唐詩の構造化に関する研究
ー Linked Data と TEI マークアップを用いて ー

筑波大学
図書館情報メディア研究科
2019年3月
叢 艶

目 次

表 目 次

図 目 次

第1章　はじめに

1.1　研究背景

唐詩は中国古典文化資源の一部として、現在までは千年以上の歴史がある。唐詩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り、日本の文学にも大きな影響を与えた。日本では唐詩と宋詞をあわせて、漢詩と呼び、日本の中学校と高等学校の古典と国語の教科書に掲載され、国内の全ての生徒が学ぶものとなっている。

日本の中学校と高等学校における国語と古典のシラバスにおいて、唐詩の学習は、唐詩に関わる文体、詩体、作者、唐詩作品のコンテンツの理解や、唐詩の時代背景、地理情報、歴史情報などの情報を理解する学習目標と学習活動例がある [1]。これらの学習目標と学習活動例は、教育学習ニーズによって、生徒らは唐詩に関わる基本的な文体や、詩体などの基本的な知識の理解だけではなく、作者 (詩人) の人となりを知ること、唐詩の本文フルテキストに関わる訓点情報、ルビ情報の利用、唐詩における全文を描いている情景や作者の心情を捉えるような学習目標も達成する必要がある。そのためには、唐詩作品における多様な関連情報を簡単に扱え、数多くの関連情報を自由に利用することが必要である。

唐詩作品は中国では常に学ぶ作品であり、唐詩作品に関わる情報を利用できるデータベースはあるものの、データベースがばらばらに各所で作られ、データベース内容は、公開させられず、研究者や、教員、学生などが使えないという問題があり、学習のための利便性も低いと考える。具体的には、北京大学数拠分析研究中心による「全唐詩分析系統」オンライン DB [2] が唐詩作品の全文検索を提供し、研究および教育の活用が期待されている。千田 [3] は唐詩情報に関わるデータベースは中国の書籍店による取り扱いで入手が難しく、高額で販売され、検索インターフェイスも扱いにくいという問題を提言した。したがって、このような問題を解決するに対して、唐詩情報を公開共有でき、自由に扱える学習環境が望ましい。

また、漢文教育の課題について、荒井 [4] は日本の中学校と高等学校の教科書に、

含まれる唐詩作品において、同一作品であっても訓読が違うものの、詩の意味の捉え方に大きな影響を与えて、文法が一致しない点を提言した [4]。中川 [5] は同じ唐詩作品でも、文体や、文字などの差異が存在し、文法によって訳す本文フルテキストの内容も、矛盾を生じる可能性がある [5] と述べた。それらの問題点の解決のためには、漢詩に関わる学習内容を一致させることが望ましい [5] と述べた。そのため、教育ニーズに応じて、唐詩情報は、基本的な情報の習得だけではなく、本文フルテキストによって、違う訓読情報や、ルビ情報をつけていても、唐詩作品に関わる情報を統一的に扱える学習環境を提供したいと考える。

このような漢詩情報の共有や学習のための課題を解決するためには、漢文教育における唐詩作品の構造化が望ましいと考える。唐詩作品の構造化については、唐詩における多様なニーズに応じて、多様な関連情報を簡潔に扱え、自由に利用できることが必要である。まず、唐詩作品に関わる基本的なデータを扱い、関係がある一連リソースを全て関連できるように構築したいと考える。そのため、本研究では、LOD 技術を用いて、唐詩作品に関わるメタデータを構造化し、整備する。

また、本文フルテキストにおいても、同一作品であっても訓読などの情報に差異が存在するため、それらを統一的に扱い、唐詩作品に関わるデジタル化を推進するため、本文フルテキストの標準化が必要となる。さらに、構造化の方式は、多くの研究者やシステムと共有できるよう標準化も求められる。標準化されたマークアップでーたは長期の保存や、利用に多く重要な役割を果たすと考える。

本研究では、漢詩の一部である唐詩作品を研究対象として用いて、唐詩の構造化に関する研究を進める。唐詩の構造化に対して、膨大な情報の中からデータを的確に抽出できるとして注目を集めている LOD 技術 [6] [7] を用いて、唐詩のメタデータと本文フルテキストを構造化し、標準的に利用できるようにしたいと考える。これらに基づいて、唐詩情報を公開共有でき、自由に扱える学習環境が入手でき、唐詩に関わる学習内容を一致させることが可能になると考える。そのため、ここでは、標準的な TEI マークアップ [8] を用いて、唐詩作品に関わる本文フルテキストの整備を行いたい。

1.2 研究目的

本研究では文化資源に注目し、唐詩作品を研究対象として、唐詩の構造化を目指す。多様な関連情報を簡単に扱え、自由に利用でき、教育学習のニーズに応え

て、異なる表現を統一的に扱う学習環境を提供したいと考える。

本研究の研究手法として、唐詩の Linked Open Data (LOD) 化 [6] [7] と TEI マークアップ [8] を用いる。具体的には以下の通りになる。

(1) 日本の中学校と高等学校で学習する唐詩作品、唐詩と関連がある基本的なデータと関連付けて、それらと掲載される教科書の関連関係や外部リソースとも繋げて、LOD 化することを試みる。LOD 化の構築にあたって、BIBFRAME Model [9] に基づいて、新たなメタデータモデル化も行う。

(2) 唐詩の本文フルテキストのうち、訓読文、書き下し文を中心に、訓点やルビなどの要素を対象として、TEI マークアップ [8] を用いて、標準的にマークアップする手法を提案する。

第2章　関連研究

2.1　漢詩のデータベース

中国における古典データベースは、大きな発展がある。例えば、台湾の中央研究院の「漢籍電子文献」翰典全文検索系統 [10] は、漢籍全文資料庫、古漢語語料庫、台湾文献叢刊、近代史全文資料庫、清代経世文編、中華民国史事日誌など 15 か所以上の資料庫を組み合わせて、編集したものである。台湾の中央研究院の「漢籍電子文献」翰典全文検索系統 [11] は、この資料庫に関わる編集した一部リソースを台湾の小学校から中学校、高等学校および大学教育の学生らが無料で利用し、教育学習ニーズにおけるデータベースの有効化を高めるようにする利用している。

二階堂 [12] は 1996 年からそれらに関わる「史記」、「漢書」、「三国志」から「清史稿」に至る約 2 千万字に登る中国歴史書文献をインターネットに公開してから、ネットにアクセスしているものが誰でも利用でき、全ての用例を検索し、表示できた [12] という例を挙げた。また一方で、出版社がこれらのデータを CD − ROM に収録して、販売していることもあり、多くの人が利用できるようになった [12]。

「中国大百科全書」[13] は中国大百科全書出版社の編集発行物であり、百科全書全 74 巻および CD − ROM4 枚の形で編集したものである。中国や、日本の書店でも 2,500 円くらいで販売されており、利用できる。

古典籍データの利用では、台湾師範大学の陳郁夫は「寒泉」古典文献全文検索資料庫 [14] を作成し、中国における代表的な古典籍を含まれる全唐詩や、漢書などの有名な中国の古典を中心としてデータを公開した。「寒泉」古典文献全文検索資料庫による作成した公開サイト内のデータ [14] も続々と増えており、数多くの中国古典籍が使用できる。

北京大学数拠分析研究中心による「全唐詩分析系統」オンライン DB [2] が唐詩作品を対象とした全文検索を提供し、研究および教育の活用が期待されている。しかし、唐詩情報に関わるデータベースは中国の書店により、高額で販売され [3]、検索インターフェイスが扱いにくい [3] という問題がある。

中国の企業のサービスにおける「超星数字図書館」[15] は膨大な数の画像データを収めて、貴重なデータ資料が多く収録され、一方では日本では入手しにくい文献も含んでいる。ただし、これらのデータを利用する条件として、読書カードという会員カードが必要であり、著作権への配慮から閲覧には専用のソフトが必要とされる。

そのほか、台湾の「中華電子佛典協會 CBETA」[16] と日本の SAT 大正新脩大藏經テキストデータベース「SAT」[17] が協力して、『大正新脩大藏經』の第 1 巻から第 85 巻までの全テキストをデータベース化し、IIIF [18] などの研究も進めている。二階堂 [12] は、これらの研究がほかの電子テキストの方法に影響を与えると述べている。

現在の日本に関わる漢文・漢詩の教育研究の現場で、これらのデータは十分に利用されているとは言い難く、それらのデータはいろいろなところがあるが、それらの情報を関連されていないという問題がある。その一方で、漢詩に関わるデータベースがばらばらに各所で作られ、学習のための検索の利便性が低いと考える。したがって、漢詩情報を公開共有でき、教育学習のニーズに応じて、自由に扱える環境が望ましい。

2.2 漢詩教育

日本では中学校と高等学校のシラバスに含まれる学習活動と学習目標において、漢詩教育が行われている。日本の中学校と高等学校に利用する教科書における国語と古典のシラバスに [1]、唐詩の学習は、唐詩に関わる文体、詩体、作者、唐詩作品のコンテンツの理解や、唐詩の時代背景、地理情報、歴史情報などの情報を理解する学習目標と学習活動例がある [1]

荒井 [4] は日本の中学校と高等学校の教科書に、漢詩に関わる訓読が違うものの、詩の意味の捉え方に大きな影響を与えて、文法と一致しない問題点がある [4]。中川は教育の現場に、漢詩に関わる文法によって、訳す本文フルテキストの内容も、矛盾を生じる可能性がある [5] と述べている。

今村 [19] は小学校から高等学校までの訓読情報の難しさ、レ点、一・二点の用法と書き下し文に関わる情報を述べた。また、漢文訓読学習の系統性について考察した上で、漢文教材の適切性を検証した。

林 [20] は漢詩における訓読の不統一に関わる整理が難しいと考え、例えば、「 以

(テ) 、 以 (ツ テ) 」、「 於 (テ) 、 於 (イテ) 」など、訓読の違い事例が数多くことを示した。その一方で、送り仮名の付け方では、漢詩の送り仮名も、原則はどちらの学習資料における訓読や書き下し文などのルールを従うことになるが、全ての差異がある漢詩はそれぞれの学習資料によって、どちらかの資料に統一しなければならないのではない [20] と述べた。

それらの問題点に基づいて、漢詩に関わる学習内容を多く参照して、違う内容によって、学習資料を一致に調べさせ、唐詩作品に関わる数多くの学習資料を簡単に扱え、参照できることが望ましい。

2.3 Linked Open Data に関する研究

情報技術が発達するとともに、 提供される情報も段々と増加している。これに対して、大量の情報の中からデータを正確に抽出できる Linked Open Data (LOD) [6] [7] と呼ばれる方法が注目を集めている。研究者は LOD 技術を用いて、図書館や、博物館など様々な文化領域に文化情報の普及に向けた活動を目指す。

まず、欧米の大学図書館や国立図書館が行っている先駆的な事例として、Europeana [21] の取り組み事例は、そのコレクションの利用範囲としてはオープン化したデータの利活用の促進を含むという特徴をもつ。

OCLC Worldcat [22] は典拠データや各種コードを LOD 化し、様々なサービスとして実行している。

英国全国書誌 (British National Bibliography) [23] の Linked Data 化は 1950 年以降の全データ約 300 万件に関わる各種コード類や典拠データを無料で提供し、公開する。それは Linked Data 化の取り組みの代表例であり、英国全国書誌は今後も英国図書館のデータを追加続いている。

また、米国議会図書館は BIBFRAME という書誌フレームワークを 2012 年 11 月に公表した [24]。

日本でも、様々な図書館や、博物館などもそのメタデータなどを公開利用する事例がある。江草ら [25] は国立教育政策研究所教育図書館や教科書研究センター附属教科書図書館が長年かけて組織化してきた書誌情報をまとめて、LOD 化した教科書 LOD [26] を公開している。1992 年施行の学習指導要領以降の検定教科書を対象として、書誌事項と教科等の関連情報を LOD 化し、それらに関わるデータを公開した。

橋詰 [27] は国立国会図書館における出版物に関するメタデータと LOD の取り組みを述べて、図書館に所蔵する書籍などの書誌情報の LOD 化の可能性を紹介し、そのざまざまな領域での利活用に関わる情報を紹介した。

嘉村ら [28] は LODAC Museum を事例として、博物館情報における芸術・文化情報の LOD 普及に向けた現状を述べた。それは、博物館により、メタデータの設計および情報の相互運用性を考察し、情報を共有した上で、多様な利用ができるという利便性を述べた。そのほか、日本の横浜市は美術館所蔵品データ [29] の LOD 化を公開し、利活用されている。

図書館 Linked Data [30] によって、様々な研究アプローチがある。例えば、研究者や、学生などは、LD 技術の原則によって、研究者や学生などは、ウェブ上で存在する論文や、データを直接利用可能になる利益として目立つ [30]。

また、中国でも、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」[31] を始めた。このデータベースは LOD 技術を利用して、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、LOD 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 [31] 上で公開した。このサイト [31] 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。

LOD 技術は図書館に限定で利用できるではなく、代表的なウェブサイト Wikipedia 百科事典 [32] を元にした LOD のオントロジーをよく知られている。

加藤 [33] [34] は DBpedia [35] が Wikipedia 百科事典 [32] として、かつ、DBpedia 日本語版 [36] が含まれている。 DBpedia [35] に基づいて、構造化された多言語の知識を抽出して、Resource Description Framework (RDF) [6] データセットを生成するものであり、LOD としてウェブ上に再公開しているコミュニティプロジェクトとなる。このプロジェクトでは、いろいろな事物に関わるものに関連付けされ、LOD 技術の原則によって、それぞれの関連付けはグラフとして表現され、認識がしやすいため、RDF というウェブと親和性の高いグラフモデルを用いている [33] と紹介した。

玉川ら [37] は日本語 Wikipedia オントロジーの構築と利用を紹介した。その中にメインとしては、日本語 Wikipedia オントロジーの概略図を紹介し、それぞれ

の関連関係を構築した。また、関係がある情報の間の関係を定義した。

本研究では文化資源の一部を用いて、唐詩作品を対象として、LOD に基づく BIBFRAME Model [9] によって、新たな唐詩情報モデルを構築し、唐詩情報の LOD 化を目指す。

2.4 TEI マークアップに関する研究

デジタル化の推進とともに、全文資料のデジタル化がだんだん増えてきた。それについて、全文資料の本文フルテキストの構造化が非常に重要な一環として様々な試みが進められてきた。本研究ではその中で、テキスト資料に対して、標準的な構造化を行いたいと考える。

本文フルテキストのマークアップ手法はいくつかがある。李ら [38] は一切経音義全文データベースを構造化し、そのデータの項目構造の表示と効率的な検索のため、注文テキストをマークアップする方法を用いる。それは掲出語を用いて、独自のタグセットを作ったものである。具体的には第一段階では注文テキストにある「、」「。」のような区切り要素を音注、字体注、および義注としてタグを付与する。第二段階ではタグの属性を用いて、第 3 段階では連続している「,」で終わる同様なタグを合併するように分けて、マークアップする [38]。

高橋 [39] はサンスクリット語で「中辺分別論疏」という仏教文献を用いて、それらと関わる写本と校訂テキストの本文フルテキストおよび注釈情報を混在するテキストの XML によるマークアップをした。

また、山口 [40] は中国語原元および和刻本における文字情報のデータベース化およびウェブブラウザ等を文書表示できることを用いて、漢文表記の特徴 (両側ルビと訓点の組み合わせ) をブラウザ上で表記することを試みた。その上で、漢文表記のマークアップを行なったと指摘した。

上で述べたような様々なマークアップ手法が提案されてきたが、それぞれのテキストごとに異なるマークアップ手法であり、これらの標準化が課題であると考える。一方、標準化されたマークアップ手法として、TEI マークアップのような手法がある。

TEI マークアップは、人文学資料におけるテキスト資料のデジタル化するための標準的なマークアップ方法である。それは Text Encoding Initiative (以下、TEI) [41] による人文学資料をデジタル化を行うとともに、TEI ガイドライン [8] に従って、

デジタル媒体上で機械可読な形で記述し、特定のハードウェアやソフトウェアに依存せず効率的・効果的に共有することを目指し、基盤技術として XML を採用する方法論である。永崎 [42] は、TEI を欧米の人文学向けデジタルテクスト資料の構造化ではデファクト標準になっており、デジタル・ヒューマニティーズ分野の基幹技術の一つであると述べている。

典型的な TEI マークアップの事例によって、アメリカ詩プロジェクト [43] や、オックスフォード大学が Eighteenth-Century Poetry Archive [44] を行なっている事例がある。

後藤ら [45] は日本古代の資料である延喜式に対して金属加工や食品などのデータの長期保存するために、延喜式における 1 巻から 50 巻までのテキストファイルの TEI マークアップを行なった [45]。

上原らは「水滸伝」諸版本を用いて、TEI と IIIF という国際的な標準化を用いて、2 つの方法論を一緒に利用し、デジタル化した内容をマークアップする [46]。

松田ら [47] は TEI ガイドラインの適用としては、フランス語仏教辞典「法寶義林」目録のデジタル化を試みた。それはフランス語で、仏教辞典として、現版面を尊重した上で、「法寶義林」の第一分冊から第四分冊まで目録を用いて、著者や訳者などの色々な本文と関係がある情報のマークアップを行なった。

河瀬ら [48] は近世口語資料の洒落本に含まれる『傾城買二筋道』 (1798 年刊行) の版本を事例として、TEI マークアップによるタグセットを考え、その構造化を試みたうえで、文書構造化を検討した。

高橋ら [49] は転写テキストと楔形文字粘土板資料における T 文字研究・言語研究のためのプラットフォームの TEI マークアップを行なった。

日本における SAT 大正新脩大藏經テキストデータベース「SAT」[17] は、『大正新脩大藏經』の第 1 巻から第 85 巻までの全テキストをデータベース化し、全文資料を無料で公開されている。その中の電子本文フルテキストを TEI マークアップし、標準的な全文資料を構築した [42]。

日本で訓読文などで掲載される唐詩情報に関わる TEI マークアップの研究は見当たらないため、本研究では TEI マークアップ手法を試みたい。

一方、TEI マークアップ以外では、本文フルテキストを用いて、電子版の資料として、データベースや画像などとして提供する例もある。田中 [50] はジャパンナレッジをネットアドバンスから提供され、日本の学術市場における日本語の有料データベースサービスとして紹介した。それは辞書を中心として、「e-リソース

