Integration of variation data through SPARQL Micro-Services

Frederic Metereau[1], Franck Michel[1], Pierre Larmande[2,3,*], Guilhem Sempere[3,4] and Catherine Faron[1]

[1]Université Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
[2]DIADE, IRD, Univ. Montpellier, CIRAD, Montpellier, France
[3]French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier, France
[4]Intertryp, CIRAD, INRAE, IRD, Montpellier, France

Abstract

Integrating genetic variations data is essential to understand the interactions involving multiple genes in complex diseases. However, managing and extracting meaningful information from a large volume of genotyping data is challenging. This work aims to interconnect efficiently a MongoDB database with an RDF database through SPARQL Micro-Services. We first developed an RDF Model reusing existing ontologies and implemented it. Then, we evaluated some examples of queries interconnecting two applications Gigwa (MongoDB) and AgroLD (SPARQL endpoint).

Keywords

Knowledge Graphs, MongoDB, FAIR data, Genetic variations, Bioinformatics

Genetic variation refers to discrepancies in the DNA sequence among individuals. This variability in the genome accounts for distinctions in traits like eye colour and blood group, as well as a person's susceptibility to certain diseases. While specific traits and diseases can be attributed to variants in single genes, common conditions such as diabetes, heart disease, various cancers, Alzheimer's disease, and Parkinson's disease to name a few, result from intricate interactions involving multiple genes and environmental factors. Over 80 million variant sites in the human genome have been identified, encompassing single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and other structural variants. The processing of genetic variation data can reach several Gigabytes to several Terabytes. Indeed, each genome of individuals is stored and compared to the reference genome of the species. Thus, the analysis and exploration of this data is a real challenge. A solution to this problem is to use NoSQL databases tailored to manage large volumes of data with low latency. However, they lack semantics when the data must be extracted and compared with other data types such as phenotypes, diseases or gene function. The Semantic Web provides an answer to this

SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

*Corresponding author.
frederic.metereau@etu.univ-cotedazur.fr (F. Metereau); fmichel@i3s.unice.fr (F. Michel); pierre.larmande@ird.fr (P. Larmande); guilhem.sempere@cirad.fr (G. Sempere); faron@i3s.unice.fr (C. Faron)
0000-0001-9064-0463 (F. Michel); 0000-0002-2923-9790 (P. Larmande); /0000-0001-7429-2091 (G. Sempere); 0000-0001-5959-5561 (C. Faron)

CEUR Workshop Proceedings (CEUR-WS.org)

problem, as RDF enables data to be interconnected between several databases. This work aims to find a way to interconnect efficiently a MongoDB database with another RDF database.

As a proof of concept, we decided to use the Gigwa [1] and AgroLD [2] database applications to demonstrate the benefits of leveraging data semantics on a high volume of genomic data. Gigwa is a web application designed to store large volumes of genotypes (up to tens of billions), initially imported from VCF or other file formats, in a MongoDB [3] database, and to provide a straightforward interface for filtering these data. It makes it possible to navigate within search results, visualize them in different ways, and re-export subsets of data into various common formats. AgroLD is a knowledge graph that exploits Semantic Web technologies to integrate data of interest for the plant science community. AgroLD is built incrementally spanning vast aspects of plant molecular interactions. The current phase covers information on genes, proteins, predictions of homologous genes, metabolic pathways, plant trait associations and genetic studies.
For this work, we first developed an RDF model based on existing ontologies and inspired by DisGeNET [4]. We extended it with some features needed for the Gigwa data model which integrates gene annotation information. Then we developed some SPARQL Micro-Services [5] using the Gigwa RESTFul API. Finally, we developed and evaluated some queries interconnecting Gigwa and AgroLD through SPARQL query examples.

References

[1] G. Sempéré, A. Pétel, M. Rouard, J. Frouin, Y. Hueber, F. De Bellis, P. Larmande, Gigwa v2—Extended and improved genotype investigator, GigaScience 8 (2019). doi:10.1093/gigascience/giz051.

[2] A. Venkatesan, G. T. Ngompe, N. E. Hassouni, I. Chentli, V. Guignon, C. Jonquet, M. Ruiz, P. Larmande, Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy, PLOS ONE 13 (2018) e0198270. doi:10.1371/journal.pone.0198270.

[3] A. Kamsky, Adapting TPC-C benchmark to measure performance of multi-document transactions in MongoDB, Proc. VLDB Endow. 12 (2019) 2254–2262. doi:10.14778/3352063.3352140.

[4] J. Piñero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, L. I. Furlong, DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes, Database 2015 (2015). doi:10.1093/database/bav028.

[5] F. Michel, C. Faron, O. Gargominy, F. Gandon, Integration of Web APIs and Linked Data Using SPARQL Micro-Services—Application to Biodiversity Use Cases, Information 9 (2018) 310. doi:10.3390/info9120310.