マテリアルインフォマティクスにおけるFAIR「知識」原理の実現

Realizing FAIR “Knowledge” Principles in Materials Informatics

〇物材機構，木野日織

〇NIMS, Hiori Kino

E-mail: kino.hiori@nims.go.jp

材料・プロセス情報学は学際的な分野であり、共同作業を行う際、データベースを扱うデータサイエンティスト、機械学習手法に取り組む科学者、理論、実験物質・材料科学者などが関わるため、互いの研究を理解することが難しい場合がよくあります。また、例えば、二年で卒業する大学院学生の研究を開始するにあたり迅速な知識の伝達が必要です。これらの問題に対するアプローチの一つとして、分類に基づいた関係記述法を用いて最低限の理解を図るオントロジーの利用があります。

過去に保存されたデータは、そのままでは数値解析に使用できない場合が多くあります。例えば、人がブラウザで閲覧するために最適化されている NIMS Atomwork データベースは数値データ解析には適していません。数値データに変換するにはデータクレンジングが必要ですが、その目的は場合ごとに異なるため、それぞれの目的に応じた変換を行う必要があります。また、FAIR データ原理を満たすためにはそれらの変換過程を明示する必要があります。また、多くの変換を経ると関係性が分かりにくくなるため、ショートカットを記載する必要もあります。更に重要なのは、それらの利用・検索方法が容易に理解可能なことです。

このデータベースを例に、Protégé とその推論エンジンを用いてオントロジーに基づいたデータ構造を定義し、neo4j グラフデータベースに変換する仕組みを開発しました。このデータベースには、以下の 3 種類の関係が記述されます。1 つ目は、科学的知識に基づく階層構造です。例えば、has_crystal_structure は物質が結晶構造を持つことを示します。2 つ目は、個別のデータ変換、いわゆる実績グラフです。例えば、validated_to は記録されたデータから、数値解析可能なデータに変換したデータクレンジングの過程を示します。3 つ目は、検索を容易にするための関係です。例えば、has_descriptor は物質が何かの特徴量を持つことを示します。

この手法により、データの追跡可能性、明確性、理解しやすさが大幅に向上し、データだけでなく知識に対しても FAIR データ原則を達成することができることが期待されます。[1]

[1] Hiori Kino, et al. in preparation.