一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SC2019-9(2019-06)

NIMS 高分子データベース PoLyInfo の RDF 化
－セマンティックなデータベース統合への試行報告－

石井 真史† 　竹村 太郎‡ 　谷藤 幹子‡

†物質・材料研究機構　統合型材料開発・情報基盤部門 〒305-0047 つくば市千現 1-2-1
E-mail: †ISHII.Masashi@nims.go.jp, ‡{TAKEMURA.Taro, TANIFUJI.Mikiko}@ nims.go.jp

あらまし　物質・材料研究機構の高分子データベース PoLyInfo の一部を試験的に RDF 化し，有機化学でよく知られた JST の日化辞や NIH の PubChem データベースとの統合をシミュレートした．重合反応を接点にした情報共有の実証と併せ，本格的なセマンティックなデータ統合を進める上で PoLyInfo の構造上の課題を調査した．一般的に RDB（Relational Database）から RDF は比較的容易に生成できるが，PoLyInfo の場合，データベースの列名の設定やデータ記述の正規化において，RDF 化にあたり幾つかの課題があることが明らかになった．独立した大規模な Web 上の データベースやデータセットを関連付ける Linked Data の手法を，データ駆動の材料科学に適用するにあたり，現状と解決すべき課題を報告する．

キーワード　RDF, トリプル，統合データベース，PoLyInfo，RDB

Trial of remodeling of NIMS polymer database, PoLyInfo in RDF
－The first report of semantic connection to unspecified databases－

Masashi ISHII† 　Taro TAKEMURA‡ 　and 　Mikiko TANIFUJI‡

†National Institute for Materials Science, MaDIS　1-2-1 Sengen, Tsukuba, Ibaraki, 305-0047 Japan
E-mail: †ISHII.Masashi@nims.go.jp, ‡{TAKEMURA.Taro, TANIFUJI.Mikiko}@nims.go.jp

Abstract　We partly rearranged NIMS polymer database PoLyInfo in RDF, and demonstrated integration with Nikkaji and Pubchem databases which are well used among organic chemists. The demonstration of data sharing through polymerization path indicated feasibility of RDF in materials science, and so we investigated technical barriers involved in conventional PoLyInfo prior to full-scale construction of “PoLyInfo RDF”. Although RDB (Relational Database) can be simply converted into RDF in general, we revealed that symbolized column name in the PoLyInfo RDB tables and unnormalized data form in each table were still technical issues. We discussed problems which we should to solve for the Linked Data and data-driven science using PoLyInfo RDF.

Keywords　RDF, Triple, Database-integration, PoLyInfo, RDB.

1. 序

1.1. 化学データベースにおけるセマンティック web の動き

データを発見することを目的としたセマンティック web の考えが提唱されて以来，web 形式の標準化を牽引する W3C（World Wide Web Consortium）[1]の指針の下，アプリケーション，企業，コミュニティの境界を越えたデータの共有と利用の動きが高まっている．このアイディアは RDF（Resource Description Framework）を技術的基盤としており[2,3]，化学分野においても，これまで各機関が個別に構築し運用してきたデータベースを統一規格によって繋ぎ，知識の統合を進める方向に進んでいる．日本国内の化学物質辞書としてよく知られた日化辞の RDF データ「日化辞の本体の RDF データ」と「日化辞の InChI の RDF データ」が 2015 年に公開され，その翌年には「日化辞と他の DB のリンク情報の RDF データ（UniChem 由来）（PubChem 由来）」が公開され，異なる概念体系（データベース）が skos（Simple Knowledge Organization System）を用いてマッピングされた[4,5]．例えば，日化辞番号 J1.000.001G「α-フェナシルベンジルマロノニトリル」が PubChem SID 272692933 “alpha-Phenacylbenzylmalononitrile” に完全一致（skos:exactMatch　*has exact match*）し，PubChem SID 2807482 “2-(3-Oxo-1,3-diphenylpropyl)malononitrile” にほぼ一致する（skos:closeMatch　*has close match*）というように，二つのデータベース，日化辞と PubChem[6]が ID を介して概念的に繋がり，情報が共有できるようになった．更新は逐次行われており，現在 330 万件以上の化学物質が RDF 化されている[3].

