Creating, Querying and Publishing an RDF Graph

Semantic Data Explotation

MSc in Bioinformatics

Noelia Moreno González

2023/2024

Índice

1. Introduction

Understanding the involvement of various genes in different types of cancer is fundamental for unraveling the underlying mechanisms of the disease.

In this context, the **APC**, **KRAS**, **TP53**, and **BRAF** genes, closely associated with the pathogenesis of **colorectal**, **pancreatic**, and **lung cancers**, will be examined. These genes are key players in understanding the pathophysiology of each cancer type and represent potential targets for targeted therapy and important biomarkers for early detection and prognosis. Notable studies have highlighted the impact of mutations in TP53 and KRAS on colorectal cancer (Chee et al, 2024; Zhu et al, 2021), while comprehensive analyses have verified KRAS mutations in various cancer types, including lung and pancreatic cancers (Li et al, 2019). Additionally, the role of TP53 mutations in modulating antitumor immunity has been significantly documented (Li et al, 2020).

In addition, they are compared to their orthologous genes in common mouse (*Mus musculus*), allowing a deeper understanding of the function of these genes in an evolutionary and preclinical context. This cross-species comparison provides additional insight into the conservation of molecular pathways and their relevance to the study of cancer in humans.

Exploring the interactions between these genes and various types of cancer provides a comprehensive understanding of the intricate molecular network underlying cancer development. Such an expanded approach yields critical insights for cancer research and opens up new avenues for developing more efficacious and personalized cancer therapies.

2. Objectives

1. To **construct an RDF graph** representing the relationships between various genes in humans (*Homo sapiens*) and their associated proteins in different types of cancer, establishing connections between the genes studied and their orthologues in the common mouse (*Mus musculus*) for a more complete understanding of their function and evolutionary relevance.

2. To **develop queries** in Blazegraph and R to extract specific information from the RDF graph.

3. To **publish** the RDF knowledge graph of genes involved and its relationshipwith cancer on the Pubby platform.

3. Files location

Todos los ficheros mencionados en este informe se encuentran ubicados en la siguiente ruta del servidor dayhoff: **/home/alumno22/ESD/Entrega**.

4. RDF graph creation

First of all, the prefixes for each URI used, shown in **Table 1**, must be defined. Once this is done, start creating the graph.

Tabla 1: Prefixes and their corresponding URIs

PREFIX	**URI**
rdfs	`http://www.w3.org/2000/01/rdf-schema#`
rdf	`https://www.w3.org/1999/02/22-rdf-syntax-ns#`
owl	`http://www.w3.org/2002/07/owl#`
afo	`http://purl.allotrope.org/ontologies/property/#`
obo	`http://purl.obolibrary.org/obo/`
ncit	`https://ncit.nci.nih.gov/ncitbrowser/#`
gene	`https://www.ncbi.nlm.nih.gov/gene/`
protein	`https://www.ncbi.nlm.nih.gov/protein/`
plength	`https://www.ncbi.nlm.nih.gov/protein/length/`
up	`http://www.uniprot.org/uniprot/`
ensembl	`http://www.ensembl.org/id/`
pub	`https://pubmed.ncbi.nlm.nih.gov/`
dcterms	`http://purl.org/dc/terms/`
cd	`https://cancerdata.org/`
low_exp	`https://cancerdata.org/low_expression/`
high_exp	`https://cancerdata.org/high_expression/`
chr	`https://cancerdata.org/chromosome/`

4.1. Turtle file creation

The Turtle format (**Terse RDF Triple Language**) is a notation for representing data as triples in the Semantic Web model known as RDF (**Resource Description Framework**). It uses prefixes to abbreviate long URIs, allows the representation of literals, and includes comments, making data more readable and understandable.

These triplets follow the structure shown in the **Figure 1**, with **Node 1** being an entity (such as, in this case, a *gene*) that is linked to **Node 2** (a *protein*) by a continuous line, representing a relationship between the two (Node 1 encodes Node 2). On the other hand, Node 1 presents another relationship with a dashed line to a **Datatype**, which could be a *comment of the gene*.

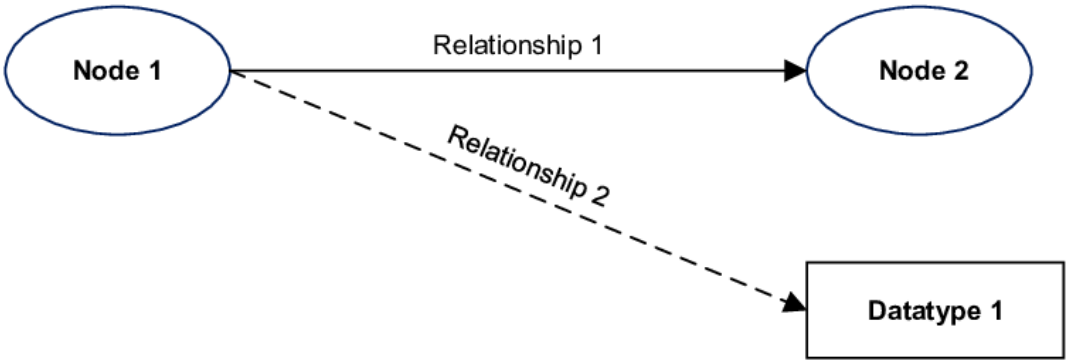

Figura 1: RDF graph, consisting of two nodes and one relation (Sahlab et al, 2022).

To establish the connections within this graph, the subjects have been categorized into 5 groups: Specie, Gene, Protein, Cancer, and Publication.

4.1.1. Specie

As specified above, we work with two species (*Homo sapiens* and *Mus musculus*), with two properties each:

Property	**Prefix**	**Object**	**Prefix**
type	rdf	**General Specie NCIt Code**	ncit
label	rdfs	**Specie Name**	-

4.1.2. Gene

In the case of genes, the 4 genes that have been found to be involved in the cancers studied for humans are taken as subjects, as are their orthologues in the common mouse. In the case of humans, there are 9 properties, while in mice, there are 8, as the latter does not exhibit the ortholog property.

