情報処理学会論文誌 Vol.59 No.2 257–266 (Feb. 2018)

古典籍書誌データ構造に対応したLinked Dataへの半自動変換

吉賀 夏子[1,a)] 只木 進一[1,b)]

受付日 2017年5月9日, 採録日 2017年11月7日

概要：Web 上には，江戸時代以前の書籍である古典籍の書誌および画像データが多数公開されている．これらのデータの活用には，古典籍を構成する序や跋などの情報を含めて，Linked Data 形式などへ構造化することが有効である．古典籍の書誌データは専門用語が多く，構造化のためには，一般には専門家の手作業が必要であった．本稿では，形態素解析と注記の記述様式に特化したスクリプトを活用することで，半自動でこうした作業を実施する方法を提案する．また，具体的なコレクションへの適用を通じて，手法の有効性を検証する．提案手法は，コレクションごとに軽微な調整で適用できる手法であり，古典籍の書誌データを構造化するために有効である．

キーワード：Linked Data，固有表現抽出，非構造化データ，オントロジ，古典籍書誌

Semi-automated Conversion of Bibliographic Data for Japanese Historical Rare Books into Linked Data

Natsuko Yoshiga[1,a)] Shin-ichi Tadaki[1,b)]

Received: May 9, 2017, Accepted: November 7, 2017

Abstract: A large amount of bibliographic data and images of Japanese historical rare books, published during the Edo period and before, have been accessible recently on the Web. For utilization of those data, it is effective to convert information including book parts such as preface and epilogue into well-formed data format like Linked Data. Such conversion, in general, has required handworks by persons with special knowledge. This paper proposes a new semi-automatic method for this task, where the method employs morphological analysis tools and scripts specialized to formats in bibliographic notes. The method is evaluated by applying it to some collections. The method can be applied to a collection with small amount of tunings, and is effective to formalize bibliographic data of Japanese historical rare books.

Keywords: linked data, named entity extraction, unstructured data, ontology, bibliographic data

1. はじめに

Web 上には，近世以前の書籍である古典籍の書誌情報および原本画像が，国内外の様々な組織により多数公開されている [1], [2]．古典籍の研究者にとって，研究対象とする書籍のみではなく，その書籍に関係する周辺の書籍あるいは文化財とその成立に影響を与えた人，人の役割，時，場所などの単語（以下，これらを固有表現 [3] と呼ぶ）について注記事項を含めて精査し，その書籍の位置付けを知ることは重要である [4]．たとえば，その書籍を構成する表紙，序文，跋文（後書き），挿絵といった「部品」の著者，作成年，作成地などの来歴である．

このような事物の関連を分析可能にする手段として，図書館，博物館などの書誌データを発行する組織（以下，書誌データ提供者と呼ぶ）は，書誌から固有表現を抽出し，関連情報へ導く URI [5] を紐付ける Linked Data [6] 化を試みている [7]．これにより，書誌利用者は．Web 上の関連情報を横断的かつ機械的に分析することが可能になる．

[1] 佐賀大学大学院工学系研究科
Graduate School of Science and Engineering, Saga University, Honjo, Saga 840–8502, Japan
[a)] natsukoy@cc.saga-u.ac.jp
[b)] tadaki@cc.saga-u.ac.jp

© 2018 Information Processing Society of Japan

一方，従来の古典籍に対する書誌データは，人が読んで理解することを前提としており，注記事項として記載されることの多い序文や跋文などの部品の著者，作成年など，古典籍の分析に必要な情報を機械可読なデータとして登録する枠組みが整備されていない [8]．対象書籍の序文や跋文の著者といった書籍成立の手がかりになる情報は，凡例の略号を付加した値あるいは非構造的な注記事項として記述される．これら注釈的記述方法は，通常メタデータ作成機関ごとに定められている [9], [10]．

古典籍書誌データの Linked Data を作成するには，固有表現抽出，人名や件名の名寄せ，データ項目および専門用語の整備，内容確認の大半を，古典籍自体の内容に精通した専門家の助けを借りて手作業で行う必要があるため，多大なコストがかかる [11]．特に，部品情報や書籍の来歴などは注記に記載され，データ化対象から残される可能性が高い．

これらの課題を解決するために，既存の書誌情報から，部品情報や来歴を含む古典籍の書誌記述に適応した構造的データに変換し，Linked Data として機械可読にする半自動的手法を提案する．

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．本研究では，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 から収集した研究対象周辺の予備知識を利用する．また，実際の書誌データを構造化する枠組みとして，古典籍を構成する概念およびそれに紐づく固有表現やデータ項目の関係に関する機械可読な構造的知識（以下，オントロジ [12] と呼ぶ）を用いる．提案手法では，対象コレクションに応じた手作業による若干の調整と確認が必要なため，提案手法を「半自動変換」と呼ぶ．

古典籍の専門家によって内容確認済みのオントロジおよび Linked Data は，Web で共有および再利用可能な知識データとなり，類似分野の書誌データへの活用が容易となる．また，部品情報までが構造化された Linked Data を用いて，古典籍研究者は対象古典籍に関連する人脈，時空間をはじめとした様々な書誌情報を，URI を通じ，部品単位で高精度かつ機械的に取得することが可能となる．

本稿の構成は以下のとおりである．Linked Data への半自動変換および書誌データ構造化手法について 2 章で述べる．3 章では，提案手法の有効性について述べる．4 章でまとめおよび議論とする．

2. Linked Data への半自動変換手法

2.1 概要

図 1 に提案手法の概要を示す．目的の Linked Data を出力するには，大別して固有表現抽出（2.2 節），外部機関で作成された標目 URI（以下，外部リンクと呼ぶ）の整備

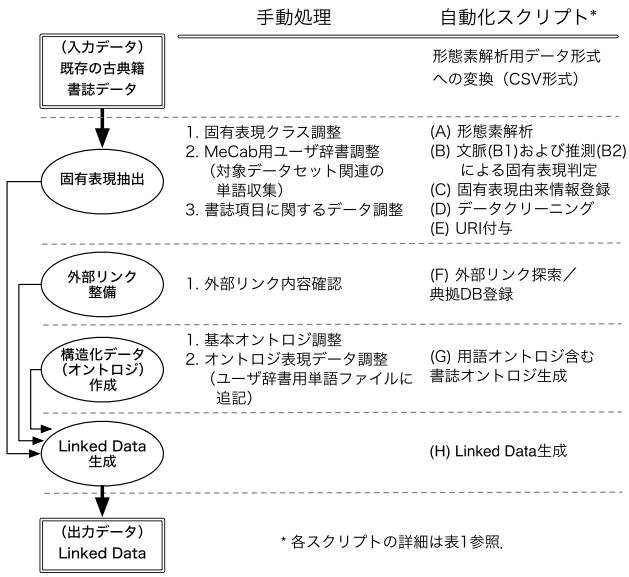

図 1 Linked Data 生成半自動化の手順

Fig. 1 The process of semi-automated generation of Linked Data.

