

PoLyInfo RDF による MI のための知識基盤の構築

Knowledge base construction for MI using PoLyInfo RDF

物材機構 坂本 浩一, °石井 真史

NIMS, Koichi Sakamoto, °Masashi Ishii

E-mail: ISHII.Masashi@nims.go.jp

【序】 マテリアルズ・インフォマティクス (Materials Informatics, MI) が近年盛んになっているが、数理的なデータ駆動が先行する中で、データの収集・構造化・統合については多くの課題が残っている。我々は、Linked data の考えを取り入れ、Resource Description Framework (RDF)によるデータ統合を試みている[1]。RDF による分野を越えたデータの統合はバイオ関係で 10 年以上前から進んでおり[2]、材料分野は大きく後れを取っている。

今回、我々は NIMS の高分子データベース PoLyInfo を RDF 化し、“PoLyInfo RDF”を作成したので、その概要と新しい MI への展開を紹介する。

【作成】 RDF は、全ての情報をトリプルと呼ばれる主語・述語・目的語の三語の組み合わせで表現する。この考えはデータテーブルが意味する「高分子 A の特性 B の値は C である」と整合しているために、従来の Relational Database (RDB)を機械的に変換することができる。実際、我々は RDB である PoLyInfo から PoLyInfo RDF を D2RQ [3]にて直接作成した。ここでは D2RQ Mapper [4]により RDB から RDF へのマッピングファイルを作成しているが、そのデータ構造に Protégé [5]によって作成した PoLyInfo オントロジーが反映されている。

【構成と特徴】 作成した PoLyInfo RDF はトータルで 13,417,272 トリプルを有する。大きな構成としては、図 1(a)に示すように、サンプル情報「Identification (SID)」を出発点にして、合成「Fabrication」と形成「Formation」のトリプルが続いた後、最後に特性「Property」がリンクする直列構造を持っている。高分子種を表す「Master (PID)」は、「Identification (SID)」に並列で繋がっている。実際に使ってみると高分子種と特性の距離が長く、MI の目的では使い

勝手が良くないように思える。現在は、同図(b)のような、高分子種の下位概念に新たなサンプル ID「Identification (SIDP)」を配置し、そこから構成物「Components」、Fabrication 及び Formation、property を放射状に配置する構造への改訂を検討している。

【展開】 Linked data の考え方は、連携先のデータ空間が知識ベース (Knowledge Baes, KB) を形成しているならば、知識の統合となる。知識の統合による MI は、原理原則を組み合わせた網羅的材料探査を意味する。その範囲を超えた未踏領域に対しては、類推や仮説推論に基づく開発指針を、PoLyInfo の蓄積データから確度を高めて提供する推論 MI 基盤の実装を進めている。

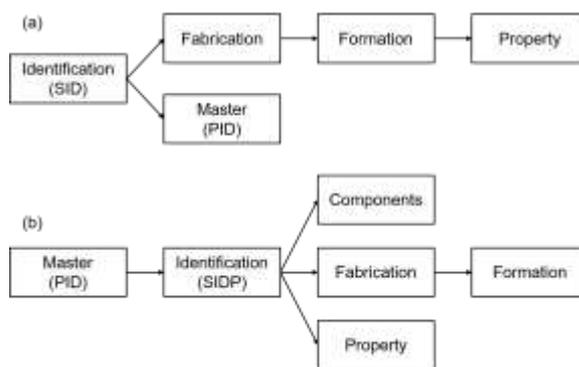


図 1 (a) PoLyInfo RDF の構成と(b)改訂方針

【参考文献】

- [1] M. Ishii et al., Proc. ISWC (2019), http://ws.nju.edu.cn/conf/iswc2019_ceur_auth_or_check/paper18.pdf
- [2] 解説としては例えば G. Fu et al. J Cheminform (2015) 7:34
- [3] <http://d2rq.org/>
- [4] <http://d2rq.dbcls.jp/>
- [5] <https://protege.stanford.edu/>