PoLyInfo RDF: A Semantically Reinforced Polymer Database for Materials Informatics

Masashi Ishii[1], Taro Takemura[1], and Mikiko Tanifuji[1]

[1] National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan
ISHII.Masashi@nims.go.jp

Abstract. For materials database integration, we introduce a semantic web technology into a polymer database, called PoLyInfo. A resource data framework (RDF) was used to create a semantic description of the polymer formation process (polymerization). The polymerization correlates polymers with their source monomers, and then the monomers in the PoLyInfo RDF were conceptually linked to those in other chemical substance databases, such as Nikkaji. Although common ontology was not used in the PoLyInfo and Nikkaji RDFs, 94.3% of the monomers in PoLyInfo were assigned a Nikkaji substance ID, and the established information path provided probable polymerization information for monomers in the Nikkaji RDF.

Keywords: RDF, Polymer Database, Monomer.

1 Introduction

A semantic web technology using resource data framework (RDF) [1] for linked open data (LOD) is widely accepted among life science societies in which open science has been historically approved. However, in material science, a sub-domain of physics and chemistry, industrial importance and confidentiality of material development are not always consistent with open science, and hence, semantic web technology has never held a position of major importance. In spite of the traditional marketing mechanism, recent material development, accompanied by ecology and economy, accountability for products, and social contribution, necessitates knowledge-sharing. LOD satisfies this demand and is considered the leading player in open-data-driven materials informatics (ODD-MI) [2]. In this study, we design an RDF for a material (polymer) database, for ODD-MI, and demonstrate the linking of data with other international databases.

2 RDF Design for a Polymer Database

The materials database “PoLyInfo”, managed by the National Institute for Materials Science (NIMS)[1], is a unique database for basic polymer science and has accumulated 334,738 properties for 116 headings (as of April 2019) over a period of fifteen years [3]. Since the transferal of PoLyInfo from the Japan Science and Technology Agency

[1] https://www.nims.go.jp/eng/index.html, [2] https://www.jst.go.jp/EN/

(JST)[2] to NIMS in 2003, the database has been unconcerned with semantic web technology, that is, the database has been functioning as an unlinked data source. Due to the skillful identification of polymer naming in academic articles, PoLyInfo has proved expensive owing to an increase in polymer data that are not easily obtainable from other chemical databases. On the other hand, the uniqueness of PoLyInfo means it contains few common terms in comparison to other related elements, resulting in a technical barrier to the linked data. From a survey of PoLyInfo aiming to realize linked data, we deduced that the most important and linkable term in the database is the "monomer name". The grounds of this deduction were established as follows:

- Polymers are essentially synthesized from monomers.
- The synthesis (polymerization path) represents the primary information for polymer developers.
- There are several monomer databases linked to life science with RDF.

Fortunately, as PoLyInfo has managed polymer names, polymerization paths, monomer names with original identification numbers (IDs), and a protocol between these IDs, we were able to design an RDF-connected graph for PoLyInfo, as shown in Figure 1, without utilizing frontier tools [4], whereby we defined ns1 and ns2 as prefixes:

@prefix ns1: <https://polymer.nims.go.jp/rdf/> . (tentative prefix for development)
@prefix ns2: <http://www.w3.org/2000/01/rdf-schema#> .

The triple graph in Fig. 1 shows an example of an industrially important polymer, "polystyrene", and its semantic links can be described as follows:

1) The polymer with composition unit identified by CU010001 has the name "Polystyrene" (CU010001 ns2:label "Polyethylene"@en.).
2) CU010001 has a polymerization path with the ID number J000002 (CU010001 ns1:pHasPolymerizationPath J000002.).
3) J000002 has the name of "Addition polymerization" (J000002 ns2:label "Addition polymerization"@en.).
4) J000002 has source monomers with ID numbers M0301021 and M0101001 (J000002 ns1:pHasMonomer M0301021, M0101001.).
5) M0301021 has the name "Buta1,3-diene", and M0101001 has the name "Ethene" (M0301021 ns2:label "Buta1,3-diene"@en. M0101001 ns2:label "Ethene"@en.).

Consequently, the polymer CU010001 is correlated to the monomers M0301021 and M0101001.

In this study, we performed data linking with an organic chemical database created by JST, called Nikkaji (Japan Chemical Substance Dictionary) [5], which primarily has monomer information comprising 3,459,747 records. Nikkaji has previously been published in RDF (Nikkaji RDF) [6], in 2015. The Nikkaji RDF was designed by the National Bioscience Database Center (NBDC)[3] with the aim of linking data to the life science domain using a monomer. The Nikkaji RDF was standardized with an ontology that is commonly used in the ChEMBL database of the European Bioinformatics Institute, the European Molecular Biology Laboratory (EMBL-EBI)[4] in the UK, and the PubChem database [7] of the National Center for Biotechnology Information (NCBI)[5] in the US. The latest Nikkaji RDF provides 146,220,942 triples.