表 1 自動化スクリプトおよびその入出力の概要．図 1 に実行位置を示す

Table 1 Scripts and their inputs and outputs. The generation process is shown in Fig. 1.

略称	用途	入出力
A	形態素解析	（入）書誌データ
		（出）形態素解析結果
B	文脈(B1)および推測(B2)による固有表現判定	（入）同上
		（出）固有表現クラス由来情報（CSV）
C	固有表現由来情報登録	（入）同上
		（出）固有表現クラス由来情報（関係データベース）
D	データクリーニング	（入）同上
		（出）不要データを除去した由来情報
E	固有表現への URI 付与	（入）同上
		（出）基底 URI と由来情報から整形した URI
F	外部リンク探索および典拠データベース*登録	（入）固有表現表記名および英語名
		（出）外部リンクの URI および統制語
G	書誌オントロジ生成	（入）基本オントロジおよびオントロジ表現データ
		（出）書誌オントロジ
H	Linked Data 生成	（入）A の書誌データ，D の固有表現由来情報，F の典拠情報および G の書誌オントロジ
		（出）Linked Data

CSV：コンマ区切りのデータ（Comma Separated Values）．
*外部組織が蓄積する一意に識別可能な単語情報を格納するデータベース．

（2.3 節）および構造化データ（オントロジ）生成（2.4 節）の 3 つの手順で取得したデータ（**表 1** の A，D，F および G の出力データ）を Linked Data 生成スクリプト（表 1 H，

© 2018 Information Processing Society of Japan

2.5 節）に入力することで得られる．各手順には，調整や確認などの若干の手動処理がある．これらは，データ提供者によって実行し，固有表現抽出状況，外部リンクおよびオントロジ表現データの内容確認は古典籍の専門家が手作業で行う．

提案手法は，佐賀大学附属図書館所蔵の市場直次郎コレクション [13] の書誌データセット [14]（以下，*Ichiba* と記す）を対象に開発してきたものである [15], [16]．提案手法により Linked Data 化したデータセット[*1]および SPARQL エンドポイント[*2]を公開している．

2.2 古典籍の構成を表す固有表現の抽出

古典籍の情報を構造化するとは，書誌中のメタデータ [17] を表現する概念の集合，つまり，固有表現の集合を構造化することである．古典籍の場合，現在流通している書籍の書誌に準じ，古典籍の題名，著者名，刊行年など，一部の項目が独立して記述される一方，その他の詳細項目は注記に自然文で記述されている（**表 2**）．

古典籍に特徴的な詳細項目の記述方法は，国立国会図書館の日本目録規則 [9]，国文学研究資料館 [1] の日本古典籍書誌レコード作成要領 [18]，アメリカ議会図書館の目録規則 [10] にまとめられている．たとえば，国立国会図書館の日本目録規則の「古.0.3.2　各書誌的事項の情報源」および「古.0.4　記述すべき書誌的事項」によると，古典籍書誌の主な情報源は，タイトル，責任表示（著者，書写者，書肆（しょし）名など），出版・頒布情報，形態，シリーズおよび注記とされる．特に，記述すべき書誌的事項とされるタイトル，責任表示，出版・頒布に関する情報源は，表紙，見返し，序，跋（ばつ）（後書き）など，書籍の一部とすることが明示されている．本研究では，これらの書籍の一部を示す用語を「部品」と呼ぶ．「古.7.3.1（タイトルに関する注記）」「古.7.3.2（責任表示に関する注記）」においても，データ提供者は，これらの部品情報を，別途定められた略号を用いて注記することが推奨されている．部品の情報を記述する理由は，古典籍では，破損や汚れなどで，書籍名や著者，刊行年，現代書籍で必須となっている事項が確定しない場合に，書名の候補としていくつかの部品にある文字列を採用することになるからである．このように，古典籍の分析に必要な事項の多くが部品名に対応して記載される．

そこで，目録規則に用いられている部品名を抽出する．同時に，書誌データの内容，形状などの客観情報および成立状況など，研究者が求める周辺情報に関連する，部品名以外の固有表現クラスを特定する．固有表現クラス名と特定した固有表現の一部を**表 3**（文献 [16] の表 3 に専門用語および分類の項を追加）に示す．本稿では，部品名およびロール名以外の専門用語および書籍の分類名も，固有表現クラスとした．

注記などの非構造化テキストに対し，句読点などの記号で 1 文単位に区切った後，MeCab [19] による形態素解析を行う（表 1 A）．その際，MeCab 用基本辞書の UniDic [20] に加えてユーザ辞書を用いる．ユーザ辞書作成の準備として，古典籍の読解に必要となる，時節用語，人名，題名，部品，ロール，それ以外の専門用語群および書籍の分類名を，DBpedia [21] や NDLA [22]，専門用語サイトなどの Web 資源からあらかじめ収集し，固有表現クラスに分けた単語リストに登録する [16] とともにユーザ辞書とする．辞書用単語リストには，形態素解析時に固有表現クラス名を識別

表 2　既存の *Ichiba* 書誌データの例
Table 2　An example of bibliographic data of *Ichiba* in the original form.

書誌項目	値
整理番号	45
分類	5-1-(2)-(6)-(6)
分類 1	文学
分類 2	国文
分類 3	小説
分類 4	-
分類 5	洒落本
書名	青楼心得艸
読み	セイロウココロエグサ
書型	-
巻冊	-
編著者	（二世）蓬莱山人作
刊行・書写年次	安政四年序
西暦	1857
巻写	写
注記	安政四年一月蓬莱山人序。 天明五年刊「息子部屋」ノ改題本ナリ。
印記	1
文庫	市場
他	-

表 3　古典籍書誌データの固有表現クラスおよび固有表現の例
Table 3　Classes of named entities, their definitions and examples for bibliographic data of Japanese historical rare books.

