Phenopackets for the Semantic Web

Rajaram Kaliyaperumal[1][0000−0002−1215−167X], Gurnoor Singh[2][0000−0003−1615−4197], Núria Queralt-Rosinach[1][0000−0003−0169−8159], Jumamurat Bayjanov[2][0000−0002−0637−9950], Peter-Bram 't Hoen[2][0000−0003−4450−3112], and Marco Roos[1][0000−0002−8691−772X]

[1] Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands
[2] Radboud University Medical Center, 6525 GA Nijmegen, The Netherlands

Abstract. The GA4GH Phenopackets standard facilitates integrated analysis of genomics and phenomics from patients. Specifically, it allows the representation of phenotypic profiles in a computable and machine-readable exchange format. However, opportunities for integration with resources not represented in GA4GH standards, are limited due to its lack of compatibility with semantic web standards. Here, we present Semantic Phenopackets (RDF schema) for the phenopackets schema which are interoperable with semantic web technologies. Using an approach based on ontological modelling driven by a use case, we show how to represent and query Phenopackets described as RDF graphs.

Keywords: Phenotypes · GA4GH · Phenopackets · FAIR · Ontologies · Semantic Web · RDF · SPARQL

1 Introduction

Integrative analysis of phenotypic and genomic data facilitates the understanding of genotype:phenotype relationships. Standard genomic exchange data formats exist for some years and are widely used in genomics research. However, phenotype information is usually captured in different data format standards and ontologies. There are vocabularies and ontologies to represent phenotype information in clinical contexts (eg., SNOMEDCT [8]) and in research contexts (eg., HPO [12]). Furthermore, there is not a standard way to link genotype to phenotype information for discovery. To create a common representation and exchange format for the clinical and research settings, the Global Alliance for Genomics and Health (GA4GH) initiative established Phenopackets as the phenotypic standard [4]. Importantly, it allows linking this rich phenotypic description with files containing genomic data.

Phenotypes are of special importance for the rare disease community. A detailed description of computable phenotypes benefits research by allowing computational mining and new discoveries. For efficient research on the rare disease field, the European Joint Programme on Rare Diseases (EJP RD) [2] is developing a virtual platform of relevant clinical and biomedical tools and data resources that adopt the FAIR principles [13]. Federated data discovery is a cornerstone of

the project, and hence, a semantic interoperability layer for data and metadata based on ontologies, RDF, Linked Data and Semantic Web technologies is also a key part of the FAIR strategy. Rare disease patient registries are relevant clinical data resources for rare disease research and can potentially solve complex problems in biomedicine. These registries describe a set of well defined common data elements (CDEs) that thanks to a collaborative effort among European Reference Network (ERN) data managers, data stewards and FAIR experts, are being translated into Linked Data by means of ontological models [1].

While the GA4GH Phenopackets standard allows the use of ontologies, the schema per se is not interoperable with the Semantic Web. Therefore, our research question was, how to make Phenopackets interoperable with the EJP RD virtual platform and thus with the Semantic Web. Here, we present Semantic Phenopackets which is the 'ontologized' version of the GA4GH Phenopackets schema. Our approach relies on ontological modelling and the use of Semantic Web technologies. We provide Semantic Phenopackets (RDF schema) for phenotypic representation and for analyses leveraging Linked Data technologies.

2 Methods

2.1 Data

We used a rare disease dataset related to the congenital anomalies of the kidney and urinary tract (CAKUT) disease [10] as a driven use case for the semantic modelling of phenopackets blocks. CAKUT involves a broad spectrum of renal and urinary tract malformation phenotypes ranging from complete renal agenesis (the most severe) to renal hypodysplasia and multicystic kidney dysplasia. This dataset contains clinical data of 178 bilateral CAKUT individuals. The dataset includes personal information, sample information, disease information, phenotypic features and pathology reports of each individual.

2.2 Semantic modelling

We created semantic models for the GA4GH Phenopackets schema version 1 [3]. Our models are based on the Semanticscience Integrated Ontology (SIO), which is an upper-level ontology [9]. It contains various ontological classes and properties to describe entities and their attributes and specifies simple design patterns to uniformly represent them. We used the entity-attribute subpattern within the measurements design pattern to semantically model the phenopackets blocks. To represent the semantic models of the phenopackets blocks we used Shape Expressions (ShEx) [11]. For each phenopackets block model, we also provide example RDF files serialized in Turtle (Terse RDF Triple Language) format [7].

2.3 RDF creation and query federation

To create RDF graphs of the CAKUT dataset we used OpenRefine [6], a software application that is used to perform data wrangling activities. OpenRefine's

RDF extension provides functionalities to transform the content of an OpenRefine project to RDF. For the CAKUT dataset transformation we used the OpenRefine software and its RDF extension version 3.4.1. To demonstrate federated querying of Semantic Phenopackets models with EJP RD virtual platform, we used example RDF Turtle files provided in the EJP RD CDE model GitHub page [1]. The CDE model describes patients and their attributes in RDF according to the definition of CDEs for rare disease registries by the Joint Research Council [5].

3 Results

3.1 Semantic Phenopackets

Out of 18 phenopackets blocks from version 1 of the Phenopackets schema [3], we modelled the following 9 Phenopackets blocks; 'Individual', 'Biosample', 'Disease', 'Sex', 'KaryotypicSex', 'Age', 'File', 'Procedure' and 'PhenotypicFeature'. We chose these 9 blocks since they are the most relevant to the content of the CAKUT dataset. For these chosen phenopackets blocks we created 21 atomic semantic models in total. These semantic models were created based on the SIO entity-attribute pattern and are publicly available on GitHub [3]. For each of these 21 semantic models we provide a separate GitHub markdown file where we specify its ShEx shapes to describe the structure of the RDF graph; an example RDF file and a graphical representation. We show in Figure 1 an example RDF instance for the phenopackets block 'Sex'. We used the generic *sio:is about* (sio:SIO_000332) object property to describe a specific attribute of the entity. We used the data property *sio:has value* (sio:SIO_000300) to describe the value of the attribute. We used ontologies recommended by the phenopackets schema to represent the attributes' values.

