別紙第１号様式　　No.1

博　士　論　文　の　要　旨

専攻名　システム創成科学専攻

氏　名（本籍）　　吉賀　夏子（佐賀県）　印

博士論文題名

非構造化記述を含む文化財書誌に対する Linked Data 化手法の開発

–古典籍書誌データへの適用とその評価–

要旨

Web 上には、文化財に関する膨大な数の書誌情報および画像を含むデジタルアーカイブが国内外の文化財関連組織により公開されている。これらアーカイブの書誌の記述には、対象文化財そのものに関する情報のみでなく、文化財を構成する部品の詳細、書籍の成立、他の書籍および人物との関係などを示す、文化財の研究者にとって重要な周辺情報が含まれている。

これらの書誌データのほとんどは、人が読んで理解することを目的としており、簡素化された自然言語で記載されている。加えて、各書誌項目に対応する値の記述に凡例記号が付記されている場合や、あらかじめ設定された基本事項に収められない詳細な事項が注記に自由文で記録されている場合が散見される。また、人名あるいは専門用語の名寄せが行われていない場合が多い。そのため、プログラムを使って他のオンライン公開された書誌データを含めて横断的に検索することは困難である。

プログラムを使って横断的に検索できるデータを、機械可読と呼ぶ。既存の書誌データを機械可読にするためには、書籍の部品や来歴を記述するデータ構造を定義するとともに、コレクションおよび対象分野で使用される語彙辞書を整備し、横断検索にふさわしいデータ形式で記述する必要がある。現状では、情報資源をグラフ形式で記述可能な RDF（Resource Description Framework）を採用する Linked Data に文化財書誌を変換することを、図書館、博物館など多くの書誌提供機関で推奨しており、実際に、Linked Data 化を試みている。

書誌データの Linked Data 化とは、従来の書誌データから書誌の読み解きに必要な人名、時、地名、専門用語などのキーワードである固有表現を抽出し、固有表現に対し Web で一意の事物を指す URI（Uniform Resource Identifier）を紐付けた正規化データに変換するとともに、RDF グラフ構造にすることである。ここでの正規化とは、各属性に適切な型を設定し、注釈などの補足的な記述を値から分離することを示す。また、書誌属性とその値も全て URI で表現することで、属性同士、固有表現と属性の関係を知識構造のデータとして利用可能とする。さらに、URI で定義された属性や固有表現には同義の外部組織の URI を紐付ける。したがって、Linked Data 化した書誌データでは、元書誌の持つ情報に加え、書誌利用者にとって有用な外部情報へのアクセスが可能となる。こうして文化財書誌が Web 空間で共有および再利用可能なデータとなるため、書誌利用者は、Web 上の関連情報を活用し、書誌データを横断的かつ機械的に分析することが可能となる。

しかし、こうした作業は、知識科学、プログラミング知識を必要とする上に、対象領域の文化財そのものの知識も必要とされる。また、大量の書誌に対して Linked Data 化を実行する具体的道筋も確立されていない。そのため、Linked Data として公開されている書誌情報には、正規化が不完全なものが散見される。したがって、文化財書誌の Linked Data 化を促進するには、実在する書誌データから、文化財の理解に必要な語の抽出および定義を容易に行えるような技術開発が必要である。

本論文では、文化財書誌の具体例として、江戸時代以前の書籍である古典籍の書誌データを採用する。そして、実在する古典籍書誌から Linked Data に変換するプロセスを、古典籍の読み解きに必要な固有表現の抽出、非構造化データから構造化データへの

博　士　論　文　の　要　旨

専攻名　システム創成科学専攻
氏　名　吉賀　夏子

変換、固有表現と外部リンクの紐付けに分け、これら一連のサブプロセスを可能な限り自動化する手法について提案する。

従来、Linked Data 化を目的とした固有表現抽出では、形態素解析ツールや文脈パターン認識を用いた一般的な手法が用いられてきた。本論文では、それらの手法を、古典籍で使用される単語や書誌学用語に対して拡張する。そのために、あらかじめ、目録記述規則および Web から研究対象周辺の予備知識を収集する。これらの予備知識は、固有表現抽出プロセスおよび古典籍を構成する概念に関する機械可読な構造的知識（オントロジー）の自動構築に用いる。また、Web 上で、固有表現と等価の外部リンクを自動探索し、データベースで管理する。対象文化財分野の専門家は、各サブプロセスで生成したデータの内容確認に主に注力することができる。

また、提案手法の有効性および汎用性を二つの規模の異なる古典籍コレクション書誌データで比較検証した。コレクションへの対応の差は、書誌項目および注記の記述規則に応じたスクリプトの変更と、収集した予備知識の追加である。その結果、語彙情報の正確性は、古典籍の専門家の協力によって担保されるが、提案手法による Linked Data 化およびその生成にかかるコストは非常に小さく、古典籍書誌を現実的コストで Linked Data 化できることを示した。

さらに、提案手法の実行により作成されたコレクションの予備知識、書誌オントロジーおよび Linked Data は機械可読データのため、Web での共有および再利用が可能である。すなわち、提案手法は、軽微な追加作業を通じ、他の分野、時代、文化財種類の類似するコレクションに対しても応用可能である汎用性の高い手法であることを示した。