固有表現クラス	定義	固有表現の例
時	時を表す語	安政，四年，春，卯月
場所	場所を表す語	江戸，京都，大阪
人名	人名・団体名を表す語	梅暮里谷峨，栄久堂
題	作品名を表す語	息子部屋，青楼心得艸
部品	書誌記述対象となる語	表紙，序，跋，見返し
ロール	人名の役割を表す語	作，画，翻訳，編，写
専門用語	上記以外の専門用語	解題本，折本，後印
分類	分類名を表す語	文学，（分類名の）洒落本

[*1] http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki/tensekiInstances.rdf
[*2] https://www.metabridge.jp/sparql?lu=cwuboswnkghofpjoizlg

© 2018 Information Processing Society of Japan

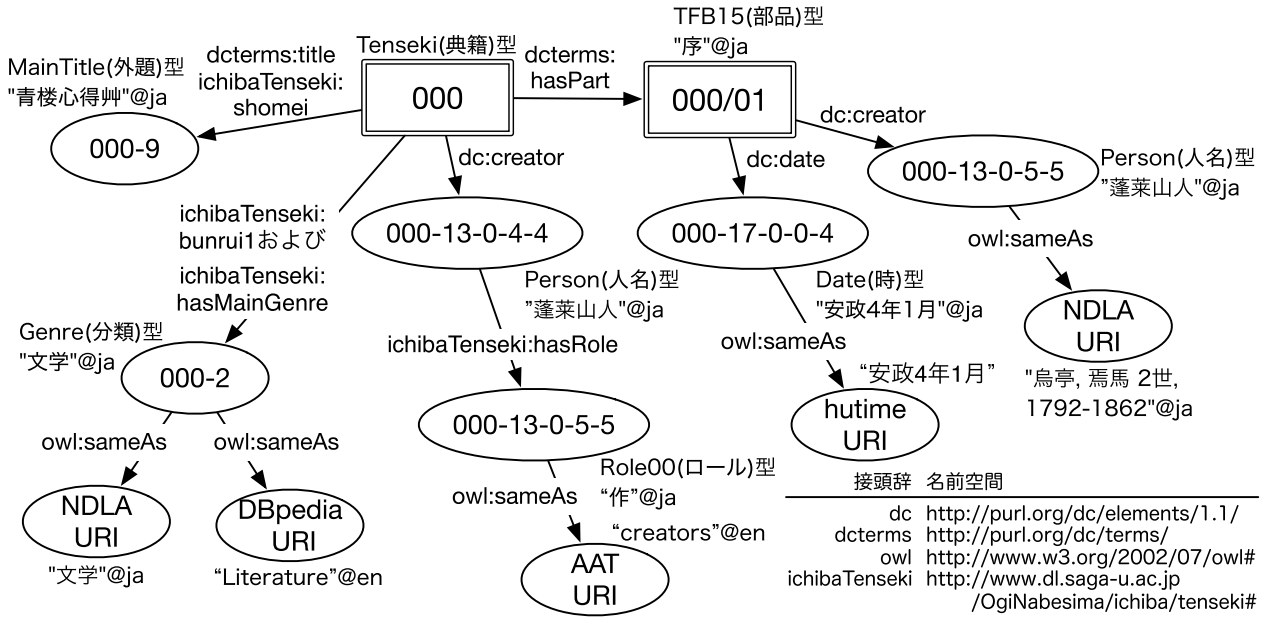

図 2 Linked Data 化された書誌データ（表 2）の一部．「000」「000-13-0-4-4」などを基底 URI（http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki/）末尾に付けることで URI となる．また，ダブルクォート内の文字列は URI に対応する識別用ラベル名をそれぞれ表す

Fig. 2 A part of bibliographic data (Table 2) converted into Linked data. Numerals such as "000" and "000-13-0-4-4" express URIs by suffixed after the base URI (http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki/), and double-quoted strings show corresponding labels respectively.

可能とする解析用フラグを付加する．加えて，辞書に登録する単語が基本辞書上の単語より優先的に判定されるように，各単語に設定されている MeCab 累積コスト値を適宜調整する．部品，ロールおよび専門用語クラスの単語リストはオントロジ表現データの入力にも流用する（2.4 節）．

こうした準備の後，形態素解析を実行し，同定された単語に対して固有表現クラスの判定を行う．まず，各単語の解析結果に出現する固有表現クラス名のフラグで判定する．フラグで判定不可能な単語群に対して，固有表現クラス別に簡易な文脈（B1）および推測（B2）判定スクリプト（表 1 B）を実行する．スクリプト B 群については，コレクションの書誌内容あるいは記述形式に対応する必要がある．*Ichiba* の B1 スクリプトの場合，時クラスでは，年号など時節用語，もしくは，数字と年月日など時の単位の組合せ，「同年」などの省略的記述が文中に出現した場合に該当単語群を「時」と判定する．題クラスでは，鉤括弧に括られた文字列を「題」と判定する．人名クラスでは，事前に対象分野および時代に使用された屋号（屋，堂，亭など）あるいは呼称（氏，山人，主人など）が文中に出現した場合に直前の単語を含めて「人名」と判定する [15]．その他のクラスについては，スクリプト B 群を実行しない．

人名については，書誌利用者ユーザ辞書に登録可能な情報が少なく，固有表現抽出精度が他クラスより低い [16]．そのため，パターン解析で判定できない単語に対し，人名を推測する B2 スクリプトを利用して可能な限り抽出する．たとえば *Ichiba* 注記では，1 文に含まれる単語数が通常 10 語以下と簡潔で，1 文に付き 1 つの出来事を含む．また，表 2 注記「安政四年一月蓬莱山人序。」のように，部品名の直前に人名が並ぶ文体が多い．そこで，1 つの注記文から，時，部品，ロールなどの判定済み単語および助詞，記号類などの除外可能な MeCab の品詞を差し引き，残った単語群が部品名の直前にあればそれを人名と判定する [16]．

スクリプト B 群を実行後，固有表現クラス名，レコード番号，項目番号，文に割り振った番号，文中の単語位置情報，単語および URI の項目で構成される固有表現の由来情報を，固有表現由来情報データベースに自動登録（表 1 C）する．このとき，由来情報データベースに格納されている単語位置情報を利用して，単語の重複や連続，誤りなどを検出し，1 つの単語に再構成するデータクリーニング（表 1 D）[*3] を実行する．