(電子資料)」という形で、流通している。コンテンツ内容としては、日本大百科全書 (小学館) や日本歴史地名大系 (平凡社) などの幅広い辞書をインターネット上のデータベースとして、検索可能になる。

さらに、前述の通り、画像データベースと TEI マークアップを組み合わせる事例として、日本の SAT 大正新脩大藏經テキストデータベース「SAT」[17] は、『大正新脩大藏經』の第 1 巻から第 85 巻までの全テキストをデータベース化し、全文資料を無料で公開されている。それは電子版本文フルテキストの TEI マークアップの取り組みとする事例というだけではなく、その中に含まれる画像データを IIIF [18] で画像検索可能にし、ほかの電子テキストの方法に影響を与えている [42]。

第3章　研究対象

3.1　対象データ

本研究では文化資源に注目し、唐詩作品を研究対象として、唐詩の構造化に関する研究を行う。

唐詩作品は典型的な古典作品として、中国だけではなく、日本の国語と古典の学習者や、各国の研究者も研究対象として利用する。唐詩に関わるデジタルリソースや紙媒体は様々がある。2.1 節の通り、北京大学数拠分析研究中心による「全唐詩分析系統」オンライン DB [2] が唐詩作品に扱う全文検索を提供し、研究および教育の活用が期待されている。唐詩情報に関わるデータベースは中国の書籍店により取り扱われ、高額で販売され [3]、検索インターフェイスが扱いにくい [3] という問題があることをわかった。

本研究では、 唐詩データは手元で取得しやすいデータを利用、研究を進んでいきたいと考える。手軽に扱え、唐詩は典型的な古典作品として、日本の古典や国語の教科書に採用、教員や生徒により利用するため、まず、日本の中学校と高等学校の国語と古典の教科書に含まれる唐詩作品を研究対象として、データを取得すると考える。

また、教育の学習ニーズに対して、学習指導要領によって、唐詩作品を含む教科書の内容は改訂されつつ, 唐詩作品の利用状況も変化するなどの原因があるため、本研究では、平成 28 年度使用の中学校と高等学校の国語と古典の教科書に含まれる唐詩作品を研究対象とする。

筑波大学附属中央図書館に所蔵されている平成 28 年度の日本の中学校と高等学校で使用する教科書を調べ、掲載される唐詩作品の数は表 1 に示す。中学校使用の教科書 5 冊に、唐詩作品は延べ 12 首、異なり 6 首が含まれる。高等学校の教科書は、国語の国語総合の教科書 23 冊、古典 A 教科書 6 冊、古典 B 教科書 19 冊を調査して、唐詩作品を含む数は延べ 374 首、異なり 53 首がある。異なる唐詩作品 53 首に対応して、作者は異なり 21 名がいた。これらの唐詩作品の延べ数は、同じ

作者で、コンテンツも同様な唐詩作品についても、違う教科書に重なって掲載される全ての唐詩作品を計数した数である。異なり数は、同一の唐詩作品は、違う教科書に何度も重なって掲載されても、唐詩作品一首とし、全体に異なる唐詩作品がいくつあるかを数えた数である。

表 1: 教科書に掲載される唐詩作品の数

	唐詩作品			
調査対象	延べ数	異なり数	教科書 (冊数)	作者 (異なり数)
中学校	12 首	6 首	5 冊	4 名
高等学校	362 首	53 首	48 冊	21 名
合計	374 首	53 首	53 冊	21 名

3.2 構造化の適用範囲

本研究では、LOD と TEI マークアップという 2 つの方法から考えて、本節では、対象データにおける構造化の適用範囲を述べる。つまり、唐詩作品に関わるメタデータ情報および本文フルテキストの情報はどのような技術を用いるのかそれぞれの技術ごとの適用範囲を解説したいと考える。

LOD のデータセットの範囲としては、前節 3.1 の通り、唐詩作品が異なり 53 首、延べ 374 首、作者が 21 名というデータ全体を用いる。このデータセットについて、唐詩作品に関わる基本的なメタデータとして、唐詩作品のタイトル、唐詩作品の別名、唐詩作品の詩体情報、コンテンツの文体を一緒に記述して、LOD 化を用いる。

唐詩の LOD 化は外部リソースとの関連付けも用いる。その作品は、唐詩作品と関係がある情報として、作者や、掲載される教科書と関連づける上で、外部リソースとして Wikidata [51]、Chinese Text Project (CTEXT) [52]、DBpedia [36] のリソースと一緒に記述する。

外部リソースの適用範囲は TEI マークアップを行なった中学校の教科書に含まれる唐詩作品 6 首のみを用いて、外部リソースとの関連付けを試みる。調べたところ、現時点では、唐詩作品異なり 6 首に対応する外部リソースが 4 首を存在し、2 首が存在しないことがある。ここでは、外部リソースが存在する際の情報源を適用範囲として、繋げた。

TEI マークアップの範囲としては、平成 28 年度に利用する中学校の教科書に含まれる唐詩作品異なり 6 首を用いて、それに対応する高等学校の教科書に掲載さ

れる任意の唐詩作品 6 首を選択して、TEI マークアップを行うこととした [53]。

マークアップ対象としては、本文フルテキストに関わる白文、訓読文、書き下し文と翻訳文の 4 つの文体 [53] だけではなく、訓読文に関わる返り点、送り仮名などの訓点情報、書き下し文におけるルビ情報、そのほか、タイトル、作者、ルビを一緒に構造化する。

第4章　唐詩のLOD化

4.1　唐詩のLOD化の意義

唐詩作品は中国の文化資源として、現在まで千年以上の歴史がある。唐詩作品は中国古典文学研究に欠かせない基本的な文献である。それらの編集は同時代の人々によって集成が行われていたが、後は現在に至るまで改訂が多くの利用者によって積み重ねられている。とりわけ唐詩作品の違いを補うことに見るべきものがある。

唐詩作品は日本に伝わった際、日本の研究者や学習者を理解できるために、研究者は各自の理解によって、中国語版の唐詩作品を日本語の文法にあわせて、唐詩作品の本文フルテキストに訓読情報や、ルビ情報などを原文資料に記述した。日本の利用者は、それぞれの訓読情報などによって、唐詩作品を翻訳して、唐詩作品を学んでいる。このように、中国に既存の唐詩作品のみならず日本に伝わる資料及び文献の違いは日本と中国における漢詩の読解方法の文化の違いに起因すると考えられる。

3章の通り、日本の中学校と高等学校が利用する教科書53冊に、掲載される唐詩を調査したところ、同じ唐詩作品に対してタイトルの差異、表現の差異があることが分かった。

同じ唐詩作品は、掲載されるタイトルや表現の差異がある。例えば、表2の「送元二使安西」という唐詩作品は、中学校の教科書では訓読文で「元二の安西に使ひするを送る」とのタイトルを使用している。

表 2: 同じ唐詩作品「送元二使安西」における表現の差異がある事例

No.	タイトル	原文資料
1	「送元二使安西」	送元二使安西　王維 渭城朝雨浥軽塵　客舎青青柳色新 勧君更尽一杯酒　西出陽関無故人 【三体詩】 [54]
2	「元二の安西に使ひするを送る」	元二の安西に使ひするを送る　王維 渭城の朝雨軽塵を浥し　客舎青々柳色新たなり 君に勧む更に尽くせ一杯の酒　西の方陽関を出づれば故人無からん 渭城朝雨浥軽塵　客舎青青柳色新 勧君更尽一杯酒　西出陽関無故人 [55]

表 3: 同じ唐詩作品「静夜思」による掲載される表現の差異がある事例

No.	タイトル	原文資料
1	「静夜思」	静夜思 牀前看月光 疑是地上霜 舉頭望山月 低頭思故郷 李白 [56]
2	「静夜の思ひ」	静夜の思ひ　李白 牀前看月光 疑是地上霜 舉頭望山月 低頭思故郷 牀前月光を看る 疑ふらくは是れ地上の霜かと 頭を舉げて山月を望み 頭を低れて故郷を思ふ [80]

また、同じ唐詩作品のタイトルに対して、異なる語句を用いる表現がある。例えば、表 4 の「秋夜寄丘二十二員外」という唐詩作品は「秋夜寄丘員外」というタイトルを採用する場合があって、表 5 の「山中対酌」は「山中与幽人対酌」がタイトルとして掲載される例である。

表 4: 同じ唐詩作品「秋夜寄丘二十二員外」の異なるタイトル表現の例

No.	タイトル	原文資料
1	「秋夜寄丘二十二員外」	秋夜寄丘二十二員外 懷君属秋夜 散歩詠涼天 山空松子落 幽人応未眠 [58]
2	「秋夜寄丘員外」	秋夜寄丘員外 懷君属秋夜 散歩詠涼天 空山松子落 幽人応未眠 【唐詩三百首】 [59]

表 5: 同じ唐詩作品「山中与幽人対酌」に対してタイトルの差異がある事例

No.	タイトル	原文資料
1	「山中対酌」	山中対酌　李白 両人対酌山花開，一盃一盃復一盃 我醉欲眠君且去　明朝有意抱琴来 （『古文真宝』前集） [60]
2	「山中与幽人対酌」	山中与幽人対酌　李白 花に囲まれた山中で、心の通い合う隠者と酒をくみかわした時の作である。なにものにもとらわれない自由で奔放な心情がよく表れている。 両人対酌山花開，一杯一杯復一杯。 我醉欲眠卿且去　明朝有意抱琴来 （李太白集） [61]

これらの事例を通じて、教科書に掲載される唐詩作品における訓読文などの違

いも存在することをわかった。日本の中学校と高等学校の教科書に、含まれる唐詩作品における訓読が違うものの、詩の意味の捉え方に大きな影響を与えて、文法と一致しない [4] という問題があり、同じ唐詩作品でも、文体や、文字などの差異が存在し、文法によって訳す本文フルテキストの内容も、矛盾を生じる可能性がある [5] と指摘されている。それらの問題点に基づいて、教育現場のニーズに基づき、唐詩における訓読の不統一に関わる整理が難しく、どちらかに統一しなければならないのではないと考えた。例えば、唐詩作品における訓読文の違いによって、書き下し文や翻訳文の内容が違う可能性があり、生徒らは、本文フルテキストの理解にも差異が起きる可能性がある。それらの問題点に基づいて、漢詩に関わる学習内容を一緒に参照して、関連がある情報に多くの参照を加えられるようにできることが望ましい。

これらの理由から、本研究ではLOD技術を用いて、唐詩を統一的に扱える環境を提供したいと考える。日本の中学校と高等学校で学習する唐詩作品、作者、それらを掲載する教科書との関連関係をLOD化することを試みる。LOD化の枠組みにあたって、BIBFRAME Model [9] に基づいて、新たな唐詩作品モデルの構築、データセットの公開およびLOD化に関わるウェブサイトの構築を報告する。

4.2 唐詩の対象データ

本研究では文化資源に注目して、唐詩作品を研究対象として、唐詩の構造化を目指す。唐詩情報に関わる資料を自由に公開共有可能にし、教育学習のニーズに応えて、異なる表現を統一的に扱う学習環境を提供したいと考える。

手元に日本の中学校と高等学校で学習する唐詩作品、唐詩と関連がある基本的なデータと関連付けて、それらと掲載される教科書の関連関係や外部リソースとも繋げて、LOD化することを試みる。LOD化の構築にあたって、BIBFRAME Model [9] に基づいて、新たなモデル化も行う。その関連関係に関わるデータセットを公開共有できるようにしたいと考える。

教育学習ニーズに応えて、手元で取得しやすいデータから、筑波大学附属中央図書館に所蔵されている中学校と高等学校の国語と古典の教科書を対象として着手したい。本研究では、平成28年度用の教科書を研究範囲とする。3章の通り、唐詩作品は延べ374首があり、異なり53首がある。それぞれに対応する唐詩作品の作者は21名に対応することが分かった。

唐詩作品に関わる基本的な対象データとしては、唐詩作品のタイトル、唐詩作品の別名、詩体の定義の解説、本文フルテキストの詩体が基本的な唐詩情報を一緒に記述する。また、直接的に関係がある情報として、唐詩作品を作った作者情報、唐詩作品を含む教科書との関連関係を対象データとして収集する。作者情報は作者氏名、性別、字（あざな)、生年月日、没年月日と作者の紹介などがあると考える。教科書情報は、教科書 LOD [26] における関連関係を外部リソースとして繋げる。 江草ら [25] による教科書 LOD は、国立教育政策研究所教育図書館や教科書研究センター附属教科書図書館が長年かけて組織化してきた書誌情報をまとめて LOD 化したものであり、1992 年施行の学習指導要領以降の検定教科書を対象として、書誌事項と教科等の関連情報を LOD 化し、2018 年 3 月までの 7,257 タイトルの教科書情報、RDF データとして 157,297 トリプルを公開している [26]。本研究に用いる教科書は 53 冊あり、唐詩作品を異なり 53 首、延べ 374 首を含み、それらと教科書の関係をすべて記述すると考える。

そのほか、DBpedia [35] [36]、Wikidata [51] と Chinese Text Project (CTEXT) [52] などの外部リソースも対象データとして用いる。

DBpedia [36] は多言語の Wikipedia 百科事典 [32] に基づいて、構造化された多言語の知識を抽出して、Resource Description Framework (RDF) [6] データセットを生成するものであり、LOD としてウェブ上に再公開しているコミュニティプロジェクトである。そのほか、DBpedia [36] はフリー百科辞典 Wikipedia かたの構造化コンテントの抽出を目的とするプロジェクトである。それは構造化情報として、ウェブ上で自由に利用できる。ウィキペディア日本語版の記事を利用した DBpedia Japanese は、国立情報学研究所によって 2012 年 5 月 9 日に公開された [36]。

Wikidata [51] は、様々な情報を構造化したデータとして組み合わせるデータベースである。構造化データは、Wikipedia [32] も利用できて、誰でも利用できるデータベースになる。

CTEXT [52] は Donald Sturgeon 氏を代表として、開発されて、近年有名な中国における電子書籍のプロジェクトである [62]。それは北京大学とハーバード大学を連携したプロジェクトであり、中国の古典籍をデジタル化して、無料で公開される電子図書館である。その中のデータは、多く含まれて、唐詩に関わる情報も無料で公開されている。そのため、本研究もこの電子図書館のプロジェクトを用いて、その中に含まれるデータを利用し、関連付けをしたいと考える。これらの情報について、唐詩作品を主語として、重要な外部リソースとの関連づける試み

であると考える。

4.3 提案モデル

LOD 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する方法である。それはセマンティックウェブを指向したデータ共有の枠組みで、様々な領域のデータをその記述の粒度を問わずに扱え、自由に利活用できる特徴を持っている [6] [7] 。

唐詩の LOD 化は、これらの特徴を用いて、唐詩作品に関わるデータを対象として、関連情報を URI で表して、標準的技術を使用し、唐詩に関わる作者、タイトル、詩体などの有益な情報と、掲載される教科書などの外部リソースを関連付けて、メタデータを公開共有できるものである。本研究では、LOD 化の構築にあたって、BIBFRAME Model [9] に基づいて、新たなモデル化も行う。その関連関係に関わるデータセットを公開共有するウェブサイトを設計し、公開する。

唐詩情報を記述するための枠組みである LOD 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 [6] [7]。また、扱う BIBFRAME Model [9] に基づき、訓読文として異なる唐詩作品を創作作品 Work として、教科書に掲載された唐詩作品延べ 374 首を Instance として採用する。

そのほか、教科書に掲載された唐詩作品の Instance とそれぞれに対する関連付ける教科書と繋げている。同様の唐詩作品の作者は統一であるため、ここでは唐詩作品の作者は唐詩作品 Work と関連する。このようにして、それぞれの関連付けを BIBFRAME Model [9] に基づいて、唐詩作品の枠組みを構築した。図 1 は、唐詩作品の枠組みの概要である。

図 1 唐詩作品の枠組みに基づいて、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。枠組みの情報は下の節で説明したいと考える。

具体的に説明すると、唐詩作品 Work を主体として、 基本的な唐詩作品のタイトル、詩体、外部リソース、唐詩作品の作者と関連づけている。また、Work は教科書に掲載された唐詩作品 Instance も関連づけている。

教科書に掲載された詳しい唐詩作品 Instance は唐詩 Instance のタイトル、教科書に掲載されたページ数、唐詩作品の本文フルテキストと唐詩作品の本文フルテ

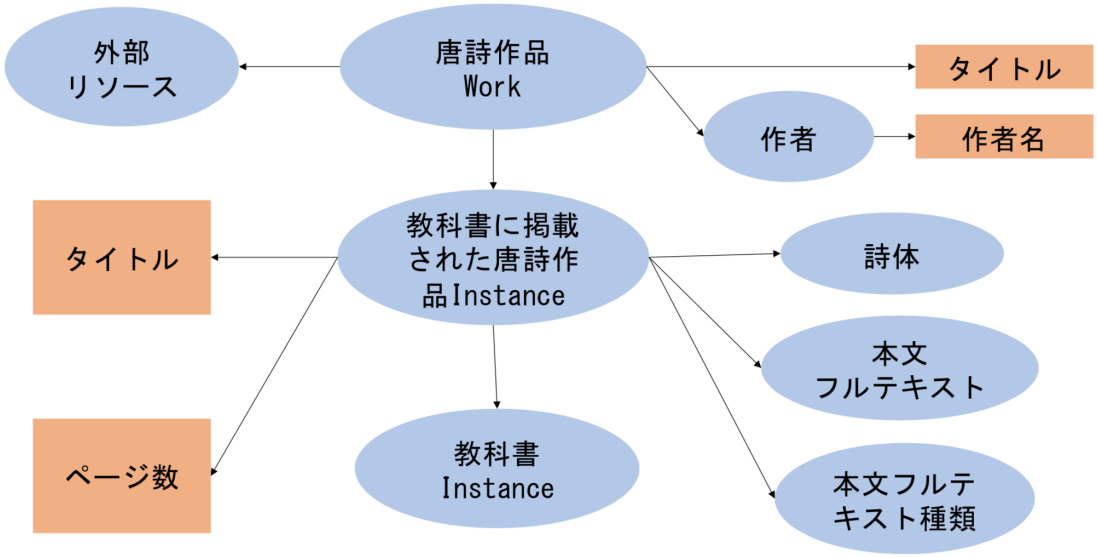


図 1: 唐詩作品の枠組みの概要

キストの種類と関連づけている。また、唐詩作品 Instance は、それぞれに対する教科書 Instance との関連付けも行う。

作者情報は作者氏名、性別、 字 (あざな)、生年月日、没年月日と作者の紹介などの情報がある。作者の情報については、唐詩作品 Work と唐詩作品 Instance を両方とも関連づけている。

