1N3-OS-10a-4in

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

Developing a Bibliographical Ontology and a Visual Search System for Historical Documents with Semantic Web Technologies

吉賀 夏子[*1] Natsuko YOSHIGA
渡辺 健次[*2] Kenzi WATANABE
只木 進一[*3] Shin-ichi TADAKI

[*1] 佐賀大学大学院工学研究科 Graduate School of Science and Engineering, Saga University
[*2] 広島大学大学院教育学研究科 Graduate School of Education, Hiroshima University
[*3] 佐賀大学総合情報基盤センター Computer and Network Center, Saga University

Digital archives for historical documents have been rapidly growing and those bibliographical data have been also available through web services. On the contrary to bibliographies for modern books, mapping bibliographical properties of historical documents to Dublin Core Application Profiles (DCAP) is inadequate to describe unique characteristics of historical objects. We propose a proper ontology for historical documents by combining DCAP and partially modified $FRBR_{OO}$ (FRBR [Functional Requirements for Bibliographic Records]-object oriented). $FRBR_{OO}$ provides a conceptual reference model in both cultural heritage and bibliographical documentation. In addition, we construct a visual search system to use the proposed ontology effectively. Bibliographical data of "*Ichiba Naojiro Collection*" owned by Saga University are used to show an actual example of searching bibliographical information by referring headings of authors' names through the SPARQL endpoint, constructed by Web National Diet Library Authorities (NDLA).

1. はじめに

ウェブ上には貴重書(概して江戸期以前に作られた歴史的価値の高い書物)や美術品のデジタル画像の閲覧および書誌情報の検索が可能な，美術館や図書館のサイトが多数存在する．しかし，歴史・人文学の研究者が実際にキーワードを検索フォームに入力して，本格的に貴重書書誌の検索を行うと，様々な困難に直面する．例えば，ある著者に関する情報を検索する場合，その著者に複数の呼び名があるため，単純にキーワードを入力するだけでは適切な結果を得られない事がある．また，ある著者の作品スタイルに影響を受けた後世の人々を探し，その作品目録を一括して取得するような検索も，現状の検索システムでは一般には困難である．

一方，目録作成者が貴重書研究において重視する属性そのものを選択し，実際のデータを確定する際，高い書誌学的スキルが求められる[山中 08]．例えば，複数の候補を有する書名および刊写(印刷したか書写したか)の違いや，「刊・印・修」と呼ばれる主に印刷工程に関する属性の取捨選択である．そのため，ある著書の版を調べ，地図やタイムライン上に表示することにより，その著作がどのように伝播していったのかというような検索結果の自動表示を行うことは困難である．さらに，デジタルコンテンツと同様に，貴重書には写本の類が著しく多いため，著作単位での名寄せが必要である．

近年，利用者の利便性を向上するため，Semantic Web 技術の導入を見据え，著作や版，個別資料等利用者にとって必要な概念とその関連を明示化し，利用者の資料発見，識別，選択および入手を支援する新しい目録の枠組みが作成された．現在，国際図書館連盟(IFLA) は「書誌レコードの機能要件」(FRBR)[和中 04]を策定し，国立国会図書館(NDL)を含む世界の主要図書館で，従来の目録記述を FRBR に対応させる動きがある[谷口 08]．貴重書の世界においても FRBR の考えを導入し，著作同定を行う試みが行われている[Tokita 12]．

しかし，現時点では，目録記述を変更することで，実際にどのような効果があるのか，目録の作成者にもそれを利用する人々にも明確に示されているとは言いがたい．

こうした問題を解決するため，実在する目録データを参考に，貴重書目録の検索に適したオントロジーを提案する．このオントロジーは，現代の書誌データを記述する Dublin Core のメタデータスキームと，文化財と書誌のオントロジーを融合させた $FRBR_{OO}$ を結合したものである．

提案したオントロジーの有効性を示すため，佐賀大学所蔵の「市場直次郎コレクション」の目録データを対象とした検索システムを構築した．提案オントロジーに基づき，目録データを RDF 形式に変換し，SPARQL サーバに格納した．この SPARQL サーバを用いて Web から属性による検索を行い，結果を可視化するシステムを構築した．

その際，NDL Authority(NDLA)の SPARQL エンドポイントを利用した名寄せ機能を追加した．すなわち，利用者が *ichiba* に存在する書物の著者名の断片を検索フォームに入力すると，システムは NDLA の SPARQL エンドポイントから典拠 ID を取得し，著者名の候補をブラウザに表示する．次に，利用者が候補者を選択すると，システムは候補者の典拠 ID に紐付いた NDL 所蔵の書名と，*ichiba* 内の書名を，ブラウザのタイムラインに併せて表示可能にした．