1.2. 高分子データベース PoLyInfo とセマンティック

web

高分子データベース PoLyInfo [7]は，2003 年 4 月 1 日に当時の科学技術振興事業団（JST）から移管されて以来，物質・材料研究機構にて管理運用している．2019 年 4 月 12 日現在，コポリマー数 5,871，ポリマーブレンド数 2,005，コンポジット数 2,351，モノマー数 17,825，物性ポイント数 334,738 のデータを公開している．しかし，これまでセマンティック web に準じた取り組みはなされておらず，日化辞の他，PubChem や ChemSpider [8]などにみられる世界的な RDF 化には同調していなかった．

しかしながら，各所にあるデータ集約し研究に利用するデータ駆動科学が昨今盛んになる中で，セマンティック web を導入して，データを共有する方向を検討する必要があろう．PoLyInfo は高分子のデータベースであり，低分子を中心にしている他の化学系データベースとは異なる稀有な存在であることを考えると，高分子・低分子の両分野にとって RDF による統合の波及効果は高いと期待される．

オントロジーのような高度な統合に先立ち，PoLyInfo の高分子重合情報にある原料モノマーの情報を統合の共通概念に使うことが考えられる．例えば，PoLyInfo には「ポリエチレン（PoLyInfo ポリマーID P010001）は，エテン（PoLyInfo モノマーID M0101001）と 1,3-ブタジエン（PoLyInfo モノマーID M0301021）の付加重合/高分子反応で形成できる」という重合情報が収蔵されているが，これらの原料モノマーは日化辞番号 J1.939I と J4.043F に紐づいている．更に前者は日化辞 RDF で PubChem SID 273486715 に skos で繋がっている（後者は SID がないが CID7845 には対応している）．

こうしたデータベース統合のストーリーが描ける中で，今回 PoLyInfo の RDF 化を試行し，本格実施に入る前の技術的課題を検討した．

2. PoLyInfo の RDF 化とデータベースの統合

今回 PoLyInfo に収蔵されている以下の三種のポリマーを RDF 化した：ホモポリマー P010001（polyethene），コポリマーP900001（poly[(acrylic acid)-co-(butyl acrylate)]），ポリマーブレンド BD000011（polyethene//polystyrene//polybutadiene）．これらの違いはここでは述べないが，高分子科学では最も大きなポリマーの区分から一つずつサンプルを取り出している．作成したトリプル数は 919 であり，極めて小規模であるが，以下の通り PoLyInfo RDF を本格作成するための見通しが得られる．

作成した RDF をトリプルストア（Apache Jena, Fuseki [9]）にアップロードし，1.2 節で思い描いた他の有機化学系データベースとの統合をシミュレートしてみる．例題となった PoLyInfo の重合情報のトリプルグラフは，図 1 のように「ポリエチレン（PoLyInfo ポリマーID P010001）は，構成単位としてポリエチレン（CU010001）を持ち，それはエテン（PoLyInfo モノマーID M0101001）と 1,3-ブタジエン（PoLyInfo モノマーID M0301021）の付加重合/高分子反応という重合（J0000002）で形成できる」となる．日化辞番号 J1.939I，すなわち PubChem SID 273486715 に繋がっているエテンモノマー（PoLyInfo モノマーID M0101001）に端を発する，次の SPARQL[12]によるクエリ

```
SELECT ?monomer ?pathname ?name
WHERE {
  ns1:rdfM0101001 ns2:label ?monomer
  filter(LANG(?monomer) = 'en') .
  ?polymerpath ns1:rdfpHasMonomer ns1:rdfM0101001.
  ?polymerpath ns2:label ?pathname
  filter(LANG(?pathname) = 'ja') .
   ?polymerpath ^ns1:rdfpHasPolymerizationPath /
  ns2:label ?name filter(LANG(?name) = 'ja') .
}
```

に対して

	monomer	pathname	name
1	"ethene"@en	"エテン_&_1,3-ブタジエン_付加重合/高分子反応"@ja	"CU-ポリエチレン"@ja
2	"ethene"@en	"エテン付加重合"@ja	"CU-ポリエチレン"@ja

の応答が得られ，エテンモノマーが二種の重合反応を介してポリマー（ポリエチレン）を形成することが分かる．ここで二種の重合反応が提示されるのは，図 1 に示していないトリプルからの応答（「エテン付加重合」）が含まれるためである．こうして期待通り，重合反応を介して，他のデータベースと情報を共有できる可能性が示された．

3. PoLyInfo の本格的 RDF 化における課題

3.1. データベース構造の観点から

PoLyInfo のデータベース構造について全体にわたって言えることは，現在の緻密なデータ構造が GUI（Graphical User Interface）側で実現されている点である．すなわち人の目に触れない RDB 内部のテーブルについては，読み解きやすいとは言い難い．一般的に RDF は RDB（Relational Database）から比較的容易に生成できるとされており，実際ツールも公開されている．しかしそれは，解釈容易な整理されたテーブルが用意されていることが前提となる．この観点での PoLyInfo のデータベース構造上の問題を二つ挙げる．