4.3.1 BIBFRAME モデルに基づく唐詩情報モデル

本研究では、LOD モデルを参照した上で、BIBFRAME (Bibliographic Framework) Model [9] の応用も目指す。BIBFRAME Model [9] は 2012 年に米国議会図書館が公表した書誌情報のためのデータモデルとボキャブラリであり、最新版は 2016 年 4 月に公開された BIBFRAME 2.0 Model [9] である。これは LOD のモデルに沿って、新たな文献目録の仕組みを作成する基準である。2016 年 4 月の BIBFRAME 2.0 [9] の概要によって、このモデルは Work、Instance と Item の 3 つの核心的なエンティティから構成される。

3.1 節の通り、日本の中学校と高等学校が利用する教科書 53 冊に、違う教科書によって、同じ唐詩作品を重複して、利用することがあるため、掲載される唐詩作品は異なり数と延べ数と分けて計算した。ただし、同じ唐詩作品中には、タイトルの差異、表現の差異が存在する。例えば、「送元二使安西」[54] という唐詩作

品は、中学校の教科書では訓読文で「元二の安西に使ひするを送る」を使用したり、「静夜思」は「静夜の思ひ」などを用いる事例がある。

それより、同じ唐詩作品を Work レベルにして、重複した唐詩作品は、各 Instance として、記述すれば、対象データの構造が簡潔になるため、ここでは、BIBFRAME Model [9] を採用して、データモデルを設計した。

BIBFRAME Model [9] における唐詩作品は、訓読文として異なる唐詩作品を創作作品 Work として、唐詩作品の延べ数を Instance として採用する。また、Instance と関連する教科書の情報を関連付けている。唐詩作品の作者は Work と関連すると考える。このようにして、それぞれの関連付けを BIBFRAME Model [9] で構築する。

例えば、唐詩作品「送元二使安西」[54] に関する基本的な枠組みを図 2 に示す。唐詩作品「送元二使安西」[54] は 2015 年発行された中学校で使用する国語教科書に「元二の安西に使ひするを送る」[55] というタイトルで含まれるという関係で、抽象的な唐詩作品「送元二使安西」のクラスは Work として使用し、それと唐詩作品「元二の安西に使ひするを送る」を Instance として関連付けている。そのほか、唐詩作品の Instance はまた掲載される教科書と関連付ける。

4.3.2 提案モデルのリソース URI と語彙

本研究では唐詩作品を研究リソースとして、唐詩作品の基本的な情報と教科書の関連関係の枠組みを考え、LOD 化する研究を行う。唐詩作品の Work、唐詩作品の Instance、唐詩作品の詩体、作者情報のベース URI が https://w3id.org/tangpoem/ を共通としてを採用する。さらに、ベース URI はプレフィックス tangpoem:または tp:として表現する。この Web サイトは W3C Permanent Identifier Community Group [63] 提供しているサービスで、目的としては、安全に永続的な URLs を提供している。

前節の通り、唐詩作品異なり 53 首は Work として、作品は作者の順によって数字 1 から 53 までに至る ID を並べて振る。唐詩作品延べ 374 首は教科書目録に掲載する教科書の番号順で国語、国語総合、古典 A と古典 B と分類し、数字 1 から 374 までに至る番号で並んでいる。同一教科書の中に、唐詩作品は教科書に含まれるページ数の順で並んでいる。そして、唐詩作品のタイトルは Work の異なり 53 までの ID と Instance の延べ 374 まで番号の ID を区別して付ける。そのため、唐詩作品 Work は tangpoem:1、…、tangpoem:53 までなどとして利用する。唐詩作

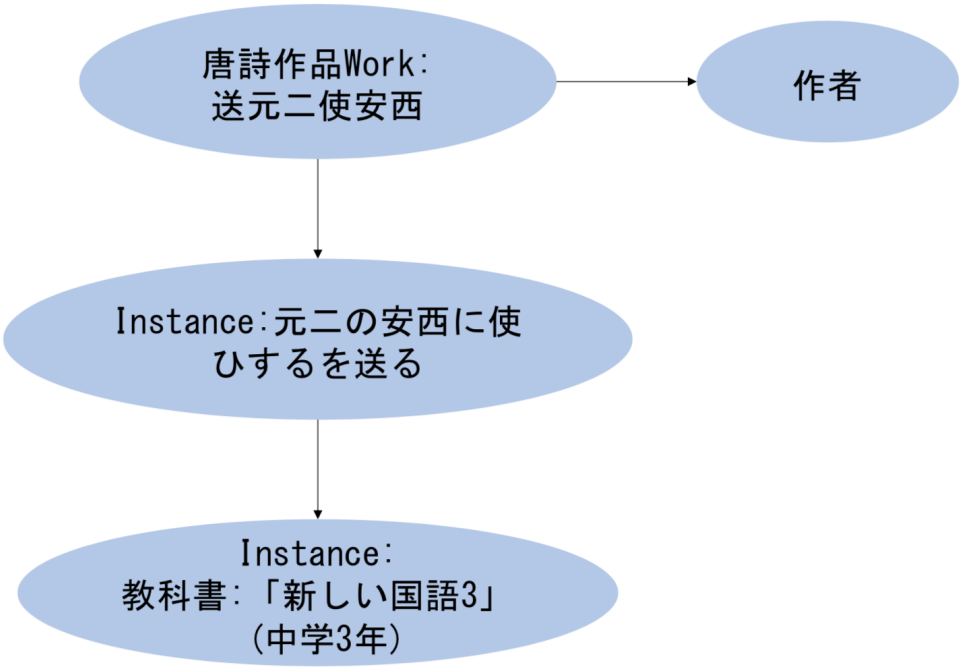


図 2: BIBFRAME Model を唐詩作品に適用した例

品 Instance は tangpoem:instance/1、..., tangpoem:instance/374 までなどとする。唐詩作品の詩体は律詩、絶句と分けて、五言と七言の 4 種類にするため、URI は tangpoem:style/詩体と採用する。

更に、唐詩作品の作者は各自の氏名を識別し、URI は tangpoem:author/氏名を用いる。そのほか、Instance として、教科書とも関連付けられる。教科書リソースの URI には、学校種別（高等学校など)、教科書の検定年、文部科学省が発行している教科書目録に使われている「教科書の記号」（主に教科や科目を由来に命名される）と「教科書の番号」（3 桁であらわされる）を使用し、`<jp-textbook:高等学校/2014/古 B/326>` などとして識別する [25] 。

外部へリンクする URI として Wikidata と CTEXT の詳しい情報と各自に対応するウェブサイトのリンクと関連づける。

本研究の URI の設計に基づいて、情報の検索を向上できるように RDF 枠組みの属性を選択する。それは RDF 枠組みの検索が詳細化できるようにするために付与する。唐詩作品の語彙は LOD ボキャブラリには既存のものを採用し、BIBFRAME ボキャブラリ [64]、Dublin Core [65]、Schema.org [66]、RDF Schema [67] と FOAF

ボキャブラリ [68] を用いて、これらに該当しないものについては、唐詩作品の独自ボキャブラリも作成する。

本研究の全体像を把握するため、唐詩作品 Work、Instance、作者情報と唐詩作品の本文フルテキストの内容を用いる。Schema.org の語彙の利用として唐詩作品の作者は、schema:Person で使い、唐詩作品と教科書が schema:CreativeWork を用い。プロパティとしては、schema:author、唐詩作品と教科書の関係は schema:isPartOf として利用した。RDF schema では外部リソースをリンクされる URI を rdfs:seeAlso を用いる。Dublin Core の語彙は、唐詩作品タイトルが dc:title を用いた。

具体的に Dublin Core [65] の語彙から説明すると、唐詩作品の詩体などの解説は dc:description を採用する．RDF Schema のボキャブラリ [67] は唐詩作品のタイトルのラベル rdfs:label、外部へリンクする URI として rdfs:seeAlso を用いる。Schema.org [66] の利用は各リソースの種別として、唐詩作品のクラス schema:CreativeWork、作者のクラスは schema:Person を用いる。作者の生年月日 schema:birthDate/deathDate を採用、作者の性別 schema:gender とする。唐詩作品を含む教科書のページ数が schema:pageStart を採用する。

唐詩作品 Work と Instance の言語表現は中国語と日本語を区別して、schema:inLanguage の zh と ja で扱う。唐詩作品の詩体などの解説の言語表現は文字列に日本語言語タグ@ja を付与する。

また、唐詩作品の作者情報について、字は FOAF Vocabulary から foaf:nick を用いる。唐詩作品の独自プロパティは特に唐詩作品の詩体への関連付けで使って、tangpoem:style として表現する。

BIBFRAME Vocabulary [64] について、唐詩作品を創作作品として、Work のクラスは bf:Work で、Instance は bf:Instance で表現する。それに基づき、詳しい属性プロパティを与えると、唐詩作品 Work のタイトルは bf:workTitle、Instance のタイトルは bf:instanceTitle で採用する。Work から Instance への関連付けは bf:hasInstance として定義する。また、Instance としての唐詩作品 と教科書の関係は bf:partOf で扱う。

この形式を用いることで、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。唐詩作品の LOD 化の枠組みの詳細は以下の表 6、表 7、表 8 にそれぞれ示している。ただし、モデルとして唐詩のスタイルや本文コンテンツリンクの部分はプロパティ定義と値を定めておらず、これらのデータモデル構築は今後の課題として研究を

進んでいきたいと考える。

表 6: 唐詩作品 Work に関わる語彙定義とプロファイル

No.	プロパティの説明	プロパティ名	プロパティ値の例
1	唐詩のクラス	rdf:type	bf:Work
2	唐詩の言語	schema:inLanguage	”zh”
3	唐詩の Work 名	bf:workTitle	唐詩のタイトル
4	唐詩 Work の詩体	tangpoem:style	https://w3id.org/tangpoem/style/詩体
5	唐詩の作者	schema:author	作者の名前

表 7: 唐詩作品 Instance に関わる語彙定義とプロファイル

No.	プロパティの説明	プロパティ名	プロパティ値の例
1	唐詩のクラス	rdf:type	bf:Instance
2	唐詩の言語	schema:inLanguage	ja
3	唐詩の Instance 名	bf:instanceTitle	唐詩のタイトル
4	唐詩 Instance の詩体	tangpoem:style	https://w3id.org/tangpoem/style/詩体
5	教科書に含むページ数	schema:pageStart	ページ数
6	教科書との関係	bf:partOf	教科書の外部リソース
7	唐詩 Work との関係	bf:instanceOf	唐詩 Work のリンク

表 8: 唐詩作品の作者に関わる語彙定義とプロファイル

No.	項目	プロパティ名	プロパティ値の例
1	作者のクラス	rdf:type	schema:Person
2	作者名	rdfs:label	名前
3	性別	schema:gender	女
4	字	foaf:nick	字
5	生年月日	schema:birthDate	具体的な生年月日
6	没年月日	schema:deathDate	具体的な没年月日
7	作者の紹介	dc:description	作者の紹介

図3に唐詩作品の全体像 (抜粋) として、唐詩の LOD モデルに基づき、BIBFRAME Model によって新たに構築したものを示す。具体的に唐詩作品1を事例として、tp:1 の Work を中心に、唐詩作品インスタンスと関連付け、プロパティを付与し、関連関係を表示した。また、唐詩作品インスタンスはその作品を含む教科書などの外部リンクと関連付けて、その他、唐詩作品の詩体や、作者などの情報も関連付けた。このようにして、唐詩作品の LOD データモデルを構築した。

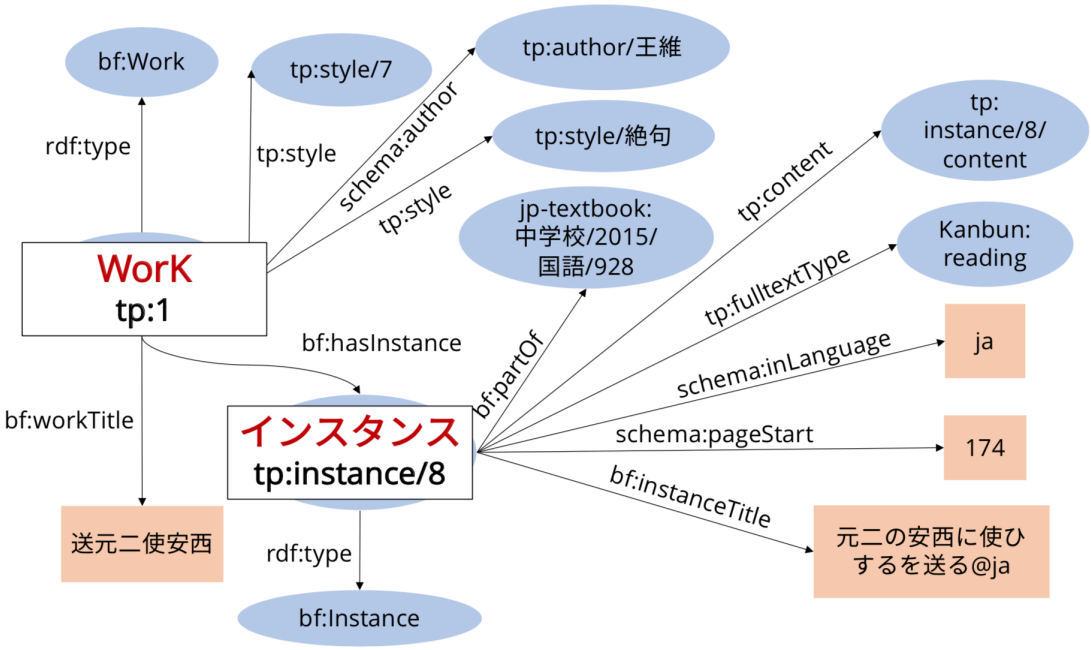


図 3: 唐詩作品の全体像 (抜粋)

4.3.3 外部リンクの利用

前節までの通り、唐詩の LOD 化は唐詩作品を研究対象として、LOD 技術を用いて、日本の中学校と高等学校で学習する唐詩作品、作者、それらを掲載する教科書との関連関係を LOD 化することを試みる。LOD 化の枠組みにあたって、BIBFRAME Model [9] に基づいて、新たな唐詩作品モデルの構築を用いる上で、唐詩の LOD 化の枠組みは、唐詩作品の基本的なデータだけではなく、外部リソースの情報との関連付けも注目する。

まず、外部リソースの情報を紹介したいと考える。唐詩作品を含まれる教科書情報は教科書 LOD [26] における関連関係を外部リソースとして繋げる。本研究に

用いる教科書情報は唐詩作品を異なり 53 首、延べ 374 首を含まれ、それらと教科書の関係を用いて繋げると考える。

外部リソースとの関連付けは唐詩の LOD 化の枠組みの一部として試みるが、全ての唐詩のデータの利用ではなくて、ここでは中学校に利用する唐詩作品異なり６首を選択し、それらと Wikidata [51] と CTEXT [52] の関連付けを表現した。表 9 は外部リソースの情報源の確定をした。表 9 に示すように、「黄鶴楼送孟浩然之広陵」[89] という唐詩作品の情報を調べて、CTEXTcitectext にはこの唐詩作品のウェブページがあるが、Wikidata のデータベースにこの唐詩作品のエントリーが存在しない。「黄鶴楼送孟浩然之広陵」[89] だけではなく、「絶句」[94] の Wikidata 情報も存在しないことがわかった。これらの存在しない情報はどうするのかを今後の課題として、唐詩情報のデータベースをどう増やすのかを考えたい。

表 9: 外部リソースの情報源の確定

No.	教科書			Wikidata	CTEXT
	唐詩作品	中学校	高等学校		
1	黄鶴楼送孟浩然之広陵 (黄鶴楼にて孟浩然の広陵に之くを送る)	現代の国語 2 [89]	精選国語総合 [90]	NA	https://ctext.org/text.pl?node=138509&if=en
2	春望	現代の国語 2 [85]	精選国語総合 [91]	https://www.wikidata.org/wiki/Q18024342	https://ctext.org/text.pl?node=149538&if=en
3	春暁	現代の国語 2 [92]	精選国語総合 [93]	https://www.wikidata.org/wiki/Q11088809	https://ctext.org/text.pl?node=136295&if=en
4	絶句	国語 2 [94]	精選国語総合 [95]	NA	https://ctext.org/text.pl?node=150936&if=en
5	送元二使安西 (元二の安西に使ひするを送る	中学校国語 3 [55]	精選国語総合 [54]	https://www.wikidata.org/wiki/Q17368400	https://ctext.org/text.pl?node=129844&if=en
6	静夜思 (静夜の思ひ)	中学校国語 3 [80]	高等学校国語総合 [56]	https://www.wikidata.org/wiki/Q4391398	https://ctext.org/text.pl?node=137062&if=en

DBpedia 日本語版 [36] も外部リソースとして利用しているんが、調べたところ、DBpedia 日本語版に唐詩作品として、情報が多く存在してないため、外部リソースの利用としては、唐詩作品との関連付けではなく、作者情報とつなげている。対象データの数によって、作者が 21 名がいたことによって、DBpedia との関連をそれぞれの作者と関連づけている。例えば、外部リソースにおける DBpedia との関連事例は表 10 に示すようになる。

表 10: 外部リソースにおける DBpedia との関連事例 (抜粋)

No.	作者情報	DBpedia
1	杜甫	http://ja.dbpedia.org/resource/杜甫
2	王維	http://ja.dbpedia.org/resource/王維
3	李白	http://ja.dbpedia.org/resource/李白

4.4 データセットの構築

4.2 節の通り、唐詩作品に関わる対象データは、(1) 唐詩作品の基本データ、(2) 作者情報、(3) 教科書情報、DBpeida、Wikidata と CTEXT などの外部リソースの 3 つを用いる。

本研究では、唐詩作品を主体として、 基本的な唐詩作品のタイトル、詩体、詩体の定義の解説、教科書に含まれるページ数、外部リソース、唐詩作品の作者と教科書の関連付けを行う。その中、作者情報は作者氏名、性別、 字 (あざな)、生年月日、没年月日と作者の紹介 などがある。これらの情報について、唐詩作品を主語として、唐詩作品は詩体や作者などの情報としている。

