Semi-automated Generation of Linked Data from Unstructured Bibliographic Data for Japanese Historical Rare Books

Natsuko Yoshiga
Graduate School of Science and Engineering,
Saga Univesity
Saga 840-8502, Japan
natsukoy@cc.saga-u.ac.jp

Shin-ichi Tadaki
Graduate School of Science and Engineering,
Saga Univesity
Saga 840-8502, Japan
tadaki@cc.saga-u.ac.jp

ABSTRACT

A large number of bibliographic data and images of Japanese historical rare books have been published on the Web. For constructing structured bibliographic data, such as in a format of Linked Data, the data providers need to extract structured information from notes of the bibliography, which are written in Japanese natural texts with domain-specific terms. These jobs have been usually performed by persons with special knowledge. In this paper, the authors propose a semi-automated method to convert natural texts in real bibliographic data into Linked Data. As a part of the method, a simple ontology of key elements (named entities) in bibliographic data is constructed along general bibliographic rules for historical heritage. The ontology also has a capability to describe relations between a book and its parts. This allows mechanical access to information such as a creator of a cover picture, preface and so on. Finally, a script creates connections from named entities to URIs which describe headings or glossaries provided online by public organizations.

CCS CONCEPTS

•**Applied computing** →**Digital libraries and archives;** •**Information systems** →*Semantic web description languages; Multilingual and cross-lingual retrieval;* •**Computing methodologies** →Ontology engineering;

KEYWORDS

unstructured data, named entity, Linked Data, ontology, conceptual model, bibliographic description

ACM Reference format:
Natsuko Yoshiga and Shin-ichi Tadaki. 2017. Semi-automated Generation of Linked Data from Unstructured Bibliographic Data for Japanese Historical Rare Books. In *Proceedings of International Conference on the Preservation and Long-term Management of Digital Materials, Kyoto, Japan, September 2017 (iPRES2017),* 5 pages.

1 INTRODUCTION

Historical rare books, published before Edo period (until 1868) in Japan, have been collected by various organizations worldwide and those bibliographic data have also been published on the Web. For instance, one of the biggest collections of historical rare books in Japan [1], containing about 555,300 sets of bibliographic data, has been available online since February 2017.

Most of conventional bibliographic data are in tabular form. Some important information for researchers, such as creators of cover pictures and book owners, are written in notes with Japanese natural texts using domain-specific terms. As a result, those data are human readable, but not structured. In other words, those data are not suitable for analyses by software applications.

Recently, some organizations [2, 3] have tried to transform conventional bibliographic data written in Japanese into Linked Data. These attempts require a large amount of cost because Japanese natural texts are converted into structured data by specialists familiar with domain-specific terms.

In this paper, we propose a semi-automated method for transforming bibliographic data, especially texts in notes, into Linked Data form. We call the method as *semi-automated* because the method partly requires inspection by human for improving accuracy. The method consists of the following steps: 1) extracting named entities which are keywords with semantic meanings in the data, 2) giving ontological structure to the data for showing the structure of a book, and 3) creating connections from the extracted named entities and values to other URIs such as headings created by NDLA [2], VIAF [4], AAT [5], DBpedia [6] and HuTime [7].

The proposed method is applied to an actual online collection: one of digital rare book archives at Saga University Library [8], Ichiba Naojiro Collection. Its bibliographic data (*Ichiba*) are originally described in tabular form (Table 1). The fields of *Ichiba* basically obey standard bibliographic rules proposed in [9–11]. Finally, the data are transformed into structured RDF format, which are accessible as a SPARQL endpoint [12].

There are preceding studies on automatic transformation from natural texts into Linked Data [13–15]. In addition to methods for word extraction employed in those studies, we use a novel concept to construct structural information by referring to standard bibliographic rules of historical rare books and conceptual models for cultural heritage resources which are recommended in CIDOC CRM [16], FRBR_{OO}[17], EDM [18] and so on. These conceptual models are employed to represent relations between a book and named entities which are essential keywords for describing the historical rare books.

The proposed model connects a historical rare book with its contributors and establishing processes as structured bibliographic data. Thus, those structured bibliographic data help researchers to study historical rare books through background information. In addition, Linked Data enables the data to be shared through the Web.