3.2 Query Semantic Phenopackets

To demonstrate the simplicity of our semantic model we created two SPARQL queries to retrieve information about individuals. The query [4] shows how to retrieve all individuals and all their attributes, whereas the query [5] shows how to retrieve all individuals and only their date of birth attribute. These two queries show how to query Semantic Phenopackets RDF graphs to perform generic and specific data retrieval with only some minor modifications to the queries. To demonstrate interoperability with EJP RD virtual platform we created a federated SPARQL query [6]. The query matches on the diseases of individuals in the

[3] https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema/wiki
[4] https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema/blob/master/example-queries/query1.rq
[5] https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema/blob/master/example-queries/query2.rq
[6] https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema/blob/master/example-queries/query3.rq

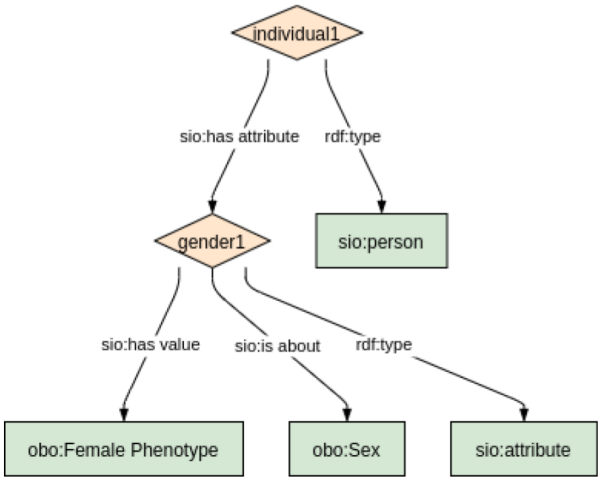

Fig. 1. Example RDF instance for the phenopacket block 'Sex'. The diamond represents an RDF instance and the rectangle represents an IRI value.

Semantic Phenopackets RDF and patients from the RDF graphs of the CDE model. Further, the query lists biobanks from the CDE model RDF graph for the matched patients.

4 Discussion

Semantic Phenopackets is a more machine readable and interoperable version of the GA4GH Phenopackets schema. It aims to capture, for machines, what the elements in a phenopacket mean. It can be used directly in semantic web queries or as a reference for other phenopacket schemas. With our approach of using a simple entity-attribute ontological design pattern we can represent different Phenopackets blocks in a uniform way, which also facilitates data retrieval, and it enables interoperability with the EJP RD virtual platform and with the semantic web by means of the SPARQL query language. Moreover, the reuse of ontological design patterns is a knowledge-engineering recommended good practice. We provided the community with a first set of 21 atomic Semantic Phenopackets models in ShEx, RDF and graphical files open and publicly available on GitHub. Furthermore, using semantic models to represent Phenopackets makes some of the blocks obsolete in the sense that there is no need to explicitly model the attributes since they are already described when resolving the IRI (if the RDF description follows semantic web best practices). For instance, the Phenopackets block 'OntologyClass' only requires the identifier (as a CURIE-style string) and the label (as string).

We developed Semantic Phenopackets as the 'ontologized' version of the GA4GH Phenopackets schema that is interoperable with the semantic web. A

rare disease driven use case was useful to prioritize the set of Phenopackets blocks to model. As future work, we envision updating and modelling the full newly released Phenopackets schema version 2. Moreover, we will make a tool to automate the conversion of Phenopackets to RDF and the translation of SPARQL query results into YAML serialization to facilitate interoperability with other GA4GH Phenopackets clients and tools.

Acknowledgements

Our work is supported by funding from the European Union's Horizon 2020 research and innovation program under the EJP RD COFUND-EJP N° 825575.

References

1. EJP RD CDE model GitHub page. https://github.com/ejp-rd-vp/CDE-semantic-model/tree/develop, last accessed 27 September 2021
2. EJP RD Homepage. https://www.ejprarediseases.org/, last accessed 2020/08/24
3. GA4GH Phenopackets schema version 1. https://phenopacket-schema.readthedocs.io/en/1.0.0/, last accessed 27 September 2021
4. GA4GH Phenopackets standard. https://www.ga4gh.org/news/phenopackets-standardizing-and-exchanging-patient-phenotypic-data/, last accessed 27 September 2021
5. JRC CDE homepage. https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en, last accessed 27 September 2021
6. OpenRefine homepage. https://openrefine.org/, last accessed 27 September 2021
7. Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G.: Rdf 1.1 turtle. World Wide Web Consortium pp. 18–31 (2014)
8. Donnelly, K., et al.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics **121**, 279 (2006)
9. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., et al.: The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. Journal of biomedical semantics **5**(1), 1–11 (2014)
10. Klein, J., Buffin-Meyer, B., Boizard, F., Moussaoui, N., Lescat, O., Breuil, B., Fedou, C., Feuillet, G., Casemayou, A., Neau, E., et al.: Amniotic fluid peptides predict postnatal kidney survival in developmental kidney disease. Kidney International **99**(3), 737–749 (2021)
11. Prud'hommeaux, E., Labra Gayo, J.E., Solbrig, H.: Shape expressions: an rdf validation and transformation language. In: Proceedings of the 10th International Conference on Semantic Systems. pp. 32–40 (2014)
12. Robinson, P.N., Mundlos, S.: The human phenotype ontology. Clinical genetics **77**(6), 525–534 (2010)
13. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3** (2016)