貴重書書誌の注記から抽出したメタデータによるオントロジー構築および書誌・美術関連 Linked Data と連携した検索システム構築

吉賀夏子[*1], 渡辺健次[*2], 只木進一[*3]

[*1]佐賀大学大学院工学研究科, 12573015@edu.cc.saga-u.ac.jp
[*2]広島大学大学院教育学研究科, wtnbk@hiroshima-u.ac.jp
[*3]佐賀大学総合情報基盤センター, tadaki@cc.saga-u.ac.jp

概要

歴史的文献等の貴重書の書誌は、現代書誌に用いられている Dublin Core などの汎用メタデータでは十分に記述できない。それは貴重書の利用者が現代書の書誌データ項目に加え、文化財としての書物の成立過程を含む書誌学的なデータ項目により強い関心があるためである。このような情報は、しばしば「注記」と呼ばれる項目に記載されている。そこには、利用者が参考にする書肆や序作成者などのデータが多数含まれている。また、貴重書書誌データ全般に、客観的なデータだけでなく、データ登録者の主観的判断や推論による情報も含まれている。以上のことから、注記に移動している主に書誌学関連の情報を記載するために、Dublin Core 及び FRBRoo を拡張したオントロジーを構築した。特に、主観的判断や推論に対応可能なアノテーションの付加による正規化を提案する。このオントロジーの有効性を確認するために、佐賀大学所蔵の貴重書コレクションのデータを格納し、外部の書誌・美術関連 Linked Data と連携した検索システムを構築した。このような方法により、貴重書類の書誌情報の正規化を可能とし、オンラインデータ構築と利用を促進することができる。

キーワード: 貴重書, 書誌, オントロジー, FRBRoo, Linked Data, 注記, 正規化

Constructing Ontology and a Visual Search System for Bibliography of Historical Documents and Art Objects by Extracting Metadata from Miscellaneous Fields

Natsuko YOSHIGA[*1], Kenzi WATANABE[*2], Shin-ichi TADAKI[*3]

[*1]Graduate School of Science and Engineering, Saga University, 12573015@edu.cc.saga-u.ac.jp
[*2]Graduate School of Education, Hiroshima University, wtnbk@hiroshima-u.ac.jp
[*3]Computer and Network Center, Saga University, tadaki@cc.saga-u.ac.jp

Abstract

Generic bibliographical metadata sets, such as Dublin Core for modern literatures, are insufficient to describe characteristics of historical documents and art objects. The reasons include that researchers of those historical objects are interested in properties relating to, for example, cultural backgrounds of those objects. Those properties have been described in note or remark sections. Those sections often contain useful information, such as publishers and writers of prefaces. Besides, bibliographical data for historical objects contain subjective judgments and inferences by editors of data. For normalized descriptions of information written in note and remark sections, we propose a new ontology, which is the extension of Dublin Core and FRBRoo. In particular, we introduce alternative annotations, which describe subjective judgments and inferences. For evaluating effectiveness of the proposed ontology, we construct a data system for storing metadata of historical documents owned by Saga University Library, where the system provides searching functions of those documents with cooperation of external Linked Data on historical and art objects. The methodology promotes construction and utilization of on-line database systems of historical and art objects, by normalized descriptions of annotations not described in generic metadata sets.

Keywords: historical Documents, Bibliography, Ontology, FRBRoo, Linked Data, Miscellaneous Field, Normalization

1. はじめに

従来、研究者が江戸時代以前の貴重書を検索する場合、図書館・博物館等の目録を調べ、必要だと思われる書誌を発見した場合、現地まで出かけて実物を確認していた。近年は貴重書の分野でもデジタルアーカイブが多数構築され、文字での情報だけでなく、いくつかの参考画像を頼りに確認することができるようになった。つまり、このようなデジタルアーカイブは、現地に赴いて実物を確認する必要の有無の判断にある程度は役立ってきた。しかし、貴重書の書誌情報には、以下のような問題点がある。