データクリーニング後，書誌データを図 **2** に示す Linked Data にするために，元の書誌データの項目および値に加え，抽出したすべての固有表現および固有表現間の関係，固有表現クラスに対し，一意の URI を付与する．そこで，各固有表現に対し，表 1 E のスクリプトで基底 URI [*4] およ

[*3] 文献 [16] におけるクリーニングスクリプト．
[*4] *Ichiba* の基底 URI は http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki/である．

© 2018 Information Processing Society of Japan

び固有表現由来情報に基づく URI（図 2 参照）を生成し，由来情報データベースに登録する．固有表現クラス自体の URI，固有表現クラス間の関係を表す URI，書誌データ項目を表す URI は，基本オントロジ作成時にオントロジ作成ツールで作成する．部品，ロールおよび専門用語については，2.4 節で用語オントロジを生成する際（表 1 G）に，各固有表現クラス名とリスト上の番号を組み合わせた URI を自動生成する．これらの対象コレクションに関する URI が外部情報資源の URI と等価であればスクリプト G 実行時に紐付ける（表 1 F）．

また，書誌データの各項目に対応する，項目名，固有表現クラス名，Dublin Core [23] などの標準メタデータスキーマなどをリストにした書誌項目に関するデータを作成する．書誌項目に関するデータは，2.5 節の Linked Data 生成時に利用する．

2.3 固有表現と外部情報資源の接続

固有表現として抽出した単語に対して，DBpedia，The Getty Vocabularies [24]，VIAF [25]，NDLA などの外部組織で定義されている Linked Data 形式の単語情報に紐付ける．Linked Data 形式の単語情報には，Web で一意となる識別子として URI（外部リンク）が紐づく．そこで，事前に用意した典拠データベースに，抽出した固有表現，固有表現の出典元，スクリプト（表 1 F）で自動探索した外部リンクおよび表記名を蓄積する．その後，典拠データベース上の単語定義情報を手作業で確認する[*5]．

2.4 書誌オントロジの定義および構築

2.2 節で特定した固有表現クラスは，古典籍書誌の内容を構成する概念の集合である．これらは，オントロジ言語 OWL [27] を用いて，それらの関係を構造化することで，機械的処理が可能となる．

古典籍の読み解きを目的とした周辺情報の検索においては，書籍の構成部品に紐づく情報が求められる [28] ため，本研究では，目録規則における書誌記述対象である部品を単位として構造化する．そのため，EDM [29]，CIDOC CRM [30]，$FRBR_{OO}$ [31] など，作品構成を作品成立にまつわる出来事として記述可能な文化財用形式オントロジの考え方を基に，古典籍世界での書誌学のルールに従って簡易な書誌オントロジを定義する．これにより，たとえば，古典籍自体の著作情報とは別に，部品の表紙に表紙画の作者名，作者のロール，題名，作成日，作成地など，部品自体の著作情報を書誌データとして登録可能となる．

[*5] 探索自体は容易であるが，人名，地名などの固有名詞には表記揺れや別名が多く，特定の自動化は困難であり [11]，専門家による同定が必要である．適切な典拠データの同定を支援するために，たとえば，Open Calais [26] のように，候補となる典拠に対して関連する周辺情報から確率を求めるような機械学習的仕組みの導入も選択肢の 1 つとなる．

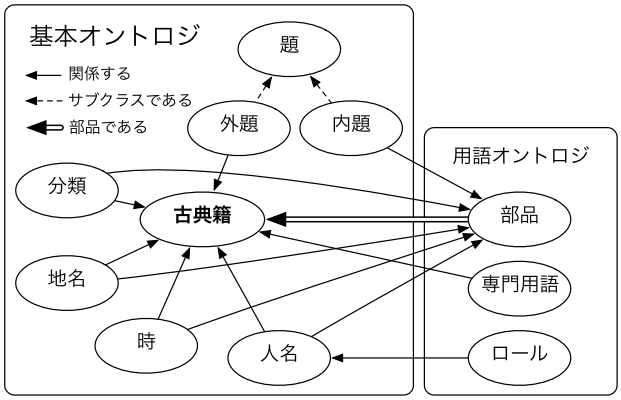

図 3 簡易な古典籍書誌オントロジ

Fig. 3 The simple bibliographic ontology for Japanese historical rare books.

提案書誌オントロジ（**図 3**）は，古典籍と固有表現クラスおよび各メタデータ項目の関係を記述する基本オントロジと，部品，ロールおよび専門用語について詳細な関係を記述する用語オントロジを組み合わせた構成とする．基本オントロジは，古典籍とその部品の関係のほかに，代表書名となる外題と題名の候補である内題の関係のような，古典籍分野特有の関係を考慮しながら，オントロジ構築ツール [32] を用いて作成する．

次に，表 3 に示す部品，ロール，専門用語の固有表現クラスが内包する固有表現群について，その定義に加え，等価および包含の関係を用語オントロジとして定義する．その際，各固有表現にオントロジ表現データと呼ぶ，固有表現の関係を表現した略記号を便宜的に付加したデータを，2.2 節で作成した MeCab ユーザ辞書用単語リストに追加する．オントロジ表現データは，固有表現表記，関係記号，検索用英語表記からなる．なお，固有表現表記および検索用英語表記については，等価な語彙定義を持つ外部リンクを探索し，典拠データベースに格納するスクリプト（表 1 F）で利用する．

図 4 にオントロジ表現データの例を示す．たとえば，固有表現「序」では，あらかじめ序の URI を TFB15[*6] と定義する．この TFB15 は，固有表現「序中」と等価であるため，等号「＝」を付加する．後に，等号が付加された固有表現は，等号の左側の固有表現，すなわち序と等価として登録する．序と題の複合語である「序題」の場合は，序と題の複合語と等しいことを等号とコロン「：」で＝ (序：題) と表現する．序題は，書誌学のルールで「内題」のサブクラスとなるため，包含記号「⊂」を用いる．

オントロジ表現データおよび外部リンク情報を作成後，スクリプト（表 1 G）を実行する．スクリプト G に基本オントロジおよびオントロジ表現データを入力すると，各固

[*6] TFB は Terminology for bookbinding（製本用語）の略．序の URI は http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB15 に示す．