テーブルにおける列名の問題：通常，RDB から RDF への変換は，テーブルの行名を主語，列名を述語，行と

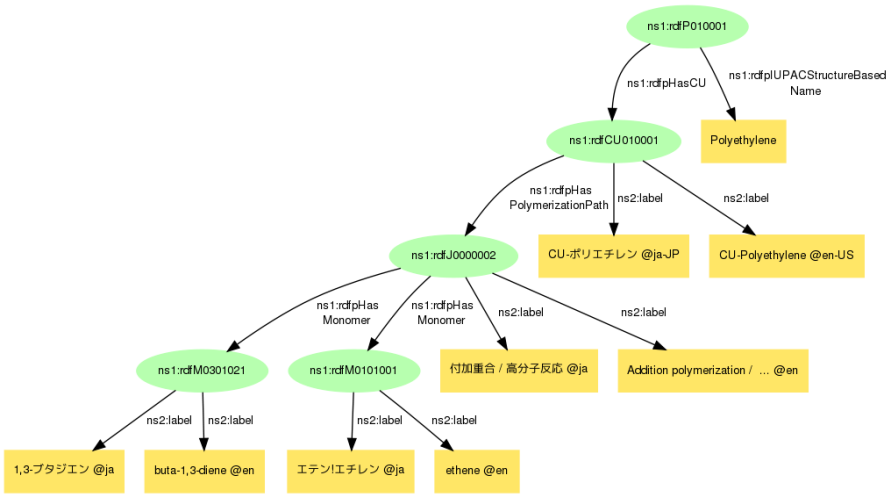


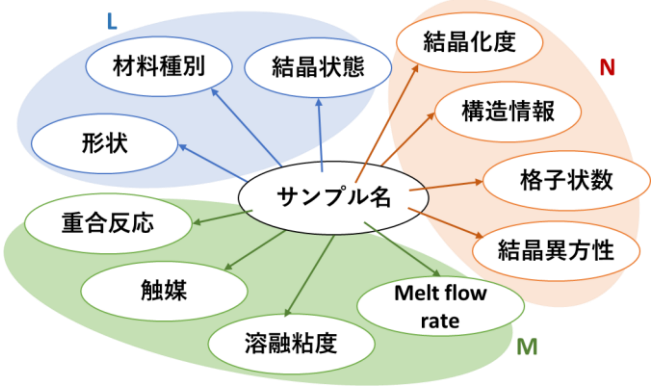
図 1 PoLyInfo のポリエチレン（PoLyInfo ポリマーID P010001）重合情報のトリプルグラフ．作成した RDF を選別して ARC2[10]をベースにしたオープンツール[11]でグラフ化した．

列の交点を目的語としてトリプルを作成する前提に立っている．PoLyInfo のテーブルの列名は，pm_01，jm_03，jm_06 のように記号化されており，セマンティックな表記になっていない．これらの列名は本来の意味である，ポリマーID，モノマーID，重合パス名を指し示す表記にする必要がある．PoLyInfo 設計当時に，RDB 内部のテーブルが人目に触れ，RDF としてデータベース統合に使われることは想定されていなかった結果であるが，新たなステージに立って編纂を要する．

情報の散在の問題：PoLyInfo では，設計当時の技術水準から，テーブル構成は検索機能を優先している．そのためサンプルの種類・形状，熱処理条件など合成過程や結晶状態・結晶化度などの材料情報が複数の RDB テーブルに散在している．図 2 に，この様子を示している．L，M，N が実際に使われているテーブル名であり，プロセス情報が一つにまとまっていないことが分かる．RDF 変換に先立ち，テーブルの整理がまず必要になる．

3.2. 人可読と機械可読の観点から

2.では，他のデータベースへの共通概念として重合情報を使うことを，ポリエチレンを例に示した．しかし，トリプルグラフを良く読み解くと「付加重合/高分子反応という重合（J0000002）」という重合パスのラベル（ns2:label）は，本来分けるべき二つの反応「付加重合」と「高分子反応」を"/"を使って一つにまとめて記載している．PoLyInfo は機械可読の考えが浸透するよりもずっと昔に構築されたため，人が見れば理解できる（人可読）情報は，多くの場合慣用的表現で表記されてきた．PoLyInfo の人可読のデータ表記は，この例のように「/」で分てば，機械可読用に容易に正規化できるものだけではない．以下は図 2 のテーブル M に記載されている，サンプル ID M01007_2_1 の重合に関する情報を，Turtle 形式のトリプルで表した例である．



https://polymer.nims.go.jp/rdfM01007_2_1
ns0:rdfpolymzInfo_PolymerReactionCondition
"Exposured to difluoromethylene generated by pyrolysis of sodium chlorodifluoroacetate for 8min, pyrolysis temp. 385C" .

