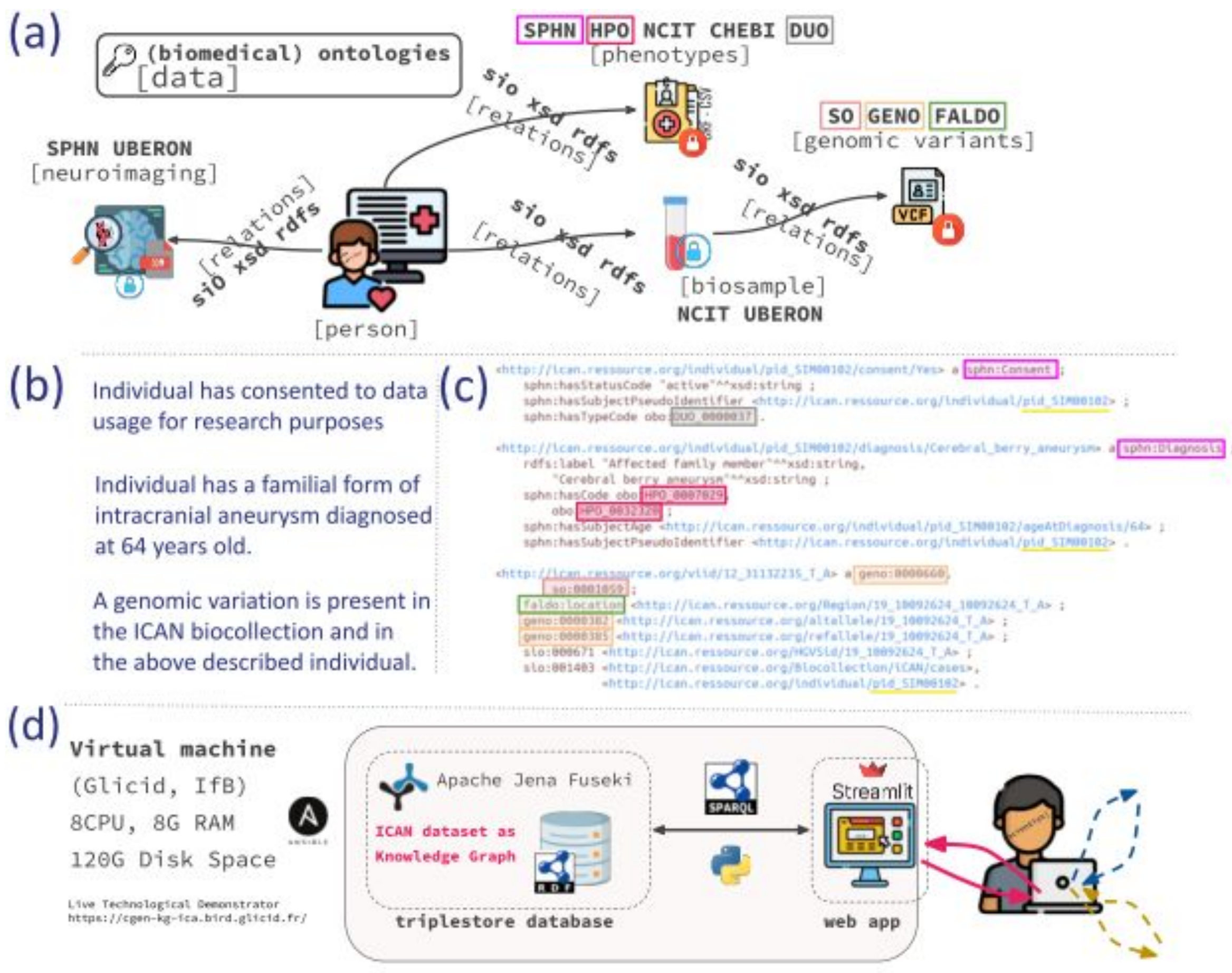
INFO POINT: BIOMEDICAL KGs

FIGURE 2: MOTIVATION USE CASE SCENARIO

Answering our motivational question requires drawing on knowledge from several separate sources. We consider the ICAN dataset as a single integrated source built to be interoperable with existing biomedical KGs through common controlled vocabularies and data models.

A FULL SEMANTIC APPROACH

We investigated broadly used open source domain expert ontologies. We chose the **Swiss Personalized Health Network** (SPHN) [4] framework to semantify our pheno-clinical data following their data model. We added descriptions using the **Human Phenotype Ontology** (HPO) [5]. We used several ontologies to represent genomic variation (SO, GENO) and genomic location (FALDO) [6], and developped our **own genomic variation data model** tailored to our needs.