LOD データセット構築の流れを図 4 で示す。データセット構築は Poorman's LOD Toolkit [72] の手法を用いて構築した。まず、唐詩作品に関するメタデータなどの情報を人手で収集した。収集したデータセットを Excel などで項目に関するファイルを作成した。それらのデータを洗練されて、データセットを作成し、データセットの Excel のファイルを RDF へ変換、トリプル形式で記述する。検索するために、記述したデータセットを検索サーバに入れる。記述した LOD データセットの一連のファイルは自分で作った公開用のウェブサイト [70] に公開して自由に利用可能になった。このようにして、唐詩作品のデータセットの枠組みを構築した。今後は唐詩作品の情報を公開・検索できるように用意したいと考える。

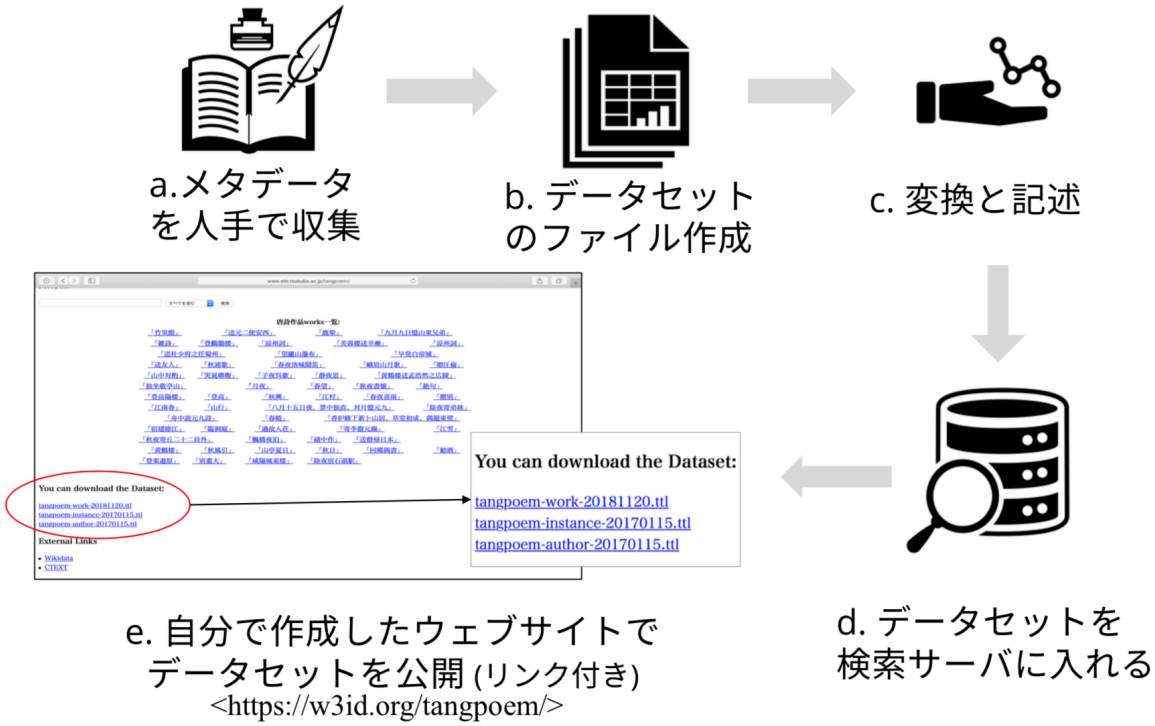
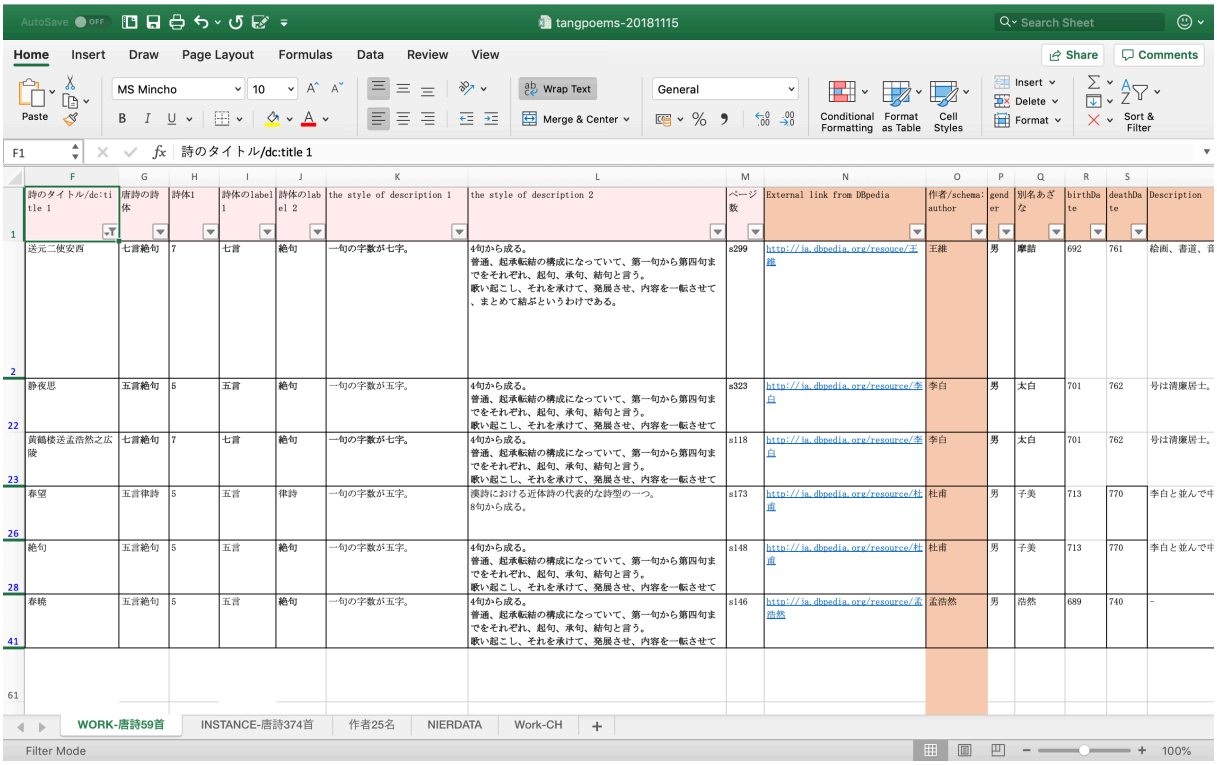


図 4: LOD データセット構築の流れ

唐詩作品は異なり 53 首 (Work)、延べ 374 首 (Instance) があるため、Excel に Work のシートと Instance のシートを別に分けて、唐詩作品に関わるメタデータ情報を各項目単位として、一列ずつ入力する。その項目は唐詩作品のタイトル、唐詩作品の別名、詩体、詩体の定義の解説、本文フルテキスト、本文フルテキストの詩体などとなる。具体的には、図 5、6 に示す。

また、作者情報は作者氏名、性別、字（あざな）、生年月日、没年月日と作者の紹介などがある。ここでは、作者情報も独自の対象データとして、Excel に独自のシートファイルに、分別に各自の項目を設計する。具体的には、作者情報は独自のデータセットとして設計した上で、唐詩作品の異なり数シートと延べ数シートにも作者情報と入力すると考える。作者に関わる具体的な項目の設計は、図 7 のように示す。

そのほか、一部の教科書の情報は、唐詩作品の延べ数のシートに含まれて、Wikidata [51] と Chinese Text Project (CTEXT) [52] などの外部リソース情報は唐詩の延べ数に設計すると考える。Wikidata [51] と Chinese Text Project (CTEXT) [52] などの外部リソースも対象データとして Excel に入力された。また、唐詩作品と関連づける外部リソースの項目の設計の事例は図 8 のように示す。



	詩のタイトル/dc:title 1	唐詩の詩体	詩体1	詩体のlabel 1	詩体のlabel 2	the style of description 1	the style of description 2	ページ数	External link from DBpedia	作者/schema:author	gender	別名あざな	birthDate	deathDate	Description
2	送元二使安西	七言絶句	7	七言	絶句	一句の字数が七字。	4句から成る。 普通、起承転結の構成になっていて、第一句から第四句までをそれぞれ、起句、承句、結句と言う。 歌い起こし、それを承けて、発展させ、内容を一転させて、まとめて結ぶというわけである。	s299	http://ja.dbpedia.org/resouce/王維	王維	男	摩詰	692	761	絵画、書道、音
22	静夜思	五言絶句	5	五言	絶句	一句の字数が五字。	4句から成る。 普通、起承転結の構成になっていて、第一句から第四句までをそれぞれ、起句、承句、結句と言う。 歌い起こし、それを承けて、発展させ、内容を一転させて	s323	http://ja.dbpedia.org/resource/李白	李白	男	太白	701	762	号は清廉居士。
23	黄鶴楼送孟浩然之広陵	七言絶句	7	七言	絶句	一句の字数が七字。	4句から成る。 普通、起承転結の構成になっていて、第一句から第四句までをそれぞれ、起句、承句、結句と言う。 歌い起こし、それを承けて、発展させ、内容を一転させて	s118	http://ja.dbpedia.org/resource/李白	李白	男	太白	701	762	号は清廉居士。
26	春望	五言律詩	5	五言	律詩	一句の字数が五字。	漢詩における近体詩の代表的な詩型の一つ。 8句から成る。	s173	http://ja.dbpedia.org/resource/杜甫	杜甫	男	子美	713	770	李白と並んで中
28	絶句	五言絶句	5	五言	絶句	一句の字数が五字。	4句から成る。 普通、起承転結の構成になっていて、第一句から第四句までをそれぞれ、起句、承句、結句と言う。 歌い起こし、それを承けて、発展させ、内容を一転させて	s148	http://ja.dbpedia.org/resource/杜甫	杜甫	男	子美	713	770	李白と並んで中
41	春暁	五言絶句	5	五言	絶句	一句の字数が五字。	4句から成る。 普通、起承転結の構成になっていて、第一句から第四句までをそれぞれ、起句、承句、結句と言う。 歌い起こし、それを承けて、発展させ、内容を一転させて	s146	http://ja.dbpedia.org/resource/孟浩然	孟浩然	男	浩然	689	740	-

図 5: 唐詩作品 Work データ構築のための Excel ファイル

データセットにおける Work、Instance と作者情報は RDF で構築した上で、Turtle 形式で記述されている。Turtle 形式 [6] [73] は RDF を簡便に記述するために利用する形式であり、RDF の枠組みによって、主語、述語と目的語を表示し、URI を<>で囲んでいるものである。具体的には、以下の通りに、Work 情報、Instance 情報、作者情報の書き方を紹介したい。

また作者情報と、外部リソースとを関連づけると考える。その書き方は以下の表 11、12、13 のようにデータセットにおける Work 情報、Instance 情報、作者情報の書き方の事例である。

このようにデータセット情報を書いて、唐詩作品異なり 53 首、延べ 374 首のデータに関わるデータセットにおけるファイルを構築した。これらのファイルは work ファイルとインスタンスのファイルで別にした。つまり、唐詩作品の work のデータセットは 491 トリプルの情報を作成し、唐詩作品のインスタンスのデータセットはデータ形式は Turtle 形式とした 3,000 トリプルで作成した。そのほか、作者情報も 233 トリプルで作成した。

これらのデータセット情報を一緒に記述した上で、RDF データサーバーとして、Apache Jena Fuseki [74] を用いて、検索サーバーに入れた。クエリ言語としては、

表 11: データセットにおける Work 情報の Turtle データ (抜粋)

Work 情報の書き方

```
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX schema:<http://schema.org/>
PREFIX tangpoem:<https://w3id.org/tangpoem/>
PREFIX vcard:<http://www.w3.org/2006/vcard/ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX bf:<http://bibframe.org/>
PREFIX foaf:<http://xmlns.com/foaf/0.1/>

<https://w3id.org/tangpoem/1>
a schema:CreativeWork;
a bf:Work;
schema:inLanguage "zh";
bf:workTitle "送元二使安西"@ja;
tangpoem:style <https://w3id.org/tangpoem/style/7>;
tangpoem:style <https://w3id.org/tangpoem/style/絶句>;
schema:author <https://w3id.org/tangpoem/author/王維>;
rdfs:seeAlso <https://www.wikidata.org/wiki/Q17368400>;
rdfs:seeAlso <https://ctext.org/text.pl?node=129844&if=en>.
<https://w3id.org/tangpoem/21>
a schema:CreativeWork;
a bf:Work;
schema:inLanguage "zh";
bf:workTitle "静夜思"@ja;
tangpoem:style <https://w3id.org/tangpoem/style/5>;
tangpoem:style <https://w3id.org/tangpoem/style/絶句>;
schema:author <https://w3id.org/tangpoem/author/李白>;
rdfs:seeAlso <https://www.wikidata.org/wiki/Q4391398#sitelinks-wikipedia>;
rdfs:seeAlso <https://ctext.org/text.pl?node=137062&if=en>.
```

表 12: データセットにおける Instance 情報の Turtle データ (抜粋)

Instance 情報の書き方
PREFIX dc:<http://purl.org/dc/elements/1.1/> PREFIX schema:<http://schema.org/> PREFIX tangpoem:<https://w3id.org/tangpoem/> PREFIX vcard:<http://www.w3.org/2006/vcard/ns#> PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#> PREFIX bf:<http://bibframe.org/> PREFIX foaf:<http://xmlns.com/foaf/0.1/> <https://w3id.org/tangpoem/instance/1> a schema:CreativeWork; a bf:Instance; bf:instanceTitle "黄鶴楼にて孟浩然の広陵に之くを送る"; schema:inLanguage "ja"; schema:pageStart "142"; bf:partOf <https://w3id.org/jp-textbook/中学校/2015/国語/827>; bf:instanceOf <https://w3id.org/tangpoem/22>. <https://w3id.org/tangpoem/instance/2> a schema:CreativeWork; a bf:Instance; bf:instanceTitle "黄鶴楼にて孟浩然の広陵に之くを送る"; schema:inLanguage "ja"; schema:pageStart "123"; bf:partOf <https://w3id.org/jp-textbook/中学校/2015/国語/829>; bf:instanceOf <https://w3id.org/tangpoem/22>.

表 13: データセットにおける作者情報の Turtle データ (抜粋)

作者情報の書き方

```
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX schema:<http://schema.org/>
PREFIX tangpoem:<https://w3id.org/tangpoem/>
PREFIX vcard:<http://www.w3.org/2006/vcard/ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX bf:<http://bibframe.org/>
PREFIX foaf:<http://xmlns.com/foaf/0.1/>

<https://w3id.org/tangpoem/author/杜甫>
a schema:Person;
rdfs:label "杜甫";
schema:gender "男";
foaf:nick "子美";
schema:birthDate "713";
schema:deathDate "770";
dc:description "李白と並んで中国を代表する詩人。「詩聖」と称される。";
rdfs:seeAlso <http://ja.dbpedia.org/resource/杜甫>.
<https://w3id.org/tangpoem/author/王維>
a schema:Person;
rdfs:label "王維";
schema:gender "男";
foaf:nick "摩詰";
schema:birthDate "692";
schema:deathDate "761";
dc:description "絵画、書道、音楽にも秀で、「詩仏」と称される。";
rdfs:seeAlso <http://ja.dbpedia.org/resource/王維>.
```

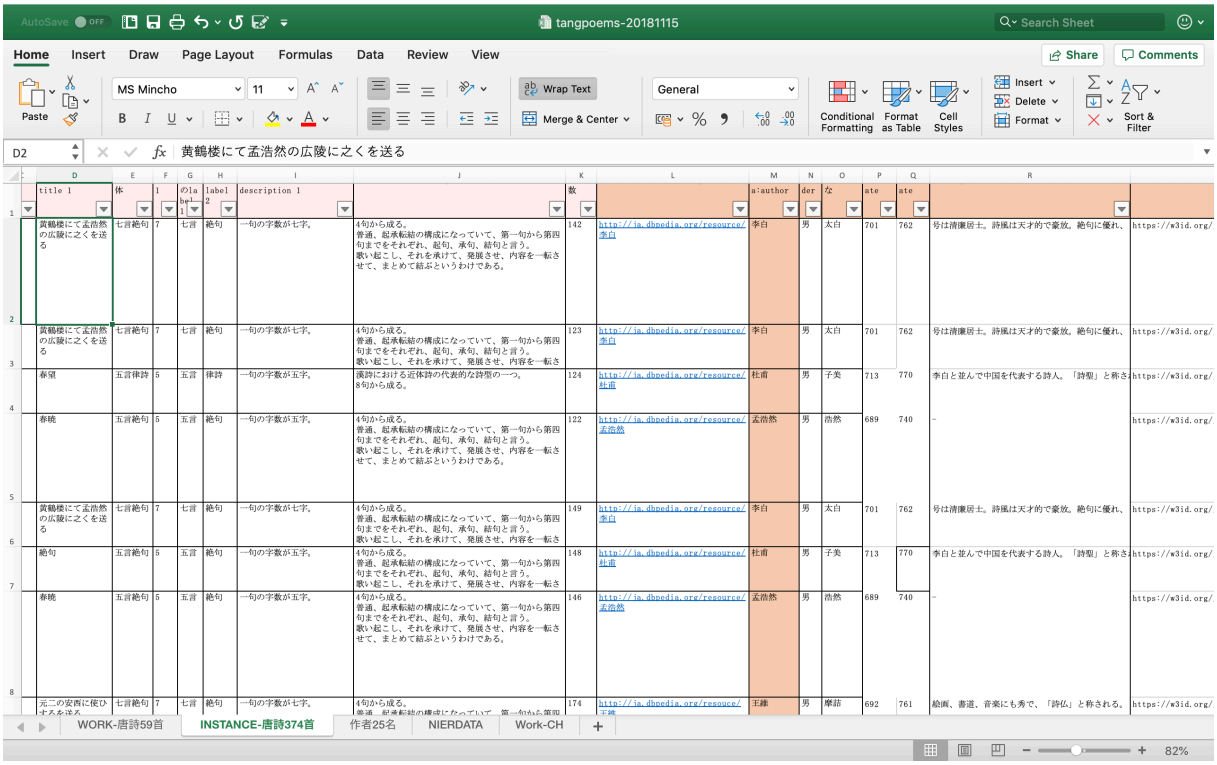


図 6: 唐詩作品 Instance データ構築のための Excel ファイル

SPARQL エンドポイント [6] で利用し、関連情報を検索できるようにした。

4.5 LOD データセットの公開

Tim Berners-Lee [6] が提唱した Linked Data Principle [75] は、LD の基本原則として、URI を参照した時は、標準の技術 (RDF や SPARQL 等) を使用して、関係する有用な情報を利用できるようにすることとしている。この原則に基づいて、本研究では、図 4 に示すように、唐詩の LOD 化に関わるモデルを構築した上で、アクセスとする公開用のデータウェブサイトを構築した。