貴重書の書誌情報は、現代書誌の目録記述ルールに従って書誌情報を記述することは容易ではない。例えば、貴重書の場合、代表となる題名そのものが欠落しており、文章の冒頭を引用してその代用とすることが多い。また、資料の性格上、入手可能な客観的情報が少なく、データ編集者の主観的な分析や推論が必要な場合が多い。一方で、データの質および信頼性を担保するため、データ編集者の分析や推論の入力そのものが控えられることもある。データ編集者の主観的分析や推論が記述されている場合には、値そのものにデータ編集者の推定を示す記号類が付加されていることがあり、データの正規化が困難となる。こうした要因が、貴重書のデジタルアーカイブの構築とその連携の遅滞を起こしている。

加えて、多くの機関は互換性を考慮して、貴重書目録に関するメタデータ群として、現代書誌を対象とした Dublin Core Metadata Initiative（DCMI）が提唱しているアプリケーション・プロファイル（DCAP）[1]を採用している。しかし、利用者が意図する検索結果にできるだけ近づくには、貴重書の成立過程に関連する書誌学および歴史学的アプローチが可能な表現を含むオントロジーが必要である。そのため、貴重書を書籍として扱うだけではなく、文化財として扱えるオントロジーとして、DCAP に加えて国際博物館会議（ICOM）ドキュメンテーション委員会（CIDOC）が現在策定している FRBRoo [2] を基に貴重書の特徴を表す概念セットが参考となる。

書誌の記述ルールでは、[3]のように、あらかじめ決められたデータ項目に当てはまらなかった内容は、注記に記述しておくように定められている。結果的に、注記には貴重書の表現に必要な内容が羅列され、全文検索によってのみ参照できる状態になっている。実際に佐賀大学貴重書コレクション[4]に含まれる「市場直次郎コレクション」 [5]の書誌データ（*ichiba*）の注記の内容を分析した結果、一冊の書籍および扇面が出来上がる過程で序や跋（後書き）の執筆者、外題と異なる内題および見返し題、出版書肆名等、貴重書利用者にとって重要な項目が抽出できた。

これらの情報をもとに、DCAP および FRBRoo のモデルに沿ってオントロジーの構築を試みた。データ編集者の主観的判断や推論の記述内容に対しては、データ編集者の推定を考慮したアノテーションを設置することで個々の値から分離し、正規化を行った。

提案オントロジーの有効性を確認するために、データを RDF 形式に変換し、SPARQL[6]サーバに格納することにより、属性による検索が容易に行える。また、国立国会図書館の NDL Authority（NDLA）および LODAC[7]といった、書誌および文化財に関連する、より規模の大きい SPARQL サーバに格納されているデータと *ichiba* のデータで共通する属性を組み合わせて、検索結果に表示することが可能となる。

2. 貴重書のためのオントロジー構築

2.1. 対象とする貴重書書誌の概要

本研究では、貴重書書誌データとして「佐賀大学貴重書コレクション」の典籍および扇面の書誌目録を利用した。図 1 のような作品が含まれる同コレクションの一部は書誌目録化されており、データはリレーショナルデータベースに保存されている。このデータは Web 上でも公開されている。このコレクションは個人が発掘、蒐集したものであるため、比較的規模の小さなものである。

図 1 市場直次郎コレクションの作品例：典籍「大津ゑぶし」（上）、扇面「花卉図」（下）

2.2. 貴重書目録の注記に収められているデータの特徴

データ作成者が対象となる作品の特徴を記述する上で重要な情報であっても、書誌情報の項目として記述できないものは、注記に記載される。表 1 に例を示す。*ichiba* の典籍データの注記に記載された内容から、貴重書の書誌項目として抽出可能であろう情報を手作業で抽出した（表 2）。その結果、なんらかの情報が含まれている注記データは典籍全 233 レコードに対し、90%を占めていた。その内、書肆名（出版者）、序作成者、序作成年、合本中の題に関する情報は、同じく全レコードに対し、それぞれ 54%、42%、25%、13%含まれていた（図 2）。これらの結果から、データ編集者は、本文以外の本の成立に関わる人名や書肆名等、主に書籍の成立に関わる情報を注記に多くまとめていることが明らかになった。