Individual has consented to data usage for research purposes

Individual has a familial form of intracranial aneurysm diagnosed at 64 years old.

A genomic variation is present in the ICAN biocollection and in the above described individual.

```
<http://ican.ressource.org/individual/pid_SIM00102/consent/Yes> a sphn:Consent ;
    sphn:hasStatusCode "active"^^xsd:string ;
    sphn:hasSubjectPseudoIdentifier <http://ican.ressource.org/individual/pid_SIM00102> ;
    sphn:hasTypeCode obo:DUO_0000037 .

<http://ican.ressource.org/individual/pid_SIM00102/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_0032329 ;
    sphn:hasSubjectAge <http://ican.ressource.org/individual/pid_SIM00102/ageAtDiagnosis/64> ;
    sphn:hasSubjectPseudoIdentifier <http://ican.ressource.org/individual/pid_SIM00102> .

<http://ican.ressource.org/vlid/12_31132235_T_A> a geno:0000660,
        so:0001059 ;
    faldo:location <http://ican.ressource.org/Region/19_10092624_10092624_T_A> ;
    geno:0000382 <http://ican.ressource.org/altallele/19_10092624_T_A> ;
    geno:0000385 <http://ican.ressource.org/refallele/19_10092624_T_A> ;
    sio:000671 <http://ican.ressource.org/HGVSid/19_10092624_T_A> ;
    sio:001403 <http://ican.ressource.org/Biocollection/ICAN/cases>,
            <http://ican.ressource.org/individual/pid_SIM00102> .
```

FIGURE 3 : Integrating genomic health data as a KG

Existing biomedical ontologies (a) can be used to add semantics to health genomic data. This approach permits the precise representation of the dataset's concepts and relationships (b) in a machine readable RDF format (c). The RDF file can then be loaded in a triplestore database (as a KG), allowing automated reasoning and querying of multimodal health data ((d) - pink arrows), and integration with public KGs ((d) - blue and yellow dotted arrows).

Take away message

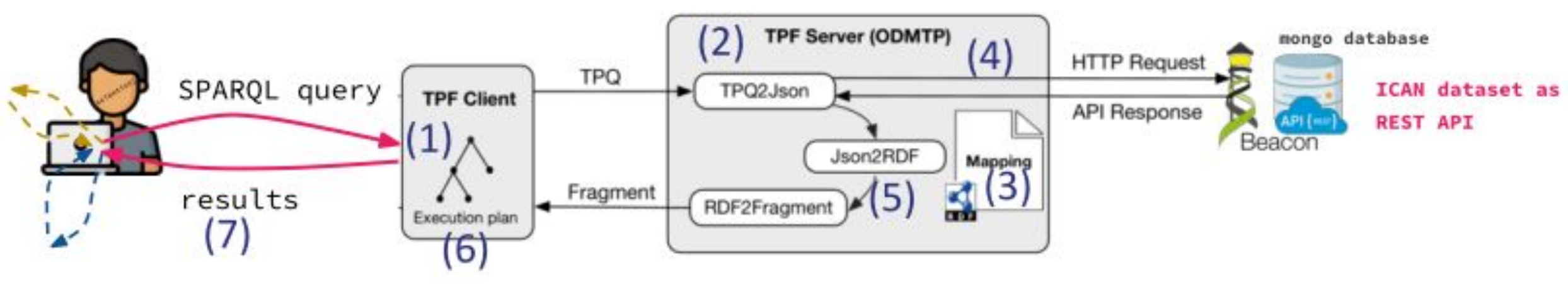
This approach responds to our 2nd and 3rd challenge. It **semantifies**, **integrates** and makes **multi-modal health data queryable** thanks to the **SPARQL query language**. Identical queries could be **distributed** to several organisation using the same framework *i.e.* Swiss Hospitals and Universities. Moreover, thanks to the **federation features** of SPARQL, the ICAN KG becomes **interoperable with UniprotKB and Wikidata**, making the **FIGURE 2** use case question possible.

SEMANTIC BEACON FRAMEWORK

We first deployed a Beacon API following an existing implementation [7]. This required data transformation to make the ICAN dataset follow the Beacon standard data model. A Beacon API deployer can configure the API **response granularity** : **Boolean** responses, **Aggregated** responses or **Record level** responses. This allows making the data discoverable by signaling its existence (**Boolean**) or sharing statistics (**Aggregated**), without giving detailed (**Record level**) access to unauthorized requests. This is a **powerful feature for sensitive data sharing** and adresses our 1st challenge. We then developped the **Semantic Beacon framework** [8] to make the communication possible between the Beacon REST API and existing biomedical knowledge graphs.

