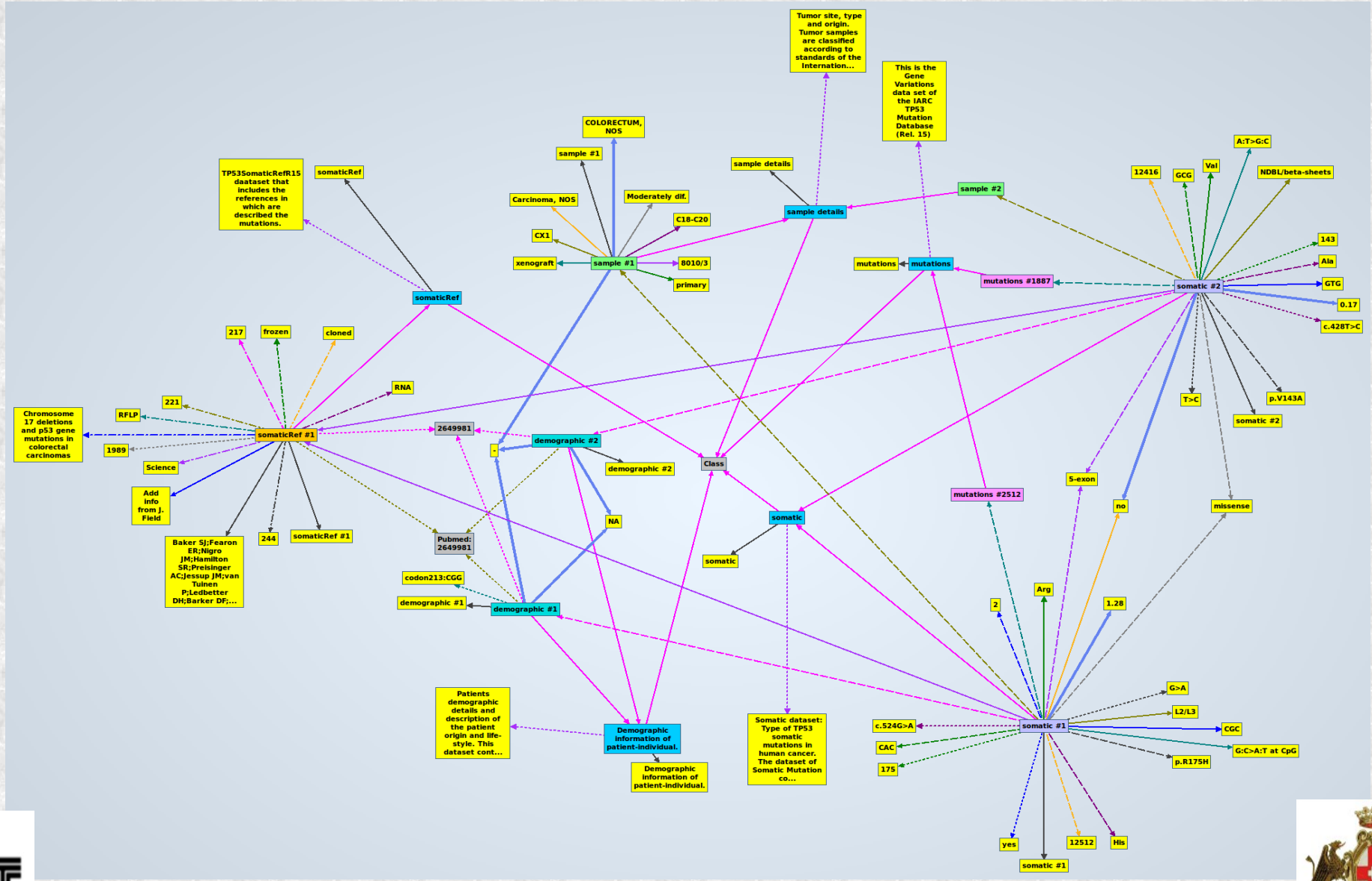
Towards Linked Open Gene Mutations Data



Introduction

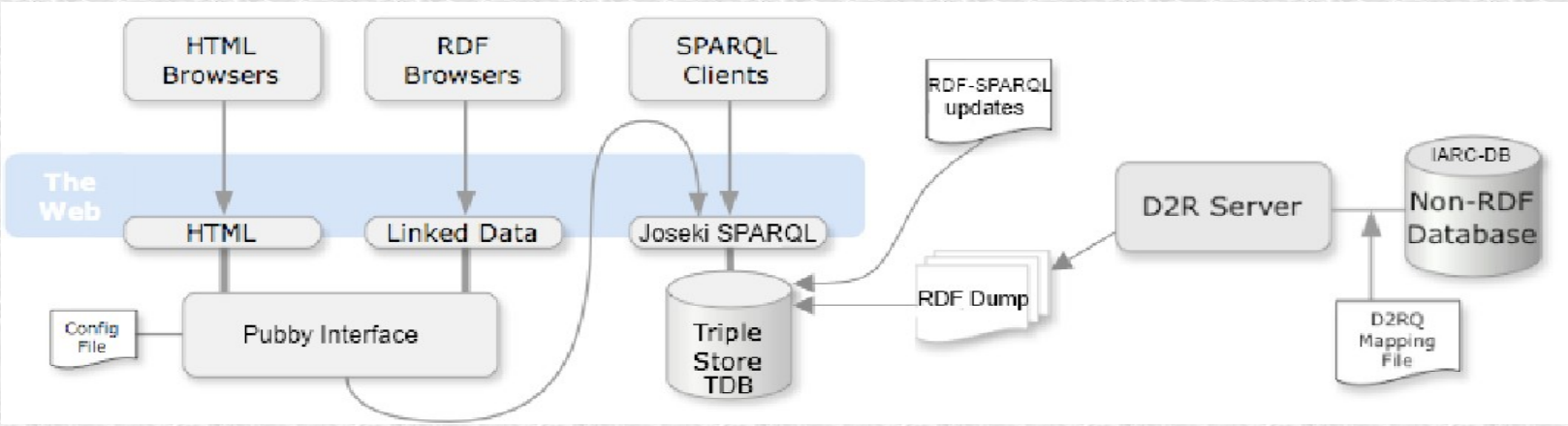
- A great number of information resources are being made available in RDF and exposed on the LOD (Linked Open Data) cloud.
- Biological information is also being converted and it constitutes a major component of the cloud.
- The Human Genome Variation Society and Human Variome Project have produced recommendations 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, have not yet been adequately taken into account.
- This project is a first attempt in this direction.

IARC TP53 Database

- The IARC TP53 mutations 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 Genoa (Italy).

LOGVD prototype

- 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.



Infrastructure

- Automatic mappings to RDF were first created by using D2RQ and later manually refined by introducing concepts and properties from domain vocabularies and ontologies, as well as links to Linked Open Data datasets,
- A RDF dump was stored into a TDB triple store,
- Once the RDF database was created, it was made accessible as a SPARQL endpoint using Joseki,
- RDF enrichment via SPARQL Update,
- Finally, we exposed the content of our TDB triple store as a Linked Data interface using Pubby server querying our SPARQL endpoint.

Triple store size	
Number of entities	85,785
Number of triples	1,002,597
Number of external URIs	
LinkedLifeData	25,094
Bio2RDF	2,244
DBpedia	23,015
Total	50,353
Number of links to external web pages	
Total	2,436

Shared properties from re-used ontologies

Ontology	**No. of shared properties**	**Involved triples**
rdf:	1	85,893
rdfs:	3	88,249
owl:	1	2,241

Thank you for your attention!

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, Harpenden, United Kingdom

Contact us : achille.zappa@istge.it