現時点では、公開した LOD データセットをウェブ上から利用できる公開用のウェブサイトを構築してみた。図 9 の示すように、TOP Page から、唐詩作品 Work と具体的な唐詩作品例 (インスタンス) とそれらを含まれる現行教科書と関連付ける順番で構築して、そのほか、唐詩作品の詩体、作者などの関連情報も付けている。この web page は「LOD チャレンジ 2016」のデータセット部門の作品 [71] として参加するために、自分で作ったもので、`<https://w3id.org/tangpoem/>`リンクからから公開にリダイレクトされ、利用できるようになった。

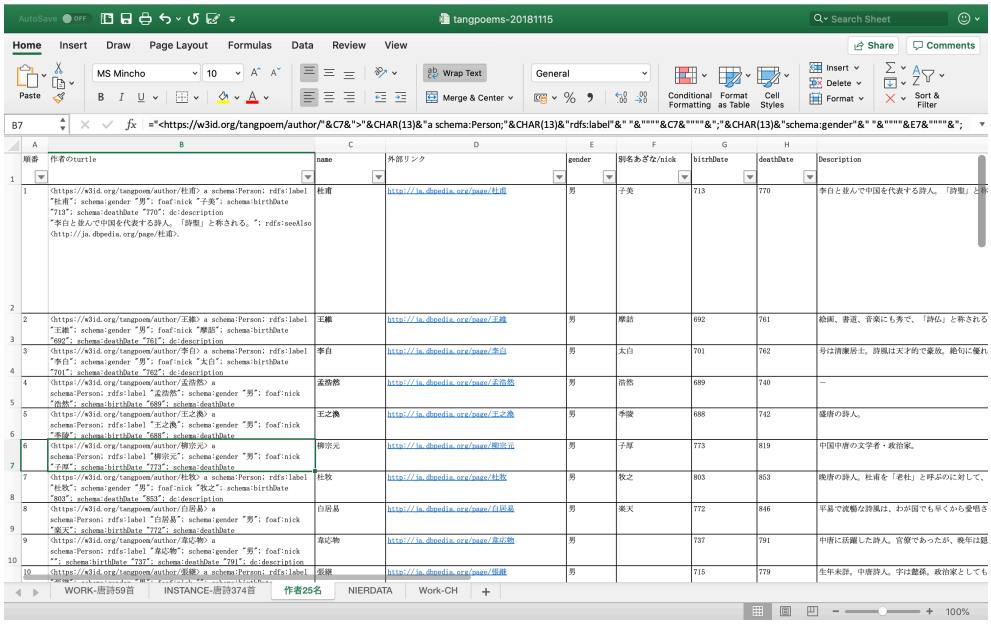


図 7: 唐詩作品の作者データ構築のための Excel ファイル

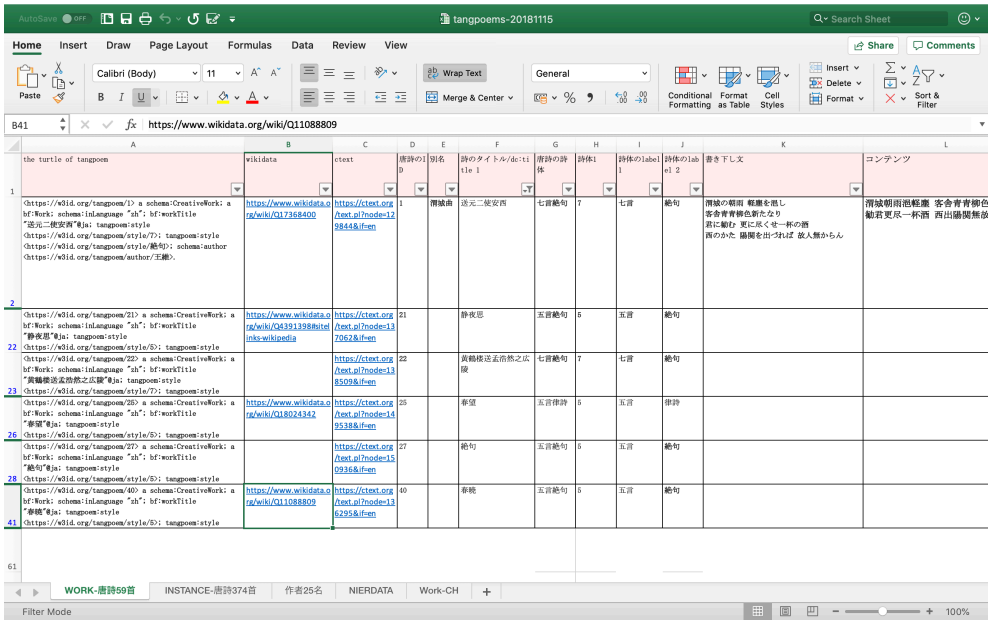
またこれらの一連のデータセットは 3 つがある。唐詩作品の創作作品 work のデータセット、インスタンスのデータセット、と唐詩作品に関わる作者情報のデータセットも一緒に唐詩の LOD 化のウェブサイト https://w3id.org/tangpoem/ [70] にアップロードした。

4.6 公開したデータの利活用

前節の通り、唐詩作品の創作作品 work、インスタンス、と唐詩作品に関わる作者情報の 3 つのデータセットが一緒に唐詩の LOD 化の公開用のウェブサイト https://w3id.org/tangpoem/ [70] に載せて、公開した。それによって、利用者は平成 28 年度の教科書に含まれる唐詩作品に関わる情報を自由にダウンロードでき、利用できる。それらのデータセットの利活用については以下の通りになる。

教科書に掲載された唐詩作品のタイトルや、本文フルテキストにおける様々な差異があるため、教育学習ニーズに応じて、まず、唐詩作品に関わる基本的な文体、本文フルテキストの種類などの情報を確認できる。また、唐詩作品に関わる作者や、掲載された教科書の情報も調べられると考える。

外部リソースの関連に対して、生徒らは、教科書に関わる情報を関連でき、日本語版の資料として、多くの情報も調べられるため、教育学習ニーズによって、便利になる。また、CTEXT のような、中国語の唐詩作品も関連でき、生徒らは、多



	A	B	C	D	E	F	G	H	I	J	K	L
1	the turtle of tangpoem	wikidata	ctext	唐詩のID	別名	詩のタイトル/dc:title 1	唐詩の詩体	詩体1	詩体のlabel 1	詩体のlabel 2	書き下し文	コンテンツ
2	<https://w3id.org/tangpoem/1> a schema:CreativeWork; a bf:Work; schema:inLanguage "zh"; bf:workTitle "送元二使安西"@ja; tangpoem:style <https://w3id.org/tangpoem/style/7>; tangpoem:style <https://w3id.org/tangpoem/style/絶句>; schema:author <https://w3id.org/tangpoem/author/王維>.	https://www.wikidata.org/wiki/Q17368400	https://ctext.org/text.pl?node=129844&if=en	1	渭城曲	送元二使安西	七言絶句	7	七言	絶句	渭城の朝雨 軽塵を浥し 客舎青青柳色新たなり 君に勧む 更に尽くせ一杯の酒 西のかた 陽関を出づれば 故人無からん	渭城朝雨浥輕塵 客舎青青柳色新君勸更尽一杯酒 西出陽関無故
22	<https://w3id.org/tangpoem/21> a schema:CreativeWork; a bf:Work; schema:inLanguage "zh"; bf:workTitle "静夜思"@ja; tangpoem:style <https://w3id.org/tangpoem/style/5>; tangpoem:style	https://www.wikidata.org/wiki/Q4391398#Links-wikipedia	https://ctext.org/text.pl?node=137062&if=en	21		静夜思	五言絶句	5	五言	絶句		
23	<https://w3id.org/tangpoem/22> a schema:CreativeWork; a bf:Work; schema:inLanguage "zh"; bf:workTitle "黄鶴楼送孟浩然之広陵"@ja; tangpoem:style <https://w3id.org/tangpoem/style/7>; tangpoem:style		https://ctext.org/text.pl?node=138509&if=en	22		黄鶴楼送孟浩然之広陵	七言絶句	7	七言	絶句		
26	<https://w3id.org/tangpoem/25> a schema:CreativeWork; a bf:Work; schema:inLanguage "zh"; bf:workTitle "春望"@ja; tangpoem:style <https://w3id.org/tangpoem/style/5>; tangpoem:style	https://www.wikidata.org/wiki/Q18024342	https://ctext.org/text.pl?node=149538&if=en	25		春望	五言律詩	5	五言	律詩		
28	<https://w3id.org/tangpoem/27> a schema:CreativeWork; a bf:Work; schema:inLanguage "zh"; bf:workTitle "絶句"@ja; tangpoem:style <https://w3id.org/tangpoem/style/5>; tangpoem:style		https://ctext.org/text.pl?node=150936&if=en	27		絶句	五言絶句	5	五言	絶句		
41	<https://w3id.org/tangpoem/40> a schema:CreativeWork; a bf:Work; schema:inLanguage "zh"; bf:workTitle "春暁"@ja; tangpoem:style <https://w3id.org/tangpoem/style/5>; tangpoem:style	https://www.wikidata.org/wiki/Q11088809	https://ctext.org/text.pl?node=136295&if=en	40		春暁	五言絶句	5	五言	絶句		

図 8: 唐詩作品と関連づける外部リソースの項目の設計の事例

言語の対応もできるようにしたと考える。

今回公開されたデータセットを利用すれば、LOD 化技術に基づき、唐詩作品のデータを簡単に検索できるようになったと考える。

また、唐詩の理解の難しさは中国時代の文化や歴史の内容と日本の文化などの違いに起因すると考える。教育学習ニーズに基づくシラバス [1] に掲載された項目 (要求) は、唐詩に関わる思想や、感情などの理解しづらい項目がある。今後の課題としては、公開したデータセットの利活用については、唐詩作品に関わる外部リソースとつなげて、これにより、関連資料を調べる時間をかからずに、外部リソースと繋げて、思想や感情などの抽象的な調べ方が簡単になり、唐詩を読みつつ、この作品の意味付けを直接見られ、唐詩の理解をしやすくなると考える。

そのほか、唐詩の LOD 化における作者情報のデータセットは、人物や事件などに関わる時・空間情報の可視化という研究に、人物や事件などに関わる時・空間情報の可視化を試みたい。例えば、唐詩作品に関わる唐代 (618 年-907 年) は初唐、盛唐、中唐と晩唐の 4 時期があって、22 人皇帝がいた。それは社会変動が大きく、非常に歴史が長い時代であった。作者時期によって、どこでどのような事件が起こるということを年表で可視化すれば、いろいろな歴史情報を地図にも行なった地理情報を表記して、時・空間情報の可視化になると考える。

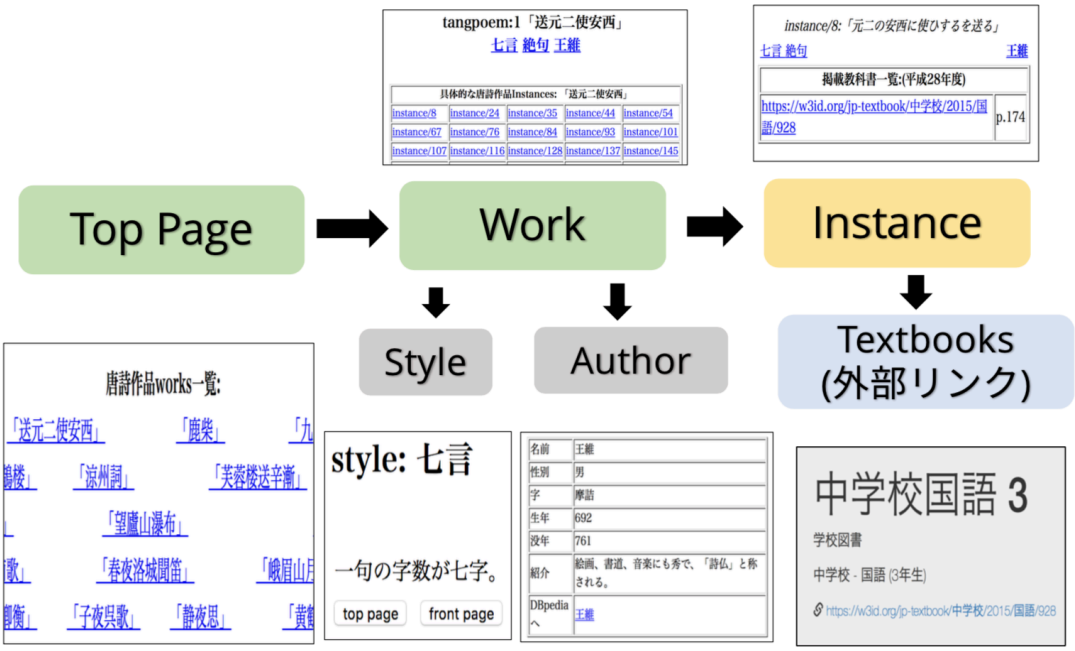


図 9: 公開用のデータウエブサイトの構築

4.7 考察

中国で唐詩作品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、重要な役割を果たしている。そのため、本研究では唐詩作品を研究対象として、日本の中学校と高等学校で学ぶ唐詩作品と掲載教科書の関連関係の日本語版を LOD 化する研究を行った。また、BIBFRAME Model [9] による唐詩情報モデルも構築できた。

今後は、唐詩作品のモデルに基づいて、多くの唐詩作品を増やしたいと考える。現時点の唐詩データは手元で取得しやすいデータを利用、研究を進んできた。手軽に扱え、唐詩は典型的な古典作品として、日本の古典や国語の教科書に採用、教員や生徒により利用するため、まず、日本の中学校と高等学校の国語と古典の教科書に含まれる唐詩作品を研究対象として、データを取得した。また、教育の学習ニーズに対して、学習指導要領によって、唐詩作品を含む教科書の内容は改訂されつつ, 唐詩作品の利用状況も変化するなどの原因があるため、本研究では、平成 28 年度使用の中学校と高等学校の国語と古典の教科書に含まれる唐詩作品を研究対象として、唐詩作品異なり 6 首、延べ 374 首、作者が 21 名に基づくデータセットを構築し、公開した。

それらに基づき、多くの唐詩作品に関わるデータセットを拡大したいと考える。例えば、今回外部リソースとして利用した CTEXT に関わる全唐詩の事例であり、この唐詩作品のデータに関わる唐詩作品データを拡大し、日本で利用する唐詩作品と合わせて、多様なニーズに応じて、多言語化にしたいと考える。

また、多くの唐詩作品を関連付けて、LOD 化できるようにして、それらに基づいて、検索や閲覧ができるアプリケーションを構築したい。現時点では公開用のウェブサイトを構築したが、それに基づく応用アプリケーションの設計を考えている。基本的な情報だけではなく、それぞれの関連できる情報を一括でリンクさせて、利用できるように設計し、それより、地理情報や、歴史情報などまでの横断検索も一緒に設計したく、今後の課題として検討する。

第5章　TEIマークアップ

5.1　TEIマークアップの意義

第1章の通り、筆者は文化資源に注目し、中学校と高等学校の国語と古典の現行教科書の教育学習ニーズに向けて、学校の生徒が唐詩作品をより簡単に理解できることを目指す。そのため、教科書に含まれる唐詩作品および作者の情報、詩体などの有益な情報を関連づけて、ウェブ上で膨大なデータの中から関連情報を正確に抽出し、リンクで直接繋がるLOD化することを試してきた [76] [77] [78] [53]。本章では、それらの研究成果を拡張して唐詩作品の本文フルテキストのデジタル化を試みる。

唐詩作品の本文フルテキストのデジタル化は重要な役割がある。まず、唐詩作品の本文フルテキストの情報を公開・利用できれば、対象とする唐詩作品の情報テキストを検索できて、学習者が漢文や唐詩作品などの情報を理解しやすくなると考える。また、教科書における唐詩作品のデジタル化に当たっては、学習に必要な教科書の原版面を尊重し、訓読文と書き下し文に基づくデジタル化する。訓読文の特有の訓点情報を機械可読のテキストにできれば、有益だと考える。したがって、テキスト検索や訓点情報の表現ができれば、唐詩作品のテキストとその関連情報は、多くの人が共有でき、簡単に使える利便性がある。

古典籍を機械可読化して共有するTEIマークアップ [41] が様々なプロジェクトで利用されている。TEIマークアップは汎用的なXMLと組み合わせて、記述できる。その特徴としては、TEI:P5ガイドライン [8] にタイトルや本文フルテキストなどの特定要素がある。そのため、本章では、唐詩作品を研究リソースとして、TEIマークアップを用いる。

現在、国語や古典の教育学習ニーズに応じて、生徒がより簡便に古典籍を読めるよう、教科書に含まれる漢詩などの本文フルテキストは、書き下し文にルビで漢字の解釈などを付与され、訓読文に関わる訓点情報も注釈として付けられる。本文フルテキストの内容に対する言葉の解説などの注釈も、番号や小さな符号など

で、その語の周りに付している。一方で、これらの訓点資料や、ルビ情報、注釈などの要素をテキスト化をするとき、基本的な作成基準が存在しない [39] との指摘がある。すなわち、古典籍における訓点情報およびルビ情報や注釈などのテキスト化は難しい研究課題である。

したがって、本研究では文化資源に注目し、唐詩作品を研究対象として、唐詩の構造化を目指す。唐詩情報に関わる資料を自由に公開共有可能にし、教育学習のニーズに応えて、異なる表現を統一的に扱う学習環境を提供したいと考える。具体的には、唐詩の本文フルテキストを、TEI マークアップを用いて、標準的にマークアップする。

5.2 唐詩作品の本文フルテキスト

本章では、平成 28 年度使用の中学校と高等学校の教科書に含まれる唐詩作品を研究対象とし、唐詩作品の本文フルテキストをメインとする TEI マークアップ手法を提案する。唐詩作品の訓読文における送り仮名や返り点などの訓点情報や、書き下し文におけるルビ情報の表現する方法を検討し、唐詩情報のウェブ上でデジタル化すること目指す。

