RDF resources for biomedical research

Mayumi Kamada[1,*], Shuichi Kawashima[2] and Toshiaki Katayama[2]

[1]School of Frontier Engineering, Kitasato University, 1-15-1 Kitazato, Minami, Sagamihara, Kanagawa, 252-0373, Japan
[2]Database Center for Life Science, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, Japan

Abstract

The integration of biomedical databases is essential for drug discovery and understanding disease mechanisms. Med2RDF addresses this by proposing a unified data model and providing tools to convert key biomedical datasets into Resource Description Framework (RDF) format. Target databases include those for disease variants, molecular interactions, cancer omics, and genomic variations. To enhance usability, Med2RDF offers Med2RDF-ontology for organizing shared concepts and the RDF-config for generating SPARQL queries and schema diagrams from YAML-based models. Converted RDF data is available through SPARQL endpoints on the DBCLS RDF Portal, enabling seamless integration with other life science data. This standardized and semantic approach facilitates efficient data utilization, advancing biomedical research.

Keywords

biomedical, integrated data utilization, structural variation, RDF

1. Introduction

Integrating life science and biomedical databases is essential in the early stages of drug discovery and in elucidating disease mechanisms in medical research. However, major biomedical databases are provided in diverse formats using various database systems. Furthermore, utilizing these databases requires a thorough understanding of domain-specific data contexts and the relationships between data items.

Med2RDF (http://med2rdf.org/) proposes a data model to integrate and semantically enrich (add semantics to) key biomedical databases. Additionally, it develops and pro- vides software tools to convert these datasets into semantic data representations in the Resource Description Framework (RDF) format (https://github.com/med2rdf).

2. Target Databases

Med2RDF focuses on major databases in the biomedical field, particularly those rel- evant to genomic medicine. It proposes data models and provides conversion programs2 for the RDF format. Table 1 lists the databases that developed the conversion program as of Dec. 2024.

ClinVar and MGeND are databases of disease-associated variants. COSMIC is a database for somatic mutations in cancer. HINT (High-quality INTeractomes) and CPDB (ConsensusPathDB) are databases providing molecular interactions information. ICGC (International Cancer Genome Consortium) and TCGA (The Cancer Genome Atlas) are large-scale consortia collecting multi-omics data for various cancers, providing omics datasets. CCLE (Cancer Cell Line Encyclopedia) and GDSC (Genomics of Drug Sensitivity in Cancer) provide profiles of cancer cell lines, including genomic, gene expression, and drug sensitivity data. dbSNP and dbVar are widely used catalogs of genomic variants. dbNSFP and dbscSNV provide predictive scores of variant effects from bioinformatics tools. HGNC (HUGO Gene Nomenclature Committee) records all human gene names and symbols approved by the committee.

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
kamada.mayumi@kitasato-u.ac.jp (M. Kamada)
0000-0002-2555-7345 (M. Kamada); 0000-0001-7883-3756 (. S. Kawashima); 0000-0003-2391-0384 (. T. Katayama)

Table 1
The converted databases of Med2RDF project (2024)

Database	URL
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
MGeND	https://mgend.ncgm.go.jp/
COSMIC	https://cancer.sanger.ac.uk/cosmic
ICGC	https://dcc.icgc.org/
TCGA	https://portal.gdc.cancer.gov/
CCLE	https://sites.broadinstitute.org/ccle/
GDSC	https://www.cancerrxgene.org/
dbSNP	https://www.ncbi.nlm.nih.gov/snp/
dbVar	https://www.ncbi.nlm.nih.gov/dbvar/
HGNC	https://www.genenames.org/
HINT	https://hint.yulab.org/
CPDB	http://cpdb.molgen.mpg.de/CPDB
dbNSFP	https://sites.google.com/site/jpopgen/dbNSFP
dbscSNV	http://www.liulab.science/dbscsnv.html

3. Ontology

Med2RDF introduces Med2RDF-ontology, which organizes common concepts across the aforementioned biomedical databases (Figure 1). Furthermore, to facilitate connectivity with other public databases, it has developed ontologies such as HCO (Human Chromosome Ontology), MCO (Mouse Chromosome Ontology), and GVO (Genome Variation Ontology), organizing shared concepts.

HCO provides simple and stable URIs for human reference genome versions to identify human chromosomes semantically. In medical research, linking human and mouse data is often necessary. MCO similarly provides URIs for mouse versions. GVO is an ontology that systematically describes various genomic variations, including complex structural variations in genomes.

4. Conversion Programs

The RDF conversion programs are publicly available on the Med2RDF GitHub re- pository (https://github.com/med2rdf). The programs are implemented in multiple lan- guages, including Ruby, Python, and Rust. Schema diagrams are provided for each da- tabase to facilitate the use of converted RDF data. However, understanding a data model from scratch requires domain knowledge, pos- ing a time cost for users other than the original developers. To address this, The Data- base Center for Life Science (DBCLS) has developed a tool called "RDF-config"(https://github.com/dbcls/rdf-config), which provides practical features for using knowledge graphs. RDF-config can automatically generate SPARQL queries and schema diagrams from data models described in YAML format. Figure 2 shows an example of a schema generated by RDF-config. The YAML definition file includes the RDF graph structure, key elements and their attributes with cardinalities, variable names for retrieved values, and example values for inferencing datatypes. These fea- tures help new users understand data models more quickly.

The benefits of RDF-config extend beyond data model comprehension. It standard- izes RDF conversion methods and reduces dependency on programming languages and toolkits, thereby enhancing maintainability. We developed RDF-config to create RDF and JSON-LD data for databases including TCGA, GDSC, dbNSFP, and dbscSNV. Going forward, Med2RDF plans to unify the conversion process for future databases using RDF-config.