注記に含まれていた情報は、利用者が貴重書を書籍としてのみではなく、文化財として書籍の成立過程を分析する場合に必要な情報である。これらの情報は、DCAP および FRBR$_{oo}$ のモデルに沿ってオントロジーを構築する際に、新たな項目となる。

表 1 市場直次郎コレクション典籍の部書誌データにおけるレコードの例（一部項目は省略）

項目	値
書名	大津ゑふし弐編
読み	オオツエブシニヘン
書型	中
巻冊	1 巻 1 冊
編著者	春風亭主人柳絲述
刊行年	嘉永 7 年？ ← 推定を表す記号
西暦	1854
刊写	刊
注記	見返し題「大津画ぶし／二編」。「寅の春春風亭のあるし柳絲なり」。奥付「嘉永六丑春／書肆　大坂心斎橋南本町北エ入　河内屋平七板」。他に「（新製）大津絵ふし／春風亭主人柳絲述／従初編至十篇／近日出板」の広告を付す。刷り表紙で「布袋と女」図。

表 2 表 1 の注記項目から抽出した新たな属性と値の組

項目	値
見返し題	大津画ぶし／二編
書肆	大坂心斎橋南本町北エ入　河内屋平七板
広告	（新製）大津絵ふし／春風亭主人柳絲述／従初編至十篇／近日出板
表紙画	「布袋と女」図

2.3. 貴重書目録における主観的なデータ入力の扱い

利用者が求める書誌の内容は、貴重書と現代書では大きく異なる。貴重書の利用者は、現代書のように必ず入手できるタイトルや著者名、出版者名、出版年といった情報を、実物から確実に得られるとは限らない。データ編集の段階で高度な専門的知識が必要であり、主観的な分析内容を汲み入れざるを得ない場合も多い。図 1 の例では、「刊行年」という項目のデータに「？」が付されることで、データ編集

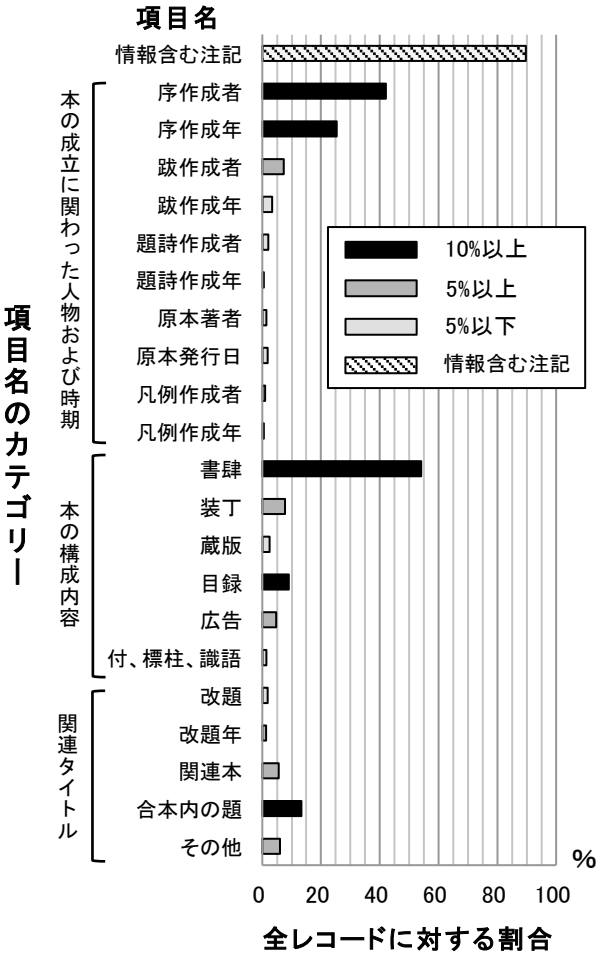

図 2 *ichiba* 全レコードに対する注記から抽出した新たな項目の占める割合

者の主観的推論が記入されている。こうした情報は、研究者にとって有用である一方で、非正規データであるために、年代の絞り込みなどに利用することができないという問題がある。こうした主観的記述は、「注記」にも含まれている。