唐詩作品の本文フルテキストをデジタル化する際の観点として、(1) テキスト表現の文体、(2) 訓点情報やルビ情報の表現に関わる 2 点を挙げて、以下でそれぞれについて述べる。

本研究では、平成 28 年度の中学校と高等学校に利用する教科書に含まれる唐詩作品の本文フルテキストを研究対象として、教科書に含まれる唐詩作品の利用状況の調査を行った。調査したところ、唐詩作品の本文フルテキストは 4 つの文体に分類できる。これら 4 つの文体は「白文」、「訓読文」、「書き下し文」、「翻訳文」と呼ぶ [53] 。この分類は唐詩作品の本文フルテキストにおける表現の違いに起因するものである。

中学校の教科書では、生徒の学習をより容易にするため、多数の作品が書き下し文も訓読文も両方が同時に掲載される。一方、高等学校の教科書に含まれる唐詩作品はすべて訓読文のみで掲載されている。翻訳文は唐詩作品の本文フルテキストの全文の意味の解釈として、中学校の教科書に時々掲載される。

唐詩作品などの漢文における作品の読み順は、常に縦書きで教材の上から下まで 1 行、右から左の順番で読んでいる。4 つの文体に関わる詳しい内容は以下の節

に紹介する。

5.2.1 白文の概要

白文は唐詩作品の本文フルテキストのみ、返り点、送り仮名などの訓点情報が付いていないそのままの原文資料である。この文体は主に中国の古典籍および漢文のコンテンツに利用するが、日本語の文法と合わずにそのまま学ぶと、読解での違いがあり、基本的な理解が難しい。日本の中学校と高等学校の教科書には白文が含まれない。

ただし、白文は中国語などの原文資料に基づくものであり、訓読文や書き下し文などの文体は白文に基づいて、訓点情報、音読などのルビ情報、注釈要素を記入して、日本語の語順に合わせて利用しているものであるため、白文は日本で使える古典籍および漢文の原文資料として、重要な役割を果たす。

5.2.2 訓読文の概要

訓読文は白文の原文資料に基づき、古典籍および漢文などの白文に対して、日本語の文法に従って、返り点や送り仮名などの訓点情報を白文内の各文字の四隅に書き入れたものである。訓読文は漢文を日本語の語順で読めるようにしたものである。したがって、日本語話者の生徒や研究者などの利用者が原文資料の意味を理解しやすくなる。そのため、訓読文は古典籍および漢文における生徒の読解力を上げる利益をもたらす。訓読文は日本の中学校と高等学校の教科書に両方含まれている形になる。

図 10 は、第一学習社が出版された国語の教科書「高等学校国語総合」[79] に掲載された李白の唐詩作品「静夜思」[56] である。このような訓読文を読むとき、返り点や送り仮名などの訓点情報の読み方に従って、助詞、助動詞などを補い、日本語の文法に合わせることができる。

5.2.3 返り点と送り仮名の概要

訓点情報は広い意味で言えば、返り点、送り仮名と句読点である。返り点は縦書きの場合、漢字や単語の左下のところに付けて、下から上へ戻って読むことを

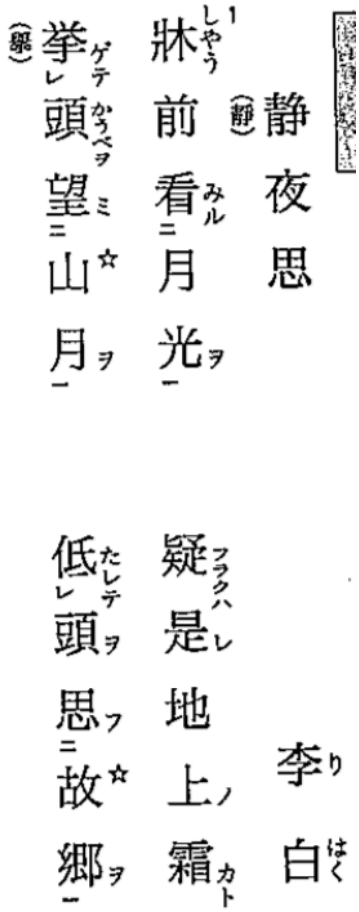


図 10: 訓読文の文体事例

示す記号である。中学校と高等学校の教科書に扱う返り点の記号種類は主にレ点、一・二・三点である。

送り仮名は文章の単語を解釈する場合、その漢字や単語の読み方を表すために、その四周に加える小さな文字である。平成 28 年度の教科書における唐詩作品の本文フルテキストはほぼ送り仮名と返り点だけを含んだものだが、高校教科書 [79] には図 11 のように「八月十五日夜、禁中独直、対月憶元九」という唐詩作品 1 首のみはタイトルに句読点を含む。

5.2.4 書き下し文の概要

書き下し文は訓読文に関わる返り点や送り仮名などの訓点情報の要素に基づき、漢文を日本語の語順に合わせて、書き直した文章である。図 12 は「中学校国語 3」の教科書 [80] に掲載の「静夜の思ひ」[80] という唐詩作品の訓読文 (上) と書き下

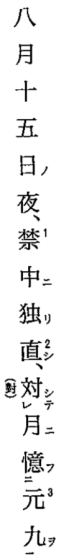


図 11: 句読点の事例 [79]

し文 (下) を一緒に掲載した本文フルテキストの事例である。

書き下し文は訓読文に表記する訓点のルールに基づいて、日本語の語順で読みやすい文で読解し、訓読文を日本語として読み取りやすい利益がある。書き下し文は日本語の語順に合わせたものであるため、生徒は訓読文より理解しやすいため、中学校の教科書のみに含まれている。

5.2.5 翻訳文の概要

ここでは漢文を興味が持つ日本の方が日本語の文法、語順などに合わせて、現代文のような形で翻訳したものが翻訳文と呼ばれる。それは、日本で古典に興味を持つ研究者らが日本語の語順で、個人的な理解や、考えによって、日本語の語順で翻訳したものである。

翻訳文は時々中学校の教科書にできくる。本研究では、翻訳文に関わる構造化は研究の範囲に含まれない。

5.2.6 ルビの情報および対象

ルビ [81] [82] [83] は現代文や古典籍の文章のある文字に対し、振り仮名や文字の説明、音読、異なる読み方などを原文資料の親文字より小さな文字で付与され

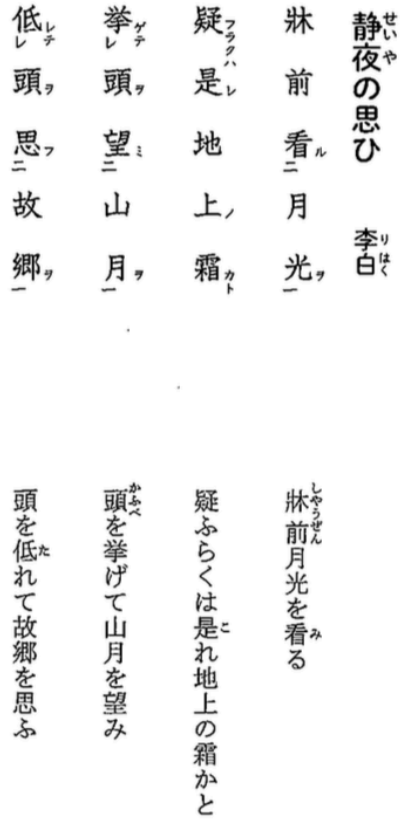


図 12: 訓読文 (上) と書き下し文 (下) が一緒に掲載された事例 [80]

るものである。それは日本語の文章では、漢字の発音や、漢字の説明などに役に立つ。ルビはモノルビ、グループルビと熟語ルビの 3 つの種類 [84] がある。

現代文や古典籍に文章は縦書きする場合、音読や注釈などのルビの位置は基本的に原文資料の任意の漢字の右側に仮名で付与する。教科書におけるルビ情報では、通常のルビに加えて、両側ルビを配置する場合。親文字の左側にカタカナでこの親文字の振り仮名や、文字の説明、音読、異なる読み方などの情報を小さな文字で表記する。

ここでは、親文字の片側にカタカナでこの親文字の振り仮名や、文字の説明、音読、異なる読み方などの情報を小さな文字で表記するものが片側ルビと呼ばれて、本文によって、両側にルビを配置する場合が両側ルビと呼ばれる [81] [82] [83]。

中学校の教科書には、書き下し文も訓読文も掲載されているため、ルビ情報は、書き下し文だけで付けられている。高等学校の教科書における全ての唐詩作品は訓読文だけで掲載されるため、ルビ情報は、振り仮名や、文字の説明、音読、異な

る読み方などの注釈情報をそれぞれの語の周りに対応して利用する。なお、縦書きの場合、ルビ文字列の位置に基づいて、ルビの処理は片側にのみ配置することと両側に配置する場合の処理として議論する。

平成28年度の教科書における唐詩作品の書き下し文および訓読文におけるルビの利用状況を表14でまとめた。教科書に含むルビの利用状況は、ルビの位置によって、書き下し文と訓読文に関わる教科書に含むルビ付けの事例である。

表14において、第1番から5番までが片側ルビである。第1番と2番は、ルビの位置における基本的な片側ルビとして、親文字の右側に平仮名や左側にカタカナで表記する。第3番と4番は親文字の片側に括弧で音読や、漢字の解説を示す事例である。第5番は基本的な片側ルビと違って、親文字の片側に、2列並べて置いている事例である。また、5番目の親文字から2列目のルビは、括弧で囲んでいる。

次に、6番目と7番目は両側に配するルビの事例である。6番目の事例は基本的な両側ルビとして配置するが、5番目事例の唐詩作品と同一であり、教科書の違いによって、表現も違うことが分かる。7番目の事例は両側ルビに基づき、左側のルビが括弧で囲む場合もある。

5.2.7 注釈情報

注釈とは既存の文章における任意の言葉の解説である。それは読者が文章を理解できるように、文書の出典や、言葉の意味などを示す。それは、日本および中国の書物でも付与する。唐詩作品の注釈は本文内容の言葉や、作者の紹介、本文フルテキストに関連する背景などの内容を補足する。

中学校の教科書では、注釈番号は本文に付けられず、まとめて同じページの下部に置いている。高等学校の教科書に唐詩作品は全て訓読文で掲載され、番号や、符号が本文に付けられ、対応する説明は同じページの下部に置いている。例えば、表15の例として注釈情報の事例を説明する。

教科書に掲載される唐詩作品の注釈は番号や符号で付ける。番号を付ける場合は、この言葉の解釈などを示す。例えば、表15の1番目の注釈事例は、番号で付記して、唐詩作品の本文フルテキストに関わる地名情報を解説する。また、符号を付する場合は表15の2番目の事例であり、教育学習ニーズに応じて、学習者により唐詩作品の本文フルテキストを理解しやすくなるために、読解などの問題も示す。

表 14: 教科書におけるルビの利用状況

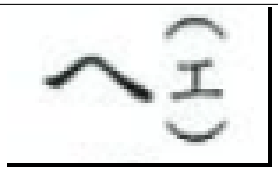
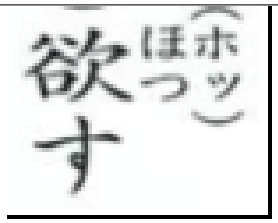
No.	ルビの位置	教科書に含む事例	文体
1	片側ルビ	 [80]	書き下し文
2	片側ルビ	 [85]	訓読文
3	片側ルビ	 [86]	書き下し文
4	片側ルビ	 [56]	訓読文
5	両側ルビ	 [86]	書き下し文
6	片側ルビ	 [85]	書き下し文
7	両側ルビ	 [87]	訓読文

表 15: 注釈情報の事例

No.	注釈の事例	対応する注釈内容
1	今夜鄜州ノ月[1ふ] [87]	1鄜州　今の陝西省富県。当時、杜甫の妻子は、安史の乱（七五五―七六三）を逃れて、そこに避難していた（⇩三三九ページ注3）。 [87]
2	低レ頭ヲ思フニ☆故郷ヲ [56]	問「山月」と「故郷」とは、どのように関係しているか。 [56]

5.3 対象データ

本研究では、平成 28 年度に使用する教科書を研究対象として、唐詩作品を含む数は延べ 374 首、異なり 53 首に対応する本文フルテキストのコンテンツをメインとして用いる。ただし、現時点の対象データとしては、3 章の記述する教科書に掲

載される唐詩作品 53 首の中から、中学校に利用する唐詩作品の異なり 6 首を用いて、それらに対応する高等学校の同様の唐詩作品任意 6 首をサンプルとして、TEI マークアップ手法を試みる。対象データは以下の表 16 の通りになる。

表 16: TEI マークアップを行った唐詩作品

No.	唐詩作品	中学校	高等学校
1	黄鶴楼送孟浩然之広陵 (黄鶴楼にて孟浩然の広陵に之くを送る)	現代の国語 2 [89]	精選国語総合 [90]
2	春望	現代の国語 2 [85]	精選国語総合 [91]
3	春暁	現代の国語 2 [92]	精選国語総合 [93]
4	絶句	国語 2 [94]	精選国語総合 [95]
5	送元二使安西 (元二の安西に使ひするを送る)	中学校国語 3 [55]	精選国語総合 [54]
6	静夜思 (静夜の思ひ)	中学校国語 3 [80]	高等学校国語総合 [56]

その上で、唐詩作品の本文フルテキストのデジタル化 (ウェブ化) は唐詩作品に関わるタイトル、作者や本文を一括で TEI マークアップする。TEI マークアップにおいて、詳細に扱う唐詩情報は以下の通りである。

1) 唐詩作品のタイトルおよび作者名。書き下し文と訓読文の両方の文体を含む場合はそれら両方に対応するメタデータになる。

2) 本文フルテキストに関するコンテンツの内容。主に訓読文と書き下し文を中心とする。訓読文は、親文字を中心として、四周に付ける返り点や送り仮名などの訓点情報が付く。書き下し文は唐詩作品の内容をそのままマークアップし、ルビなどの情報も付くことがある。

3) 本文フルテキストに関わる文体。

4) 唐詩作品の本文フルテキストに関わる行番号。

5.4 唐詩作品のマークアップ手法

本研究では、平成 28 年使用の中学校と高等学校の教科書に含まれる唐詩作品の本文フルテキストを研究対象として、TEI マークアップ手法を用い、唐詩作品のデジタル化を試みる。訓読文における唐詩作品の特有の要素として、送り仮名や返り点などの情報を表現する方法を検討する。

TEI (The Text Encoding Initiative) [41] は人文学系の文章を対象とするデジタル化を促進するために基本方針を定める組織である。TEI ガイドラインは 1994 年から図書館や博物館など、または個人的な利用者が研究、教育、保存を進むために、幅広く利用されてきた。

本研究の研究手法では、TEI マークアップの標準的な TEI:P5 ガイドライン [8] の基準を利用し、平成 28 年度使用の中学校と高等学校の教科書に掲載される唐詩作品の本文フルテキストを研究データとして、唐詩作品の本文フルテキストの TEI マークアップを試みる。

5.4.1 TEI マークアップ手法

唐詩作品の本文フルテキストの XML に基づく TEI マークアップを行う。5.3 節の通り、TEI マークアップにおいて、詳細に扱う唐詩情報は以下の通りである。

1) 唐詩作品のタイトルおよび作者名。
2) 本文フルテキストに関するコンテンツの内容。
3) 本文フルテキストに関わる文体。
4) 唐詩作品の本文フルテキストに関わる行番号。

まず、唐詩作品のタイトルは書き下し文の形のタイトルと訓読文の形のタイトルの 2 種類がある。唐詩作品のタイトルの表現には要素`<head>`を用いる。次に、唐詩作品に関わる作者名に扱うタグは要素`<persName>`を用いる。この要素`<persName>`は書誌情報における著作者 (個人・団体) の名前を示す要素である。

唐詩作品の本文フルテキストの全文は要素`<lg>`を用い、要素`<lg>`は line group の省略名であり、全文の詩節などのまとまりを示す要素である。唐詩作品の全文の特徴として、type 属性の値は verse(韻文) としてそれぞれの唐詩作品`<lg>`要素の属性に対応する。

唐詩作品の本文フルテキストにおける 4 つの文体は「白文」、「訓読文」、「書き下し文」、「翻訳文」と呼ぶ。白文は unpunctuated text、訓読文は punctuated text、書き下し文は reading text、翻訳文は translation text で文体を区別するための名前として利用する [53]。唐詩作品の本文フルテキストにおける文体の種類は独自の属性を用いて表現する。この属性は prefix:tp として、名前空間 URI は https://w3id.org/tangpoem/を用いる [76] [77] [78] [53]。その属性名は tp:fulltextType を用いる。属性値は、4 つの文体の英文名を利用して、全唐詩作品の本文フルテキ

ストに関わる文体の種類 (unpunctuated、punctuated、reading、translation) を表現できる。

また、本文フルテキストの詩節の各行は要素<l> を用い、行番号はその属性 n=1,2,…8 のようにする。<l>要素は verse line であり、詩節全体<lg>要素に含まれる。つまり、4 句の唐詩作品は 4 つの要素<l> で、8 句の唐詩作品は 8 つの要素<l>を組み合わせて、1 つの要素<lg>に含まれる。

表 17 は、「静夜の思ひ」[80] の TEI マークアップ例である。

表 17: 「静夜の思ひ」[80] に関わる TEI マークアップ例

原文資料	TEI マークアップ
牀前月光を看る 疑ふらくは是れ地上の霜かと 頭を挙げて山月を望み 頭を低れて故郷を思ふ [80]	<head>静夜の思ひ</head><persName>李白</persName> <lg type="verse" tp:fulltextType="reading"> <l n="1">牀前月光を看る</l> <l n="2">疑ふらくは是れ地上の霜かと</l> <l n="3">頭を挙げて山月を望み</l> <l n="4">頭を低れて故郷を思ふ</l></lg>

これらに基づいて、表 17 に「静夜の思ひ」[80] という唐詩作品の基本的なマークアップ例を示す。これは作者が李白の「静夜の思ひ」の唐詩作品が任意の verse(韻文) とし、書き下し文 (reading) の文体を用い、本文フルテキストの第 1 行目の内容は「牀前月光を看る」というものである。

訓読文と書き下し文における本文フルテキストの TEI マークアップは全体とする唐詩作品の詩の全体要素<lg>や、各行の要素<l>を利用する。訓読文にある任意の漢字は訓点を付ける場合、送り仮名や返り点などの訓点情報は本文の注釈として見なし、要素<seg>を用いる。要素<seg>は句の任意一部のテキスト部分を解説するという意味を持つ TEI 要素である。