Figure 1: The Schema of Med2RDF-ontology

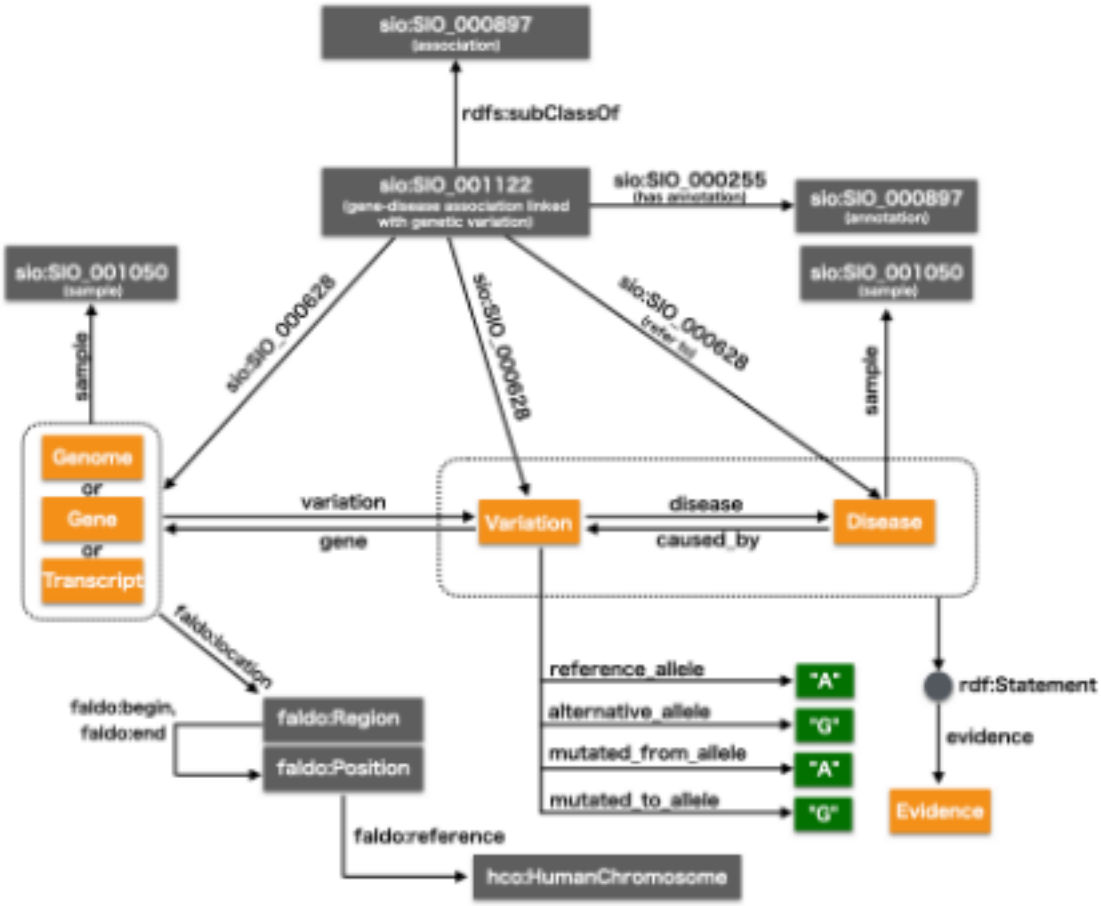
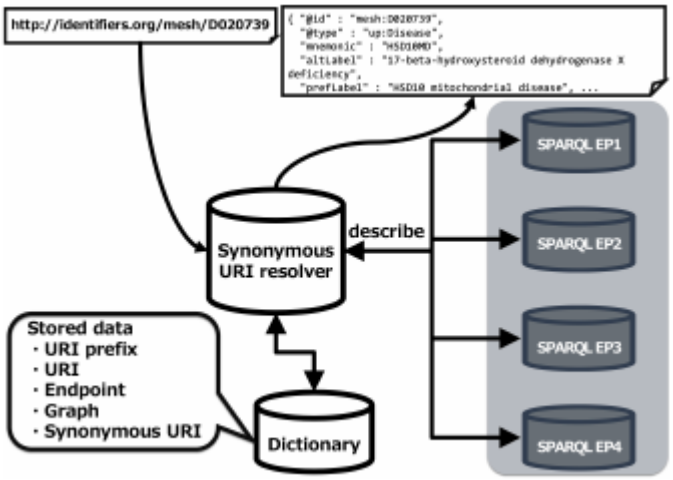


Figure 2: An example of a schema of dbscSNV generated by RDF-config



5. Endpoint

The RDF data for each database converted by Med2RDF programs is made available through SPARQL endpoints on DBCLS RDF Portal (https://rdfportal.org/) [1]. The RDF Portal regularly updates the endpoints to accommodate new databases and changes to source data. It provides an environment for

SPARQL querying and data download. Users can leverage this endpoint to combine data with other life science databases for integrated analyses.

Acknowledgments

This research is supported by a Grant-in-Aid for Transformative Research Areas (A) “Latent Chemical Space” [JP23H04880 and JP23H04886] for MK from the Ministry of Education, Culture, Sports, Science and Technology, Japan

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Kawashima, T. Katayama, H. Hatanaka, T. Kushida, T. Takagi, Nbdc rdf portal: a comprehensive repository for semantic data in life sciences, Database 2018 (2018) bay123. doi:10.1093/database/bay123.