唐詩情報の Linked Data 化の試み

Development of Linked Data Application for Tang poems.

叢　艷[1]，　　高久雅生[2]
Yan CONG[1]，　　Masao TAKAKU[2]
[1]筑波大学大学院図書館情報メディア研究科，
[2]筑波大学図書館情報メディア系
[1] Graduate School of Library, Information and Media Studies, University of Tsukuba,
[2] Faculty of Library, Information and Media Science, University of Tsukuba

あらまし:近年、政府、科学などのデータの公開共有するための研究が盛んだが、文化資源について様々な情報をウェブ上に公開共有の試みはさほど多くない。本研究では唐詩を対象として、教科書の関連を Linked Data 化を目指す。そのため、データの構造や公開共有の枠組みを議論し、教科書や出版社などの関連するリソースを Linked Data 化、その公開と利活用にチャレンジする。

キーワード：唐詩、教科書、Linked Data、Schema.org Vocabulary、BIBFRAME Model

1. はじめに

近年、情報技術が発達するとともに、膨大な情報も段々と増加している。これに対し、大量の情報の中からデータを精確に抽出できる方法が注目を集めている。Tim Berners-Leeは Linked Data[1]と呼ばれる(以下LDと呼ぶ)ウェブ上に様々な情報をリンクで関連できる方法を提唱した。政府、銀行などの機関では各自のデータを、誰でも利用できるような形でデータを公開する活動が活発化している。2016年2月中国文化部指定公共デジタル文化研究拠点である上海図書館は、重点研究プロジェクト「系譜ナレッジベース」を始めた。このデータベースはLD技術を利用したもので、自機関データを公開共有するため、利用者は資料を閲覧するだけではなく、LDの関連を通して、様々な情報を自由にシェアでき、開発も可能としている。同年5月、同館の系譜などのデジタル所蔵資料をネットで公開して、上海図書館公開サイト[2]上で公開した。このサイトでは、系譜2,500点、オンラインデジタル展示品約20点などが閲覧できる。同プロジェクトはオープンデータ化の新たな試みであり、オープン化と再利用によるデータの価値向上を目指すもので、誰でも使える情報が提供されている。

唐詩は中国古典文化資源の一部として、今までは千年以上の歴史がある。それは中国の文化遺産として、中国古典文学研究に欠かせない基本的な文献である。

ここから、本研究では唐詩を対象として、日本の中学校と高校で学習する唐詩、それを掲載する教科書と出版社の関連性をLinked Data化することを試みる。

2. 対象と方法

2.1 現行教科書について

本研究では、唐詩を対象として日本の中学校と高校の古典編の教科書に掲載されている唐詩の利用、唐詩を含む教科書と出版社の関係付けを考えている。唐詩作品の使用状況を知るために筑波大学附属中央図書館に所蔵されている中学校と高校の古典編の現行教科書を調べてみたところ、教科書81冊に唐詩は延べ371首、異なり68首があった。教科書を出している出版社は全部で10社あった。所蔵されている教科書に掲載されている371首の唐詩の中か

ら頻繁に使われている唐詩6首を選択した。それらについての作者、教科書、出版社の情報を一緒に記述した。教科書から選択した６首の具体例を表1に示す。

表1:唐詩作品6首

唐　詩（ID）	作　者	教科書/出版社
鹿柴 (tangpoem:1)	王維	新編国語総合言葉の世界へ/教育出版，新編古典B/東京書籍，古典B漢文編/数研出版,ほか14冊
早発白帝城 (tangpoem:2)	李白	中学生の国語学びを広げる/三省堂,物語 小説 評論 漢詩思想 史伝/右文書院,ほか13冊
登高 (tangpoem:3)	杜甫	古典B/桐原書店，古典B漢文編/筑摩書房,ほか10冊
望廬山瀑布 (tangpoem:4)	李白	精選古典B漢文編/明治書院,ほか8冊
除夜寄弟妹 (tangpoem:5)	白居易	高等学校古典B/第一学習社,ほか2冊
秋浦歌 (tangpoem:6)	李白	新編古典B/大修館書店,ほか2冊