FIGURE 4: Semantic Beacon framework

The Semantic Beacon framework uses an ODMTP server as a mediator between the REST API and KGs. On Demand Mapping of Triple Patterns (ODMTP) framework integrates non-RDF datasets on-demand into Linked Data using the Triple Pattern Fragment (TPF) methodology [9]. Query decomposition into tripple tattern queries (TPQ) is delegated to the client (1). These are then passed to the TPF server (2) that uses a Semantic RML Mapping (3) of the genomic variants in the Beacon API to convert the TPQ into a HTTP requests (4). The API response is then mapped back to fragments (5) following the same RML mapping and the client performs the final reconstruction (6) (Execution plan) before sending back a Linked Data response (7).

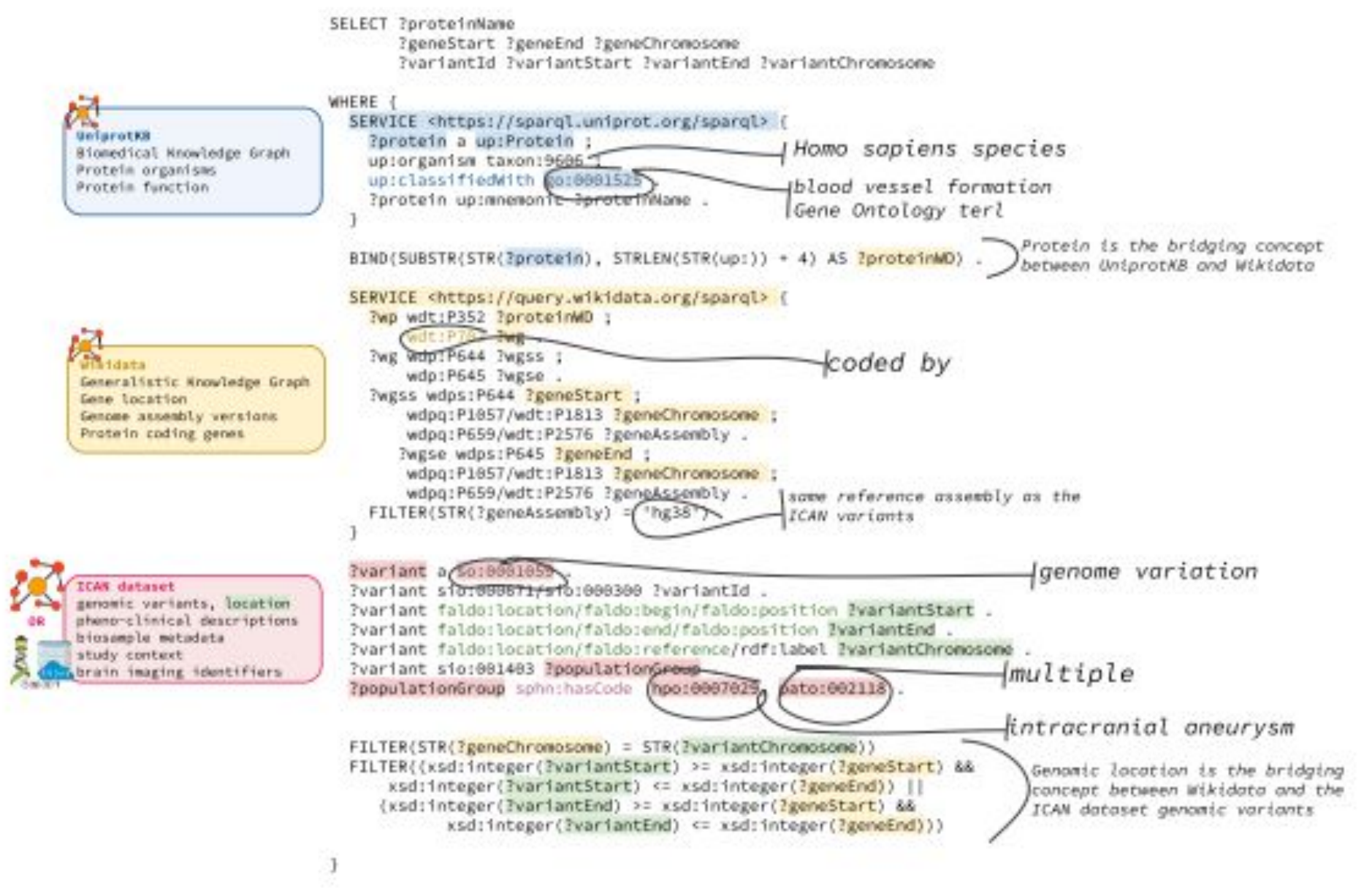


Take away message

The Semantic Beacon framework allows **on-the-fly integration of non-RDF data** *i.e.* Beacon API data. In other words genomic variants found within our Beacon API can be enriched with **fresh** public knowledge from KGs **without data duplication** or manual intergration. This framework responds to our 3rd challenge of **integration with public KGs** and offers an **innovate combination** of the biomedical research approach and the semantic web approach for sharing genomic health data knowledge. The Semantic Beacon was successfully deployed as a Proof Of Concept [8], however it is **still experimental** and has some limitations such **slow query execution**.

USE CASE SPARQL QUERY

SPARQL querying allows knowledge retrieval from KGs and from the Semantic Beacon. The question from motivating use case scenario in **FIGURE 2** can translated to such a query.


UniprotKB
Biomedical Knowledge Graph
Protein organisms
Protein function
Wikidata
Generalistic Knowledge Graph
Gene location
Genome assembly versions
Protein coding genes
ICAN dataset
genomic variants, location
pheno-clinical descriptions
biosample metadata
study context
brain imaging identifiers
OR
Homo sapiens species
blood vessel formation
Gene Ontology terl
Protein is the bridging concept between UniprotKB and Wikidata
coded by
same reference assembly as the ICAN variants
genome variation
multiple
intracranial aneurysm
Genomic location is the bridging concept between Wikidata and the ICAN dataset genomic variants

```
SELECT ?proteinName
       ?geneStart ?geneEnd ?geneChromosome
       ?variantId ?variantStart ?variantEnd ?variantChromosome

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:0007025, 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)))

}
```

FIGURE 5: Federated SPARQL querying