データの正規化と、「？」などのデータ編集者の判断をともにデータとして記載することは、データの属性としてデータ編集者による未確定な情報であることを示すアノテーションによって可能となる。本稿が対象とする *ichiba* では、このアノテーション名を仮に ichiba:editorialNote とした。

このアノテーションは、シソーラスの統制語彙である SKOS にプロパティとして含まれている、skos:editorialNote に準じている。このプロパティは、「まだ編集作業中であるというお知らせや、将来的に編集を加える際の注意点などのような、維持管理の補助となる情報を提供する[8]」と定義されている。この ichiba:editorialNote を属性のプロパティ、すなわちアノテーションとして付加しておくことで、データ自体に表記されていた不確定な値または推定を意味する括弧類およびクエスチョンマーク等（表 1 の丸枠内）の表記を値から除外できる。そのため、今後データを Linked Data を利用したソフトウェアで多用される JSON 形式のようなフォーマットに変換する場合に利便性が高まる。また、データの正規化を行いながら、データ中の個々の値に適宜注釈を付加し、より詳細な説明を行うことが可能になる。

2.4. 貴重書のためのオントロジー概要

２.２で述べたように、貴重書書誌データの利用者が必要とする属性を分析していくと、Dublin Core で定められた属性に比べ多くの複雑な属性が出現する。これらの属性一覧を書誌データ編集者で共有し、Linked Data 対応のデジタルアーカイブを構築するには、可能な限り既存のメタデータスキーマおよびオントロジーを利用し、不足する部分を補完する方式を取る方が効率的である。

そのために *ichiba* のような貴重書書誌データの属性群に対し、一般的な書誌用メタデータスキーマを組み合わせた DCAP に対応させた。次に、書誌学的特徴を持つ属性にも対応するため、書誌データから抽出した表２のような属性群を既存の属性に追加して FRBRoo に対応させた。FRBRoo は Semantic Web 技術の利用を前提に 2007 年に策定された、書誌用概念モデル FRBR と文化財用概念モデル CIDOC CRM が融合した FRBR 改良モデルである。この概念モデルを利用することで、利用者の環境や目的を考慮した書誌・文化財資料の発見、識別、選択および入手が可能となる[9]。*ichiba* 中に実在する作品である「はや口大津ゑふし」を FRBRoo に対応させると、各々の属性同士は図３のような関係となる。しかし、既存の FRBRoo のみの対応では、書誌学的分析に必要な要素の記述が明確に表現されない。例えば、写本の工程での刊・印・修の作業のうち、「修」（修正が行われた版）にあたる概念を追加する必要がある。

実際の作業でこの概念に対応するデータを見つけることは、校合等の過程を経て結果的に明らかになるため、非常に困難であるが、貴重書の世界においてポイントとなる概念として[10]、オントロジーに記述することを提案する。

3. NDLA、LODAC 等外部 SPARQL エンドポイントを利用した検索システム構築

各組織で作り上げたデータ項目で、セマンティックの等しい共通の属性があれば、それらの属性はデータを組み合わせて利用可能である。Web 上に SPARQL エンドポイントが設置されていれば、常時必要なデータをアプリケーションに組み込む事が可能である。例えば、国立国会図書館には、NDLA と

