実用材料開発のためのデータベース「PoLyInfo RDF」の構築 (II)：PubChem RDF との統合による境界なきデータ駆動

Construction of "PoLyInfo RDF" database for practical material development (II): Open data-driven by using data linking with PubChem RDF

物材機構 [1]　○石井真史 [1]，竹村太郎 [1]，谷藤幹子 [1]

NIMS[1], ○Masashi Ishii[1], Taro Takemura[1], Mikiko Tanifuji[1]
E-mail: ISHII.Masashi@nims.go.jp

【序】実用材料開発では、採算性、安全性、環境問題などが学術研究以上に重要になり、また逆にそれが研究のモチベーションとなっている。このことは、領域を越えたデータ駆動の必要性を示しており、W3C（World Wide Web Consortium）[1]の国際的枠組で展開されるセマンティックウェブ技術がその具現化の鍵と目される[2]。

この技術は、あらゆる情報を主語・述語・目的語の三種（Triple と言う）で記述する RDF（Resource Description Framework）と呼ばれるプロトコルの導入で実現できる。世界最大級の低分子データベース PubChem は 2014 年に既に RDF を取り入れており、BioAssay など生物試験データとの連携に成功している[3]（図 1 左）。昨年、我々は NIMS 高分子データベースを RDF 化した PoLyInfo RDF を試作し（図 1 右）、低分子 DB 日化辞 RDF との統合に成功した[2]。本研究では、この PoLyInfo RDF を拡張する二つの取り組み(1) PoLyInfo RDF にて高分子→低分子のブレークダウンの役を担っている重合情報の増強 (2) PubChem RDF との新規統合、を紹介する。

【統合強化の方法】

(1) 重合情報の強化　巨大データベース作成の必須事項として語彙の揺らぎの除去（類語辞書化と ID 附番）がある。PoLyInfo では原料低分子の多くは原料低分子の論文記載名で収録されてきたが、今回その揺らぎを除去し、さらに重合パスを既定の 14 種類に帰属させた。これを使いホモポリマーの重合情報 71,615 件について高分子→低分子のブレークダウンを RDF triple で記述した。

(2) PubChem RDF との統合　W3C の規格である skos（Simple Knowledge Organization System）[3]を使って実現した PoLyInfo RDF と日化辞 RDF 間の低分子の概念一致は、さらに今回 PubChem との概念一致に拡張された。具体例で示すと ns1:M4001239 skos:closeMatch rpcsid:273962184, skos:closeMatch rpccid:101536123.

ここで、
rpcsid: http://rdf.ncbi.nlm.nih.gov/pubchem/substance/SID
rpccid: http://rdf.ncbi.nlm.nih.gov/pubchem/compound/CID
としており PoLyInfo の低分子 ID M4001239 (N,N-Bis(4-bromophenyl)-4-(trifluoromethyl)aniline, $C_{19}H_{12}Br_2F_3N$) が PubChem の Substance ID 273962184 と Compound ID 101536123 の URI（Uniform Resource Identifier）に対応付けられる。この Compound ID の URI を主語、図 2 に示す has attribute を術語、23 種の物性の URI を目的語とする triple ができ、更にこの物性 URI に has value の術語で物性値がつながり、両 DB は統合される。

【結果とまとめ】(1)の重合情報の強化による PoLyInfo RDF の改良については現在校閲を行っており当日その結果を報告する。(2)の PubChem RDF との統合として、上記の skos:closeMatch を術語とする triple を 161,640 作成することに成功した。これにより BioAssay や特許など PoLyInfo RDF にとって未知領域の情報が境界なきデータ駆動の射程内に入った。

【謝辞】本研究は、内閣府「戦略的イノベーション創造プログラム（SIP）」の「革新的バイオ素材・高機能品等の機能設計技術及び生産技術開発」の支援を戴いた。

【文献】

[1] https://www.w3.org/
[2] M. Ishii, T. Takemura, and M. Tanifuji, ISWC 2019, Auckland, NZ (2019).
[3] https://pubchemdocs.ncbi.nlm.nih.gov/rdf
[4] https://www.w3.org/TR/skos-primer/

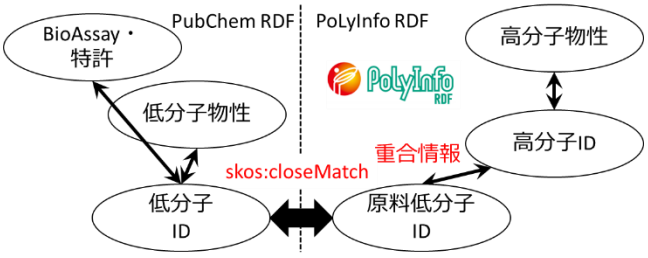

図 1 PoLyInfo RDF と PubChem RDF の統合
赤字が今回の強化部分

図 2 PubChem RDF の物性の URI（CID 101536123 を例として）