Towards Linked Open Gene Mutations Data

Achille Zappa[1,2], Andrea Splendiani[3], Paolo Romano[1]

[1] Bioinformatics, IRCCS AOU San Martino – IST National Cancer Research Institute, Genoa, Italy

[2] Department of Informatics, Systems and Telematics, University of Genoa, Genoa, Italy

[3] Rothamsted Research, Harpendem, Hertfordshire, United Kingdom

Background

The vision of the Semantic Web is to evolve the Web into a distributed knowledge base, and this vision relies on the evolution of the Web into a Web of Data. A great number of information resources are being made available in the LOD (Linked Open Data) cloud. Biological information is also being converted and it constitutes a major component of the cloud. The Human Genome Variation Society and Human Variome Project have produced recommendations for nomenclature of variations and for contents of mutation databases. They also outlined conditions for the integration of Locus Specific Data Bases (LSDB) with other biological databases. However, human variation data, despite its relevance for medicine, has not yet been adequately taken into account. This poster is a first attempt in this direction. [Fig. 1]

IARC TP53 Mutations database

The IARC TP53 Mutation Database has been maintained at the International Agency for Research on Cancer in Lyon, France, since 1994. The database compiles all TP53 mutations that have been reported in the published literature since 1989. The database includes annotations on functional impact of mutations, either predicted or experimentally assessed, clinico-pathologic characteristics of tumors and demographic and life-style information on patients. A relational implementation is available at the National Cancer Research Institute of Genova.

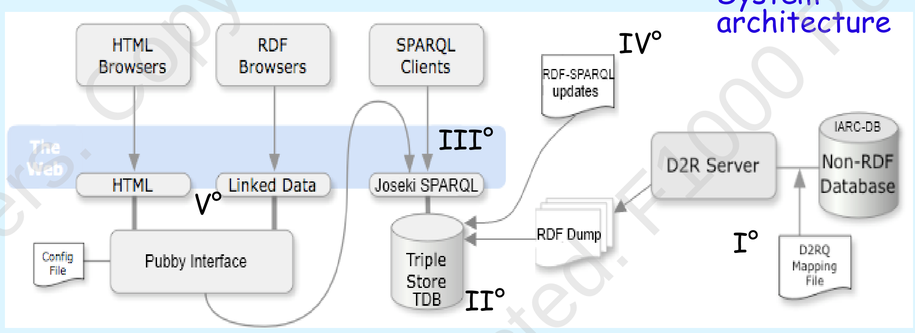

Figure 1: System architecture

First stage: RDB to RDF

In our case, automatic RDB to RDF mappings were first created by using D2RQ. The mapping file was then manually refined to improve its commitment to a shared representation and, thus, to encode in the mapping some semantics that is not expressed in the RDB schema. By customizing predicates we have been able to better represent the semantics of our data, according to shared ontologies.

Second stage: native Triple Store

The query performance of D2RQ is lower than that which can be achieved by using a devoted RDF triple store. In order to evaluate reliability and performances of an on-the-fly mapping system, such as D2RQ, compared to a native RDF framework, a dump of all triples generated by D2RQ was loaded into a dedicated Jena TDB triple store.

Third stage: SPARQL Endpoint

Once the RDF database was created, it was made accessible as a SPARQL endpoint using Joseki, a Jena tool that provides support for SPARQL queries through an HTTP engine. Joseki was configured to connect to the TDB database to create an implementation of SPARQLer, a user friendly interface to a SPARQL server.

Fourth stage: RDF enrichment via SPARQL Update

We have then enriched our datasets by adding triples connecting samples to related UMLS concepts. This mapping was implemented by means of a SPARQL Update federated query interconnecting our datasets with the Linked Life Data [LLD] endpoint.

Fifth stage: Linked Data Interface

Finally, we exposed the content of our TDB triple store as a Linked Data interface using Pubby is used to add a Linked Data interface to our SPARQL endpoint. It handles external requests by connecting to the SPARQL endpoint, issuing a SPARQL "DESCRIBE" query about the requested URI, and showing the result in a HTML or RDF page, supporting Linked Data compliant content resolution and negotiation procedures. [Fig. 2]

NM_000546.1:c.742C>T from original project at TP53/IARC LOGVD, National Cancer Research Institute, Genova, Italy
http://bioinformatics.istge.it/logvd/resource/variations/3294