The query fetches information about proteins and their function in UniportKB, about gene coding products and gene location in Wikidata and about genomic variation location and their association to phenotypes in the ICAN KG or the Semantic Beacon API. The FALDO ontology (in green) is used to precisely describe variation location. The query looks for the same graph structure as the ones represented in the RDF file in **FIGURE 3-(c)**

REFERENCES

[1] Bourcier R, Chatel S, Bourcereau E, et al. Understanding the Pathophysiology of Intracranial Aneurysm: The ICAN Project. Neurosurgery. 2017;80(4):621. doi:10.1093/neuros/nyw135

[2] Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. doi:10.1038/sdata.2016.18

[3] Rambla J, Baudis M, Ariosa R, et al. Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Hum Mutat. 2022;43(6):791-799. doi:10.1002/humu.24369

[4] Touré V, Krauss P, Gnodtke K, et al. FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network. Sci Data. 2023;10(1):127. doi:10.1038/s41597-023-02028-y

[5] Gargano MA, Matentzoglu N, Coleman B, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res. 2024;52(D1):D1333-D1346. doi:10.1093/nar/gkad1005

[6] Bolleman JT, Mungall CJ, Strozzi F, et al. FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. J Biomed Semant. 2016;7(1):39. doi:10.1186/s13326-016-0067-z

[7] Rueda M, Ariosa R, Moldes M, Rambla J. Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data. Bioinformatics. 2022;38(19):4656-4657. doi:10.1093/bioinformatics/btac568

[8] Bodrug-Schepers A, Chabane H, Montoya G, Redon R, Gaignard A, Serrano-Alvarado P. Semantic Beacons: a framework to support federated querying over genomic variants and public Knowledge Graphs. In: SWAT4HCLS 2025. 2025. Accessed June 11, 2025. https://hal.science/hal-04908530

[9] Moreau B, Serrano-Alvarado P, Desmontils E, Thoumas D. Querying non-RDF Datasets using Triple Patterns. In: 2017. Accessed October 8, 2025. https://hal.science/hal-01583518

ACKNOWLEDGMENTS

This work was partially funded by the French government, through the National Research Agency (ANR), under the "France 2030" program with reference ANR-22-PESN-0008.
We thank Glicid and IfB for providing computational ressources, Fabien Laporte for genetic synthetic data simulation, Matilde Karakachoff for the help with clinical data and Laurent Vallet for help with neuroimaging metadatadata explanations.

Alexandrina Bodrug
alexandrina.bodrug@univ-nantes.fr
l'Unité de Recherche de l'institut du Thorax
Bâtiment 06, IRS UN - 8 quai Moncousu 44007 NANTES

funded by ANR-22-PESN-0008