Property	Prefix	Object	Prefix
type	rdf	**Specie Name**	specie
label	rdfs	**Gene Name**	-
type	rdf	**General Gene Code**	ncit
sameAs	owl	**Specific Gene Code**	ncit
sameAs	owl	**Ensembl ID**	ensembl
***Located in* code**	obo	**Chromosome**	chr
***Encondes* code**	obo	**Protein Accession**	up
Comment	rdfs	**Comment**	-
***Orthologue* code**	ncit	**Orthologue Gene Name**	gene

4.1.3. Protein

On the other hand, the proteins encoded by the genes seen above are taken as subjects. These have 6 properties.

Property	Prefix	Object	Prefix
type	rdf	**General Protein Code**	specie
label	rdfs	**Protein Name**	-
sameAs	owl	**Protein Accession**	up
***Length* code**	ncit	**Numer of aa**	-
***Enconded by* code**	obo	**Gene name**	gene
Comment	rdfs	**Comment**	-

4.1.4. Cancer

This group includes the three associated cancers: colorectal, pancreatic, and lung. Each of them exhibits 8 properties.

Property	Prefix	Object	Prefix
type	rdf	**General Cancer Code**	ncit
label	rdfs	**Cancer Name**	-
sameAs	owl	**Specific Cancer Code**	ncit
***Affects* code**	afo	**Protein Name**	protein
***Low* expression**	exp	**Protein Name**	protein
***High* expression**	exp	**Protein Name**	protein
Ref	dcterms	**Publication ID**	publi
Comment	rdfs	**Comment**	-

4.1.5. Publication

This group presents as subjects some of the recent publications that have been found for these 3 types of cancer, having 4 attributes each.

Property	Prefix	Object	Prefix
class	rdf	**General Publication Code**	ncit
sameAs	owl	**Pubmed ID**	pub
title	dcterms	**Publication Title**	-
date	dcterms	**Publication Date**	-

All these subjects together with their attributes give rise to a total of **189 triplets** in turtle format, contained in the file **turtle.txt**.

4.2. RDF/XML file creation and Blazegraph repository

Once the set of triplets is in **Turtle format**, it needs to be converted to **RDF/XML format** to load the dataset into the **Blazegraph** triple store. RDF/XML (**Resource Description Framework/XML**) is a format for representing structured data in the Semantic Web model called RDF. It's an XML-based format that describes resources and their relationships using XML tags. Each RDF triple is represented as an XML element containing a subject, a predicate, and an object.

To carry out this transformation, the **EasyRDF Converter** tool is used. It verifies the Turtle syntax and converts it to the desired format, RDF/XML, in this case. The following figures show the input (**Figure 2**) and output (**Figure 3**) determined in this case.

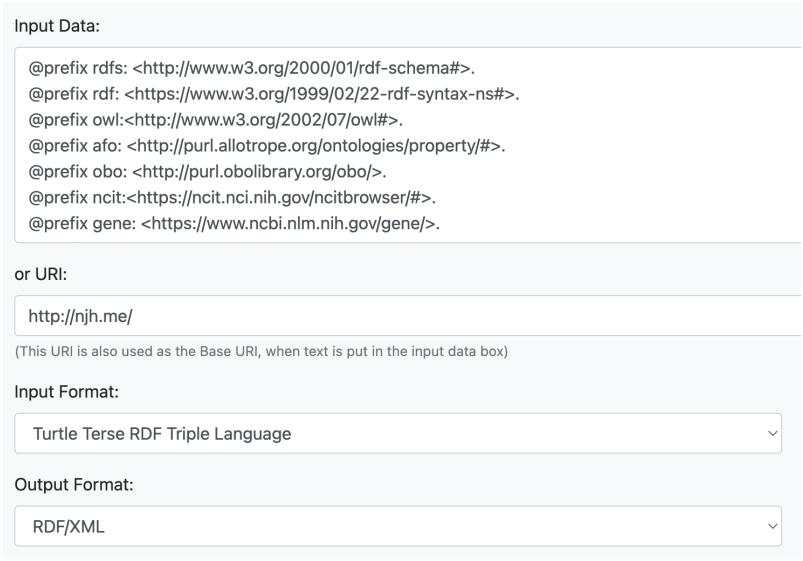

Figura 2: EasyRDF Converter input.

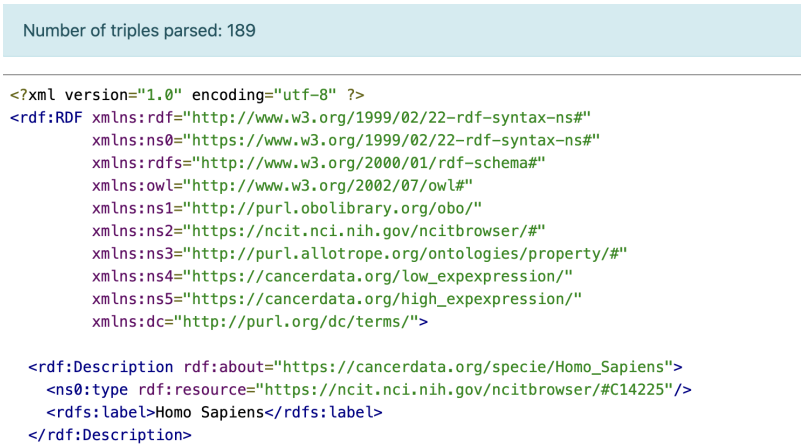

Figura 3: EasyRDF Converter output.

The RDF/XML format file is called **rdfxml.txt**. In order to ensure that the format is correct, a second check of the EasyRDF output is carried out in the **W3 RDF Validator**. Apart from checking the syntax, it displays the graph (**Figure 4**) that generates the dataset.