Property	Value
logvd:AGVGDClass	deleterious
logvd:AIP1nWT	0
logvd:BAXnWT	0,7
logvd:Codon_number	248
logvd:CpG_site	yes
logvd:DNEclass	Yes
logvd:Domain_function	DNA binding
logvd:ExonIntron	7-exon
logvd:GADD45nWT	0
logvd:Genomic_nt	13379
logvd:Grantham_deviation	101,29
logvd:Grantham_variation	0
logvd:MDM2nWT	0
logvd:NOXAnWT	0
logvd:P53R2nWT	0
logvd:Polymorphism	no
logvd:Residue_function	DNA binding
logvd:SIFTClass	deleterious
logvd:SomaticCount	707
logvd:Splice_site	no
logvd:StructureFunctionClass	non-functional
logvd:TransactivationClass	non-functional
logvd:Variation_Description	C>T
logvd:Variation_Type	G:C>A:T at CpG
logvd:Variation_motif	L3
logvd:WAF1nWT	0
logvd:WildType_codon	CGG
logvd:Wildtype_Base	C
logvd:c_description	c.742C>T
logvd:h1433snWT	0
logvd:hasEffect	missense
logvd:hasMut_rate	1,243
logvd:hasMut_rate_AA	1,243
mio:hasMutantResidue	Trp
logvd:hasMutant_codon	TGG
logvd:hasMutant_nucleotide	T
is logvd:hasVariation of	db:somatic_mutations/10020 db:somatic_mutations/1014 db:somatic_mutations/10259 db:somatic_mutations/103 db:somatic_mutations/1030 db:somatic_mutations/10322 db:somatic_mutations/10378 db:somatic_mutations/10387 db:somatic_mutations/10428 db:somatic_mutations/10433 db:somatic_mutations/863 db:somatic_mutations/9791 db:somatic_mutations/9820 db:somatic_mutations/9847 db:somatic_mutations/9856 db:somatic_mutations/9857 db:somatic_mutations/9887 db:somatic_mutations/9938
mio:hasWildTypeResidue	R
is logvd:hasgene_variation of	db:genes/TP53
rdfs:label	NM_000546.1:c.742C>T
logvd:p_description	p.R248W
rdf:type	logvd:variation

This page shows information obtained from the SPARQL endpoint at http://bioinformatics.istge.it/logvdsparql/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

samples/KIK92-MY279 from original project at TP53/IARC LOGVD, National Cancer Research Institute, Genova, Italy
http://bioinformatics.istge.it/logvd/resource/samples/847

Property	Value
logvd:Material_sequenced	direct
logvd:Morpho_code	8010/3
logvd:Name	KIK92-MY279
NCIT:Neoplasm_by_Morphology	Carcinoma, NOS
logvd:Prescreening	SSCP
logvd:Sample_source	surgery
NCIT:Stage	A-Dukes'
logvd:Start_material	DNA
logvd:Sub_topography	COLORECTUM, NOS
logvd:Tissue_processing	frozen
logvd:Topo_code	C18-C20
NCIT:Topography	COLORECTUM, NOS
logvd:Tumor_origin	primary
logvd:hasIndividual	db:individual/847
is logvd:hasSample of	db:somatic_mutations/863
logvd:hasUMLS_neoplasm	<http://linkedlifedata.com/resource/umls/id/C0007097>
rdfs:label	samples/KIK92-MY279
logvd:p53_IHC	positive
rdf:type	logvd:sample

This page shows information obtained from the SPARQL endpoint at http://bioinformatics.istge.it/logvdsparql/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

somatic_mutations/863 from original project at TP53/IARC LOGVD, National Cancer Research Institute, Genova, Italy
http://bioinformatics.istge.it/logvd/resource/somatic_mutations/863

Property	Value
logvd:Putative_stop	0
logvd:Structural_motif	L2/L3
is logvd:TP53_somatic_mutations of	db:database/IARC-TP53
logvd:c_description	c.742C>T
logvd:hasReference	db:somatic_refs/98
logvd:hasSample	db:samples/847
logvd:hasVariation	db:variations/3294
rdfs:label	somatic_mutations/863
rdf:type	logvd:somatic_mutation

This page shows information obtained from the SPARQL endpoint at http://bioinformatics.istge.it/logvdsparql/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

individual/847 from original project at TP53/IARC LOGVD, National Cancer Research Institute, Genova, Italy
http://bioinformatics.istge.it/logvd/resource/individual/847

Property	Value
NCIT:Age	59
NCIT:Country	dbpedia:Japan
logvd:Family_history	non-FAP
NCIT:Geographic_Area	Tokyo
NCIT:Population_Group	Eastern Asia
NCIT:Sex_or_Gender	F
NCIT:Tobacco_Use	non-smoker
is logvd:hasIndividual of	db:samples/847
rdfs:label	individual/847
rdf:type	logvd:individual

This page shows information obtained from the SPARQL endpoint at http://bioinformatics.istge.it/logvdsparql/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Figure 2: the Pubby interface of the prototype. You can browse among entities in different datasets that are linked each others.