© 2018 Information Processing Society of Japan

オントロジ表現	オントロジ表現データをスクリプトGで変換した結果
序,,preface	<owl:Class rdf:about="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB15"> <rdfs:subClassOf rdf:resource="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB"/> <rdfs:label xml:lang="ja">序</rdfs:label> <rdfs:label xml:lang="en">preface</rdfs:label> <owl:sameAs rdf:resource="http://vocab.getty.edu/aat/300055032"/> <owl:sameAs rdf:resource="http://www.wikidata.org/entity/Q670787"/> </owl:Class>
序中,=序,	<owl:Class rdf:about="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB20"> <rdfs:subClassOf rdf:resource="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB"/> <owl:equivalentClass rdf:resource="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB15"/> <rdfs:label xml:lang="ja">序中</rdfs:label> </owl:Class>
序題, =(序:題)|⊂内題, title from preface	<owl:Class rdf:about="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB21"> <owl:equivalentClass> <owl:Class> <owl:intersectionOf rdf:parseType="Collection"> <owl:Class rdf:about="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TFB15"/> <owl:Class rdf:about="http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki#TitleOfWork"/> </owl:intersectionOf> </owl:Class> </owl:equivalentClass> <rdfs:label xml:lang="ja">序題</rdfs:label> <rdfs:label xml:lang="en">title from preface</rdfs:label> </owl:Class>

接頭辞	名前空間
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
owl	http://www.w3.org/2002/07/owl#
xml	http://www.w3.org/XML/1998/namespace/

図 4 オントロジ表現の例
Fig. 4 Examples of the ontological expressions.

有表現に対し，クラス型，基本オントロジで設定したクラス間の関係定義，対象の固有表現と等価である外部リンクについての情報を読み込み，OWL で記述した用語オントロジを自動生成する．その後，基本オントロジと合わせて，書誌オントロジ全体を出力する[*7]．これにより，元の書誌データから，提案書誌オントロジ中に記載されているデータ構造に準拠した Linked Data に変換可能となる．

2.5 提案オントロジに対応する Linked Data への変換

最後に，表 2 の書誌データに 2.4 節の提案オントロジを適用し，RDF 形式のデータに自動変換する（表 1 H）．スクリプト H に書誌データを入力すると，提案オントロジで定義した古典籍どうし，古典籍と固有表現，固有表現どうしなどの関係定義，スクリプト E および F で生成した URI 情報，書誌項目に関するデータ上に記載されている標準メタデータを参照して，書誌データ項目と値の関係を示すデータの組（トリプル）を出力する．

3. 提案手法の有効性

本章では，小規模な古典籍書誌データ *Ichiba*（佐賀大学附属図書館所蔵の市場直次郎コレクション，典籍の部）および大規模な古典籍書誌データの *Kokubun*（国文学研究資料館所蔵の日本古典籍データセット [33]）に提案手法を適用し，Linked Data 化についての有効性を検証する．解題は，古典籍の解説であり，注記と記載内容が異なるため，本稿では形態素解析の対象としない．

検証は，以下の項目について，人手のみで Linked Data 化を行う場合（以下，人手と呼ぶ）と比較した場合に本手法による Linked Data 化が妥当かどうかの観点から行う．

(1) 提案手法の利点
(2) 注記項目から固有表現抽出を行う意義
(3) MeCab による単語切り出し精度
(4) 固有表現抽出精度
(5) スクリプト群の汎用性
(6) 抽出単語の外部リンクへの接続の容易さ
(7) 生成した Linked Data の評価
(8) 提案手法の適用可能性

なお，半自動化処理にあたり，MeCab ユーザ辞書（2.2 節）およびオントロジ表現データ（2.4 節）は，*Kokubun* と *Ichiba* 間で共通の語彙が多いため，*Ichiba* のデータを基に追加して作成した．

3.1 提案手法の利点

古典籍書誌から Linked Data を人手で作成する作業では，Linked Data 化作業に携わる人が古典籍の書誌学，目録記述規則および周辺知識に加え，データ変換に関連する情報処理技術にも精通することが前提となる．提案手法では，人手で管理していた専門的知識をデータ化し，煩雑な情報処理を可能な限り自動化することで，一定の質を保つ Linked Data を生成できる．その結果，古典籍の専門家は確認作業に注力できる．

[*7] 生成結果は http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki/index.rdf から取得可能．

© 2018 Information Processing Society of Japan

3.2 注記項目から固有表現抽出を行う意義

Ichiba および *Kokubun* の書誌項目で，固有表現抽出を行う項目を**図 5** に示す．これらの書誌項目は，記録すべき項目に書誌学的な慣例が存在するため，類似している．

分類以外の固有表現クラスに注目する（図 5）．*Kokubun* の項目は，*Ichiba* より粒度が細かく，題（内題に相当する記載書名），ロールおよび部品に相当する項目が独立して設けられている．加えて，*Kokubun* の注記項目にも，*Ichiba* と同様に様々な固有表現が記載されている．また，*Ichiba* および *Kokubun* ともに，登録された書誌の半数以上に注記の記載がある（**表 4**）．

表 5 に，全項目中に対する注記中の固有表現出現数の割合を書誌データおよび固有表現クラスごとに示す．*Kokubun* には，場所，題（記載書名），部品，ロールに対する独立した項目があるため，*Ichiba* に比べて，注記に記載されている固有表現は少ない．また，専門用語および分類を除く固有表現クラスごとに固有表現出現割合の平均をとると，

Ichiba	**固有表現クラス**	***Kokubun***
刊行・書写年次 西暦 注記	時	書写年 注記
注記	場所	書肆（書肆名に続いて記載） 注記
編著者 注記	人名	記載著者名表記 記載著者名よみ 記載著者名他等 注記
書名 読み 注記	題	統一書名 統一書名よみ 記載書名表記 記載書名よみ 注記
注記	部品	記載書名種別 注記
注記	ロール	記載著者名役割 注記
注記	専門用語	注記
分類 分類1 分類2 分類3 分類4 分類5	分類	オープンデータ分類 書誌種別

図 5 固有表現抽出を行う項目
Fig. 5 Entries for extracting named entities.

表 4 *Ichiba* および *Kokubun* における全レコード数および注記ありレコード数．カッコ内は全レコード中の出現割合
Table 4 The numbers of the records and those with note fields, for *Ichiba* and *Kokubun* collections. Their ratios are shown in brackets.