2. 貴重書目録の検索に適したオントロジーの構築

ウェブ空間に利用価値の高いデジタルアーカイブを構築するためには，Semantic Web 技術の活用が有効である．貴重書や美術品のアーカイブを対象としたオントロジーを提案する．

2.1 貴重書のためのオントロジーの概要

貴重書は，利用者の観点から見て，現代の書籍とは大きく異なる性質を有している．貴重書の利用者は，書物としての性格だけでなく，外観や成立過程等，より文化財的な視点で貴重書

連絡先：吉賀夏子，佐賀県佐賀市本庄町本庄 1 番地，佐賀大学大学院工学研究科知能情報システム学専攻，12573015@edu.cc.saga-u.ac.jp

を分析している．したがって，現代の書籍の書誌を対象とした 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）[DCMI 04]を用いるだけでは，貴重書の持つ特徴を網羅するのは難しい．このため，貴重書を文化財として扱える，あるオントロジーが必要である．

文化財用オントロジーとしては，国際博物館会議（ICOM）ドキュメンテーション委員会（CIDOC）が策定した，概念参照モデル（CIDOC CRM）[CIDOC 11] がある．さらに，CIDOC では，IFLA の FRBR と連携させた，FRBR Object Oriented（$FRBR_{OO}$）という，派生概念モデルを近年策定している[FRBR 12]．

本稿では，$FRBR_{OO}$ に対応するメタデータのマッピングを，[Bœuf 12]を参考に行い，DCAP のみで表現が困難な貴重書の特徴を $FRBR_{OO}$ で補った．さらに，$FRBR_{OO}$ での補完でも不十分だと考えられる属性を補充した．つまり，DCAP および一部クラスを補充した $FRBR_{OO}$ を組み合わせた概念セットを「貴重書のためのオントロジー」として提案する．

2.2 DCAP による貴重書記述の課題

DCAP は，実際のデータが有する属性を別のメタデータにマッピングする際の利便性を考慮して，あらかじめ利用が推奨されるメタデータの種類や符号化方法を，使用用途別にフレームワークとしてまとめたものである．フレームワークとして定義することで，データ変換作業の手間を軽減し，相互運用性を高めることができる．既に NDL での運用例やガイドラインなど情報が豊富に公開されている[NDL 11]．

貴重書目録のデータにおいても，NDL をはじめとした外部の書誌データと連携するためには，DCAP への対応は有効である．ただし，NDL では DCAP で推奨されたメタデータセットに当てはまらなかった属性に対して，独自のメタデータスキーマを新たに設定することで対応している．NDL の独自メタデータスキーマを含めたアプリケーション・プロファイルは DC-NDL と呼ばれている．

実際に，DC-NDL に準じて，表 1 のように，*ichiba* の属性項目に外部メタデータスキーマをマッピングすると，分類，書名読み，巻冊の項目が独自メタデータスキーマでの対応となった．また，各項目に対応したメタデータに重複が多く，リテラル値で判断するしかない属性がいくつか存在した．そのため，機械的な対応を作成することが困難である．

表 1 *Ichiba*（典籍）属性項目から DC-NDL へのマッピング一覧
dcndl: http://ndl.go.jp/dcndl/terms/

ichiba 属性項目	DC-NDL の対応メタデータ
整理番号	dcterms:identifier, rdfs:seeAlso
分類	dcndl:materialType
分類 1	dcndl:materialType
分類 2	dcndl:materialType
分類 3	dcndl:materialType
分類 4	dcndl:materialType
分類 5	dcndl:materialType
書名	dc:title, dcterms:title
読み	dcndl:transcription
書型	dcterms:extent
巻冊	dcndl:volume
編著者	dc:creator
刊行・書写年次	dcterms:date
西暦	dcterms:date
刊写	dcterms:publisher
注記	dcterms:description
印記	dcterms:transcription
文庫	dcterms:transcription
他	dcterms:transcription

2.3 貴重書目録データにおける $FRBR_{OO}$ の役割

IFLA で策定した FRBR では，図書目録利用者が関心を持つ概念を分析し，図 1 に示すように，4 つの基本概念クラスとその関係を表現する．一方，CIDOC CRM は，ある人または団体が知的アイデアを思いつき，それを表現するための製作作業を経て，誰の目にも見える形になった「作品」を完成させるという，著作の成立過程を概念参照モデルで表現したものである．