Table 1: An Example of Bibliographic Data in *Ichiba* (with Translation)

Fields	Values
Serial number	45
Genre	5-1-(2)-(6)-(6)
Genre1	文学 (literiture)
Genre2	国文 (Japanese literiture)
Genre3	小説 (novels)
Genre4	(no data)
Genre5	洒落本 (Sharebon)
Title	青楼心得艸 (Seirokokoroegusa)
Pronunciation of the title	セイロウココロエグサ
Size	(no data)
Volume	(no data)
Authors and editors	（二世）蓬莱山人**作** (Houraisanjin the Second *created*)
Date of publication (the Japanese era)	安政四年**序** (Ansei 4, *a preface was created*)
Date of publication (the Christian Era)	1857
Printed or handwriting	写 (handwriting)
Notes	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。 (Houraisanjin *created a preface* in January of Ansei 4. It is a retitling of "Musukobeya" *published* in Tenmei 5.)
Seal of ownership	1
Collections	市場 (Ichiba)
Miscellaneous	(no data)

Table 2: Named Entity Classes and Examples of Instances in *Ichiba*

Classes	Samples of Instances
Date	安政四年 (Ansei 4 nen, 1857), 卯月 (April)
Place	江戸 (Edo), 京都 (Kyoto)
Person	蓬莱山人 (Houraisanjin), 梅暮里谷峨 (Umebori Kokuga)
Role[1]	作 (authors), 画 (painters), 編 (editors)
Title	息子部屋 (Musukobeya), 青楼心得艸 (Seiroukokoroegusa)
Genre	文学 (literature), 洒落本 (Sharebon)
TFB[2]	表紙 (covers), 序 (prefaces), 跋 (endnotes)
Term[3]	改題本 (a retitling book), 後印 (reprints)

[1] Forms of contributions.

[2] TFB is a class of words which represent parts in a book. This class name comes from "terminology for bookbinding".

[3] A class of technical terms except for ones of TFB and Role.

2 NAMED-ENTITY EXTRACTION AND ASSIGNMENT OF URI

Named entities [19] are characteristics of a book, such as publication dates and places, authors, roles, titles, genres and predefined technical terms (Table 2). Some are in standard fields of bibliographic data in tabular formats and others are described as natural texts in notes. The predefined technical terms consist of terms from terminology for bookbinding, types of publishers, and other domain-specific technical terms, which are defined obeying cataloging guidelines for libraries [10, 11]. Those named entities help researchers to explore related information.

For describing historical rare books, it is important to refer to some parts of the target book using a word in terminology for bookbinding. A bibliographic description for a historical rare book usually contains, in addition to its main text, historical information, such as creators of parts of the book, relations to other books, owners, publishers. Such information referring to parts of the book with words in terminology for bookbinding is very important for researchers to study origins and establishing processes of the book. Moreover, some historical rare books lack explicit titles because of damages or loss. In those cases, those books are called with some texts, such as the first sentence of the main text or titles in other parts of the book. In consideration of these backgrounds, descriptions on parts of books with words in terminology for bookbinding are very important.

For extraction of named entities from natural texts, Japanese language has specific difficulties because it is an agglutinative language which has no delimiters between words. For overcoming the difficulties, a morphemic analysis tool called MeCab [20] and pattern search programs are used in the proposed method. In notes of bibliographic data, Japanese texts consist of around 10 words per sentence because the texts in data are required to be in laconic style. The performance of MeCab is improved if a user prepares user dictionaries suitable for target texts. In the proposed method, we prepare a dictionary containing words relevant to historical rare books collected through digital archives and web databases [21].

In the previous experiment on *Ichiba* [21], 951 sentences in notes from 222 records were analyzed by MeCab. An evaluation tool of MeCab with 4449 prepared correct answers was employed to measure the accuracy of the analysis. The F-measure as an accuracy was 0.863 with the user dictionary, comparing to 0.594 without it. The result shows that use of a user dictionary relevant to historical rare books is very effective. The rest of words which were not properly analyzed need to be corrected by hand.

In addition, the extracted words are classified by categories into named entities. Finally, the words tagged by one of the named entities are assigned to URIs generated by location of the words in the sentence and registered in a database [21].

