Constructing a synonymous URI dictionary in life sciences

Yasunori Yamamoto[1,*], Takatomo Fujisawa[2,*]

[1]Database Center for Life Science (DBCLS), ROIS-DS, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, JAPAN
[2]Bioinformation and DDBJ Center, National Institute of Genetics (NIG), 1111 Yata, Mishima, Shizuoka 411-85

Abstract

Many life science data have been published in RDF. In general, life science data are diverse and store knowledge about various relevant concepts and relationships among them. These concepts include proteins, genes, compounds, and diseases, and are represented by identifiers in databases. To understand biological phenomena, it is crucial to extensively investigate their characteristics and relationships among them, and it is ideal to use one and only identifier for a concept over the databases However, multiple identifiers are often used for a concept in reality. Database Center for Life Science (DBCLS) constructs or collects life science RDF data and provides them at RDF Portal. Here, we have investigated the synonym URIs in it and examined the challenges and future works. Keywords

Keywords

RDF data, URI synonyms, Database integration

1. Introduction

Life science data are diverse and have complex structures. The vast array of molecules and compounds that constitute life, such as genes, proteins, and glycans, has led to the development of numerous related databases. For example, UniProt[1], a globally renowned and used protein database, contains over 250 million protein identifiers. Each protein entry includes detailed descriptions of its biological functions and other related information. Furthermore, various chemical reactions and molecular interactions occur within living organisms, and these phenomena have also been recorded in databases. Consequently, data pertaining to a specific protein can be found across multiple databases.

data pertaining to a specific protein can be found across multiple databases. With advancements in experimental techniques and research in the life sciences, the volume and complexity of acquired knowledge have increased significantly. As a result, the scale and variety of databases have expanded, posing challenges for efficiently retrieving the necessary data. To address this issue, a technical solution leveraging the Resource Description Framework (RDF) to represent life science knowledge was proposed in the mid-2000s, leading to the development of RDF databases worldwide.

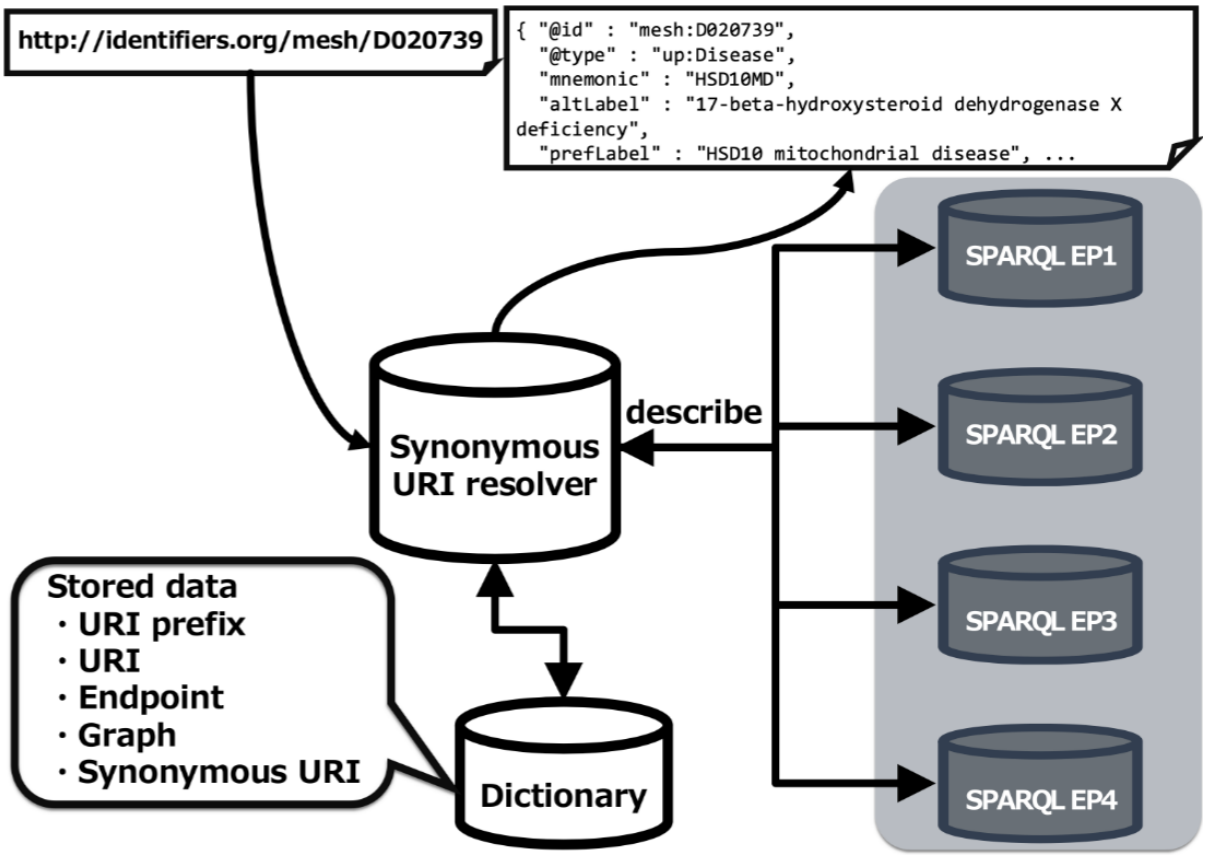
In RDF, concepts are identified using Uniform Resource Identifiers (URIs), and when a single URI is assigned to a concept, even if multiple entities independently construct their databases, it becomes easier to integrate their knowledge and build new insights. However, in reality, multiple URIs often exist for a same concept, undermining the value that can be derived from integrating multiple databases. This challenge remains a significant barrier to fully leveraging the potential of RDF in the life sciences.