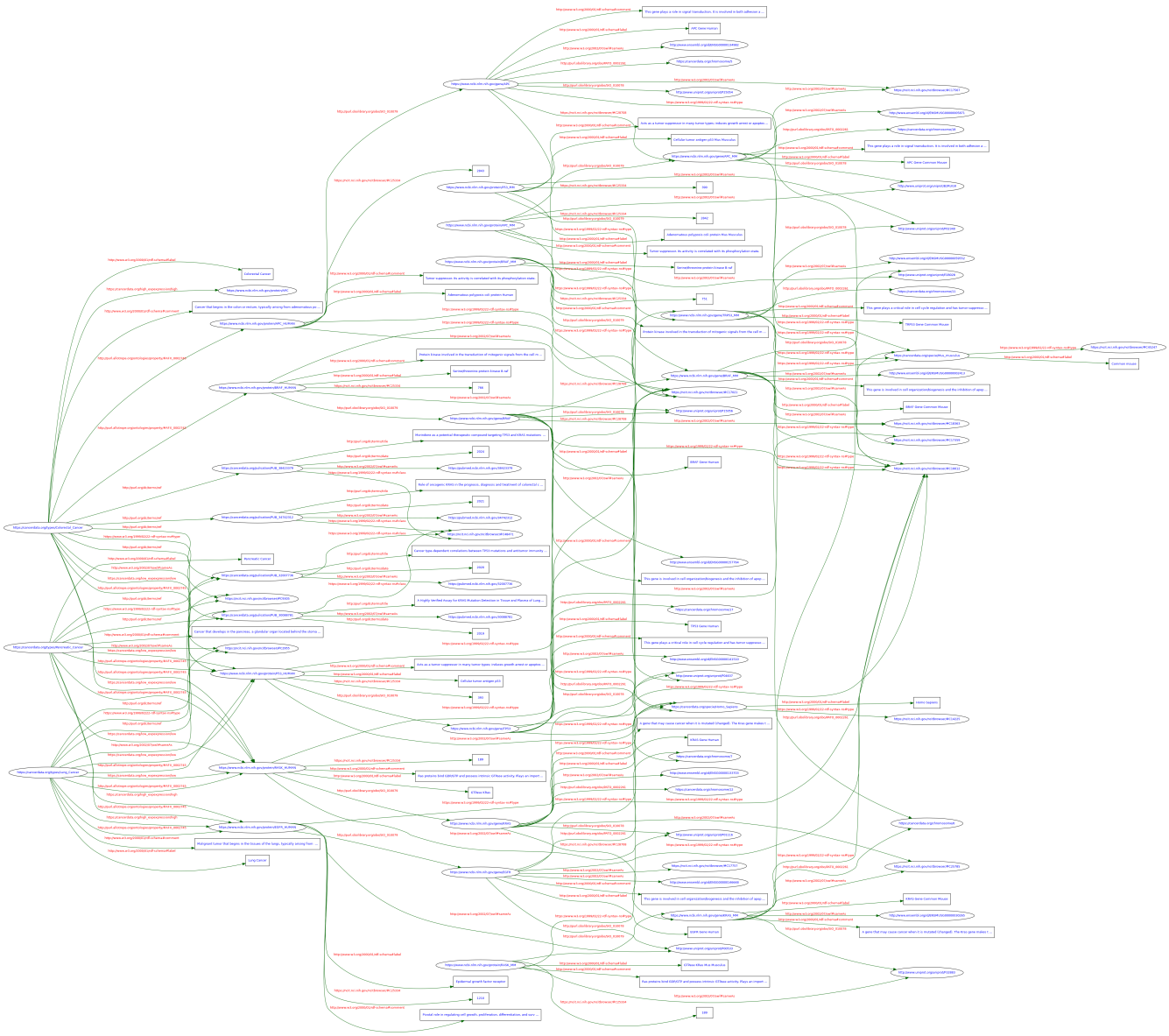

Figura 4: Graph representation.

For a more detailed view, below is the graph representation generated by a gene with its species and the protein it encodes (**Figure 5**).

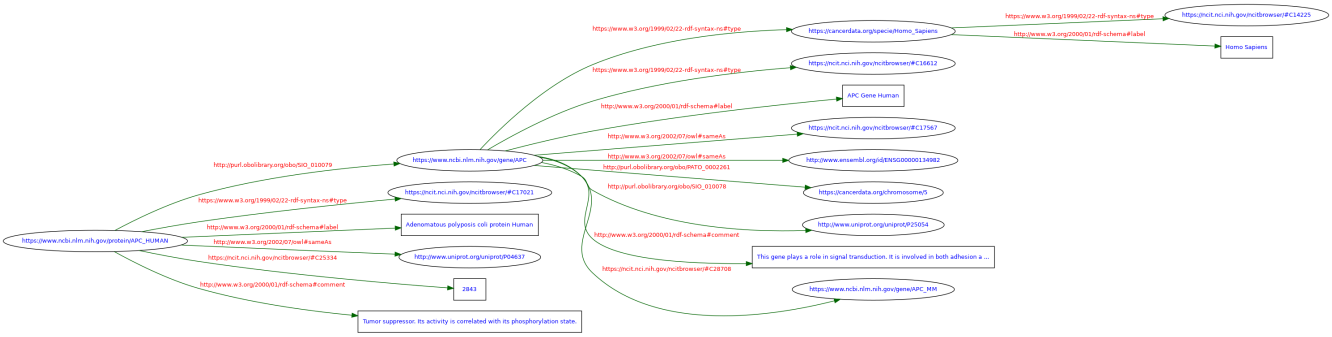

Figura 5: Partial graph representation.

Once the syntax has been verified, the file is uploaded to the **Blazegraph** repository, creating a **Namespace** beforehand that will contain the dataset.

5. Metadata

The **metadata.txt** file organizes and provides detailed metadata about the Cancer dataset along with links to various formats of the dataset distributions. Here is an overview of each section within the file:

1. **Prefixes**: The file includes several prefixes that simplify the referencing of URIs for various vocabularies and schemas used within the metadata.