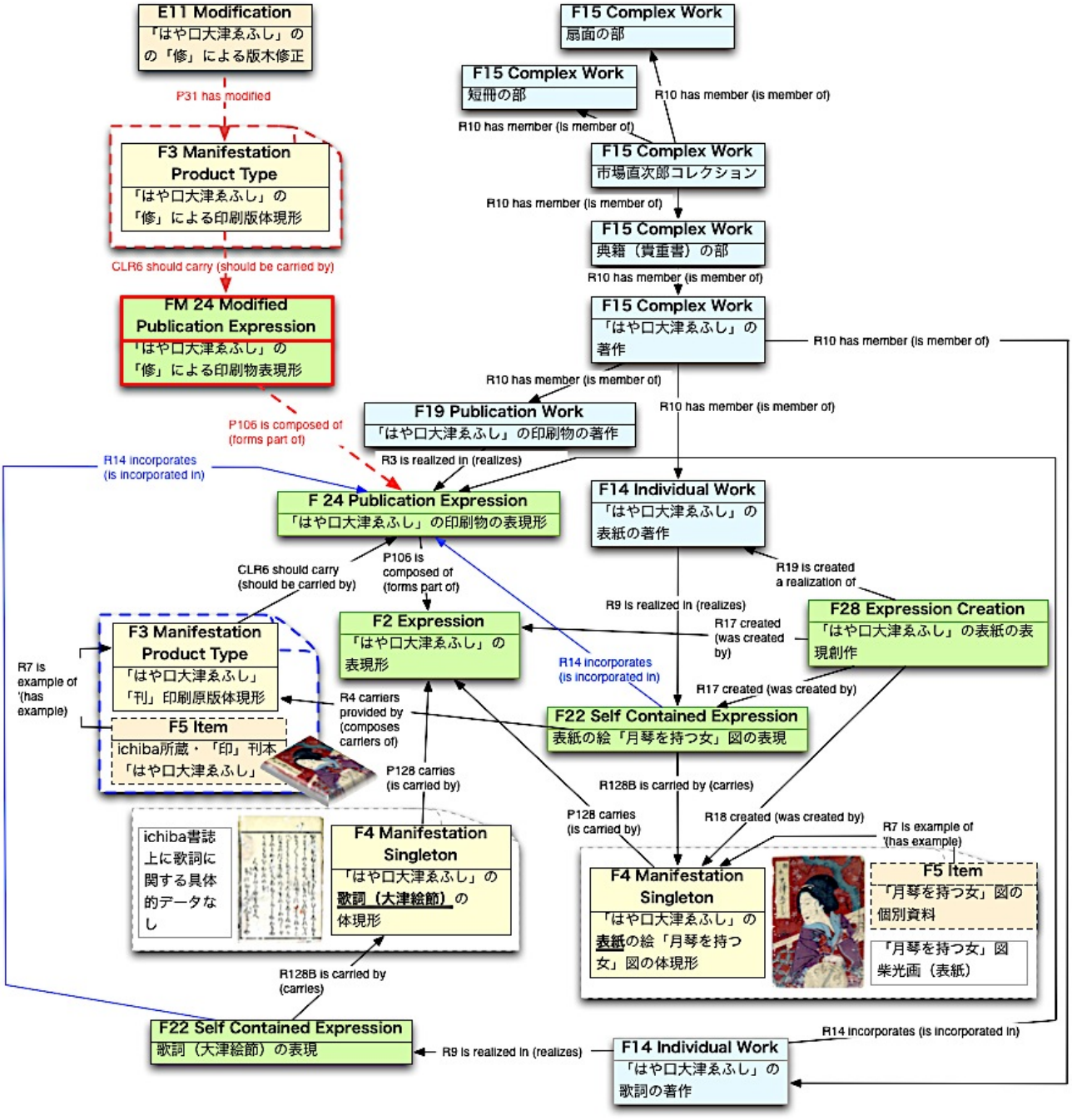

図 3 手刷りによる印刷物と印刷物に含まれる概念の FRBRoo によるモデル化（[吉賀 13]より引用）

呼ばれる、著者および件名の標目検索が可能な SPARQL エンドポイントが設けられている[11]。また、LODAC Project では、Web 空間に存在する日本の博物館、美術館および生物種に関する学術情報を集めて Linked Data 化し、SPARQL エンドポイントから利用可能にしている。他にも世界中の Linked Data のハブである DBPedia との連携も可能である。主に江戸期の典籍と美術品のコレクションである *ichiba* には、専門機関が構築した書誌および美術品データが集まっている日本の NDLA および LODAC が特に有用である。

