Towards Linked Open Gene Mutations Data

ZAPPA A (1, 2), SPLENDIANI A (3), ROMANO P (1)

(1) National Cancer Research Institute, Genoa, Italy
(2) Department of Informatics, Systems and Telematics DIST - University of Genoa, Genoa, Italy
(3) Rothamsted Research, Harpenden, United Kingdom

Motivation
The Semantic Web (1) is gaining momentum as a framework for the development of next-generation bioinformatics data integration tools since its standards and technologies have now reached enough maturity to be considered a viable solution for data integration challenges. 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 (2). A relevant contribution to this evolution comes from the conversion into a viable representation of data stored in Relational Databases (RDB). The Resource Description Framework (RDF) is able to describe semantic rich information on the web. It is based on the composition of simple predicates such as ‘Subject’ ‘Predicate’ ‘Object’ (3). Many research works have been focused on the mapping of data from RDB to RDF and have led to the implementation of both mapping tools and domain specific applications. Some mappings are automatically generated via a simple conversion where the name of the relational table is mapped to an RDF class node and the names of the columns are used as RDF predicates: entities and relations, and their meaning, reflect the RDB schema. Usually, knowledge of these schema is needed to understand the exported information. In other mappings, relations and entities of the original databases are converted to a representation which is based on a shared conceptualization. Even though simple mappings do not capture complex semantics, they can serve as a starting point to create customized, domain specific mappings. Here, we present the development of a mapping between the IARC TP53 Somatic Mutation database (IARCDB) to RDF. Although several mutation and variation databases exist, to our knowledge this kind of information is still scarce in the Web of Data. The IARCDB has been developed and maintained at the International Agency for Research on Cancer in Lyon since 1994 (4). The database compiles all TP53 mutations that have been reported in the published literature since 1989. We have therefore implemented a prototype server publishing TP53 mutation data in RDF with the aim of studying all issues related to the publication of mutation data on the Linked Open Data cloud.

Methods
IARC DB datasets are made available as flat files. A relational version of the IARCDB has been made available since many years as a basis for its SRS implementation (5). Automatic RDB to RDF mappings were created by using D2RQ, a platform for treating rela-
tional databases as virtual RDF graphs (6). This tool allows on-the-fly generation of RDF triples from the database. It also allows to browse the generated RDF triples through a standard web interface. It finally allows to query the database through a SPARQL endpoint. Mapping is the process of making explicit correspondences or relationships between entities in the relational database and the RDF graph. The first mapping file was manually refined to improve its commitment to a shared representation and, thus, to encode more ‘meaning’ in the mapping than it is expressed in the RDB schema. 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 exists. By customizing predicates we have been able to better represent our data semantics, according to shared ontologies. We used the Bibliographic Ontology (bibo) (7) to map the properties of bibliographic references, the NCI Thesaurus (8) for some cancer specific terms, and other frequently used vocabularies such as dc:, rdf:, rdfs:, and owl:. Where shared relations were not available to express the content of our database, we have used ad-hoc defined properties. Our export is deployed according to Linked Data criteria. URIs (identifiers of nodes in RDF) have been made compatible, where possible, with resources such as Bio2RDF (9) and PubMed.

Results
A prototype D2R server is available on-line. It currently presents three datasets: Somatic mutations, that includes each mutation described in literature, Gene variations, holding all different known mutations (without a reference to a single description), and related bibliographic references. 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, like Marbles (10). It is a building block for data integration solutions incorporating mutation data. This prototype demonstrates how an RDF representation of relational database contents can be easily provided. However, the main difficulty and value lies on the identification of a shared, semantically meaningful, ontology-based representation of variation information. Relying on existing predicates and ontologies allows our system to be a part of the growing Web of Data. A revised version of our prototype, including further shared concepts and all TP53 mutation data provided by IARC, is under development.

Availability
http://bioinformatics.istge.it/logvd
References

[1] Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American, May 2001.
[2] 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.
[3] RDF: http://www.w3.org/RDF/
[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] Marra D, Romano P. Integrating mutation data of the TP53 human gene in the bioinformatics network environment. Proc. First International Conference on Bioinformatics Research and Development BIRD ’07, March 12-14, 2007, Berlin, Springer Lecture Notes in Bioinformatics LNBI 4414, pp. 453-463, 2007, Springer Verlag Berlin Heidelberg 2007.
[6] D2RQ: http://www4.wiwiss.fu-berlin.de/bizer/d2rq/index.htm
[7] Bibo: http://bibliontology.com/
[8] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiud W-L, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics (2007) 40(1):30-43.
[9] 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):706-16.
[10] Marbles: http://marbles.sourceforge.net/

Contact email
achille.zappa@istge.it