A first RDF implementation of the COSMIC database on mutations in cancer

Achille Zappa[1]✉, **Paolo Romano**[2]

[1]Department of Informatics Bioengineering Robotics and Systems Engineering (DIBRIS), University of Genoa, Genoa, Italy
[2]Bioinformatics Laboratory, IRCCS AOU San Martino - IST, Genoa, Italy

Motivation and Objectives

Within a living organism, genome and proteome variations may influence many molecular interactions and biochemical pathways, leading to deleterious effects in the proper activity of cells, tissues, and organs; ultimately, this may be the cause of many syndromes and diseases. It is now well known that tumors may arise as a result of a series of DNA sequence abnormalities and mutations. It is then not surprising that there is a vast amount of information available in the scientific literature and that a lot of information systems devoted to the management of related data exist. Among these, of particular interest are the many Locus Specific Data Bases (LSDB) and the COSMIC (Catalogue of Somatic Mutations in Cancer) database (Forbes et al., 2011). Such data, however, are not yet sufficiently integrated with other molecular, biomedical, and clinical databases. New efforts are therefore needed in this direction.

Data retrieval, search and integration solutions in bioinformatics are increasingly making use of a set of standards and technologies which are the basis of the Semantic Web (Berners-Lee et al., 2001) framework. This framework is intended to evolve the web into a distributed knowledge-base and a first step in this evolution is the generation of a Web of Data (Bizer et al., 2009). In this view, we can see Linked Data as an approach to data integration that employs ontologies, terminologies, Uniform Resource Identifiers (URIs), and the Resource Description Framework (RDF) to connect pieces of data, information and knowledge on the Semantic Web (Belleau et al., 2008). In particular, RDF describes semantic rich information on the web through a composition of simple triples (predicates), such as ('Subject', 'Property', 'Object'), that link entities through relations which are expressed by using ontologies, and are defined by using URIs. See the RDF reference site: http://www.w3.org/RDF/, last accessed on October 3, 2012). A relevant contribution to this vision comes from the conversion of data stored in relational databases (RDB) into RDF. There is a vast amount of information on human variation in the literature and several mutation and variation databases, but, to our knowledge, this kind of information is still scarce in the Web of Data. Various motivations can be depicted for using Semantic Web technologies and publishing Linked Data life sciences datasets; this allows to improve data and information integration, share ability of openly accessible data through standard and programmatic interfaces, semantic normalization, data discoverability and query federation from distributed sources.

A first work carried out by our group led to the implementation of an RDF version (Zappa et al., 2012) of the IARC TP53 Somatic Mutation database (IARCDB) (Petitjean et al., 2007). Here, we present the initial development of an RDF version of the COSMIC (Catalogue of Somatic Mutations in Cancer) database by means of Semantic Web technologies.

Methods

COSMIC was developed, and is currently maintained, at the Wellcome Trust Sanger Institute. It is designed to gather, curate, and organize information on somatic mutations in cancer and to make it freely available on-line. It combines cancer mutation data, manually curated from the scientific literature, with the output from the Cancer Genome Project (CGP). Genes are selected for full literature curation using the Cancer Gene Census. COSMIC datasets are freely available as common CSV flat files. However these files don't contain all the available information and they don't reflect the original schema and table contents of the database. For this reason, and also due to the huge amount of data, we started from a basic automatic RDB to RDF mapping of a relational version of COSMIC. Many research works have been focused on mapping data from RDB to RDF. They have led to the implementation of both mapping tools and domain specific applications. The structure of an RDB database may provide a partial characterization of semantics of the domain it refers to. Some tools rely on this property to generate an approximate mapping to RDF, which can then be manually tuned and thus brought to be in line with a shared conceptualization.

Mapping is the process of making explicit correspondences or relationships between entities in the relational database and the RDF graph. In our case, the mapping was first created by using D2RQ, a platform for treating relational databases as virtual RDF graphs. See the D2RQ web site: http://www.d2rq.org/, last accessed on October 3, 2012). This tool also allows on-the-fly generation of RDF triples from the database. The relational database was then published using a D2R server. D2R enabled us to publish a first SPARQL endpoint on top of the relational database, build an RDF data dump, and make it possible browsing the generated RDF triples through a standard web interface.