TEI ガイドラインでは、訓読文の四隅に付与される訓点を説明できる要素は含まれないので、ここでは独自の属性値として、訓点の種類を日本語の読み方のローマ字で命名することとした。つまり、送り仮名の<seg>要素の属性値は送り仮名の読み方 okurigana、返り点の要素は返り点の読み方 kaeriten として使い、訓読文用の TEI マークアップを行った。表 18 は高等学校に含む訓読文「静夜思」[56] を事例として、送り仮名と返り点に対して、要素<seg>の使い方である。

表 18 では、唐詩作品「静夜思」[56] の第 4 行目の詩句を例として、訓読文の TEI マークアップを行った。詩句の冒頭として、要素<l>を入れた。第 4 行目の説明とする属性値を要素<l>に書き入れ、親文字「低」の右下の送り仮名「レテ」と左下の返り点「レ点」を表記すると、前者が<seg type="okurigana">レテ</seg>とマークアップして、後者は<seg type="kaeriten">レ</seg>とマークアップする。送り仮名と返り点の表記順番は先に送り仮名、次は返り点という順を用いる。この書き方に基づき、訓読文にそれぞれの親文字および訓点情報を対応づける。

表 18: 送り仮名と返り点の事例の書き方

原文資料	使い方
低たレテ レ 頭ヲ 思フ ニ 故☆ 郷ヲ 一 [56]	<head>静　夜　の　思　ひ</head><persName>李白</persName> <lg type="punctuated">... <l n="4">低<seg type="okurigana">レテ</seg><seg type="kaeriten">レ</seg> 頭<seg type="okurigana">ヲ</seg> 思<seg type="okurigana">フ</seg><seg type="kaeriten">二</seg> 故　郷<seg type="okurigana">ヲ</seg><seg type="kaeriten">一</seg></l></lg>

5.4.2 ルビ情報のマークアップ

ルビ情報のマークアップでは、まず、ルビの要素は次の HTML の基本的な要素<ruby>、<rb>、<rt>、<rtc>と<rp>の 5 つの要素があり [81] [82] [83]、これを採用する。ルビの種類はモノルビ、グループルビと熟語の 3 つ [84] がある。要素<ruby>はルビと親文字のまとまりを表す要素であり、それは全体の親文字と小

文字を表す要素である。<rb>は親文字だけを表し、<rt>はルビを表すタグである。そのほか、<rtc>は 1 つの親文字に対して複数のルビ情報を付ける場合に、ルビを<rtc>要素に入れて使えることができる。要素<rp>はルビをサポートしないブラウザ向けのフォールバックを提供できる要素である。筆者は、ウェブ上におけるルビのマークアップの実装 [83] [88] を参考にして、ルビ要素タグを組み合わせて、書き下し文におけるルビのマークアップを行う。ルビ情報のマークアップは基本的に、文章に文字や言葉などを単位として、マークアップすると考える。

唐詩作品の欄に、唐詩作品のタイトルは書き下し文および訓読の両方を一緒に表記した。掲載された本文フルテキストは中学校と高等学校用の教科書に基づいて、選択した。このリストの情報に基づいて、TEI マークアップを行った。

表 19: ルビのマークアップの書き方

No.	原文資料	種類	書き方
1	看(み)る [56]	片側ルビ	<ruby>看<rt>み</rt>る</ruby>
2	牀前(しやうぜん) [56]	片側ルビ	<ruby>牀前<rt>しゃうぜん</rt></ruby>
3	烽(ほう)火(くわ)三月(さんげつ)に連なり [91]	両側ルビ	<ruby>烽<rt>ほう</rt></ruby> <ruby>火<rt>くわ<rtc>カ</rtc></rt> </ruby> <ruby>三月<rt>さんげつ</rt></ruby>に連なり

書き下し文「静夜の思ひ」[80] という唐詩作品をシンプルな事例として、基本的なルビのマークアップの方法は表 19 に示す。表 19 に示す 1 番目の「看る」のような単一の文字を片側ルビで表記する場合に、モノルビとしてマークアップする。表記する際に、ルビ情報を付ける親文字の冒頭に要素<ruby>を書き入れ、ルビは要素<rt>内に表記し、終了タグ</rt>と</ruby>で終わる。

親文字の数が複数の場合はグループルビとし、全体の読みを一緒にルビとして表記する。ルビ情報が片側に配置する 2 番目の場合、表記方法が 1 番目の事例の書き方と同じく、冒頭に要素<ruby>を入れ、片側ルビを表示する要素<rt>要素を書き込んで、終了タグで囲んで書く。

グループルビを両側ルビに配置する場合は、一般論としてはグループルビとして、要素<rt>と<rtc>を用いて、単語として表記するが、簡便にマークアップするために、単一の文字ごとに表記してマークアップすると考える。その書き方は表 19 の 3 番目の事例である。

5.5 マークアップと考察

5.5.1 マークアップ結果

TEI:P5 ガイドライン [8] を基準とし、主に平成 28 年の中学校と高等学校に利用する唐詩作品の本文フルテキストを着目して、TEI マークアップを行った。

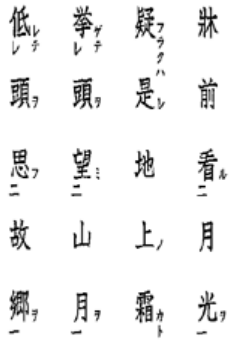
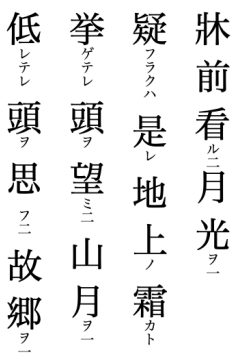
本研究では、主に訓読文と書き下し文における訓点情報やルビ情報をマークアップする方法を中心として実装した。実装したデータとしては平成 28 年度用の中学校の教科書に唐詩作品は異なり 6 首があることに基づき、任意の高等学校の教科書に含む同一の唐詩作品ごとに各 1 首を選択した。表??に示すリストは唐詩作品における TEI マークアップを行ったリストである。

唐詩作品の本文フルテキストの表現に、訓読文、書き下し文などの 4 つの文体がある。漢文および唐詩作品に関わる本文フルテキストは教科書に掲載されるとき、文体によってテキストの表現が異なる。これらの文体の役割は利用者が原文資料を理解しやすくなるために、日本語の文法に合わせて、それぞれの訓点情報や注釈などの要素が付与された。そのため、各文体は要素を付与される状況の違いに従って、教科書の原版面を尊重し、中学校と高等学校の教科書における唐詩作品の本文フルテキストの訓点情報などを確認した上で、デジタル化した。

実装環境は、MacBook Pro で主に Safari ブラウザ バージョン 11.1 を使い、マークアップの内容の表示も試みた。また、訓点情報およびルビは XHTML と CSS [96] を組み合わせて利用した。

訓読文における「静夜思」[56] のマークアップの表示例は表??に示す。原文ページの欄に原文資料を参照して、右側のデジタル化の表示欄にブラウザ表示のスクリーンショットがある。スクリーンショットを示すように大きな文字は親文字であり、送り仮名とレ点、一・二点などの訓点情報も全てを書き入れた。ただし、原文資料と比べると、現時点で返り点や送り仮名の位置は行中央に小さな文字で表示され、全て親文字の下に並べた状態となっている。

表 20: 訓読文におけるデジタル化の表示例

No.	原文ページ	デジタル化の表示 (スクリーンショット)
1	 [56]	

書き下し文におけるルビのマークアップは表 21 に示す。表 21 は書き下し文の 2 つの事例として、白色の背景で教科書に掲載された原文ページであり、右の部分はマークアップの結果である。表 21 の 1 番目は原文資料の書き下し文のルビの表現は基本的な片側ルビであるため、デジタル化は教科書の原文ページに従って同様にした。2 番目は書き下し文における唐詩作品「絶句」[95] の本文フルテキストであり、原文資料の書き下し文のルビの表現によって、親文字に基づいて、片側ルビは 2 列並べる表現もあり、括弧の利用も付けている。現時点で基本的な片側ルビが表示できるが、両列に並べるなどの表示はできていない。

表 21: 書き下し文におけるデジタル化の表示例

No.	原文ページ	デジタル化の表示 (スクリーンショット)
1	牀前月光を看る 疑ふらくは是れ地上の霜かと 頭を挙げて山月を望み 頭を低れて故郷を思ふ [80]	牀前月光を看る 疑ふらくは是れ地上の霜かと 頭を挙げて山月を望み 頭を低れて故郷を思ふ
2	江は碧にして鳥は逾よ白く 山は青くして花は然えんと欲す 今春看す又過ぐ 何れの日か是れ帰年ならん [94]	江は碧にして鳥は逾よ白く 山は青くして花は然えんと欲(ホッ)す 今春看す又過ぐ 何(ス)れの日か是れ帰年ならん

5.5.2 考察

本研究では、平成 28 年度使用の中学校と高等学校の現行教科書に掲載された唐詩作品を研究リソースとして、唐詩作品の本文フルテキストの TEI マークアップ手法を提案した。そのうち、教科書に含まれる唐詩作品の本文フルテキストに基づくマークアップを行い、ブラウザの表示を試みた。

今後の課題としては、(1) TEI マークアップを行った時、返り点や送り仮名などの訓点情報のマークアップには独自の属性値などを用いたが、これを他の古典籍を対象とする研究を共有するためには標準化を検討する必要がある。

(2) ルビの要素と訓点情報の表示の課題については基本的な要素を用いたが、デジタル化の表示にはブラウザ対応の改善や CSS [96] の調整が必要である。(3) 本文フルテキストに含まれる注釈情報もマークアップし、データの一部としたい。

第6章　考察

本研究では文化資源に注目し、唐詩作品を研究対象として、唐詩の構造化に関わる研究を目指す。教育学習のニーズに応えて、唐詩情報に関わる資料を自由に公開共有可能にし、異なる表現を統一的に扱う学習環境を提供したいと考える。

現時点では、日本の中学校と高等学校で学習する唐詩作品、唐詩と関連がある基本的なデータと関連付けて、それらと掲載される教科書の関連関係や外部リンクも繋がって、Linked Open Data 化することを試みた。Linked Open Data 化の構築にあたって、BIBFRAME Model に基づいて、新たなモデル化も行なった。また、唐詩の本文フルテキストに対して、TEI マークアップを用いて、標準的にマークアップした。

それらに対して、唐詩の構造化データを自由にダウンロードできるように、自分で作ったウェブサイト上で公開した。それは学習者が、自由に使えるように提供されているという利益がある。また、唐詩の LOD 化の枠組みによって、唐詩作品に関わるウェブサイトも公開された。唐詩作品にかかる基本的な情報や、作者情報、掲載される教科書のデータ情報を関連できる。そのほか、Wikidata と CTEXT という外部リンクと関連し、多くの情報源と関連できるようにしている。外部リソースの利用については、多くの情報の利用ができ、多言語の実現も表現できると考える。

唐詩の LOD 化 データセットにおける課題として、識別子の問題がある。唐詩作品のデータの並べ方は、掲載教科書に掲載される頻度の数を降順として、唐詩作品 Work は番号 1 から 53 までを表記し、Instance は番号 1 から 374 までを表記した。今使っているデータとして、この番号を使えるが、今後、作品の情報を追加したり、外部から作品情報を活用する場合に、この番号をどのように識別しするかが問題になる。つまり、今使っている識別番号が便利に利用できるが、その適切性は再検討する必要がある。現在唐詩作品の外部 ID として、Wikidata を使っているため、Wikidata に掲載がある作品については、これを使えて識別できる。今後の課題として、有名な DB から、番号を直接採用するのか、あるいは、自分で編

集するのかを再検討する必要があると考える。

唐詩の LOD 化におけるデータセットを公開したウェブサイトを設計したが、ウェブサイトの URI の持続可能性を考える必要がある。現在の URI としている https://w3id.org/tangpoem/は https://w3id.org [97] を利用したリンクであるが、この Web サイトは W3C Permanent Identifier Community Group [63] が提供しているサービスで、目的としては、安全に永続的な URLs を提供している Web サイトである。そのため、このウェブサイトに基づく唐詩の LOD 化の公開用 WEB サイトの URI は永続的に使える。一方で、技術側面から考えると、公開用のウェブサイトは筑波大学のサーバーを利用している。もし今後、筆者が卒業した後でも、他のサーバーに移動したとしても、リダイレクト先を交換すると、このウェブサイトの公開を容易に続けられる。

唐詩の理解の難しさは中国時代の文化や歴史の内容と日本の文化などの違いがあり、教育学習ニーズに基づくシラバス [1] に掲載された項目 (要求) に応じて、唐詩に関わる思想や、感情などの理解しづらい項目があることをわかった。今後の課題としては、公開したデータセットの利活用については、唐詩作品に関わる外部リソースとつなげて、これにより、関連資料を調べる時間をかからずに、外部リソースと繋げて、思想や感情などの抽象的な調べ方が簡単になり、唐詩を読みつつ、この作品の意味付けを直接見られ、唐詩の理解をしやすくなりたいと考える。

また、唐詩の LOD 化における作者情報のデータセットは、人物や事件などに関わる時・空間情報の可視化という研究に、人物や事件などに関わる時・空間情報の可視化を試みたい。作者時期によって、どこでどのような事件が起こるということを年表で可視化する。いろいろな歴史情報を地図にも行なった地理情報を表記して、時・空間情報の可視化になると考える。

本文フルテキストに、訓読文や、書き下し文における訓読情報とルビ情報などの差異があっても、コンテンツ情報を整備し、関連できるようにしたい。

これらに基づいて、本研究では唐詩の構造化に関するデータセットを公開する用のウェブサイトを構築したいが、今後の課題としては、応用アプリケーションを構築したと考える。

第7章　おわりに

本研究では文化資源に注目し、唐詩作品を研究対象として、唐詩の構造化に関わる研究を目指した。教育学習のニーズに応えて、唐詩情報に関わる資料を自由に公開共有可能にし、異なる表現を統一的に扱う学習環境を提供したいと考えた。

本研究の研究内容として、唐詩の LOD 化と TEI マークアップを用いた。具体的には以下の通りになった。

(1) 日本の中学校と高等学校で学習する唐詩作品、唐詩と関連がある基本的なデータと関連付けて、それらと掲載される教科書の関連関係や外部リンクも繋がって、Linked Open Data 化することを試みた。Linked Open Data 化の構築にあたって、BIBFRAME Model に基づいて、新たなモデル化も行った。

(2) 唐詩の本文フルテキストによって、TEI マークアップを用いて、マークアップした。

現時点の唐詩データは手元で取得しやすいデータを利用、研究を進んできた。手軽に扱え、唐詩は典型的な古典作品として、日本の古典や国語の教科書に採用され、教員や生徒により利用するため、日本の中学校と高等学校の国語と古典の教科書に含まれる唐詩作品を研究対象として、データを取得した。また、教育の学習ニーズに対して、学習指導要領によって、唐詩作品を含む教科書の内容は改訂されつつ, 唐詩作品の利用状況も変化するなどの原因があるため、本研究では、平成 28 年度使用の中学校と高等学校の国語と古典の教科書に含まれる唐詩作品を研究対象として、唐詩作品異なり 6 首、延べ 374 首、作者が 21 名に基づくデータセットを構築し、公開した。

データセットの利用によって、教科書に掲載された唐詩作品のタイトルや、本文フルテキストにおける様々な差異があるため、教育学習ニーズに応じて、まず、唐詩作品に関わる基本的な文体、本文フルテキストの種類などの情報を了解できる。また、唐詩作品に関わる作者や、掲載された教科書の情報も調べられた。

外部リソースの関連に対して、生徒らは、教科書に関わる情報を関連でき、日本語版の資料として、多くの情報も調べられるため、教育学習ニーズによって、便

利になる。また、CTEXT のような、中国語の唐詩作品も関連でき、生徒らは、多言語の対応もできるようにした。

今後の課題としては、まず、唐詩作品のモデルに基づいて、多くの唐詩作品を増やし、唐詩作品に関わるデータセットも拡大したいと考える。例えば、今回外部リソースとして利用した CTEXT に関わる全唐詩の事例であり、この唐詩作品のデータに関わる唐詩作品データを拡大し、日本で利用する唐詩作品と合わせて、多様なニーズに応じて、多言語化にしたいと考える。

本文フルテキストの利用としては、(1) TEI マークアップを行った時、返り点や送り仮名などの訓点情報のマークアップには独自の属性値などを用いたが、これを他の古典籍を対象とする研究を共有するためには標準化を検討する必要がある。(2) ルビの要素と訓点情報の表示の課題については基本的な要素を用いたが、デジタル化の表示にはブラウザ対応の改善や CSS の調整が必要である。(3) 本文フルテキストに含まれる注釈情報もマークアップし、データの一部としたい。

これらに基づいて、多くの唐詩作品を関連付けて、LOD 化できるようにして、それらに基づいて、検索や閲覧ができるアプリケーションを構築したい。現時点では公開用のウェブサイトを構築したが、それに基づく応用アプリケーションの設計を考えている。基本的な情報だけではなく、それぞれの関連できる情報を一括でリンクさせて、利用できるように設計し、それより、地理情報や、歴史情報などまでの横断検索も一緒に設計したく、今後の課題として検討する。

また、唐詩の LOD 化における作者情報のデータセットを利用し、作者時期によって、どこでどのような事件が起こるということを年表で可視化し、いろいろな歴史情報を地図にも行なった地理情報を表記して、時・空間情報を可視化する応用アプリケーションも今後の課題である。

謝辞

感謝なり。

いよいよ修士の修論を執筆する段階になるまで、もう日本に来てからも五年目になった。この 5 年間を思い出して、色々なことがあった。慣れるまでが大変だと思ったが、優しい先生らと会えると、幸いと思う。

まず、本論文を作成するにあたり、ご指導を頂いた指導教員の高久 雅生先生に心より感謝致します。入学してから、そろそろ四年間になって、ずっと丁寧かつ熱心にご指導を賜って、日常の議論を通じて、研究を進めていくことができましたこと、多くの知識や示唆も頂戴したことを、誠に深く感謝致したいと思います。

また、進学するために、入試を受けた審査員より、優しく面接してくれた方、各学会で会った先生らのもとでに励めたことを幸せに思います。

次、同研究室に博士後期課程の方や、卒研生らのご意見を伺い、楽しみ生活の方を教えてくれて、ありがとうございます。特に、各学会発表前に、発表練習を行い、いろいろな意見や、真剣なコメントを頂いたのような、研究室での生活を、楽しいと思います。外国人ですが、意識せずに、真面目に手伝って頂いて、ありがとうございます。

最後になりましたが、本論文および、修士課程の段階が無事に修了しましたこと、これまで温かい目で見守ってくれた家族に、深く感謝申し上げます。Daddy and Mummy, I love you all.