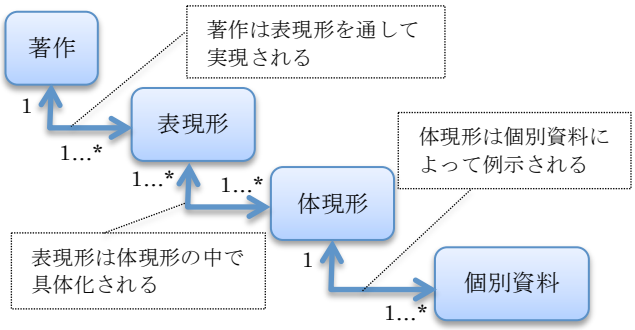

図 1 FRBR における，書誌データ利用者が関心を持つ概念クラスとその関連（[和中 04]から引用）

FRBR と CIDOC CRM が融合した $FRBR_{OO}$ では，時間や出版など，FRBR だけでは表現不足だった概念を補足している．例えば，$FRBR_{OO}$ では，著作の成立において，著者が直接製作した元々の成果物と，その元々の成果物を印刷用に編集して出版した成果物は，別個の成果物として表現される．すなわち，貴重書の成立過程の分析で重要とされる，写本の過程も記述することが可能となっている（図 2）．

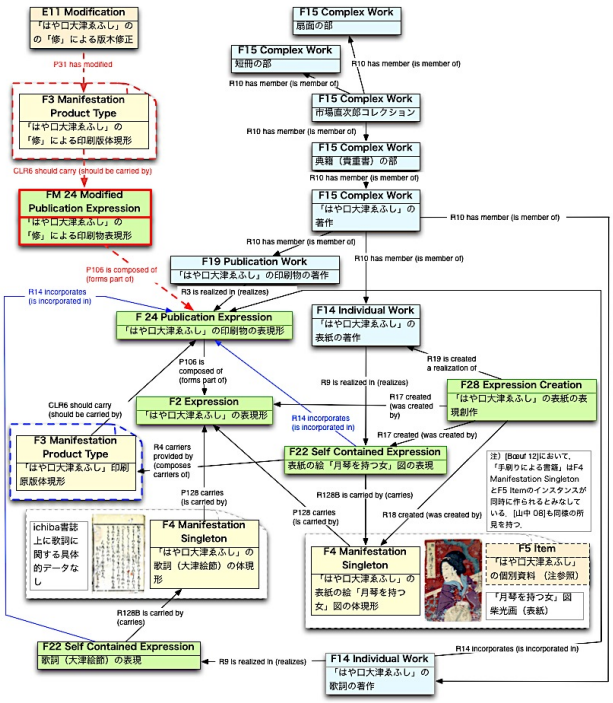

図 2 手刷りによる印刷物と印刷物に含まれる概念の $FRBR_{OO}$ によるモデル化

しかし，$FRBR_{OO}$ のみの対応では，写本の工程で版の修正が生じた場合の明確な表現が不十分であるため，「修」(修正が行われた版)にあたる概念を追加した．具体的には，印刷物の表現形である F24 Publication Expression の下位クラスとして，「修」を表現するクラス(本稿では，仮に FM24 Modified Publication Expression とする)とその関連を補充した(図 2 左上赤枠部分)．図 2 のように，$FRBR_{OO}$ を用いると，一つの書物の様々な成立過程を，様々な属性を基準に，専門家の視点でより深く分析することが可能となる．

3. 提案オントロジーの適用

3.1 市場直次郎コレクション目録

「市場直次郎コレクション」は，佐賀大学附属図書館所蔵の貴重書コレクションのひとつで，近世前期，17 世紀後半から昭和の間に作成された和書や日本および中国文人の書画類，民俗資料のコレクションである[井上 2007]．この内，主に典籍(書)と扇面の目録が電子化され，ウェブ上で公開されている(図 3)．

本稿では，この目録データ *ichiba* を，2 章で構築したオントロジーにもとづいて RDF に変換し，SPARQL サーバに格納した．さらに，4 章 1 節および 2 節において，*ichiba* の実データを Exhibit3.0 上で可視化した．

図 3 佐賀大学附属図書館:貴重書コレクション
http://www.dl.saga-u.ac.jp/OgiNabesima/

3.2 NDL の書誌目録および典拠データ

(1) NDL 書誌データ概要

NDL は，現状日本国内の書誌情報に関してハブ的位置づけとなっている組織である．NDL は 2 章 2 節で述べた通り，Dublin Core Metadata Initiative(DCMI)で定めている書誌用アプリケーション・プロファイルを参考にして，独自の DC-NDL を定め，運用している．4 章 2 節において，NDL と *ichiba* の書誌データをブラウザ上に表示する際，あらかじめ RSS 形式で引用可能な NDL 書誌データを利用した．

(2) Web NDL Authorities

NDL には，NDLA(図 4)と呼ばれる，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[神崎 11]．利用者が *ichiba* に存在する著者名の断片から，著者名の候補，すなわち典拠 ID を取得する際に，検索システムは NDLA にクエリを発行して，該当する結果を取得するために利用する (4 章を参照)．