One of the most important aspects of the RDB to RDF conversion is, however, the capability of representing the semantics that is not explicitly defined in the relational schema. After a careful analysis of the database schema, we were able to map our resources into separated well defined classes and sub-graphs and to define the relationships and properties of our statements. For instance, D2RQ generates predicate names which are based on the RDB column names: it has no way to know when a predicate refers to a property for which a shared representation (ontological concept) exists. By customizing predicates we have been able to improve the representation of data semantics, according to shared ontologies. Where shared relations were not available to express the content of our database, we have used ad-hoc defined properties.

The final RDF dataset is being deployed according to Linked Open Data (LOD) principles with external links set to datasets such as DBpedia, a system including all structured information which is present in Wikipedia pages (see DBpedia web site: http://www.dbpedia.org/, last accessed on October 3, 2012), PubMed, the Human Genome Nomenclature Committee (HGNC) database (see HGNC web site: http://www.genenames.org/, last accessed on October 3, 2012), the On-line Mendelian Inheritance in Man (OMIM) system, UniProt (Belleau et al., 2008) and Linked Life Data.

In order to improve performances, the RDF export must be imported into a native RDF triple store system. The RDF dump of COSMIC was then uploaded in a Jena TDB triple store. See the Jena and TDB web sites at: http://openjena.org/index.html and at http://jena.sourceforge.net/TDB/, last accessed on October 3, 2012). A Fuseki server was implemented to make available our data through a SPARQL endpoint. See the Fuseki web site at: http://fuseki.sourceforge.net/, last accessed on October 3, 2012).

Since one of the main use cases and aim of COSMIC is to provide somatic mutation frequencies and distributions via plots and histograms, it is then a good practice to deploy a web interface able to graphically visualize such kind of information also in a Semantic Web context. A web interface based on javascript and some graphical libraries can then display results of SPARQL queries to improve visualization of this kind of information by means of charts.

Results and Discussion

Prototype servers are available on-line. The D2R server web site is available at http://bioinformatics.istge.it/D2R_CosmicRDF_proto/. The SPARQL endpoint, that is only meant for SPARQL queries and cannot therefore be used as-is by researchers, is available at the following URL: http://bioinformatics.istge.it/CosmicRDF_protosparql/cosmic/sparql. Currently, servers present only a subset of the database, corresponding to the "full export" that may be downloaded from the COSMIC web site. This dataset, however, does not reflect the database schema, whose analysis is an undergoing effort.

A Linked Data view, an HTML view and a SPARQL endpoint are available. The latter can be explored by any Semantic Web browser or application. These are building blocks for data integration solutions incorporating mutation data. The standard web interface includes graphical visualization features of results of some specific SPARQL queries. These prototypes demonstrate how an RDF representation of relational database contents can be easily provided.

Although a great value of our system would lie on the identification of a shared, semantically meaningful, ontology-based representation of variation information, that could only be defined through a collaboration with the community of curators of variation databases, our approach already allows to carry out queries on the database, as well as some graph-analysis for validation of data and elucidation of implicit relations among data, relations that could not be exploited with the current system.

Relying on dereferenceable URIs, that is URIs that may be redirected to a unique existing Internet address usually accessible via HTTP, existing predicates and ontologies allows our system to be a part of the growing Web of Data with the aim to be in integrated and interlinked part of the Linked Open Data Cloud.

An improved and extended version of our prototypes and interfaces are under development. A new web interface with demo queries and specific use-cases in also under development, with the aim of building a prototype user-friendly interface that can be more proficiently used by such users as biologists and clinicians.

Acknowledgements

This work has been partially supported by the Liguria region (project Liguria eScience).

References

1. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics (2008) 41(5):706-716.
2. Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American, May 2001.
3. Bizer C, Heath T, Berners-Lee T. Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems (2009) 5(3):1-22.
4. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Human Mutation (2007) 28(6):622-629.
5. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucl. Acids Res. (2011) 39 (Suppl 1): D945-D950. doi: 10.1093/nar/gkq929
6. Zappa A, Splendiani A, et al. (2012) Towards linked open gene mutations data. BMC Bioinformatics 13(Suppl 4):S7. doi:10.1186/1471-2105-13-S4-S7