2. **Dataset Description**: The dataset, referred to as **cancerdata** dataset, is described using the *dcat:Dataset* class. Attributes include an *rdfs:label* with the title **Relationship between genes and cancer types**, a *foaf:primaryTopic* set to Cancer, and a *dct:License* specifying its use under the **MIT License**. The dataset's URI is also provided: `https://cancerdata.org/`.

3. **Distributions**: These distributions are classified under the *dcat:Distribution* type and includes a download URL (*dcat:downloadURL*), a specified license (*dct:license*), and a format (*dct:format*).

4. **SPARQL Access Point**: There's a distribution set up for SPARQL queries at `http://155.54.239.183:3039/blazegraph/namespace/Cancer/sparql`, enabling direct querying on the semantic dataset.

5. **RDF Graph**: The dataset includes an RDF graph named Relationship between Genes and Cancer types RDF Graph", described using the *sd:NamedGraph* class.

6. **MIT License Details**: The file also details the MIT license under which the dataset is distributed. This includes the type of permissions granted (cc:permits), the classification of the license (*cc:licenseClass*), and a URL to the legal code (*cc:legalcode*).

This structured approach ensures that all necessary information regarding the dataset's content, usage, and access is clearly communicated and easily accessible.

6. Queries

Once the data has been uploaded to Blazergraph, queries can be carried out in the **QUERY** section. On the other hand, the **R environment** will also be adapted to allow queries to be carried out there. To do so, the SPARQL library must be loaded, as well as the endpoint of the Blazergraph repository, as shown below.

Consulta R

```
library(SPARQL)
endpoint <- "http://155.54.239.183:3039/blazegraph/namespace/Cancer/sparql
    "
```

The 5 queries performed are presented below.

6.1. Query 1

First, a simple query is performed to show the list of genes present in the dataset and, if available, the proteins they encode with their respective tags, filtering by human species.

SPARQL Query

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX ensembl: <http://www.ensembl.org/id/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX specie: <https://cancerdata.org/specie/>

SELECT DISTINCT ?gene ?geneLabel ?protein ?proteinLabel
WHERE {
  ?gene rdf:type ncit:C16612 ;
        rdfs:label ?geneLabel .

  OPTIONAL {
    ?protein rdf:type ncit:C17021 ;
             rdfs:label ?proteinLabel ;
             obo:SIO_010079 ?gene .
  }

    ?gene rdf:type specie:Homo_Sapiens .
}
```

R Query

```
library(SPARQL)
endpoint <- "http://155.54.239.183:3039/blazegraph/namespace/Cancer/sparql
   "
query1 <- "
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX ensembl: <http://www.ensembl.org/id/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX specie: <https://cancerdata.org/specie/>

SELECT DISTINCT ?gene ?geneLabel ?protein ?proteinLabel
WHERE {
  ?gene rdf:type ncit:C16612 ;
        rdfs:label ?geneLabel .

  OPTIONAL {
    ?protein rdf:type ncit:C17021 ;
             rdfs:label ?proteinLabel ;
             obo:SIO_010079 ?gene .
  }

    ?gene rdf:type specie:Homo_Sapiens .
}
"
qd1 <- SPARQL(endpoint, query1)
View(qd1$results)
```

The result obtained after these queries is shown in the **Figure 6**.

gene	geneLabel	protein	proteinLabel
<https://www.ncbi.nlm.nih.gov/gene/BRAF>	BRAF Gene Human	<https://www.ncbi.nlm.nih.gov/protein/BRAF_HUMAN>	Serine/threonine-protein kinase B-raf
<https://www.ncbi.nlm.nih.gov/gene/EGFR>	EGFR Gene Human	<https://www.ncbi.nlm.nih.gov/protein/EGFR_HUMAN>	Epidermal growth factor receptor
<https://www.ncbi.nlm.nih.gov/gene/TP53>	TP53 Gene Human	<https://www.ncbi.nlm.nih.gov/protein/P53_HUMAN>	Cellular tumor antigen p53
<https://www.ncbi.nlm.nih.gov/gene/KRAS>	KRAS Gene Human	<https://www.ncbi.nlm.nih.gov/protein/RASK_HUMAN>	GTPase KRas
<https://www.ncbi.nlm.nih.gov/gene/APC>	APC Gene Human	<https://www.ncbi.nlm.nih.gov/protein/APC_HUMAN>	Adenomatous polyposis coli protein Human

Figura 6: Query 1 output.

6.2. Query 2

This query is designed to identify orthologous genes between humans and mice, an important step for biomedical research, especially in the development and testing of therapies. By studying these genes in mice, researchers can gain valuable insights into their functions and regulations in humans. It is requested to display the chromosome on which the gene is located in humans, as well as the protein encoded by both the human gene and its ortholog in mice. Lastly, a brief description of the gene is shown in order to provide more information about the context.

SPARQL Query

```
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX specie: <https://cancerdata.org/specie/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX cancer: <https://cancerdata.org/types/>
PREFIX afo: <http://purl.allotrope.org/ontologies/property/#>

SELECT ?humanGeneLabel ?humanChromosome ?humanProteinLabel ?
   mouseProteinLabel ?humanGeneComment
WHERE {
  ?humanGene rdf:type specie:Homo_Sapiens ;
             rdfs:label ?humanGeneLabel ;
             ncit:C28708 ?mouseGene ;
             obo:PATO_0002261 ?humanChromosomeUri ;
             obo:SIO_010078 ?humanProteinUri ;
             rdfs:comment ?humanGeneComment .
  ?mouseGene rdf:type specie:Mus_musculus ;
             rdfs:label ?mouseGeneLabel ;
             obo:SIO_010078 ?mouseProteinUri .
  ?protein rdf:type ncit:C17021 ;
           obo:SIO_010079 ?humanGene .