3 CONSTRUCTION OF INTERNAL DATA STRUCTURE

In section 2, we discuss that some predefined technical terms in bibliographic rules are important to indicate characteristics and structure of historical rare books. The proposed method describes the relations of these terms in OWL [22], where we define a ontology suitable for the data. Ontologies written in OWL can mechanically provide knowledge in domains. To model an ontology, it is necessary to collect classes, properties and terms for constructing relations.

First, basic relations (R1, Figure 1) between Tenseki class, a conceptually top class for historical rare books, and named entities (Table 2) are constructed with Protégé [23].

In detail, we first define TFB (terminology for bookbinding), Role (roles) and Term (other technical terms) classes according to standard Japanese bibliographic glossaries and rules [9–11]. In real operation, ontological relations between TFB, Role and Term classes (R2, Figure 2) are prepared in an original laconic style using symbols for referring to entries (Table 2 in [24]). After completion of R2, a script is employed to translate R2 into OWL format.

TFB, Role and Term classes also have domain-specific relations: variants, inclusions, intersections and other relations shown in Figure 2. For instances, Tenseki class has instances of Title class which relates to MainTitle and TitleOfWork classes (Figure 1 and Figure 2). MainTitle class indicating titles written on outer covers refers to main titles in standard bibliographic descriptions. On the other hand, TitleOfWork class describes titles taken from texts in inner parts of a book, which are used as a title if cover titles are unreadable or lost. TitleOfWork class is described with words in TFB. In Figure 2, "序題", a title on a preface, belongs to both TitleOfWork and Preface classes. In other words, this relation is described as an intersection between TitleOfWork and Preface in the ontology.

In the similar way of the relation between TitleOfWork class and classes for parts described with words in terminology for bookbinding, Role class describes roles and activities of people in Person class. Term class offers definitions and structural information of technical terms.

Finally, R2 is combined to R1 to supply the detail domain-specific ontology in addition to the basic ontology of Japanese historical rare books. The combined ontology [25] works fine on Protégé.

4 TRANSFORMATION OF REAL BIBLIOGRAPHIC DATA INTO ONTOLOGICALLY VERIFIED LINKED DATA

4.1 Applying Ontology onto Real Data

A bibliographic data, *Ichiba* is transformed into RDF format, which obeys the ontology defined in section 3. First, a main URI are assigned to the top level of a book. URIs which belong to instances of main fields are directly connected to the main URI.

Then sub URIs assigned to instances of TFB (parts of a book) class are also connected to the main URI. Other URIs for named entities extracted from each record of *Ichiba*, are assigned to the sub URIs. By this operation, each sub URI, which is a part of a book, is also able to be described with its origins. The model of the relations among the main URI, the sub URIs and named entities are shown in Figure 3.

Table 3: Ratios of the Number of Linked Entities to That of Words in Bibliographic Data

Types	Total number[1]	Linked number	Ratio
Date	241	221	91.70%
Place	123	100	81.30%
Person	634	283	44.64%
Role	318	302	94.97%
Title[2]	598	(Not applicable)	
Genre	875	692	79.09%
TFB	379	291	76.78%
Term	134	129	96.27%

[1] Total number of words appeared in *Ichiba*.

[2] No standard Linked Data services for rare books' titles are available in Japan.

4.2 Linking Instances of Named Entities to External Headings

To connect instances of named entities to external headings, a script crawls vocabularies on the Web. The script can retrieve candidates of headings relevant to the instances and store them in the prepared database. Actually, NDLA, VIAF, AAT, DBpedia and HuTime are chosen for exploration because these vocabularies have relevant words to Japanese historical rare books. For instance, "安政4年1月" in notes of Table 1 belongs to Date class. It can be automatically accessible to an entry through HuTime. For cases of person names, "（二世）蓬莱山人" in Table 1 can be linked to NDLA and VIAF. As the result of retrieval, you can get three candidates for "蓬莱山人" because descriptions in bibliographic data are often ambiguous.

In Table 3, the results of queries for each type of named entities show that most of words except for Title class can be connected to a URI in external Linked Data services.