例えば、*ichiba* の書誌情報に記載されている人名から NDLA の人名典拠 ID を探して、その ID に紐付いた関連著書を検索することができる(図 5)。その結果は Exhibit3.0[12]等のフレームワークやウェブアプリケーション等で検索者のニーズに合わせて表示可能である[13]。また、LODAC に接続することで、

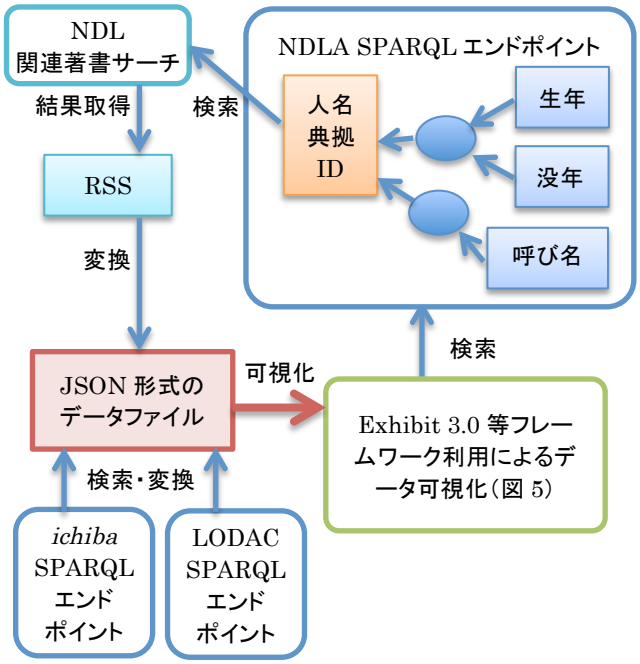

図 4 Linked Data を利用した
検索システム模式図

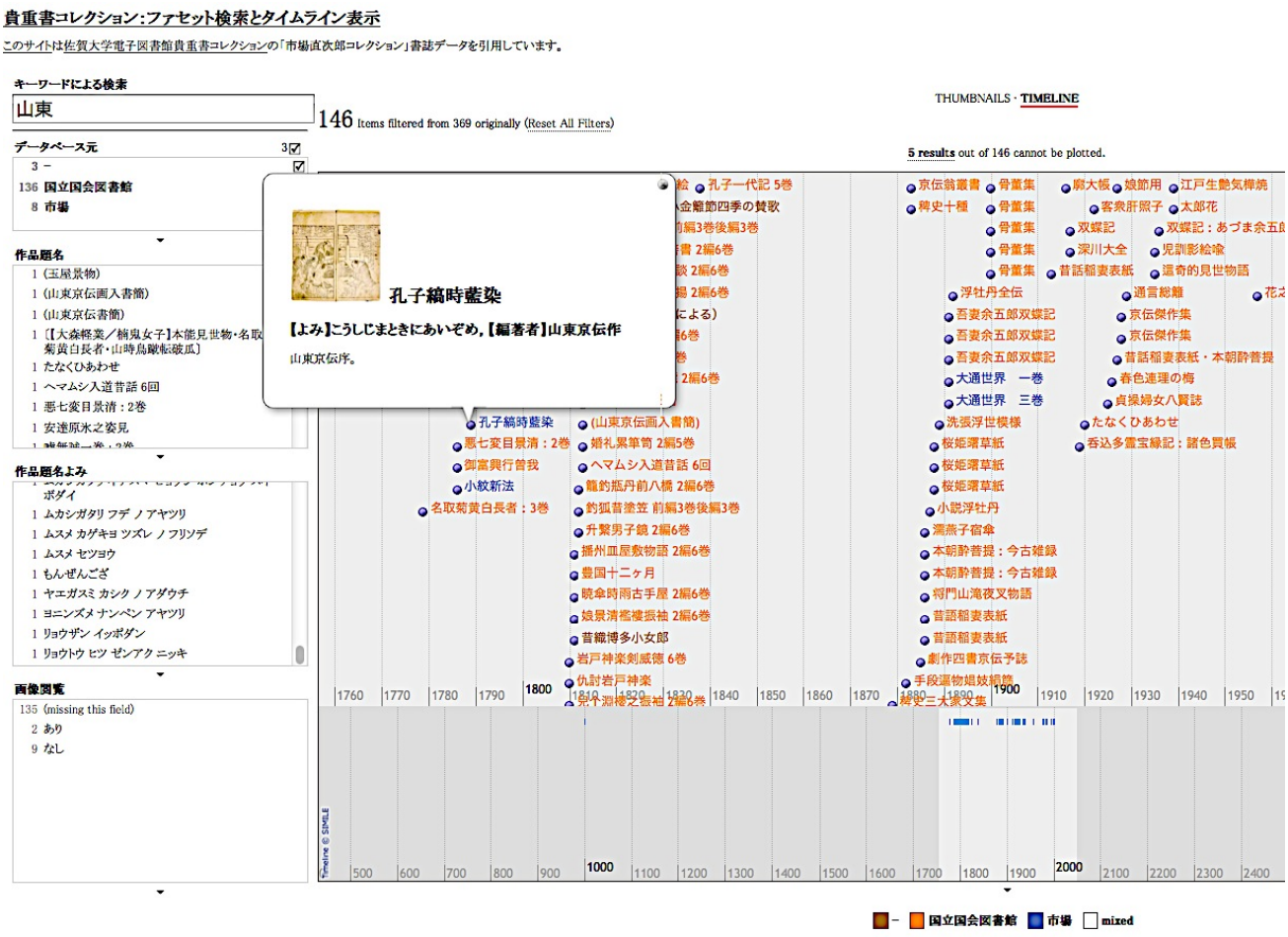

図 5 国立国会図書館および市場直次郎コレクション所蔵の山東京伝著書を
Exhibit 3.0 Timeline で検索した結果

LODAC がウェブ空間から収集した豊富な美術品および画家の経歴および人間関係、所蔵場所等の情報との連携が可能となる。

4. まとめ

貴重書書誌は現代書籍の書誌同様、Dublin Core 等書誌用メタデータを用いてデータセットを構築可能である。しかし、多くの貴重書書誌の利用者が必要としている情報は、書物のコンテンツのみでなく、書物の成立過程に関する文化財的あるいは書誌学的情報である。加えて、実物を手にとって閲覧する前に参考にする画像であることが多い。

こうした情報は、現代書誌情報に含まれず、「注記」として記述される。佐賀大学の *ichiba* を例に、「注記」での記載内容から、書肆名や序などの書物の成立に関する情報を抽出した。この分析で出現した情報を題名や著者名等の一般的な書誌項目に加えることで、貴重書利用者により適したデータが作成できる。

また、貴重書の書誌データは、高度な知識を必要とするため、入力者の主観が多少含まれていることがある。実際のデータにおいても括弧類やクエスチョンマーク等様々な記号を値自体に直接付加して推測であることを表現している。これらのデータの正規化と、推論等であることの情報を両立される記述が必要である。そのため、推測であることを表現するアノテーションをメタデータに付与することで、個々の値に対して注釈を付加することが可能となる。すなわち、客観的なデータと主観的なデータを分離して記述できるようになる。

データから注釈の記号類を除去し、データを正規化すれば、Linked Data への変換が容易となる。Linked Data に変換され、ウェブ上に外部公開および相互利用が可能となったデータセットは、様々な利用者が別のデータセットと自由に組み合わせて活用できる。その結果、入力の比較的困難な貴重書書誌のようなオンラインデータの構築と利用が促進される。