  BIND(REPLACE(STR(?humanChromosomeUri), "https://cancerdata.org/
   chromosome/", "") AS ?humanChromosome)
  BIND(REPLACE(STR(?humanProteinUri), "http://www.uniprot.org/uniprot/", "
   ") AS ?humanProteinLabel)
  BIND(REPLACE(STR(?mouseProteinUri), "http://www.uniprot.org/uniprot/", "
   ") AS ?mouseProteinLabel)
}
```

R Query

```
query2 <- "
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX specie: <https://cancerdata.org/specie/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX cancer: <https://cancerdata.org/types/>
PREFIX afo: <http://purl.allotrope.org/ontologies/property/#>

SELECT ?humanGeneLabel ?humanChromosome ?humanProteinLabel ?
   mouseProteinLabel ?humanGeneComment
WHERE {
  ?humanGene rdf:type specie:Homo_Sapiens ;
             rdfs:label ?humanGeneLabel ;
             ncit:C28708 ?mouseGene ;
             obo:PATO_0002261 ?humanChromosomeUri ;
             obo:SIO_010078 ?humanProteinUri ;
             rdfs:comment ?humanGeneComment .
  ?mouseGene rdf:type specie:Mus_musculus ;
             rdfs:label ?mouseGeneLabel ;
             obo:SIO_010078 ?mouseProteinUri .
  ?protein rdf:type ncit:C17021 ;
           obo:SIO_010079 ?humanGene .

  BIND(REPLACE(STR(?humanChromosomeUri), 'https://cancerdata.org/
   chromosome/', '') AS ?humanChromosome)
  BIND(REPLACE(STR(?humanProteinUri), 'http://www.uniprot.org/uniprot/',
   '') AS ?humanProteinLabel)
  BIND(REPLACE(STR(?mouseProteinUri), 'http://www.uniprot.org/uniprot/',
   '') AS ?mouseProteinLabel)
}
"
qd2 <- SPARQL(endpoint, query2)
View(qd2$results)
```

The result obtained after these queries is shown in the **Figure 7**.

humanGeneLabel	humanChromosome	humanProteinLabel	mouseProteinLabel	
APC Gene Human	5	P25054	B2RUG9	This gene plays a role in signal transduction. It is inv
BRAF Gene Human	7	P15056	P28028	This gene is involved in cell organization/biogenesis an
BRAF Gene Human	7	P15056	P28028	This gene is involved in cell organization/biogenesis an
KRAS Gene Human	12	P01116	P32883	A gene that may cause cancer when it is mutated (changed
TP53 Gene Human	17	P04637	P02340	This gene plays a critical role in cell cycle regulation

Figura 7: Query 2 output.

6.3. Query 3

In this query, proteins found in the dataset are retrieved along with their respective UniProt ID, sorted by their length from shortest to longest. It also identifies the types of cancer in which they are implicated. Finally, it indicates whether they are underexpressed or overexpressed.

SPARQL Query

```
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cancer: <https://cancerdata.org/types/>
PREFIX low_exp: <https://cancerdata.org/low_expexpression/>
PREFIX high_exp: <https://cancerdata.org/high_expexpression/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?proteinName (REPLACE(STR(?uniProtURI), "http://www.uniprot.org/
   uniprot/", "") AS ?uniProtID) ?cancerType ?proteinLength ?
   expressionLevel
WHERE {
  {
    ?cancer rdf:type ncit:C9305 ;
            rdfs:label ?cancerType ;
            low_exp:low ?protein .
    ?protein rdf:type ncit:C17021 ;
             rdfs:label ?proteinName ;
             ncit:C25334 ?proteinLength ;
             owl:sameAs ?uniProtURI .
    BIND("Low" AS ?expressionLevel)
  }
  UNION
  {
    ?cancer rdf:type ncit:C9305 ;
            rdfs:label ?cancerType ;
            high_exp:high ?protein .
    ?protein rdf:type ncit:C17021 ;
             rdfs:label ?proteinName ;
             ncit:C25334 ?proteinLength ;
             owl:sameAs ?uniProtURI .
    BIND("High" AS ?expressionLevel)
  }
}
ORDER BY xsd:integer(?proteinLength)
```

R Query

```
query3 <- "
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cancer: <https://cancerdata.org/types/>
PREFIX low_exp: <https://cancerdata.org/low_expexpression/>
PREFIX high_exp: <https://cancerdata.org/high_expexpression/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?proteinName (REPLACE(STR(?uniProtURI), 'http://www.uniprot.org/
   uniprot/', '') AS ?uniProtID) ?cancerType ?proteinLength ?
   expressionLevel
WHERE {
  {
    ?cancer rdf:type ncit:C9305 ;
            rdfs:label ?cancerType ;
            low_exp:low ?protein .
    ?protein rdf:type ncit:C17021 ;
             rdfs:label ?proteinName ;
             ncit:C25334 ?proteinLength ;
             owl:sameAs ?uniProtURI .
    BIND('Low' AS ?expressionLevel)
  }
  UNION
  {
    ?cancer rdf:type ncit:C9305 ;
            rdfs:label ?cancerType ;
            high_exp:high ?protein .
    ?protein rdf:type ncit:C17021 ;
             rdfs:label ?proteinName ;
             ncit:C25334 ?proteinLength ;
             owl:sameAs ?uniProtURI .
    BIND('High' AS ?expressionLevel)
  }
}
ORDER BY xsd:integer(?proteinLength)

"
qd3 <- SPARQL(endpoint, query3)
View(qd3$results)
```

The result obtained after these queries is shown in the **Figure 8**.

proteinName	uniProtID	cancerType	proteinLength	expressionLevel
GTPase KRas	P01116	Colorectal Cancer	189	Low
GTPase KRas	P01116	Lung Cancer	189	Low
GTPase KRas	P01116	Pancreatic Cancer	189	Low
Cellular tumor antigen p53	P04637	Colorectal Cancer	393	Low
Cellular tumor antigen p53	P04637	Lung Cancer	393	Low
Cellular tumor antigen p53	P04637	Pancreatic Cancer	393	Low
Epidermal growth factor receptor	P00533	Lung Cancer	1210	High
Epidermal growth factor receptor	P00533	Pancreatic Cancer	1210	High
Adenomatous polyposis coli protein Human	P04637	Colorectal Cancer	2843	High

Figura 8: Query 3 output.

6.4. Query 4

The goal is to retrieve all publications related to colorectal cancer, sorted from the most recent to the least. The query will display both the title of the publication and its PubMed ID, aswell as the year in which they were published.

SPARQL Query