Database Center for Life Science (DBCLS) has been developing life science RDF databases while also aggregating existing ones to provide an RDF portal[2]. Currently, the portal hosts 40 databases containing approximately 160 billion triples, which can be queried using SPARQL. However, there are the aforementioned synonymous URIs, and therefore we are constructing a dictionary of synonymous URIs. By enabling searches that utilize this dictionary, it will become possible to comprehensively retrieve information related to a single concept across multiple databases.

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
0000-0002-6943-6887 (Y. Yamamoto); 0000-0001-8978-3344 (T. Fujisawa)

[1]https://www.uniprot.org
[2]https:///rdfportal.org

Figure 1: Overview of our service



2. Two types of URI synonyms

Synonymous URIs are classified into two types based on their suitability for automated processing: those differing only in their prefixes and those following entirely different naming conventions. We call them the prefix type and the naming convention type, respectively. For the prefix type, many cases involve databases that existed prior to the adoption of RDF. When these

For the prefix type, many cases involve databases that existed prior to the adoption of RDF. When these identifiers were converted to RDF, a prefix was often added to align with RDF conventions. For example, UniProt uses a URI to represent a protein in RDF as follows: http://purl.uniprot.org/uniprot/P0DTC2 Here, P0DTC2 is the identifier used by UniProt to denote a specific protein, and the rest is the prefix. While many databases adopt UniProt identifiers to represent proteins, the prefix often differs depending on the database.

In contrast, the naming convention type refers to cases where entirely different URIs are used for a same concept. This is particularly evident in representations of disease concepts.

3. Our approach

Based on the above considerations, we identified the need for a basic service that, from the perspective of database users, enables seamless searching across both prefix-type and naming conventiontype synonymous URIs. Given a specific concept, the service should retrieve all associated synonymous URIs, query the RDF database for each URI, and aggregate the results for presentation. Considering the features of this service, the intended direct users are database developers and operators while the real end users are researchers. The service will be provided as a web-based platform. Its basic specification involves treating HTTP access to a specific URL as input and returning the results in JSON format.

To address the fragmentation of knowledge graphs caused by synonymous URIs, we named the service URI Resolver. The key functionalities that URI Resolver should possess are as follows: (1) Checking for Synonymous URIs: When a URI is provided, the service should verify the existence of synonymous URIs by consulting a precompiled dictionary. To limit the potentially vast search space, prefixes will also be utilized to narrow down the scope. (2) SPARQL Query Execution: Using the dictionary, the service retrieves all endpoints associated with the synonymous URIs. For each endpoint, it performs a SPARQL query by using the corresponding graph name and URI, obtaining all triples where the given URI appears as the subject. Fig.1 describes this service.

To implement these functionalities, the dictionary must include all sets of synonymous URIs defined as equivalent within the RDF portal. Additionally, it should store the endpoints and graph names where data related to each URI can be retrieved. Furthermore, existing resources such as TogoID[1] (naming convention), Identifiers.org[2] (prefix), and Bioregistry[3] (prefix), which already provides API services for synonymous URI information, will be leveraged to enhance the functionality and coverage of the URI Resolver.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Ikeda, H. Ono, T. Ohta, H. Chiba, Y. Naito, Y. Moriya, S. Kawashima, Y. Yamamoto, S. Okamoto, S. Goto, T. Katayama, TogoID: an exploratory ID converter to bridge biological datasets, Bioinformatics 38 (2022) 4194–4199.

[2] M. Bernal-Llinares, J. Ferrer-Gómez, N. Juty, C. Goble, S. M. Wimalaratne, H. Hermjakob, Identifiers.org: Compact identifier services in the cloud, Bioinformatics 37 (2021) 1781–1782.

[3] C. T. Hoyt, M. Balk, T. J. Callahan, D. Domingo-Fernández, M. A. Haendel, H. B. Hegde, D. S. Himmelstein, K. Karis, J. Kunze, T. Lubiana, N. Matentzoglu, J. McMurry, S. Moxon, C. J. Mungall, A. Rutz, D. R. Unni, E. Willighagen, D. Winston, B. M. Gyori, Unifying the identification of biomedical entities with the bioregistry, Sci. Data 9 (2022) 714.