Person class, however, is relatively difficult to be linked to external entries, because it is difficult for non-specialists to choose a proper person from candidates. Generally, a name does not uniquely refer to a person. In addition, a person holds several names particularly before Edo period. As a result, we need to infer a proper person by peripheral information relevant to the bibliographic data by hand. This is one of the remaining problems to transform bibliographic data into Linked Data. In the proposed method, after automatically extracting candidates, proper headings especially for person names are chosen from them by hand.

5 DISCUSSIONS

In this section, we discuss the efficiency of the proposed method for other bibliographic data of cultural heritage.

First, the method of extracting named entities using MeCab and pattern search programs is common for analyzing Japanese natural texts. Moreover, most of natural texts in bibliographic notes are short and written in similar patterns. For example, in notes of Table 1, named entities in one sentence line up in order of date, person and explanatory complements [21]. Those complements often

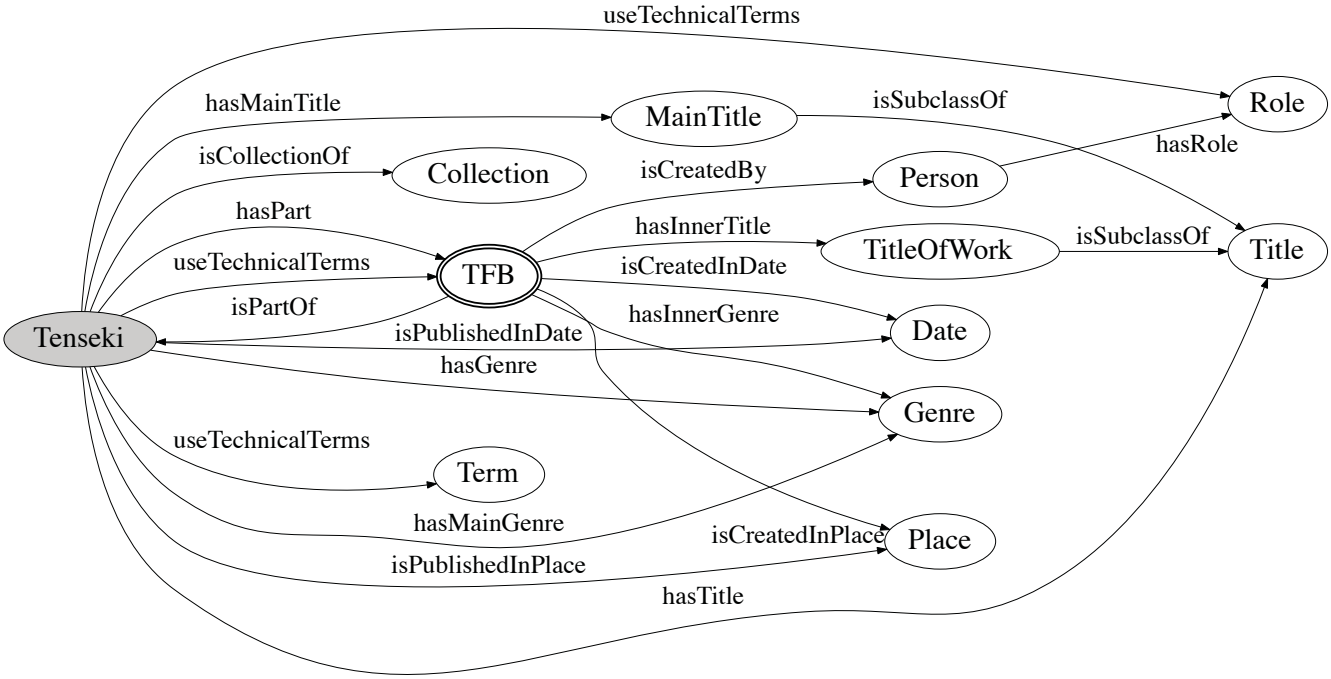

Figure 1: Basic relations (R1) between Tenseki (a rare book) and other named entity classes [24]. A relation between Tenseki and TFB classes is particularly important to describe characteristics of the book.