```
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX publi: <https://cancerdata.org/pulication/>
PREFIX cancer: <https://cancerdata.org/types/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?pubmedID ?date ?title
WHERE {
  ?cancer rdf:type ncit:C9305 ;
          rdfs:label ?cancerLabel ;
          dcterms:ref ?publication .
  FILTER (?cancerLabel = "Colorectal Cancer")
  ?publication rdf:class ncit:C48471 ;
               owl:sameAs ?pubURL ;
               dcterms:title ?title ;
               dcterms:date ?date .
  BIND(REPLACE(STR(?pubURL), "https://pubmed.ncbi.nlm.nih.gov/", "") AS ?
   pubmedID)
}
ORDER BY DESC(?date)
```

R Query

```
query4 <- "
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX publi: <https://cancerdata.org/pulication/>
PREFIX cancer: <https://cancerdata.org/types/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?pubmedID ?date ?title
WHERE {
  ?cancer rdf:type ncit:C9305 ;
          rdfs:label ?cancerLabel ;
          dcterms:ref ?publication .
  FILTER (?cancerLabel = 'Colorectal Cancer')
  ?publication rdf:class ncit:C48471 ;
               owl:sameAs ?pubURL ;
               dcterms:title ?title ;
               dcterms:date ?date .
  BIND(REPLACE(STR(?pubURL), 'https://pubmed.ncbi.nlm.nih.gov/', '') AS ?
   pubmedID)
}
ORDER BY DESC(?date)
"
qd4 <- SPARQL(endpoint, query4)
View(qd4$results)
```

The result obtained after these queries is shown in the **Figure 9**.

pubmedID	date	title
38423379	2024	Morindone as a potential therapeutic compound targeting TP53 and KRAS mutations in colorectal cancer cells
34742312	2021	Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer
32007736	2020	Cancer type-dependent correlations between TP53 mutations and antitumor immunity
30088781	2019	A Highly Verified Assay for KRAS Mutation Detection in Tissue and Plasma of Lung, Colorectal, and Pancreatic Cancer

Figura 9: Query 4 output.

6.5. Query 5

Finally, the aim is to obtain information about the protein with the highest molecular weight, including a brief description of its function, the gene that encodes it, and its ortholog in mice. Additionally, it will show in which of the studied cancers this protein is implicated.

SPARQL Query

```
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX afo: <http://purl.allotrope.org/ontologies/property/#>
PREFIX cancer: <https://cancerdata.org/types/>

SELECT ?proteinLabel ?geneLabel ?orthologGeneLabel ?cancerType ?
   maxProteinLength ?proteinComment
WHERE {
  {
    SELECT ?protein (MAX(xsd:integer(?length)) AS ?maxProteinLength)
    WHERE {
      ?protein ncit:C25334 ?length .
    }
    GROUP BY ?protein
  }
  ?protein rdfs:label ?proteinLabel .
  ?protein rdfs:comment ?proteinComment .
  ?protein obo:SIO_010079 ?gene .
  ?gene rdfs:label ?geneLabel .
  ?gene ncit:C28708 ?orthologGene .
  ?orthologGene rdfs:label ?orthologGeneLabel .

  ?cancer afo:AFX_0002745 ?protein ;
          rdfs:label ?cancerType .
}
ORDER BY DESC(?maxProteinLength)
LIMIT 1
```

R Query

```
query5 <- "
PREFIX rdf: <https://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ncit: <https://ncit.nci.nih.gov/ncitbrowser/#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX afo: <http://purl.allotrope.org/ontologies/property/#>
PREFIX cancer: <https://cancerdata.org/types/>

SELECT ?proteinLabel ?geneLabel ?orthologGeneLabel ?cancerType ?
   maxProteinLength ?proteinComment
WHERE {
  {
    SELECT ?protein (MAX(xsd:integer(?length)) AS ?maxProteinLength)
    WHERE {
      ?protein ncit:C25334 ?length .
    }
    GROUP BY ?protein
  }
  ?protein rdfs:label ?proteinLabel .
  ?protein rdfs:comment ?proteinComment .
  ?protein obo:SIO_010079 ?gene .
  ?gene rdfs:label ?geneLabel .
  ?gene ncit:C28708 ?orthologGene .
  ?orthologGene rdfs:label ?orthologGeneLabel .

  ?cancer afo:AFX_0002745 ?protein ;
          rdfs:label ?cancerType .
}
ORDER BY DESC(?maxProteinLength)
LIMIT 1
"
qd5 <- SPARQL(endpoint, query5)
View(qd5$results)
```

The result obtained after these queries is shown in the **Figure 10**.

proteinLabel	geneLabel	orthologGeneLabel	cancerType	maxProteinLength	proteinComment
Adenomatous polyposis coli protein Human	APC Gene Human	APC Gene Common Mouse	Colorectal Cancer	2843	Tumor suppressor. Its activity is correlated wit

Figura 10: Query 5 output.

7. Data Publishing

To carry out the data publication in Pubby, a new Namespace must be created in Blazergraph to store the data, with an input file of **quads** instead of Turtle format. However, before anything else, a **custom URI** for the dataset must be included in all subjects. Failing to do this will prevent specific queries as the RDF has direct prefixes to external URIs without relating to a custom URI.

Once this is done, the format can be converted. To convert from Turtle to N-Quads, the **EasyRDF Converter** tool was used again to convert the Turtle format to N-Triplets. The difference between the latter format and N-Quads is that in each line, a new URI referencing the graph containing that triplet is added, in this case *<https://cancerdata.org/graph/>*. To convert from N-Triplets to N-Quads, the following Python code was used:

```
input_file = 'ntriplets.txt'
output_file = 'nquads.txt'

with open(input_file, 'r') as file:
    lines = file.readlines()

new_lines = []
for line in lines:
    if line.strip().endswith('.'):
        new_line = line.strip()[:-1] + ' <https://cancerdata.org/graph/>
   .\n'
        new_lines.append(new_line)
    else:
        new_lines.append(line)

with open(output_file, 'w') as file:
    file.writelines(new_lines)

print("The file has been modified and saved in:", output_file)
```

Once the nquads.txt file is created, Namespace in Blazergraph is made, where both that file and the metadata file will be uploaded, the latter being in Turtle format. To verify that everything has been uploaded correctly, a simple graph, subjects, predicates, and objects query is performed, obtaining what is shown in the **Figure 11**.

Verification Query

```
SELECT ?graph ?s ?p ?o
WHERE {
  GRAPH ?graph {
    ?s ?p ?o
}
}
```

graph	s	p	
bd:nullGraph	<http://creativecommons.org/licenses/MIT/>	<http://creativecommons.org/ns#licenseClass>	<http://creativecommons.org/license/software>
bd:nullGraph	<http://creativecommons.org/licenses/MIT/>	<http://creativecommons.org/ns#permits>	<http://creativecommons.org/ns#DerivativeWorks>
bd:nullGraph	<http://creativecommons.org/licenses/MIT/>	<http://creativecommons.org/ns#permits>	<http://creativecommons.org/ns#Distribution>
bd:nullGraph	<http://creativecommons.org/licenses/MIT/>	<http://creativecommons.org/ns#permits>	<http://creativecommons.org/ns#Reproduction>
bd:nullGraph	<http://creativecommons.org/licenses/MIT/>	<http://creativecommons.org/ns#requires>	<http://creativecommons.org/ns#Notice>
bd:nullGraph	<http://creativecommons.org/licenses/MIT/>	rdf:type	<http://creativecommons.org/ns#License>
bd:nullGraph	<https://cancerdata.org/sparql/>	dcterm:license	<https://creativecommons.org/licenses/by/4.0/deed.es>
bd:nullGraph	<https://cancerdata.org/sparql/>	<http://www.w3.org/ns/dcat#accessURL>	<http://155.54.239.183:3039/blazegraph/namespace/CancerPubby/sparql>
bd:nullGraph	<https://cancerdata.org/sparql/>	rdf:type	<http://www.w3.org/ns/dcat#Distribution>
bd:nullGraph	<https://cancerdata.org/txt/>	dcterm:format	<http://publications.europa.eu/resource/authority/file-type/TXT>
bd:nullGraph	<https://cancerdata.org/txt/>	dcterm:license	<https://creativecommons.org/licenses/by/4.0/deed.es>
bd:nullGraph	<https://cancerdata.org/txt/>	<http://www.w3.org/ns/dcat#downloadURL>	<http://localhost/nquads.txt>
bd:nullGraph	<https://cancerdata.org/txt/>	rdf:type	<http://www.w3.org/ns/dcat#Distribution>
bd:nullGraph	<https://cancerdata.org/data/>	dcterm:License	<http://creativecommons.org/licenses/MIT/>
bd:nullGraph	<https://cancerdata.org/data/>	<http://www.w3.org/ns/dcat#distribution>	<https://cancerdata.org/sparql/>
bd:nullGraph	<https://cancerdata.org/data/>	<http://www.w3.org/ns/dcat#distribution>	<https://cancerdata.org/txt/>
bd:nullGraph	<https://cancerdata.org/data/>	<http://www.w3.org/ns/sparql-service-description#namedGraph>	<https://cancerdata.org/graph/>
bd:nullGraph	<https://cancerdata.org/data/>	rdf:type	<http://www.w3.org/ns/dcat#Dataset>
bd:nullGraph	<https://cancerdata.org/data/>	rdfs:label	Relationship between genes and cancer types
bd:nullGraph	<https://cancerdata.org/data/>	foaf:primaryTopic	Cancer
<https://cancerdata.org/graph/>	<https://cancerdata.org/APC>	<http://purl.obolibrary.org/obo/PATO_0002261>	<https://cancerdata.org/chromosome/5>
<https://cancerdata.org/graph/>	<https://cancerdata.org/APC>	<http://purl.obolibrary.org/obo/SIO_010078>	<http://www.uniprot.org/uniprot/P25054>
<https://cancerdata.org/graph/>	<https://cancerdata.org/APC>	<https://ncit.nci.nih.gov/ncitbrowser/#C28708>	<https://cancerdata.org/APC_MM>

Figura 11: Metadata and data query.

To finally load the dataset into Pubby, access the directory *UM-Bioinformatics-MSc-FAIR-data/LinkedDataServer/INF/* from the GitHub repository of Mikel Egaña. Once there, modify the **fair-config.ttl** file to specify the Web Base, SPARQL Endpoint, and Dataset Base, as shown in the **Figure 12**.

```
 GNU nano 2.9.8                                  fair-config.ttl

# Prefix declarations to be used in RDF output
@prefix conf: <http://richard.cyganiak.de/2007/pubby/config.rdf#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .


# Server configuration section
<> a conf:Configuration;
    # Project name for display in page titles
    conf:projectName "FAIR data publication UM MSc Bioinformatics";
    # Homepage with description of the project for the link in the page header
    conf:projectHomepage <https://github.com/mikel-egana-aranguren/UM-Bioinformatics-MSc-FAIR-data>;
    # The Pubby root, where the webapp is running inside the servlet container.
    conf:webBase <http://155.54.239.183:8179/>;
    # URL of an RDF file whose prefix mapping is to be used by the
    # server; defaults to <>, which is *this* file.
    conf:usePrefixesFrom <>;
    # If labels and descriptions are available in multiple languages,
    # prefer this one.
    conf:defaultLanguage "en";
    # When the homepage of the server is accessed, this resource will
    # be shown.
    conf:indexResource <https://github.com/mikel-egana-aranguren/UM-Bioinformatics-MSc-FAIR-data>;