発表論文一覧

査読付き国際会議発表

- Yan CONG, Masao TAKAKU. “A TEI Markup for the Contents of Tang Poems”. Japanese Association for Digital Humanities Conference 2018 (JADH2018), Tokyo, Japan, pp.80-81. (2018-09-11).
- Yan CONG, Masao TAKAKU. “Prototype of Linked Open Data Model for Tang Poems”. Japanese Association for Digital Humanities Conference 2017 (JADH2017), Kyoto, Japan, pp.50-52 (2017-09).

その他の口頭発表

- Yan CONG, Masao TAKAKU. “Expanding Tang Poems LOD Dataset with External Resources”. 10th Asia Library and Information Research Group Workshop (ALIRG2018), Fufuoka, Japan. (2018-12-15)
- Yan Cong, Masao Takaku. “TEI Markup for Tang Poems from Japanese Textbooks”. TEI 2018, Tokyo, Japan (2018-09-11).
- 叢艶, 高久雅生. 唐詩作品の本文フルテキストに対する TEI マークアップ手法の提案, 情報知識学会第 26 回年次大会. 東京, 2018 年 05 月 27 日. 情報知識学会誌, Vol.28, No.2, p.174-185. (学生奨励賞を受賞)
- 叢艶, 高久雅生, 唐詩情報の Linked Open Data 化とその利活用の試み. Code4Lib JAPAN カンファレンス 2016, 大阪, 2016 年 09 月 10 日.
- 叢艶, 江草由佳, 高久雅生. 唐詩情報の Linked Open Data 化とその利活用の試み. 人工知能学会 セマンティックウェブとオントロジ (SWO) 第 39 回研究会, 東京, 2016 年 09 月 05 日.

- 叢艶, 高久雅生. 唐詩情報の Linked Data 化の試み. 情報メディア学会第 15 回研究大会, つくば, 2016 年 06 月 25 日, p.17-20, http://hdl.handle.net/2241/00142987.

参照文献

[1] 東京書籍.「新編 新しい国語」年間指導計画作成資料. 平成 28 年度用 中学校. 東書ネット. https://ten.tokyo-shoseki.co.jp/text/chu/keikaku/kokugo/index.htm, (accessed 2018-12-25).

[2] 全唐詩分析系統. http://www.chinabooktrading.com/tang/, (accessed 2018-07-22).

[3] 千田大介. 全唐詩・全宋詩分析系統. 漢字文献情報処理研究. vol.14, 2013, p.173-176.

[4] 荒井 礼. 中・高における教育・教材を含む包括的な漢文教育の問題：大学における漢文教育の場から. 千葉大学教育学部研究紀要, Vol.66, No.2, p.23-32, 2018. http://opac.ll.chiba-u.jp/da/curator/105125/.

[5] 中川 諭. 高等学校「古典・漢文」教材としての中国古典詩の活用. 教職課程センター紀要, vol.2, p.23-30,2017.

[6] トム・ヒース, クリス・チャンバイツァー著, 武田英明 ほか 8 名訳. 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.

[7] 高久雅生. 知識をリンクする技術. 図書館情報学を学ぶ人のために，逸村裕, 田窪直規, 原田隆史 (編), 第 5 部第 19 章, pp.201-213. 世界思想社, 2017.

[8] The Text Encoding initiative:“TEI: P5 Guidelines”. http://www.teic.org/Guidelines/P5/, (accessed 2018-12-25).

[9] “Overview of the BIBFRAME 2.0 Model”. http://www.loc.gov/bibframe/docs/bibframe2-model.html, 2016-04-21, (accessed 2018-12-25).

[10] 中央研究院.「漢籍電子文献」翰典全文検索系統. http://hanji.sinica.edu.tw, (accessed 2018-12-25).

[11] 翰典相関情報. http://hanji.sinica.edu.tw/ftmsw3.html, (accessed 2018-12-25).

[12] 二階堂 善弘. コレクション構築・整備 - 中国・台湾の電子文献について-. 情報の科学と技術, vol.52, no.2, p.79-82. 2002. https://doi.org/10.18919/jkg.52.2_79.

[13] 中国大百科全書出版社.「中国大百科」, 1980, 全 74 巻。

[14] 台湾師範大学図書館.「寒泉」古典文献全文検索資料庫. http://skqs.lib.ntnu.edu.tw/dragon/, (accessed 2018-12-25).

[15] 超星数字図書館. http://zz.ssreader.com/logon.jsp, (accessed 2018-12-25).

[16] 中華電子佛典協會 CBETA. http://www.cbeta.org, (accessed 2018-12-25).

[17] SAT 大正新脩大藏經テキストデータベース. http://21dzk.l.u-tokyo.ac.jp/SAT/, (accessed 2018-12-25).

[18] International Image Interoperability Framework (IIIF). https://iiif.io/, (accessed 2018-12-25).

[19] 今村浩子. 中学校漢文教材における漢文訓読学習の系統性：小学校・高等学校の漢文教材との比較から. 全国大学国語教育学会発表要旨集. vol.127, pp.339-342, 2014-11-08.

[20] 林 教子. 漢字漢文教育の実践とその可能性：言語力の育成に資するために. 早稲田大学, 博士 (教育学) 学位論文. 2012-02. 168p.

[21] Europeana. https://www.europeana.eu/portal/en, (accessed 2018-12-25).

[22] OCLC WorldCat. https://www.worldcat.org, (accessed 2018-12-25).

[23] British Library. https://www.bl.uk/press, (accessed 2018-12-25).

[24] Library of Congress. "Bibliographic framework as a Web of data: Linked Data model and supporting services". 2012-11-21. http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf, (accessed 2014- 11-01).

[25] 江草 由佳, 高久 雅生. 教科書 Linked Open Data(LOD) の構築と公開. 情報の科学と技術. vol.68, no.7, p.361-367, 2018.

[26] 江草 由佳, 高久 雅生. 教科書 LOD. https://jp-textbook.github.io, (accessed 2018-12-25).

[27] 橋詰秋子, 福山樹里. リンクト・オープン・データの利活用: 出版物に関するメタデータと国際書誌コントロールー国立国会図書館における LOD の取り組みー. 情報処理, vol.57, no.7, p.606-611, 2016.

[28] 嘉村 哲郎, 加藤 文彦, 松村 冬子, 上田 洋, 高橋 徹, 大向 一輝, 武田 英明. 芸術・文化情報の Linked Open Data 普及に向けた現状と課題 - LODAC Museum を例に. じんもんこん 2011, vol.2011, no.8, p.409-416.

[29] 横浜の芸術文化情報のオープンデータ, http://yan.yafjp.org/lod, (accessed 2018-12-25).

[30] Library Linked Data Incubator Group Final Report -W3C Incubator Group Report 25 October 2011, https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/, (accessed 2018-12-25).

[31] 上海図書館. “上海デジタル公開サイト”. http://wrd2016.library.sh.cn, (accessed 2018-12-25).

[32] Wikipedia. https://www.wikipedia.org, (accessed 2018-12-25).

[33] 加藤文彦. DBpedia Japanese. 映像情報メディア学会誌. vol.70, no.11, 2016, p.857-861. https://www.jstage.jst.go.jp/article/itej/70/11/70_857/_pdf/-char/ja, (accessed 2018-12-25).

[34] 加藤文彦. DBpedia の現在：リンクトデータ・プロジェクト. 情報管理. 2017. vol.60, no.5, p.307-315. https://doi.org/10.1241/johokanri.60.307, (accessed 2018-12-25).

[35] DBpedia. https://wiki.dbpedia.org, (accessed 2018-12-25).

[36] DBpedia Japanese. http://ja.dbpedia.org/, (accessed 2018-12-25).

[37] 玉川 奨, 香川 宏介, 森田 武史, 山口 高平. 日本語 Wikipedia オントロジーの構築と利用. 2013 年度人工知能学会全国大会, vol.27, p.1-4, 2013. http://www.ei.sanken.osaka-u.ac.jp/sigswo/papers/SIG-SWO-A1203/SIG-SWO-A1203-01.pdf, (accessed 2018-12-25).

[38] 李 乃琦, 劉 冠偉. 一切経音義全文データベースの構造化. 人文科学とコンピュータシンポジウム, p.229-231, 2018.

[39] 高橋晃一. 論理構造と物理構造が混在するテキストの XML によるマークアップに関する考察. 研究報告人文科学とコンピュータ (CH). vol.98, no.6, p.1-5, 2013-05-04.

[40] 山口満, 三輪多恵子. 「漢文テキストの縦書き Web 表示に関する検討」. 豊橋創造大学紀要. Vol.21, p.29-36, 2017-03-30.

[41] Text Encoding Initiative: “The Text Encoding Initiative”, http://www.tei-c.org, (accessed 2018-12-25).

[42] 永崎研宣. デジタル文化資料の国際化に向けて：IIIF と TEI. 情報の科学と技術, vol.67, no.2, p.61-66, 2017.

[43] American Verse Project. University of Michigan Humanities Text Initiative. http://www.hti.umich.edu/, (accessed 2018-07-22).

[44] Eighteenth-Century Poetry Archive. University of Oxford. http://www.eighteenthcenturypoetry.org/, (accessed 2018-07-22).

[45] 後藤 真, 小風 直樹, 橋本 雄太, 小風 綾乃. 永崎 研宣.「構造化記述されたテクストの基盤整備に向けて：延喜式の TEI マークアップを事例に」. 人文科学とコンピュータシンポジウム, p.237-242, 2018.12.

[46] 上原 究一, 永井 正勝, 中村 覚, 中尾 道子, 荒木 達雄, 箕輪 顕量.「図書館における木版本のデジタル化と利活用の可能性 ― IIIF と TEI を用いた「水滸伝」諸版本のデジタル化を通じて ―」. 人文科学とコンピュータシンポジウム, p.381-388, 2018.

[47] 松田訓典, 彌永信美, 永崎研宣, 下田正弘. フランス語仏教辞典『法寶義林』目録のデジタル化とその課題—TEI ガイドラインの適用を通して—, 情報処理学会 研究報告人文科学とコンピュータ (CH). vol.87, no.7, p.1-5, 2010-07-24.

[48] 河瀬彰宏, 市村 太郎, 小木曽 智信. TEI P5 に基づく近世口語資料の構造化とその問題点. じんもんこん 2013, no.4, p.7-12, 2013-12-05.

[49] 高橋洋成, 永井正勝, 和氣愛仁. 「画像, TEI, LOD を用いた文字研究・言語研究のためのプラットフォームの構築」. 情報処理学会研究報告人文科学とコンピュータ, 2015-CH-105(5), p.1-8.

[50] 田中 政司. ジャパンナレッジの挑戦 電子レファレンスツールの可能性. 情報管理, vol.59, no.3, p.172-180. http://doi.org/10.1241/johokanri.59.172, (accessed 2018-12-25).

[51] Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page, (accessed 2018-12-25).

[52] Chinese Text Project (CTEXT). https://ctext.org, (accessed 2018-12-25).

[53] Yan CONG, Masao TAKAKU. “Prototype of Linked Open Data Model for Tang Poems”. Japanese Association for Digital Humanities Conference 2017 (JADH2017), Kyoto, Japan, pp.50-52 (2017-09).

[54] 「送元二使安西」, 精選国語総合, 三角洋一, 池内輝雄, 小町谷照彦 ほか 27 名編, 東京書籍, 2012, p.306.

[55] 「元二の安西に使ひするを送る」, 中学校国語 3, 野地潤家, 新井満 ほか 28 名編. 学校図書, 2015, p.174.

[56] 「静夜思」, 高等学校 国語総合, 東郷克美; 伊井春樹 ほか 28 名編, 第一学習社, 2012, p.336.

[57] 「静夜の思ひ」, 中学校国語 3, 野地潤家; 新井満 ほか 28 名編, 学校図書,2015, p.176.

[58] 「秋夜寄丘二十二員外」, 新編国語総合, 北原保雄 ほか 21 名編, 株式会社 大修館書店, p.314.

[59] 「秋夜寄丘員外」, 精選古典 B 漢文編, 野地潤家; 新井満 ほか 28 名編, 学校図書, 2015, p.21.

[60] 「山中与幽人対酌」, 新編国語総合, 北原保雄 ほか 21 名編, 株式会社 大修館書店, 2012, p.316.

[61] 「山中対酌」, 古典 B 漢文編, 木下資一 ほか 14 名編, 数研出版株式会社, 2013, p.16.

[62] Donald Sturgeon. Enabling digital humanities research and teaching through digital library APIs. Japanese Association for Digital Humanities Conference 2017 (JADH2017), Kyoto, Japan, pp.50-52 (2017-09).

[63] W3C Permanent Identifier Community Group. https://www.w3.org/community/perma-id/, (accessed 2019-01-31).

[64] BIBFRAME vocabulary. http://id.loc.gov/ontologies/bibframe-category.html, (accessed 2018-12-25).

[65] Dublin Core. http://dublincore.org, (accessed 2018-12-25).

[66] Schema.org. https://schema.org, (accessed 2018-12-25).

[67] RDF Schema 1.1. https://www.w3.org/TR/rdf-schema/, (accessed 2018-12-25).

[68] FOAF Vocabulary Specification 0.99. http://xmlns.com/foaf/spec/, (accessed 2018-12-25).

[69] 全唐詩. https://ctext.org/quantangshi, (accessed 2018-12-25).

[70] 唐詩 LOD 化. https://w3id.org/tangpoem/, (accessed 2018-12-25).

[71] 叢艶, 高久雅生. 唐詩情報の Linked Open Data 化. LOD チャレンジ 2016. 2017 年 1 月 15 日. http://www.slis.tsukuba.ac.jp/tangpoem/, (accessed 2018-12-25).

[72] 教科書 LOD プロジェクト. Poorman’s Linked Data Toolkit. https://github.com/jp-textbook/jp-textbook.github.io/wiki/Toolkit, (accessed 2018-12-25).

[73] David Beckett, Tim Berners-lee, Eric Prud'hommeaux, Gavin Carothers. RDF1.1 Turtle. W3C Recommendation 25 February 2014. https://www.w3.org/TR/2014/REC-turtle-20140225/, (accessed 2018-12-25).

[74] Apache Jena Fuseki. https://jena.apache.org/documentation/fuseki2/, (accessed 2018-12-25).

[75] Tim Berners-Lee. Linked Data DesignIssues. https://www.w3.org/DesignIssues/LinkedData.html, (accessed 2018-12-25).

[76] 叢艶, 高久雅生. 唐詩情報の Linked Data 化の試み, 情報メディア学会第 15 回研究大会. つくば市, 2016 年 06 月 25 日, p.17-20. http://hdl.handle.net/2241/00142987.

[77] 叢艶, 江草由佳, 高久雅生. 唐詩情報の Linked Open Data 化とその利活用の試み. 人工知能学会セマンティックウェブとオントロジ-(SWO) 第 39 回研究会, 東京都, 2016 年 09 月 05 日.

[78] 叢艶, 高久雅生. 唐詩情報の Linked Open Data 化とその利活用の試み. Code4Lib JAPAN カンファレンス 2016, 大阪市. 2016 年 09 月 10 日.

[79] 高等学校 国語総合. 東郷克美 伊井春樹 ほか 28 名. 株式会社 第一学習社. 2012 年. 402p.

[80] 中学校国語 3. 野地潤家 新井満 ほか 28 名. 学校図書株式会社. 343p.

[81] 日本規格協会:「日本語文書の組版方法 JIS X 4051」. 日本規格協会, 2004, 206p.

[82] 阿南康宏 ほか編:“日本語組版処理の要件 (日本語版)”, 2012, https://www.w3.org/TR/jlreq/ja/,(accessed 2018-04-13).

[83] Marcin SAWICKI, Michel SUIGNARD, Masayasu ISHIKAWA, Martin DÜRST. Ruby Annotation, https://www.w3.org/TR/ruby/, (accessed 2018-04-05).

[84] Robin Berjon. “ W3C HTML Ruby Markup Extensions”. https://www.w3.org/TR/html-ruby-extensions/, (accessed 2018-04-13).

[85] 「春望」, 現代の国語 2, 中洌正堯 ほか 39 名編. 三省堂, 2015, p.142.

[86] 「春望」, 国語 2, 甲斐睦朗 ほか 27 名編, 光村, 2015, p.152.

[87] 「月夜」, 高等学校 国語総合, 東郷克美; 伊井春樹 ほか 28 名編, 第一学習社, 2012, p.336.

[88] 縦書き web 普及委員会:“ルビの解説とマークアップ方法”. https://tategaki.github.io, (accessed 2018-04-13).

[89] 「黄鶴楼にて孟浩然の広陵に之くを送る」. 現代の国語 2. 中洌正堯 ほか 39 名編, 三省堂, 2015, p.123.

[90] 「黄鶴楼送孟浩然之広陵」. 精選国語総合. 三角洋一, 池内輝雄, 小町谷照彦 ほか 27 名編. 東京書籍, 2012, p.345.

[91] 「春望」, 精選国語総合, 三角洋一; 池内輝雄; 小町谷照彦 ほか 27 名編, 東京書籍, 2012, p.348.

[92] 「春暁」. 現代の国語 2. 中洌正堯 ほか 39 名編, 三省堂. 2015, p.122.

[93] 「春暁」. 精選国語総合. 三角洋一, 池内輝雄, 小町谷照彦 ほか 27 名編, 東京書籍, 2012, p.343.

[94] 「絶句」. 国語 2. 甲斐睦朗 ほか 27 名編, 光村, 2015, p.148.

[95] 「絶句」. 精選国語総合. 北原保雄 ほか 21 名編, 大修館書店, 2012, p.304.

[96] Elika J. Etemad; Koji Ishii:“CSS Ruby Layout Module Level”, https://www.w3.org/TR/css-ruby-1/,(accessed 2018-04-13).

[97] Permanent Identifiers for the web. https://w3id.org, (accessed 2019-01-31).