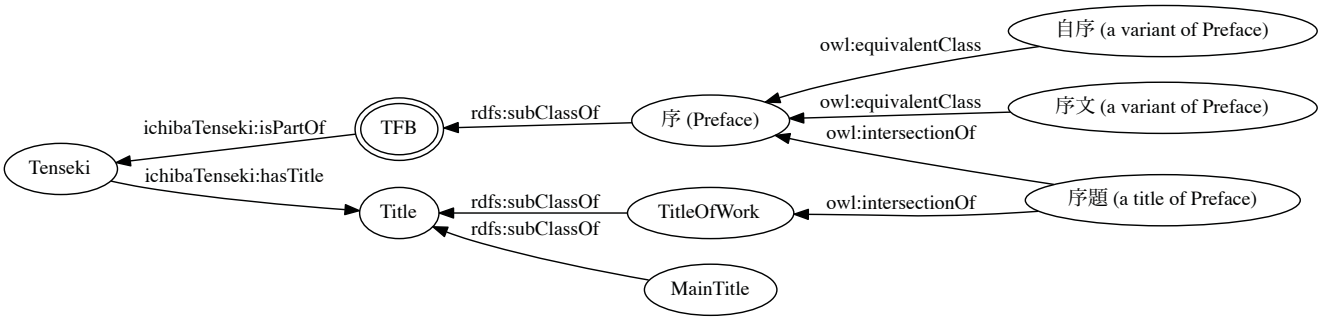

Figure 2: An example of subclasses of TFB. One of the subclasses is 序 (Preface) class shown in this figure. Variants such as 自序 and 序文 classes are equivalent to 序 class. Title class has subclasses, MainTitle and TitleOfWork (titles of TFB) such as 序題 (titles of Preface). 序題 class belongs to both Preface and TitleOfWork classes. Role and Term classes also have subclasses in the same manner of TFB class.

include predefined terms corresponding to one of instances in TFB, Role and Term classes. The rules for description in notes' fields, of course, depend on the collection itself. In any case, you can write your pattern search program for your target collection by referring to its description guidelines and real examples in notes.

Second, the proposed ontology has a simple but clear function to convey processes of contribution for a rare book by clarifying relations between a book and its parts. The ontology obeys frameworks of formal ontologies for cultural heritage such as CIDOC CRM, FRBR_{OO} and EDM. These ontologies are capable of modeling detail structures in the domain. Particularly, they can express relations between cultural objects and their peripheral events. Accordingly, the proposed ontology can describe historical rare books with peripheral events in general.

Finally, according to the result in Table 3, automated selections of proper headings to each named entity are mostly successful. In TFB, Role and Term classes, sufficiently many words have already been provided by AAT, DBpedia or NDLA in Linked Data. Person class, however, requires much handwork than other classes because it is difficult to precisely choose correct candidates. This

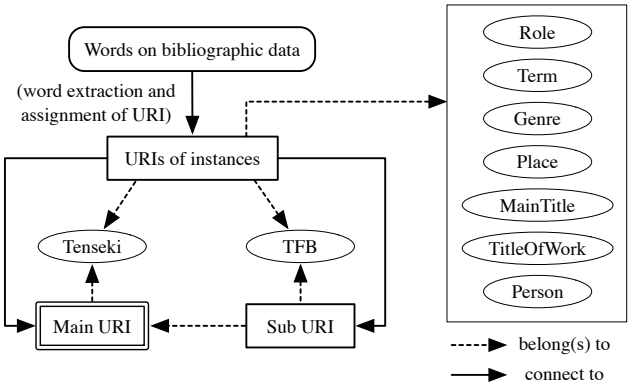

Figure 3: A data model of the relations among a main URI, sub URIs and named entity classes (ovals). A main URI is able to have some sub URIs.

situation is hoped to be improved because international organizations such as VIAF constantly construct public databases of headings on the Web. Credibility of URIs to external information should be validated by specialists of the domain.

In summary, the proposed method in section 1 is applicable to other cultural heritage not only *Ichiba*. The proposed method should be reviewed from various view points in history and linguistics.

6 CONCLUSIONS

In this paper, we proposed a semi-automated method to convert unstructured bibliographic data of Japanese historical rare books into Linked Data. For this purpose, we extract keywords relating to named entities through natural language processing and automatically assign them to URIs generated by locations of the words. The proposed method generates structured, machine-readable data, which obey a simple ontology under standard description rules. The method was applied to *Ichiba* and worked well. The method employs general characteristics in bibliographic data of Japanese historical rare books. Therefore, the method can be applied for other collections.