# Dataset configuration
    conf:dataset [
        # SPARQL endpoint URL of the dataset
        conf:sparqlEndpoint <http://155.54.239.183:3039/blazegraph/namespace/Cancer/sparql>;
        # Common URI prefix of all resource URIs in the SPARQL dataset
        conf:datasetBase <https://cancerdata.org/>;
        # Will be appended to the conf:webBase to form the public
        # resource URIs; if not present, defaults to ""
        #conf:webResourcePrefix "resource/";
    ];
    .
```

Figura 12: fair-config.tll file.

The **web.xml** file, located in the same directory, is also modified to specify the fair-config.tll file in the <context-param> section.

Finally, return to the *UM-Bioinformatics-MSc-FAIR-data/LinkedDataServer/pubby/* directory and execute the command **java -jar start.jar jetty.port=8179**. To check that it works correctly, access it with the IP and port number assigned from the browser, in this case `http://155.54.239.183:8179/`.

It is verified that the graph structure can be correctly accessed with the features specified by the metadata (**Figure 13**), as well as queries can be made by adding in the link address shown above what information is wanted. The **Figure 14** shows what Pubby displays when asking for information about the APC gene.

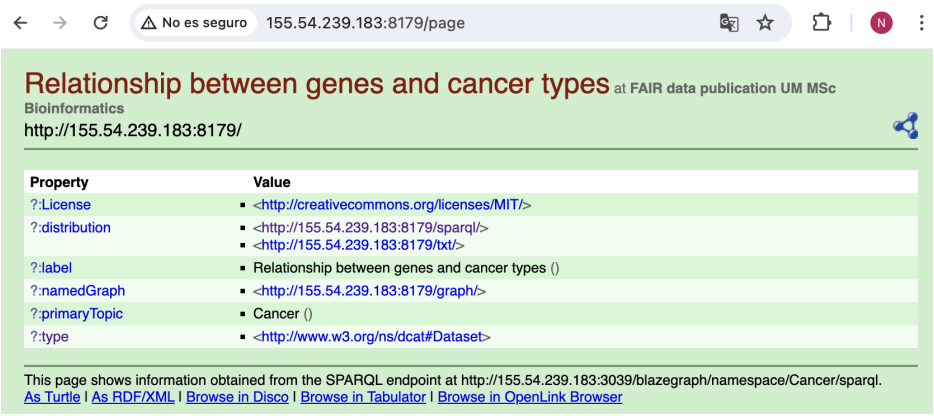

Figura 13: Pubby homepage.

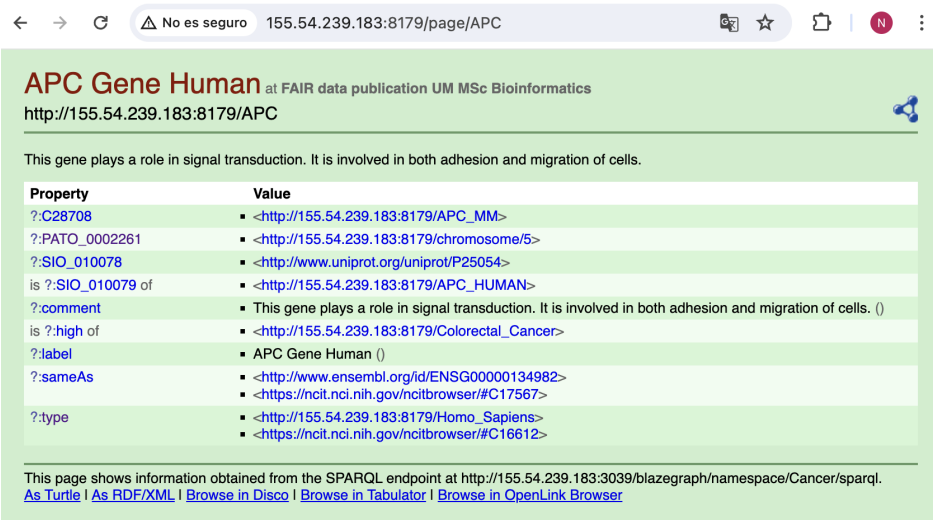

Figura 14: APC gene query in pubby.

8. Conclusions

- RDF graphs enhance interoperability, facilitate advanced data linkage, and improve the capabilities for search, inference, and knowledge discovery across varied domains and applications.
- Utilizing Blazegraph alongside R for querying purposes has proven to be very effective. This combination allows for the precise extraction of information from RDF graphs, particularly aiding in the analysis of gene behavior concerning genes links to various cancer types.
- The development, querying, and dissemination of an RDF knowledge graph emphasize the crucial role of semantic standards and the use of open data platforms in the realm of biomedical research. These technologies are pivotal in integrating diverse information sources and fostering significant discoveries in the field.

9. Bibliography

- Chee, C. W., Mohd Hashim, N., & Nor Rashid, N. (2024). Morindone as a potential therapeutic compound targeting TP53 and KRAS mutations in colorectal cancer cells. Chemico-biological interactions, 392, 110928. https://doi.org/10.1016/j.cbi.2024.110928

- Li, J., Gan, S., Blair, A., Min, K., Rehage, T., Hoeppner, C., Halait, H., & Brophy, V. H. (2019). A Highly Verified Assay for KRAS Mutation Detection in Tissue and Plasma of Lung, Colorectal, and Pancreatic Cancer. Archives of pathology & laboratory medicine, 143(2), 183–189. https://doi.org/10.5858/arpa.2017-0471-OA

- Li, L., Li, M., & Wang, X. (2020). Cancer type-dependent correlations between TP53 mutations and antitumor immunity. DNA repair, 88, 102785. https://doi.org/10.1016/j.dnarep.2020.102785

- Sahlab, Nada & Braun, Dominik & Köhler, Christian & Jazdi, Nasser & Weyrich, Michael. (2022). Extending the Intelligent Digital Twin with a context modeling service: A decision support use case. Procedia CIRP. 107. 463-468. https://doi.org/10.1016/j.procir.2022.05.009.

- Zhu, G., Pei, L., Xia, H., Tang, Q., & Bi, F. (2021). Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. Molecular cancer, 20(1), 143. https://doi.org/10.1186/s12943-021-01441-4