INEX-MED: a Knowledge Graph to explore and link heterogeneous bio-medical data

Maxime Folschette[1,2,⋆], Kirsley Chennen[1,3,4,⋆], Alban Gaignard[5], Richard Redon[5], Hala Skaf-Molli[2], Olivier Poch[3], Jocelyn Laporte[4], Julie Thompson[3] and the INEX-MED consortium

[1] CNRS UMS 3601, Institut Français de Bioinformatique, France
[2] LS2N, Laboratoire des Sciences du Numérique de Nantes, CNRS UMR 6004, Université de Nantes, France
[3] Equipe CSTB - Laboratoire iCUBE, CNRS UMR 7357, Université de Strasbourg, France
[4] Dpt Médecine translationnelle et neurogénétique, IGBMC, INSERM U1258, CNRS UMR7104, Illkirch, France
[5] Institut du Thorax, Inserm UMR 1087, CNRS UMR 6291, Université de Nantes, France

Corresponding author: maxime.folschette@ls2n.fr, kchennen@unistra.fr, alban.gaignard@univ-nantes.fr

1 Context and Motivations

Health and life sciences nowadays face the massive availability of diverse biomedical data. Providing a unified and coherent access to these large-scale multi-source data is a major challenge. The INEX-MED project aims at gathering and representing multi-disciplinary, multi-source and multi-modal data (clinical, genetic, imaging) as a Knowledge Graph, in the domain of intracranial aneurysms and congenital myopathies. The aim is to apply machine learning in order to make predictions, highlight new diagnosis variables and stratify patient populations. To this end, Linked Data principles are applied in order to integrate the data despite their initial heterogeneity, with the objective of ensuring "FAIR" data principles (Findability, Accessibility, Interoperability, Reusability) [1].

2 Knowledge Graph Creation

The acquired diverse data, come in different formats. For instance, clinical data are represented with tables (CSV format), while genetic (exome) data rely on the VCF format. Imaging data come in specific formats and are automatically processed to extract quantitative markers. All data are then represented in a directed labelled graph (RDF format). Many existing domain-specific ontologies were explored to find relevant concepts and relations.

This knowledge graph can then be queried with SPARQL, a graph-pattern based query language designed to select nodes or edges, or assemble sub-graphs. The main advantage of this approach is that such queries can relate to all parts of the data at once (clinical, genetic, imaging) without the need for explicit joins. Our implementation thus offers convenient multi-source data access: it is now possible to return, for instance, clinical features (phenotype, diagnosis, ...) of individuals having a genetic variant on a specific set of genes, with a single SPARQL query. In addition, federated queries allow to pull data from external sources (for instance, Orphanet and Uniprot) and thus dynamically enrich the knowledge graph.

3 Demonstration scenario

We propose to showcase our prototype in the form of a web application dedicated to biologists and a Jupyter Notebook dedicated to bioinformaticians.

We will demonstrate how these interfaces directly interact with the knowledge graph in the form of template SPARQL queries and how they can be used to answer biological questions. We will show results as graphical plots, for monitoring purposes, or tables, for further bio-statistics analysis or machine learning based predictive modelling.

References

[1] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

⋆. These authors contributed equally to this work