2.2 Linked Data

Linked Dataとはウェブ上のURIに基づき、リソース同士を関連付け、その関連性をグラフで表現し、検索できるようにするという方法である。本研究では、唐詩を創作作品として、教科書と出版社の三者の関連関係の枠組みを考える。唐詩の作者、詩体などの基本的なデータ、教科書の書誌情報と出版社についての関連情報をリンクし、公開共有することを検討する。唐詩情報を記述するための枠組みであるLD化のモデルはトリプル(Triple)で組み合わせられる。それは主語(Subject)、述語(Predicate)と目的語(Object)の3つの要素がある。本研究を例として具体的に説明すると、唐詩作品を主語として、述語としてプロパティを入れ、目的語としてタイトル文字列や教科書リソースなどを採用して、有向グラフで結んで表現する。この形式に変換することで、データの利用や再利用による作品の検索も可能となる。それに基づいて、唐詩作品の検索精度も向上することが期待できる。

2.3 唐詩のLinked Data化

本研究では、唐詩作品をリソースとして、LD Vocabulary には既存のものを採用し、Dublin Core、Schema.org、vCard、RDF Schemaを用いる。全体を通じて、唐詩作品のタイトルはdc:titleを利用し、教科書のISBNはdc:identifierを採用し、唐詩詩体や出版社などの解説はdc:descriptionで扱う。出版社の住所はPREFIX　vcard:<http://www.w3.org/2006/vcard/ns#>で空白ノードとして扱う。更に、Schema.orgのVocabularyを利用して、唐詩作品と教科書の関連として、詳しい属性プロパティを与える。Schema.orgのVocabularyから、各リソースの種別として、作者はschema:Person、唐詩作品と教科書を示すschema:CreativeWorkを用いる。作者の生年月日はschema:birthDate/deathDateを採用し、作者の性別はschema:genderとする。唐詩作品と教科書の関係はschema:isPartOfとして定義する。また、教科書と出版社の関係は出版社schema:publisher、教科書の編者らをschema:editor、教科書の出版日schema:datePublishedとして表現する。外部へリンクするURIとして具体的なウィキペディアへの記事にPREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>seeAlsoを

用いる。

次に、2.1節の現行教科書の調査で得られた唐詩6首の作品は数字1-6からなるIDを振り、ベースURIはtangpoem:<http://example.com/tangpoem/>とし、tangpoem:1、tangpoem:2…などとする。唐詩の詩体は全二種類:五言律詩と七言律詩があるため、ベースURIは<http://example.com/tangpoem/style/>とし、tangpoem:style/5、style/7を採用する。唐詩の作者は各自の氏名を識別し、ベースURIは<http://example.com/tangpoem/author/氏名>を用いる。教科書も唐詩作品のように数字1-10のIDを振り、ベースURIは<http://example.com/tangpoem/textbook/>とし、tangpoem:textbook/1、textbook/2などをする。出版社は正式名称を用いる。ベースURIは<http://example.com/tangpoem/publiser/>とし、tangpoem:publisher/教育出版などを使用する。唐詩作品1はtextbook/1、3と8に、唐詩作品2はtextbook/2、9に、唐詩作品3はtextbook/4と7、唐詩作品4はtextbook/5、唐詩作品5はtextbook/6、唐詩作品6はtextbook/10と関係付けられる。教科書とそれぞれの出版社を個別にリンクして、唐詩作品の関係モデルを構築できた。基本的なモデルを設計し、各関係を有向リンクで結んで、URIで構築されるモデルがグラフとして構成され、詳しい情報をリテラルやラベルで表現し、LD化できたと考える。

3. 結果と考察

3.1 結果

2.3節のLinked Data化に基づいて唐詩作品情報のグラフを作成し、図1に示す。図1は唐詩作品1のもつ関係を可視化したものである。図1において、円形のノードはリソースを表し、四角でリテラルを示し、ラベルの無い円で住所などの空白ノードを示す。唐詩のPREFIXは図中ではtp:として表現している。有向矢印にはそのプロパティを示した。