To link the PoLyInfo RDF to the Nikkaji RDF, we introduced conceptual linking, based on Simple Knowledge Organization System (SKOS) specifications [8],

[3]https://biosciencedbc.jp/en/
[4]https://www.ebi.ac.uk/
[5]https://www.ncbi.nlm.nih.gov/

published as part of the World Wide Web Consortium (W3C) recommendations. In the example illustrated in Fig. 1,

6) M0301021 (Buta1,3-diene) and M0101001 (Ethene) of PoLyInfo closely match J4.043F and J1.939I of Nikkaji, respectively. The corresponding RDF triples in SKOS can be expressed by:

@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
ns1: M0301021 skos:closeMatch nikkaji: J4.043F .
ns1: M0101001 skos:closeMatch nikkaji: J1.939I .

As these entities may not be chemically identical (e.g., they may exhibit a difference in purity), we concluded that the "skos:closeMatch" was better than "skos:exactMatch" and "owl:sameAs".

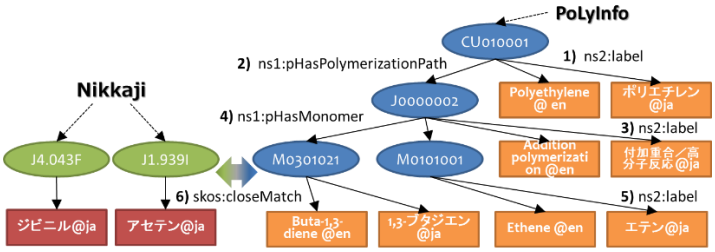

Fig. 1. Triple graph designed for the semantic description of polymerization in PoLyInfo.

3 Results and Discussions

Based on the designed RDF in Fig. 1, we made triples for the polymerization process in PoLyInfo, as listed in Table 1.

Table 1. RDF triple numbers for each polymerization step in Fig. 1.

RDF	Predicate in RDF	Triple numbers
1)	ns2:label	34,778
2)	pHasPolymerizationPath	15,884
3)	ns2:label	35,642
4)	pHasMonomer	35,018
5)	ns2:label	36,459
6)	skos:closeMatch	16,809
Others	rdf:type	52,892
	Total	227,482

This table indicates that 15,884 polymers of 2) had information concerning the polymerization path, and 35,018 monomers of 4) were related to these polymerizations. As PoLyInfo has 17,825 monomers while the RDF for skos:closeMatch of 6) was 16,809 triples, this shows that 94.3% of monomers in PoLyInfo were assigned to Nikkaji substance ID. The total number, 227,482 triples, was not significant. However, we also performed a demonstration using a query written in SPARQL. The crossover task

was to compile a list of polymers in PoLyInfo, which could be synthesized from the monomers in Nikkaji, and the response provided 39,907 polymers. This result indicates that the PoLyInfo RDF was successfully linked to the Nikkaji RDF.

Although we showed the potential of the use of linked data technology between PoLyInfo and Nikkaji, we still encountered a number of problems intrinsic to polymer science. In PoLyInfo, we normalized the polymer name to International Union of Pure and Applied Chemistry (IUPAC)[6] nomenclature. Unfortunately, the international rule sometimes conflicts with the RDF format of W3C. An example of this conflict can be seen in a triple for polymerization ID of J0018355:

<https://polymer.nims.go.jp/rdf/J0018355>
ns1:pHasMonomer <https://polymer.nims.go.jp/rdf/M4000864> ;
ns2:label "Addition_polymerization_of_9,9,9',9',9",9"-hexahexyl-7-(4-vinylphenyl)-2,2':7',2"-terfluorene"@en .

The triples indicate that J0018355 with source monomer M4000864 has a label of “Addition polymerization of 9,9,9',9',9",9"-hexahexyl-7-(4-vinylphenyl)-2,2':7',2"-terfluorene”. However, the IUPAC monomer name M4000864, including double quotes (") without an escape code, clearly conflicts with the RDF format (the escape code is undefined in IUPAC nomenclature), resulting in a syntax error in uploading process.

4 Summary

To establish linked data for the polymer database PoLyInfo, we designed RDF triples implementable in a polymerization process. The triples describing polymerization from source monomers finally bridged the entities present in the other chemical substance databases, such as Nikkaji. By using the 227,482 triples created for the purposes of this study, we found 39,907 polymers that could be synthesized from monomers registered in Nikkaji. In some cases, we found an unexpected conflict between the International Union of Pure and Applied Chemistry (IUPAC) nomenclature and the RDF format recommended by W3C.

This study was supported by the Cabinet Office, Government of Japan, Cross-ministerial Strategic Innovation Promotion Program (SIP), “Technologies for Smart Bio-industry and Agriculture” (funding agency: NARO).

5 References

1. RDF 1.1 Primer, https://www.w3.org/TR/rdf11-primer/.
2. Rajan, K.: Materials Informatics: The Materials “Gene” and Big Data. Annu. Rev. Mater. Res. **45(1)**, 153–169 (2015).
3. Polymer Database (PoLyInfo), https://polymer.nims.go.jp/index_en.html.
4. The Linked Data Integration Framework, http://silkframework.org/.
5. Nikkaji Web, https://integbio.jp/dbcatalog/en/record/nbdc01530.
6. NBDC NikkajiRDF, https://dbarchive.biosciencedbc.jp/en/nikkaji/desc.html.
7. PubChem Database, https://pubchem.ncbi.nlm.nih.gov/.
8. SKOS Primer, https://www.w3.org/TR/skos-primer/.

[6] https://iupac.org/iupac