書誌データ	全レコード数	注記ありレコード数
Ichiba	222	210 (94.59%)
Kokubun	981	645 (65.75%)

Ichiba の場合には 70.48%，*Kokubun* の場合には 24.15%の書誌が，注記に固有表現を含んでいる．

3.3 MeCab による単語切り出し精度

MeCab 搭載の評価ツールを用いて，*Ichiba* および *Kokubun* の注記文に対する単語切り出し精度を計測する（**表 6**）．*Kokubun* は，MeCab 基本辞書 UniDic [20] のみを用いた解析精度が *Ichiba* に比べて高い．その理由は，*Kokubun* の注記文が，国立国会図書館などと同様に，目録規則で採録方法が明確に示されており，*Kokubun* の表記揺れは *Ichiba* よりも少ないことから，単語切り出しが容易となるためである [18]．

3.4 固有表現抽出精度

表 1 で示した各スクリプトについて，固有表現クラスごとに抽出精度を測定した結果を**表 7** および**表 8** に示す．ただし，時クラスのクラス判定では，たとえば，「安政」「4 年」「1 月」など，連続した単語群で時の意味をなす場合，「安政 4 年 1 月[*8]」で正解とする．

A として示した MeCab ユーザ辞書の作成による抽出精度は，両コレクションのすべてのクラスで上昇した．特に，題，部品，ロールおよび専門用語クラスの抽出精度は 0.8 以上となった．一方，時，場所および人名クラスでは，時，人名および題に適用した B 群を適用することで，B 群未適用である場所クラスの判定精度も上昇した．*Ichiba* 人名クラスに適応した B1 および B2 により，判定精度は 0.429 か

表 5 各固有表現ごとの，全項目中に現れる固有表現数に対する注記中のそれの割合
Table 5 The ratios of words in the note fields to those in the records, for each class of named entities.

固有表現クラス	*Ichiba*	*Kokubun*
時	53.94% (130/241)	33.08% (171/517)
場所	98.37% (121/123)	4.39% (36/820)
人名	67.35% (427/634)	38.53% (1,040/2,699)
題	72.93% (598/820)	12.18% (902/7,405)
部品	91.29% (346/379)	27.52% (1,213/4,407)
ロール	38.99% (124/318)	29.20% (535/1,832)
専門用語	100.00% (134/134)	97.85% (683/698)
分類	0.00% (0/875)	0.00% (0/2,274)

表 6 *Ichiba* および *Kokubun* に対する MeCab の単語切り出し精度
Table 6 The numbers of words and the corresponding F values.

書誌名	コーパス単語数	F 値 ユーザ辞書無	F 値 ユーザ辞書有
Ichiba	5,406	0.594	0.863
Kokubun	21,730	0.840	0.993

[*8] 漢数字は時クラス用スクリプト B で算用数字に変換済み．

© 2018 Information Processing Society of Japan

表 7 *Ichiba* のスクリプト（表 1）および固有表現クラス別固有表現抽出精度 [1]（F 値）

Table 7 The F-values of named entity extraction for each script (Table 1) and each class of named entities in *Ichiba*.

	スクリプト名（表 1 参照）				
固有表現クラス	N[2]	A	A+B1	A+B1+B2	A+B1+B2+D
時 [3]	0.000	0.018	0.282	0.282	0.895
場所	0.579	0.685	0.768	0.768	0.768
人名	0.213	0.365	0.429	0.518	0.716
題	0.000	0.812	0.952	0.952	0.952
部品	0.869	0.911	0.911	0.911	0.911
ロール	0.698	0.800	0.838	0.838	0.838
専門用語	0.701	0.904	0.908	0.908	0.908

[1] 分類は固有表現抽出の対象外.
[2] N はユーザ辞書なしでの精度.
[3] 時クラスは複数単語の組合せを正解としている.

表 8 *Kokubun* の固有表現クラス別固有表現抽出精度 [1]（F 値）

Table 8 The F-values of named entity extraction for each class of named entities in *Kokubun*.

	スクリプト名（表 1 参照）			
固有表現クラス	N[2]	A	A+B1	A+B1+D
時 [3]	0.000	0.236	0.620	0.883
場所	0.050	0.533	0.571	0.973
人名	0.060	0.760	0.748	0.775
部品	0.609	0.990	0.990	0.990
ロール	0.618	0.981	0.981	0.981
専門用語	0.698	0.973	0.973	0.973

[1] 題および分類は注記からの固有表現抽出の対象外.
[2] N はユーザ辞書なしでの精度.
[3] 時クラスは複数単語の組合せを正解としている.

ら 0.518 に改善された．ただし，人名クラス用 B1 は *Ichiba* に対して有効である一方，*Kokubun* では精度の上昇は見られなかった．

D のデータクリーニングの実行では，時および人名クラスの単語群で，解析により形態素に分割されていた氏と名，年月日などが適切な固有表現にまとめられるため，検索精度は両コレクションともに上昇した．文字列の一括訂正についても，抽出精度を高める一助となった．たとえば，*Kokubun* の場所クラスでは 1 文字の単語を多数判定し，適合率を低下させていたが，注記文中ではすべて誤りであるため一括して削除することで，検索適合率が高まり抽出精度が上昇した．

3.5 スクリプト群の汎用性

表 1 に示すスクリプト群については，B 群および D を除き，図 1 の手順およびスクリプトの構成を変更せずに，各コレクションに対し若干の調整を必要とする．本稿で比較対象とした *Kokubun* は，コレクションの分野，時代，書名，著者など *Ichiba* と重複しているため，*Kokubun* 用ユーザ辞書およびオントロジ表現データに用いる単語リストは *Ichiba* で作成したリストに新たな単語を追加することで作成する．

時クラスに対する B1 の適用では，両コレクションとも同じスクリプトを用いる．*Kokubun* の題クラス判定用 B1 では，注記文中に内題や外題を含む題クラスに相当しない題などが記載されるため，他のクラスを判定するための材料として用いる．人名クラス判定用 B1 に対しては，コレクションに合わせて文脈判定用の記述パターンに変更を加える．B2 は，部品の前に人名が位置するパターンを探索するが，*Kokubun* の注記文には該当するものがないため適用しない．

このように，B 群は，各コレクションでの書誌記述内容および目録記述規則に合わせて固有表現の抽出を試みる場合の補助手段として用いる．仮に，時クラス以外の判定で B 群の適用を省いても，図 1 で提案するスクリプト群は実行可能である．

D は，固有表現由来情報データベースに対し，単語の不要な重複や分割の調整および訂正を行う．これらも，軽微なスクリプトの変更で対応可能である．

表 9 書誌データごとの固有表現クラス別固有表現出現数 [1] に対する接続リンク数の割合．カッコ内は接続リンク数および固有表現出現数

Table 9 The ratios of named entities linked with external resources to the total occurrences.

