

既存メタデータスキーマを用いた  
メタデータインスタンスからのスキーマ推定手法  
Metadata Schema Generation from Instances  
and Schema Published on the Web

学籍番号：201221590

氏名：田中 圭

Kei TANAKA

World Wide Web 上でデータを共有する Linked Open Data (以下 LOD) という試みが注目され、2009 年以降、政府が公開しているものをはじめとして地理情報やメディア情報など様々な分野のデータセットが急増している。しかし、LOD データセットから利用者が目的のデータを取得するためには、各 LOD データセットに蓄積されているメタデータの構造を理解し、問い合わせを行う必要がある。LOD データセットの構造や制約を定義するメタデータスキーマ (以下スキーマ) が公開されていれば、LOD 利用者が構造化問い合わせのためにメタデータの構造を理解する手間が解消される。しかし、スキーマを公開している LOD データセットは少ない。そのため、利用者は LOD データセットのデータ構造の理解が手間となり、目的のデータを取得することが困難になっている。そのため、LOD データセットを利用しやすくするために、スキーマが公開されているデータセットを充実させることが求められている。

本研究では LOD 利用者がデータセットのデータ構造を理解することを目的として、既存メタデータスキーマを用いたメタデータインスタンスからのスキーマ推定手法の提案と、手法に基づいたシステムを構築した。本手法は、メタデータインスタンスから使用されているクラスとインスタンスを抽出し、公開されているメタデータ語彙定義からクラスとインスタンスの語彙定義を取得した上でスキーマの構成要素を推定することでスキーマを推定する。このとき、推定するために不足している情報は既に公開されているスキーマを参考に補完する。

本システムを用いて LOD チャレンジ 2013 のデータセット部門のエントリー作品を対象にスキーマ推定を行い、各データセットのスキーマの定義項目をどれだけ推定できているかを評価した。その結果、メタデータが公開されているスキーマと異なるもの、データセットの公開方法が LOD の規則と異なるもの、公開されているデータ量が著しく大きいものではスキーマが推定できなかった。今後はデータセットが LOD の規則に沿っているかどうかの判断と、アルゴリズムの軽量化が求められる。

研究指導教員：杉本 重雄

副研究指導教員：永森 光晴