REFERENCES

[1] National Institute of Japanese Literature. Union catalogue of early japanese books (in japanese). http://base1.nijl.ac.jp/~tkoten/, 2006.

[2] National Diet Library. Web ndl authorities. http://id.ndl.go.jp/auth/ndla, 2011.

[3] Makoto Goto. Constructing of resource sharing method for promoting integrated studies of cultural and research resources conference. In *Jinmoncom 2016 Symposium*, volume 2016, pages 103--110. Information Processing Society of Japan, December 2016.

[4] Online Computer Library Center. The virtual international authority file. http://viaf.org, 2010.

[5] J. Paul Getty Trust. The getty vocabularies. http://vocab.getty.edu.

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. Dbpedia--a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2014.

[7] Tatsuki Sekino. Hutime (linked data). http://datetime.hutime.org/.

[8] Saga University Library. Saga university rare books collection (in japanese). http://www.dl.saga-u.ac.jp/OgiNabesima.

[9] National Institute of Japanese Literature. *Users Guide (in Japanese)*, December 2006.

[10] National Diet Library. National diet library japanese bibliographic rules for japanese rare books (tranlated title from japanese). http://warp.da.ndl.go.jp/info:ndljp/pid/9484238/www.ndl.go.jp/jp/library/data/wakosho201201.pdf, January 2012.

[11] Isamu Tsuchitani and Manae Fujishiro. Descriptive cataloging guidelines for pre-meiji japanese books. http://www.eastasianlib.org/cjm/jrb/2011_japaneseRareBooksCatalogingGuidelines.pdf, 2011.

[12] Natsuko Yoshiga. Sparql endpoint of *Ichiba* on metabridge. https://www.metabridge.jp/sparql?lu=cwuboswnkghofpjoizlg, February 2017.

[13] Takahiro Kawamura and Akihiko Ohsuga. Text2lod - development of web api for triplification of text information -. *Transactions of the Japan Society of Artificial Intelligence*, advpub, 2016.

[14] Kazuhiro Tashiro, Men Oh, Kenji Koshikawa, Satoshi Nishimura, Takeshi Morita, Shinichi Nagano, Yuich Sei, Hiroyuki Nakagawa, Yasuyuki Tahara, Takahiro Kawamura, and Akihiko Ohsuga. Experiment of social-mass media comparison with linked data. *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, (2N1-OS-10d-3):1--4, 2013.

[15] Brian Ulicny. Constructing knowledge graphs with trust. *4th International Workshop on Methods for Establishing Trust of (Open) Data*, 2015.

[16] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. Definition of the cidoc conceptual reference model. http://cidoc-crm.org/docs/cidoc_crm_version_6.2.pdf, March 2015.

[17] International Working Group on FRBR and CIDOC CRM Harmonisation. Definition of $frbr_{oo}$: A conceptual model for bibliographic information in object-oriented formalism. http://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf, November 2015.

[18] Antoine Isaac. Edm primer. http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, July 2013.

[19] Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466--471, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[20] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. http://taku910.github.io/mecab/, 2013. ver. 0.996.

[21] Natsuko Yoshiga, Kenzi Watanabe, and Shin-ichi Tadaki. Extracting peripheral information from bibliography information of historical rare books to clarify those structures and related people. *IPSJ SIG Computers and the Humanities*, 2016-CH-109(3):7, January 2016.

[22] The World Wide Web Consortium. Owl 2 web ontology language primer (second edition). http://www.w3.org/TR/owl2-primer/, December 2012.

[23] Stanford University. Protégé. http://protege.stanford.edu, April 2013.

[24] Natsuko Yoshiga and Shin-ichi Tadaki. Extracting and constructing contextual information from unstructured texts in bibliographical data of japanese historical rare books. In *Jinmoncom 2016 Symposium*, volume 2016, pages 147--152. Information Processing Society of Japan, December 2016.

[25] Natsuko Yoshiga. A downloadable ontological data of *Ichiba*'s vocabularies. http://www.dl.saga-u.ac.jp/OgiNabesima/ichiba/tenseki, 2017.