

別紙1 (博士論文の審査結果の要旨)

専攻名 システム創成科学専攻氏

名 吉賀 夏子

古典籍などの文化財の書誌情報や画像などが、大規模な収蔵館からオンラインで公開されつつある。当該分野の研究者にとっては、対象となる古典籍の収蔵場所を知る重要な手がかりである。これらの書誌の中には、本文以外の例えば序や跋の著者や書籍の発行書肆などの情報も注記欄に記載されており、対象となる書籍間や著者間の関係を知る重要な手がかりとなっている。

これらのオンライン公開された書誌情報はヒトが読むことで理解することを前提としており、コンピュータプログラムを用いて、大量に、かつ横断的に検索すること（機械可読）を想定していない。本研究は、既存の古典籍書誌情報を機械可読に自動的に変換する手法を提案し、実用的時間で現実の古典籍書籍を変換する支援をしようとするものである。

本論文は 6 章で構成されている。第 1 章では、ヒトが見て理解することを目的とした、現状の古典籍に対する書誌情報の課題と解決方向について検討している。大きな課題は二つあり、第一はデータの不十分な正規化、つまり値と型の不整合がある。また、研究者にとって書籍の来歴や他の書籍との関連性の手がかりとなる記述が注記欄に簡潔な自然文で記入されている。これらの情報を適切に保存するためのデータ構造が欠けていることが第二の課題である。

本論文では、機械可読なデータ公開の方法として Semantic Web 技術の活用を目標としている。そのための基本技術について、第 2 章で解説を加えている。

既存の書誌データでは、注記欄の自然文による記述の正規化されたデータへの変換が重要なポイントとなる。本研究では、自然文から単語を切り取る技術である形態素解析を基礎に、Semantic Web に適したデータへの変換を行う。第 3 章は、本論文の中心となる部分であり、既存の書誌データから Semantic Web 用の Linked Data と呼ばれる形式への自動変換手法を提案している。提案手法は、本学の所蔵する「市場直次郎コレクション」を対象に開発を進めた。

提案手法を実現するために、書誌中の用語を適切に分析するための準備が必要となる。この用語の中には、書誌中に出現する固有名詞のほか、古典籍の書誌で使われる専門用語なども含まれる。これらの用語の辞書を用意することで、形態素解析の精度向上が可能となる。また、Linked Data 化のためには、要素間の関係を機械可読な形で定義する必要もある。そのため、文化財を記述するためのデータ構造（オントロジー）を、国際的に標準化されている汎用オントロジーを参照しながら、対象に適した小規模なものを定義し、収集した単語との関連付けを行った。

上記の準備の下で、実質的に自動的処理により、既存の書誌情報から目的とする Linked Data 形式への変化を行うことができる。さらに、他の機械可読な Web サイトで定義されている用語との関連付けを URI (Uniform Resource Identifier) を通じて行うことができる。この URI による関連付けは、古典籍の専門家による確認が必要な部分が残る。特に、古典籍に現れる人名は、同姓同名のリストからの選択などが必要である。

第 4 章では、提案手法の精度の議論と性能に関する評価を行った。提案手法に含まれる要素毎に、精度向上への寄与を分析し、提案手法の有効性について

て確認した。また、提案手法を国文学研究資料館所蔵の「日本古典籍データセット」の一部へ適用し、その有効性を確認した。これらを通じて、現実的時間で、既存の書誌情報を Linked Data 形式へと変換できることを示した。

インターネットで公開された情報から有用な情報を機械的に得るための研究はいくつかある。また、図書館や美術館などの収蔵情報の電子的アーカイブ作成、相互参照などについても研究が展開されている。第 5 章では、関連する研究のなかで、本研究を位置づけ、今後の展開の方向を検討している。最後に第 6 章で、本研究の成果をまとめている。なお、本論文の内容は、査読付き学術論文 1 編として公表されている。

以上のように、本研究はインターネットの普及の中で文化財情報の活用を推進するうえで障害となっている書誌情報の機械可読化に大きく貢献する提案を行っているものである。このような博士論文の内容に基づき、平成 30 年 1 月 25 日に実施した学位公聴会において、手法の汎用性、結果の精度や効率など、様々な質問がなされ、いずれも著者により適切な回答・説明がなされた。以上の審査結果により、本論文は博士（学術）の学位を授与するに値すると判断し、審査委員全員一致で合格と判定した。