LODデータの構造化に着目したスキーマ設計支援手法

林坂 勇希

近年、オープンデータの普及を目的として、Web上でデータを公開して共有するための技術としてLinked Open Data（LOD）が注目されている。LODとして公開されているデータ同士を相互に結びつけることで機械処理を容易にし、データの横断的な利活用が可能となる。しかし、その利点を活かすためには、最初のスキーマ定義において、データを意味的に繋げ、機械可読のための形式に整備する「適切な構造化」が必要である。LODが注目される一方でその活用の幅が小さい原因は、上記で示した適切な構造化がLOD化のための知識や経験が少ない第三者にとって難しいことにあると考える。

そこで、本研究では、LODのスキーマ設計支援を目的として、その構造化段階に注目した支援手法を提案した。具体的な手法としては、非LODフォーマットのうち公開が比較的容易で、かつ機械可読に適したフォーマットであるExcel形式およびCSV形式を対象として、本研究で定めた制約に基づいて整形したデータを入力すると最大6種類の構造パターンが出力されるようなシステムの開発に取り組んだ。この提案手法により、第三者がLOD化のための構造化を行う際に、複数の提案構造から構造の発想をサポートしてくれるという点で役に立つことが期待できる。さらに、本研究によってオープンデータをLODで公開する過程の一部で手間が軽減されることで、LODとしての公開が加速し、オープンデータの活用可能性の拡大に貢献できると考える。

また、本研究では、システムの実装を行った後に検証を行うというフェーズを二度繰り返すことによって、より多くのケースに対応した、またより多くの構造パターンを提示できるようなシステムを目指して、実装と評価を並行しながら行った。

検証用のデータ数は、1回目はExcel形式を20件、CSV形式を10件、2回目はExcel形式を40件、CSV形式を20件に増やしてシステムの検証を行った。最終的に提案システムは、2回目の検証用データセットのうち、Excel形式の85%、CSV形式の100%に対して出力構造を提示することができた。また、Excel形式の42.5%で6種類の構造パターンを、CSV形式の55%で2種類の構造パターンを出力できることを示した。提案システムは、複数行ヘッダーを持つデータへの対応と、テーブルを転置した構造を実現したが、構造出力不可の原因には縦方向に入れ子関係が存在するデータ、他の新規パターンの構造案にはテーブル正規化の考慮、単位情報の反映など、様々な課題が残った。

今後の課題として、得られた諸課題の解決に加えて、LODに変換することを想定した、第三者がExcelやCSVのデータを公開する際の最適な統一ルールの検討と、ユーザがより容易に適切な構造の選択をできるようにするための、入力データに応じた構造のランキング・推薦が挙げられる。

（指導教員　高久 雅生）