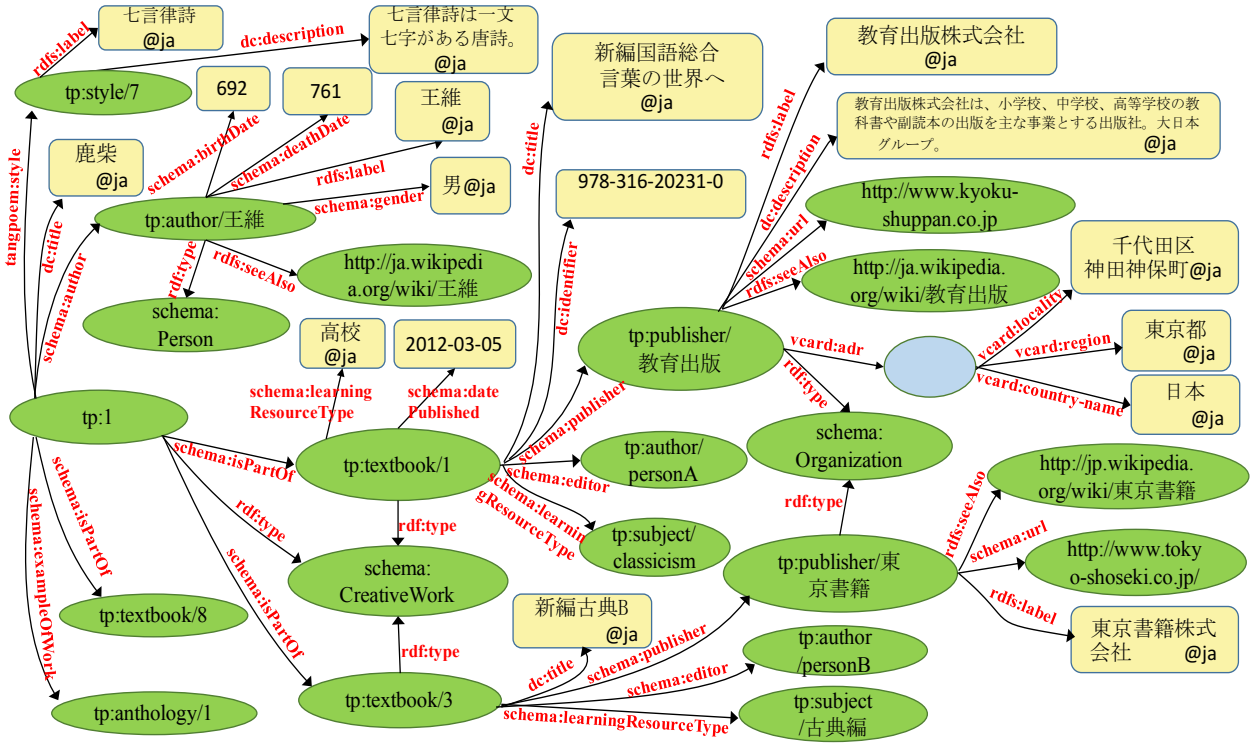

図1:LD化した唐詩情報(一部抜粋)

3.2 考察

本研究では、日本の中学校と高校で学ぶ唐詩作品、その掲載教科書とその出版社についての関係をLD化した。唐詩作品は6首を選択して、教科書10冊とそれらに対応する出版社10社を関連付けた。唐詩作品の場合、同じ唐詩作品は教科書一冊以上に含まれる場合も表現できる。それらに基づいて、唐詩作品を含む教科書を通して、出版社は1社以上を関連付けできる。また、教科書の場合、複数の唐詩作品を利用する場合はそれらの唐詩が繋がって、様々なデータと関連付けられると考えている。例えば、教科書から検索したり、作者などの基本的なデータから検索したりする。また、出版社については具体的な情報も関連付けて検索できる。

4.おわりに

中国で唐詩は古典文化資源の一部とし、中国古典文学についての研究に欠かせない文献として、非常に重要な役割を果たしている。そのため、本研究では唐詩を対象として、日本の中学校と高校で学んでいる唐詩、唐詩の掲載教科書とその出版社の関連をLinked Data化する研究を行った。独自ボキャブラリの他、Schema.org VocabularyやDublin Coreを採用した。

今後の予定として、唐詩作品を増やすことと考えている。例えば中国の全唐詩庫[4]にあるような多くの唐詩作品を関連して、Linked Data化できるようにしたい。それらに基づいて、検索や閲覧のアプリケーションを構築したい。また、更新中のBIBFRAME Model[5]による唐詩情報のモデル化も検討する。

参照文献：

[1]トム·ヒース, クリスチャン·バイツァー著. 武田英明, 大向一輝, 加藤文彦, 嘉村哲郎, 亀田尭宙, 小出誠二, 深見嘉明, 松村冬子, 南佳孝訳, Linked Data: Web をグローバルなデータ空間にする仕組み, 近代科学社, 2013, 139p. ISBN 978-4-7649-0427-9.
[2]上海図書館. “上海デジタル公開サイト”. http://wrd2016.library.sh.cn,(accessed 2016-05-30).
[3]“Schema.org Vocabulary”, http://schema.org, (accessed 2016-05-30).
[4]鄭州大学管理中心. 全唐詩庫.（中国語）, http://www16.zzu.edu.cn/qts/, (accessed 2016-05-30).
[5] “Overview of the BIBFRAME 2.0 Model”. http://www.loc.gov/bibframe/docs/bibframe2-model.html, 2016-04-21, (accessed 2016-05-30).