From VCF to RDF: RML-Based Conversion Approaches for the Semantic Representation of Variant Data

Elias Crum[1,2,*], Bart Buelens[2], Gökhan Ertaylan[2] and Ruben Taelman[1]

[1]IDLab, Department of Electronics and Information Systems, Ghent University – imec, Belgium
[2]Flemish institute for Technological Research (VITO) Mol, Belgium

Abstract

Representing Variant Call Format (VCF) data using the Resource Description Framework (RDF) offers benefits in interoperability, integration with other biomedical datasets, and selective privacy protections. Due to complexities of the data represented in VCF files, conversion of VCF to RDF poses challenges, especially concerning complex, heterogeneous data fields. Here, we propose converting VCF files to serialized RDF using the RML mapping language and established genomic data ontologies. Such a methodology will demonstrate the feasibility of an RML-based approach and inform a more FAIR, machine-actionable representation strategy for representing VCF data that is compatible with semantic data privacy policies and useful in both clinical and academic domains.

Keywords

Knowledge Representation, Genomic Data, Semantics, RML

1. Introduction

The generation and use of genomic variant data is rapidly expanding across clinical, personal, and research contexts [1, 2] and most of these applications rely on the Variant Call Format (VCF) for genomic data representation. Despite its widespread adoption, VCF is highly domain-specific and offers limited interoperability with other biomedical data systems. In practice, this constrains integration, reuse, and selective disclosure of variant data across diverse use cases.

VCF also presents technical limitations related to data sharing, linking, and querying. The format provides only weak, inconsistently structured references to external resources (e.g., via rsIDs), lacks native support for fine-grained access control, and enforces all-or-nothing data sharing, even when only subsets of variants are needed. Querying VCF data typically requires specialized tools or custom pipelines, limiting flexibility and scalability. In contrast, RDF-based representations support interoperable linking, expressive SPARQL querying [3], and the application of policy frameworks such as ODRL [4] for granular access control.

Several ontologies and schemas—such as GA4GH variant models [5], FALDO [6], and the Sequence Ontology [7]—demonstrate the potential of semantic representations for genomic variants, but robust and generalizable VCF-to-RDF transformations remain challenging. The structural heterogeneity of VCF fields complicates uniform semantic modeling, motivating declarative mapping approaches. We propose the use of RML [8] and its associated tooling to convert VCF files into RDF.

2. Implementation Design

To investigate the feasibility of representing VCF data as RDF, we developed a representative mapping workflow grounded in established genomic ontologies and executed through the RML ecosystem using pre-existing tools. We propose the use of pre-existing ontologies as well as a supplemental vocabulary

SWAT4HCLS 2026: Bridging Life Sciences and Technology, March 23-26, Amsterdam, Netherlands
*Corresponding author.
elias.crum@ugent.be (E. Crum); bart.buelens@vito.be (B. Buelens); gokhan.ertaylan@vito.be (G. Ertaylan); ruben.taelman@ugent.be (R. Taelman)
0009-0005-3991-754X (E. Crum); 0000-0001-7734-3747 (B. Buelens); 0000-0001-5602-6435 (G. Ertaylan); 0000-0001-5118-256X (R. Taelman)

for the modeling of complex fields such as the HEADER rows, and INFO, FORMAT, and GENOTYPE fields, which are the most heterogeneous and semantically dense components of a VCF file.

A main reason we chose RML as a conversion engine is its rule-based design makes it especially well-suited for representing the irregular and nested structures found in VCF files. The more challenging fields to represent using RDF frequently contain variable-length lists, optional attributes, and complex annotation encodings. RML's expressive mapping framework, including nested iterators, conditionals, and datatype coercion, can capture these structures in a declarative manner that scales well across datasets. This capability is essential for building a more robust, vocabulary-driven view of VCF where metadata definitions, variant annotations, and sample-level attributes are semantically linked.

The proposed implementation presented here highlights the potential for RML-based conversion approaches to contribute to increasingly interoperable, queryable, and privacy-aware genomic data infrastructures in the future.

Acknowledgments

During the preparation of this work, the author(s) used ChatGPT-5.2 for grammar and spelling. The author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

Project funding provided from the Research Foundation – Flanders (FWO) (SB Fellowship 1S27825N). Ruben Taelman is a postdoctoral fellow of the Research Foundation – Flanders (FWO) (1202124N).

References

[1] H. L. McLeod, Cancer pharmacogenomics: Early promise, but concerted effort needed 339 (2013) 1563–1566. doi:10.1126/science.1234139.

[2] E. Souche, S. Beltran, E. Brosens, J. W. Belmont, M. Fossum, O. Riess, C. Gilissen, A. Ardeshirdavani, G. Houge, M. Van Gijn, J. Clayton-Smith, M. Synofzik, N. De Leeuw, Z. C. Deans, Y. Dincer, S. H. Eck, S. Van Der Crabben, M. Balasubramanian, H. Graessner, M. Sturm, H. Firth, A. Ferlini, R. Nabbout, E. De Baere, T. Liehr, M. Macek, G. Matthijs, H. Scheffer, P. Bauer, H. G. Yntema, M. M. Weiss, Recommendations for whole genome sequencing in diagnostics for rare diseases 30 (2022) 1017–1021. doi:10.1038/s41431-022-01113-x.

[3] S. Harris, A. Seaborne, E. Prud'hommeaux, SPARQL 1.1 query language, 2013. URL: https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[4] R. Iannella, S. Villata, ODRL Information Model 2.2, 2018. URL: https://www.w3.org/TR/odrl-model/.

[5] A. H. Wagner, L. Babb, G. Alterovitz, M. Baudis, M. Brush, D. L. Cameron, M. Cline, M. Griffith, O. L. Griffith, S. E. Hunt, D. Kreda, J. M. Lee, S. Li, J. Lopez, E. Moyer, T. Nelson, R. Y. Patel, K. Riehle, P. N. Robinson, S. Rynearson, H. Schuilenburg, K. Tsukanov, B. Walsh, M. Konopko, H. L. Rehm, A. D. Yates, R. R. Freimuth, R. K. Hart, The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification, Cell Genomics 1 (2021) 100027. doi:10.1016/j.xgen.2021.100027.

[6] J. T. Bolleman, C. J. Mungall, F. Strozzi, J. Baran, M. Dumontier, R. J. P. Bonnal, R. Buels, R. Hoehndorf, T. Fujisawa, T. Katayama, P. J. A. Cock, FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation, Journal of Biomedical Semantics 7 (2016) 39. doi:10.1186/s13326-016-0067-z.

[7] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, The Sequence Ontology: a tool for the unification of genome annotations, Genome Biology 6 (2005) R44. doi:10.1186/gb-2005-6-5-r44.

[8] B. De Meester, P. Heyvaert, T. Delva, RDF Mapping Language (RML), 2024. URL: https://rml.io/specs/rml/.