A prototype implementation

We have therefore implemented a semantic infrastructure for TP53 variation data as a prototype for studying issues related to the publication of mutation data on the LOD cloud. It includes data on somatic mutations and related bibliographic references, bio-samples, and patients demography. We also published summary gene variations data. Specific biological ontologies on gene variation does not exist. Only a limited set of external ontologies and terminologies have been taken into account. These include the NCI Thesaurus (NCIT)for medical terminology, the Bibliographic Ontology (BIBO) and the BibTeX definition in Web Ontology Language (OWL) for bibliographic references, the Diseasome ontology and MIO. Moreover, external links were set to LOD implementations of DBpedia, a system including all structured information which is present in Wikipedia pages, PubMed, the Human Genome Nomenclature Committee (HGNC) database, the On-line Mendelian Inheritance in Man (OMIM) system, UniProt, and the Unified Medical Language System (UMLS). All links were defined to the Bio2RDF entry points of these databases, expressed by using the unified Bio2RDF URI style, with the exception of DBpedia, and UMLS, that was connected through LinkedLifeData.

```
SELECT DISTINCT ?variation_label ?neoplasm ?clinical_trial
WHERE {
  SERVICE <http://linkedlifedata.com/sparql> { ?clinical_trial relontology:hasInclusionCriteria ?umls } .
    ?sample logvd:Sub_topography "Middle third of esophagus" .
    ?sample NCIT:Neoplasm_by_Morphology ?neoplasm .
    ?sample logvd:hasUMLS_neoplasm ?umls .
    ?somatic_mutation logvd:hasSample ?sample .
    ?variation_id rdfs:label ?variation_label .
    ?somatic_mutation logvd:hasVariation ?variation_id .
}
ORDER BY ?variation_label ?neoplasm
```

?variation_label	?neoplasm	?clinical_trial
NM_000546.1:c.507G>A	Adenocarcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00001332>
NM_000546.1:c.838A>G	Adenocarcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00001332>
NM_000546.1:c.507G>A	Adenocarcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00001428>
NM_000546.1:c.838A>G	Adenocarcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00001428>
NM_000546.1:c.482C>A	Dysplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00001932>
NM_000546.1:c.482C>A	Dysplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00003076>
NM_000546.1:c.482C>A	Dysplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00003094>
NM_000546.1:c.482C>A	Dysplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00003223>
NM_000546.1:c.482C>A	Dysplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00003593>
NM_000546.1:c.469G>T	Hyperplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00001378>
NM_000546.1:c.469G>T	Hyperplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00001756>
NM_000546.1:c.469G>T	Hyperplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00001760>
NM_000546.1:c.469G>T	Hyperplasia, NOS	<http://data.linkedct.org/resource/trials/NCT00003641>
NM_000546.1:c.422G>A	Squamous cell carcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00000692>
NM_000546.1:c.451C>G	Squamous cell carcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00000692>
NM_000546.1:c.469G>T	Squamous cell carcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00000692>
NM_000546.1:c.474_475ins1	Squamous cell carcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00001450>
NM_000546.1:c.488A>G	Squamous cell carcinoma, NOS	<http://data.linkedct.org/resource/trials/NCT00001450>

Figure 3: SPARQL fed-query sample

Availability

Presently, we offer access to our database via two distinct modalities. The first interface is a SPARQL endpoint available at http://bioinformatics.istge.it/logvdsparql/sparql . It is implemented via Joseki on TDB. Interfaces to validators for SPARQL queries and for RDF data are also available. The second interface is a Linked Data representation available at http://bioinformatics.istge.it/logvd/ . It is based on the Pubby frontend.

Some perspectives

In the short time, the exploitation of SPARQL searches on mutation data and other biological databases may support some interesting and useful data retrieval presently not possible. [Fig. 3] In a longer time frame, reasoning on integrated variation data may also support discoveries towards personalized medicine. For this to happen, however, a long way is still needed, due to the relative isolation of variation data sources and lack of standardization both in terminologies and data schema.

Acknowledgements

This work has been partially supported by the Italian Ministry of Health (project RNBIO -Rete Nazionale di Bioinformatica Oncologica) and by the Liguria region (project Liguria eScience).