固有表現クラス	*Ichiba*	*Kokubun*
時	91.70% (221/241)	95.94% (496/517)
場所	81.30% (100/123)	84.15% (690/820)
人名	44.64% (283/634)	11.67% (315/2,699)
題 [2]	- (-/820)	- (-/7,405)
部品	76.78% (291/379)	31.00% (1,366/4,407)
ロール	94.97% (302/318)	85.59% (1,568/1,832)
専門用語	96.27% (129/134)	33.95% (237/698)
分類	79.09% (692/875)	42.26% (961/2,274)

[1] 全項目中の固有表現出現数.
[2] 現時点は古典籍著作典拠データにアクセス不可である.

3.6 抽出単語の外部リンクへの接続の容易さ

抽出単語の外部リンクへの接続の容易さを**表 9** に示す．2.3 節で述べたとおり，単語に対応する外部リンクの候補を Web から自動抽出することが容易な一方で，外部リンクへの接続を確定することは，書誌提供者の知識量に依存する．本研究では，古典籍に詳しくない著者が同定を行っているため，専門知識の必要な人名，部品，専門用語，分類の単語への接続リンク数が少ない（表 9）．一方，時，場所およびロールについては，hutime [34] および DBpedia などで Linked Data を取得可能な単語が増えているため，同定は困難ではない．

© 2018 Information Processing Society of Japan

3.7 生成した Linked Data の評価

提案手法によって半自動的に切り出された固有表現群は，2.4 節で構成した書誌オントロジの下，自動的に Linked Data 化することができる．つまり，部品や来歴を含む書誌情報を整合的に Linked Data とすることができる．

3.8 提案手法の適用可能性

前述のように，提案手法は *Ichiba* を対象として構築を行った．この手法を *Kokubun* へ適用する段階では，辞書用単語追加，文脈・推測による固有表現判定およびデータクリーニングスクリプトに対し，*Kokubun* の注記規則に合わせて修正を行った．このように，提案手法は対象コレクションに対する軽微な調整で適用することができる．

4. まとめと議論

古典籍書籍データの主に注記に含まれる，序や跋などの部品に関する情報を半自動的に Linked Data 化する手法を提案した．2 つの規模の異なるコレクションに対して適用することで，その有効性を示した．コレクションへの対応の差は，注記の記述規則に応じたスクリプトの変更と形態素解析用辞書の追加であった．

2.2 節の固有表現抽出手法は，非構造化テキスト文から固有表現となる単語を抽出する先行研究を参考にしている．たとえば，川村らは，形態素解析ツールの辞書に研究対象文の分野に近い単語を収集したものを利用した [35]．また，コーパスからある程度の法則性を見い出し，類似パターンを基に単語がどの固有表現クラスに該当するか分類した．Ulicny は，抽出した単語の周辺にある単語との共起性を測定し，確率で判定する機械学習の手法を提案した [26]．

2.4 節の書誌オントロジの構築では，オントロジを基にした Linked Data の参考例として，EDM [29] を文化財書誌データに適用した Europeana [36] プロジェクトがある．EDM は，文化財オントロジ CIDOC CRM [30] を基に，多様なアーカイブ資料を統合しながら，作品を構成する部品と出来事の両方についての記述を行う．また，図書資料情報を提供する WorldCat.org [37] では，schema.org [38] による語彙定義と BIBFRAME [39] を基にした資料間の関係明示を行う．

しかし，近年では，Linked Data の作成自体は数多く行われている一方で，上記汎用オントロジあるいは独自ドメインオントロジの利用例は少なく，Linked Data に使用されている語彙を構造化したデータが公開されていないことが課題となっている [40]．

そこで，これまで著者は，*Ichiba* を実例に，古典籍の構成に対応する部品情報を取得可能な Linked Data の作成を半自動化する手法を提案してきた．本稿では，研究対象として *Ichiba* に加えて，*Kokubun* を用いて半自動化を実行し，提案手法が別の古典籍書誌データでも実行可能であることを示した．その結果，提案手法が手作業によるコスト削減に有効であることを示した．ただし，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．

特に，本稿で扱ったコレクションと類似分野では，1 度作成した MeCab ユーザ辞書，OWL 形式の語彙情報の再利用が可能であるため，軽微な追加作業を通じて，他のコレクションの分析に応用することが可能である．

参考文献

[1] 国文学研究資料館：日本古典籍総合目録データベース，入手先 ⟨http://base1.nijl.ac.jp/~tkoten⟩（参照 2017-04-06）．

[2] 国立国会図書館：国立国会図書館デジタルコレクション，入手先 ⟨http://dl.ndl.go.jp/⟩（参照 2017-04-07）．

[3] Grishman, R. and Sundheim, B.: Message Understanding Conference-6: A Brief History, *Proc. 16th Conference on Computational Linguistics - Volume 1*, *COLING '96*, Stroudsburg, PA, USA, pp.466–471, Association for Computational Linguistics (online), DOI: 10.3115/992628.992709 (1996).

[4] 保立道久：歴史のひろば：歴史データベースの将来と歴史知識学：コンピュータはただの便利な道具か，歴史評論，Vol.702, pp.80–89 (2008).

[5] Berners-Lee, T., Fielding, R. and Masinter, L.: Uniform Resource Identifier (URI): Generic Syntax, available from ⟨https://tools.ietf.org/html/rfc3986⟩ (accessed 2017-04-28).

[6] Heath, T. and Bizer, C.: *Linked data: Evolving the web into a global data space*, Morgan & Claypool Publishers (2011).

[7] 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 論文集，pp.409–416, 情報処理学会 (2011).

[8] 相田　満：和漢古典学のオントロジ，勉誠出版 (2007).