5. 参考文献

[1] Coyle KA, et al. Guidelines for Dublin Core Application Profiles [internet]. US: Dublin Core Metadata Initiative; 2009 [updated 2009 May 18; cited 2013 Aug 26]. Available from: http://dublincore.org/documents/profile-guidelines/

[2] Bekiari CH, et al. FRBR object-oriented definition and mapping to $FRBR_{ER}$ (Version 1.0.2) [Internet]. Greece: International Working Group on FRBR and CIDOC CRM Harmonisation; 2012 [updated 2012 Jan; cited 2013 Aug 26]. Available from: http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.2.pdf

[3] 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～） [Internet]. 東京: 国立国会図書館; 2012 [updated 2012 Jan; cited 2013 Aug 26]. Available from: http://www.ndl.go.jp/jp/library/data/wakosho201201.pdf

[4] 佐賀大学附属図書館. 佐賀大学貴重書コレクション[Internet]. 佐賀: 佐賀大学; 2001 [updated 2012 Apr 27; cited 2013 Aug 26]. Available from: http://www.dl.saga-u.ac.jp/

[5] 井上敏幸編集. 市場直次郎コレクション目録. 佐賀: 佐賀大学附属図書館, 地域学歴史文化研究センター; 2007. 335 p.

[6] Prud'hommeaux ER, Seaborne AN. SPARQL Query Language for RDF [Internet]. Bristol: Hewlett-Packard Laboratories; 2008 [updated 2013 Mar 28; cited 2013 Aug 26]. Available from: http://www.w3.org/TR/rdf-sparql-query/

[7] 武田秀明, 大向一輝 et al. LODAC Project [Internet]. 東京: LODAC & Life Science Databases; 2010 [updated 2013 Jul 10; cited 2013 Aug 26]. Available from: http://lod.ac

[8] Issac AN, Summers ED. SKOS Simple Knowledge Organization System Primer [Internet]. W3C: W3C; 2009 [updated 2009 Jun 15; cited 2013 Aug 26]. Available from: http://www.w3.org/TR/skos-primer/.

[9] 両角彩子, 杉本重雄. 利用者の特性と環境に応じ

たリソース選択のためのメタデータスキーマモデル. ディジタル図書館[Internet]. 2005 11 [cited 2013 Aug 26]; No.29. Available from: http://www.tulips.tsukuba.ac.jp/mylimedio/dl/page.do?issueid=893411&tocid=100085965&page=3-14

[10] 山中秀夫. 現代の情報環境における和古書総合目録構築に関わる研究[dissertation on the Internet], 神奈川: 総合研究大学院大学; 2008. [cited 2013 Aug 26]. Available from: http://www.nii.ac.jp/graduate/thesis/pdf/200803/yamanaka_Dr_thesis.pdf

[11] 神崎正英, 佐藤良. 国立国会図書館の典拠データ提供におけるセマンティックウェブ対応について(<特集>典拠・識別子の可能性:ウェブ・オントロジーとの関わりの中で). 情報の科学と技術 2011; 61(11): 453-459.

[12] Huynh DA, Massachusetts Institute of Technology and Contributors. Exhibit 3.0 [Internet]. US-MA: MIT; 2010 [updated 2013 July 30; cited 2013 Aug 26]. Available from: http://www.simile-widgets.org/exhibit3/

[13] 吉賀夏子, 渡辺健次, 只木進一. 貴重書デジタルアーカイブの書誌オントロジーおよび Semantic Web 技術を活用した検索システムの構築. In: 第 27 回人工知能学会全国大会 (JSAI2013) [Internet]; 2013 Jun 4; 富山. 東京; 人工知能学会; 2013 [cited 2013 Aug 26]. 1N3-OS-10a-4in. Available from: https://kaigi.org/jsai/webprogram/2013/pdf/812.pdf