この" "で囲まれた自由記述の反応情報オブジェクトは人可読には十分であるが，機械可読を考えると構造化が必要である．しかしながら，試料ごとに表記が異なる情報をどのようにまとめるかは，今後検討課題である．

4. 今後の展開

本稿で議論した問題点以外にも課題は多く残されている．PoLyInfo に掲載されている約 100 種の物性値を表すのに単位の定義は欠かせないが，DBCLS（Database Center for Life Science）で推奨されている適切なオントロジーUO（Units of measurement ontology）での指定[13]でカバーしきれない単位（例えば比容積として使われている cm^3/g）は別途定義する必要がある．また単位に限らず，語彙の違いなど分野を越えて知識を交換する為に，PubChem RDF で導入されている Chemical Information Ontology（CHEMINF）や Semanticscience Integrated Ontology（SIO）などの共通のオントロジーを用いてデータの RDF 化を進めることは PoLyInfo のビジビリティを高める上で大切になる．

データベース内部に解決すべき課題が残されている一方で，RDF 自体を最大限に活用するための更なる技術開発は必要であろう．PubChem RDF のホームページ[14]には，“How can PubChemRDF help your research?”として

・ダウンロード可能なデータベースの利便性
・目的の RDF 形式のデータファイルをダウンロードしてトリプルストアにインポートすれば，SPARQL で必要情報に行き着くアクセシビリティ
・ツールを組み合わせれば，非構造データベースへのアクセスとクエリが可能になる拡張性

が挙げられており，データ発信側から見た RDF の有用性が端的に表現されている．一方でデータの受け手側からすると，手軽な技術が次々と現れ利用できる状況に比べて，オープンデータや先進的なセマンティック web 技術に追随するための労力は少なくないであろう．RDF，トリプルストア，SPARQL の三技術を乗り越えて初めて得られる情報のアクセシビリティを如何にサポートするかは今後の課題である．

5. まとめ

データベースの統合を目指し，NIMS 高分子データベース PoLyInfo の部分的な RDF 化を試行し，それによって明らかになった技術的課題を議論した．重合過程に RDF を適用することにより，低分子を主に扱う日化辞 RDF 及び PubChem RDF とシームレスな結合が図れることをデモンストレーションした．その一方で，本格的に PoLyInfo の RDF 化を行う前に，RDB テーブルの整理と，各テーブルの列名の編纂，各データの正規化が必要なことが明らかになった．

長年続けてきた PoLyInfo のサービス向上の過程で，データベース構造の複雑化は不可避であった．オープンサイエンス，データ駆動による材料開発が盛んになる中で，PoLyInfo のデータの構造化と機械可読性の向上は喫緊の課題と言え，物質・材料研究機構　材料データプラットフォームセンター（DPFC, Materials Data Platform Center）では，2019 年から本格的に再構築事業を開始している．

6. 謝辞

PoLyInfo の RDF 化に際して，国立研究開発法人 科学技術振興機構（JST）バイオサイエンスデータベースセンター（NBDC）の櫛田達矢研究員をはじめとする JST の方々，大学共同利用機関法人 情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター（DBCLS）五斗進教授，川島秀一特任助教をはじめとする DBCLS の皆様には，多くのアドバイスを頂いた．現段階でその全てを RDF 開発に反映出来てはいないが，順次改善したい．

本研究の一部は，内閣府「戦略的イノベーション創造プログラム（SIP）」の対象課題「スマートバイオ産業・農業基盤技術　革新的バイオ素材・高機能品等の開発を加速するインフォマティクス基盤技術の開発」において実施された．

文　献

[1] https://www.w3.org/ (accessed 2019-05-01).
[2] https://www.w3.org/TR/rdf-syntax-grammar/ (accessed 2019-05-01).
[3] S. Kawashima, T. Katayama, H. Hatanaka, T. Kushida, and T. Takagi, NBDC RDF portal: a comprehensive repository for semantic data in life sciences, Database, Vol. 2018, Article ID bay123, December 2018.
[4] https://dbarchive.biosciencedbc.jp/jp/nikkaji/download.html (accessed 2019-05-01).
[5] 木村考宏，櫛田達矢“日化辞 RDF データの公開と化合物情報の統合”情報管理, 58, 3, pp.203-212, June 2015.
[6] https://pubchem.ncbi.nlm.nih.gov/ (accessed 2019-05-01).
[7] https://polymer.nims.go.jp/ (accessed 2019-05-01).
[8] http://rdf.chemspider.com/rdf.html (accessed 2019-05-01).
[9] https://jena.apache.org/index.html (accessed 2019-05-01).
[10] https://github.com/semsol/arc2/wiki (accessed 2019-05-01).
[11] https://www.kanzaki.com/works/2009/pub/graph-draw (accessed 2019-05-01).
[12] https://www.w3.org/TR/rdf-sparql-query/ (accessed 2019-05-01).
[13] http://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline (accessed 2019-05-01).
[14] https://pubchemdocs.ncbi.nlm.nih.gov/rdf (accessed 2019-05-01).