[9] 国立国会図書館：『国立国会図書館「日本目録規則 1987 年版改訂 3 版」和古書適用細則』（2012 年 1 月），入手先 ⟨http://warp.da.ndl.go.jp/info:ndljp/pid/9484238/www.ndl.go.jp/jp/library/data/wakosho201201.pdf⟩（参照 2017-04-07）．

[10] Tsuchitani, I. and Fujishiro, M.: *Descriptive Cataloging Guidelines for Pre-Meiji Japanese Books*, available from ⟨http://www.eastasianlib.org/cjm/jrb/2011_japaneseRareBooksCatalogingGuidelines.pdf⟩ (accessed 2017-04-07).

[11] 山田太造，遠藤珠紀，荒木裕行，井上　聡，久留島典子：前近代日本史史料から人名を集める，じんもんこん 2016 論文集，Vol.2016, pp.159–164 (2016).

[12] Gruber, T.R.: Ontology, *Encyclopedia of Database Systems*, Liu, L. and Özsu, M.T. (Eds.), Springer-Verlag (2008).

[13] 井上敏幸（編）：市場直次郎コレクション目録，佐賀大学附属図書館・地域学歴史文化研究センター (2007).

[14] 佐賀大学附属図書館：佐賀大学貴重書コレクション，入手先 ⟨http://www.dl.saga-u.ac.jp/OgiNabesima/⟩（参照 2017-04-06）．

[15] 吉賀夏子，只木進一：古典籍データ中の非正規記述からの周辺情報抽出とその構造化，じんもんこん 2016 論文集，Vol.2016, pp.147–152 (2016).

[16] 吉賀夏子，渡辺健次，只木進一：古典籍書誌注記文からの作品構造および関連人物のつながりを明らかにする周辺情報抽出，人文科学とコンピュータ研究会（CH），Vol.2016-CH-109, No.3, 情報処理学会 (2016).

© 2018 Information Processing Society of Japan

[17] 谷口祥一：メタデータの「現在」：情報組織化の新たな展開，初版 edition，勉誠出版 (2010).
[18] 増井ゆう子，喜多妙子：国文学研究資料館和古書目録データベースの作成，入手先 ⟨https://www.nijl.ac.jp/pages/event/seminar/images/H26-kotenseki09.pdf⟩（参照 2017-04-21）.
[19] 工藤 拓：MeCab: Yet Another Part-of-Speech and Morphological Analyzer (ver. 0.996), 入手先 ⟨http://taku910.github.io/mecab/⟩（参照 2017-04-07）.
[20] 国立国語研究所：UniDic（バージョン 2.1.2），入手先 ⟨http://pj.ninjal.ac.jp/corpus_center/unidic/⟩（参照 2017-04-06）.
[21] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia, *Semantic Web*, Vol.6, No.2, pp.167–195 (2015).
[22] 国立国会図書館：Web NDL Authorities, 入手先 ⟨http://id.ndl.go.jp/auth/ndla⟩（参照 2017-04-07）.
[23] DCMI: Dublin Core® Metadata Initiative, available from ⟨http://dublincore.org⟩ (accessed 2014-04-28).
[24] The Getty Research Institute: The Getty Vocabularies, available from ⟨http://vocab.getty.edu⟩ (accessed 2017-04-07).
[25] OCLC：バーチャル国際典拠ファイル，入手先 ⟨http://viaf.org⟩（参照 2017-04-07）.
[26] Ulicny, B.: Constructing Knowledge Graphs with Trust, *METHOD 2015: 4th International Workshop on Methods for Establishing Trust of* (*Open*) *Data* (2015).
[27] W3C: OWL 2 Web Ontology Language Primer, 2nd Edition, available from ⟨http://www.w3.org/TR/owl2-primer/⟩ (accessed 2017-04-07).
[28] 橋口侯之介：江戸の本屋と本づくり 続和本入門，平凡社 (2011).
[29] Isaac, A.: EDM Primer, available from ⟨http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf⟩ (accessed 2017-04-10).
[30] Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M.: Definition of the CIDOC Conceptual Reference Model, available from ⟨http://cidoc-crm.org/docs/cidoc_crm_version_6.2.pdf⟩ (accessed 2017-04-10).
[31] Bekiari, C., Doerr, M., Bœuf, P.L. and Riva, P.: Definition of $FRBR_{OO}$: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism, available from ⟨http://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf⟩ (accessed 2017-04-10).
[32] Stanford Center for Biomedical Informatics Research: Protégé, available from ⟨http://protege.stanford.edu⟩ (accessed 2017-04-07).
[33] 国文学研究資料館：日本古典籍データセット，入手先 ⟨http://codh.rois.ac.jp/pmjt/⟩（参照 2017-04-07）.
[34] 関野 樹（総合地球環境学研究所）：HuTime，入手先 ⟨http://www.hutime.org⟩（参照 2017-04-24）.
[35] 川村隆浩，大須賀昭彦：TEXT2LOD，人工知能学会論文誌, Vol.advpub（オンライン），DOI: 10.1527/tjsai.LOD-21 (2016).
[36] 欧州委員会：Europeana, available from ⟨http://www.europeana.eu⟩ (accessed 2017-04-10).
[37] Online Computer Library Center: WorldCat.org, available from ⟨http://www.worldcat.org⟩ (accessed 2017-05-01).
[38] W3C Schema.org Community Group: schema.org, available from ⟨http://schema.org/⟩ (accessed 2017-05-09).
[39] アメリカ議会図書館：Bibliographic Framework Initiative, available from ⟨http://www.loc.gov/bibframe/⟩ (accessed 2017-05-01).
[40] 高間康史，矢部彩佳，石川 博：サンプリングに基づく構造推定を用いた LOD 視覚的分析支援システム，人工知能学会論文誌，Vol.32, No.1, pp.WII–B_1–11（オンライン），DOI: 10.1527/tjsai.WII-B (2017).

吉賀 夏子 （学生会員）

1993 年佐賀大学大学院農学研究科修士課程修了．2005 年佐賀大学総合情報基盤センターでウェブ関連システムおよび電子図書館構築業務に従事．2015 年佐賀大学大学院工学研究科博士前期課程修了．現在，同大学院工学研究科博士後期課程に在籍中．人工知能学会学生会員．

只木 進一 （正会員）

1987 年東北大学大学院理学研究科博士後期課程修了．理学博士．1990 年佐賀大学理工学部助教授．2000 年佐賀大学学術情報処理センター教授．2006 年同センター長．2013 年佐賀大学大学院工学系研究科教授．学術情報システムの管理運用技術，計算科学を専門とする．日本物理学会，応用数理学会各会員．

© 2018 Information Processing Society of Japan