# LOD データセット生成の自動化のための宣言的記述による半構造化データの抽出とその統合手法

# Extracting and Unifying Semi-Structured Data Based on the Declarative Description for Automated Generation of LOD Datasets

学籍番号：201721682
氏名：豊田　将平
Toyota Shohei

構造化されたデータを Web 上で公開および共有するための仕組みとして Linked Open Data(LOD)がある．LOD は Resource Description Framework（RDF）と呼ばれる標準に従ってデータを記述し，すべてのリソースを International Resource Identifier（IRI）を用いて識別する．これよって Web 上でデータ同士を参照可能にできるため，データセットの公開方法の一つとして，LOD データセットの利用が望まれる．

LOD データセットは人手によるデータ入力で作成されることもあるが，大規模な LOD データセットの多くは既存のデータセットを RDF データに変換することで生成される．また，Web 上では Web ページに記述された情報が大半を占めるが，このような情報から半構造化データを抽出することで LOD データセットを生成する試みも存在し，代表的な事例には DBpedia がある．DBpedia は Wikipedia という特定の Web サイトを対象としているが，Web 上には様々な情報が多数の Web サイトに分散して記述されている．そこで本研究では，DBpedia のアプローチを応用し，複数の Web サイトに記述された情報を抽出および統合することでデータセットを生成できると考えた．このとき，同一の実体に関する記述が複数の Web サイトに含まれる可能性を考慮する必要があり，抽出したデータの同一性の解決という課題がある．また，Web サイトのコンテンツは更新される可能性があり，更新を反映するためには LOD データセット生成を反復して実行しなければならない．LOD データセット生成の一連の手続きは，可能な限り自動化することが求められる．

本研究では，Web ページからの半構造化データの抽出，それらの同一性の判定，RDF データへの変換，といった一連の LOD データセット生成の処理内容を宣言的に記述する規則を提案した．また，それらに基づいて自動的な LOD データセットの生成を実現する手法を提案した．提案手法を実装したシステムを用いて実際の Web サイトを対象に実験を行い，規則に基づいて各 Web サイトの記述を統合した LOD データセットが生成されたことを検証した．また，この実験の考察から，データクレンジングや，規則作成および修正の簡易化が今後の課題として挙げられる．

研究指導教員：杉本　重雄
副研究指導教員：永森　光晴