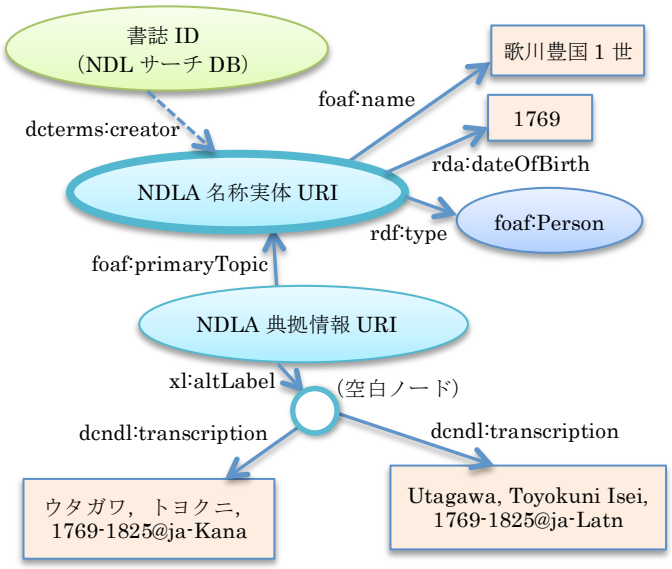

図 4 NDLA メタデータモデル部分図
(**全体図** URL:
http://www.ndl.go.jp/jp/aboutus/standards/meta/pdf/damamodel.pdf)

4. 検索システムの構築

4.1 Exhibit3.0 フレームワークを用いた書誌ブラウザの構築

まず，*ichiba* を既存のリレーショナルデータベースから取り出し，書名などのメタデータとその値の関係を Google Refine を用いて，提案オントロジーに基づき RDF データにマッピングした．その RDF ファイルを Sesame SPARQL サーバに保存した．

利用者はウェブアプリケーションを通じ，注目するメタデータを選択して必要なデータを呼び出す．SPARQL に対するクエリの実行およびデータの可視化には，Exhibit3.0(JavaScript によるデータ可視化フレームワーク)および PHP5.3 を用いた．Exhibit3.0 でのデータを表示において，JSON/Exhibit 形式のデータに変換する必要があるので，SPARQL サーバからデータを取り出し JSON 形式に変換した(図 5 ①から⑤)．

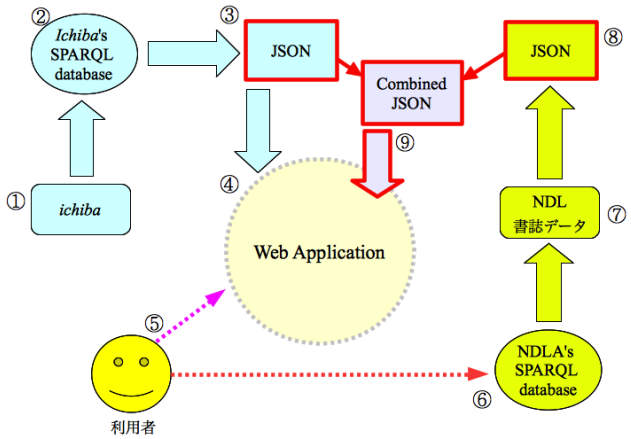

図 5 検索の流れ

Exhibit3.0 の活用により，*ichiba* データ書誌をサムネイルまたはタイムライン(図 6)として可視化しながら閲覧できる．また，調べたい属性と値を選ぶことで書誌データをファセット検索できる．

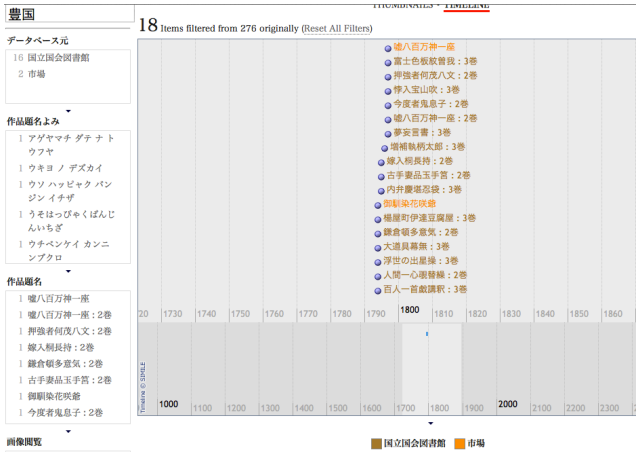

図 6 Exhibit3.0 によるファセット検索画面外観（一部）

4.2 著者名の典拠 ID を用いた書誌目録の表示

さらに Exhibit3.0 の検索画面上で，著者名に対して典拠データを検索できる機能を付加した．利用者はこの典拠データ検索機能を通じて NDLA の SPARQL データベースにアクセスできる．NDLA のように SPARQL に対応したデータベースは，ある属性のリテラル値から容易にルートである典拠 ID を参照することができる（図 4）．すなわち，ある人物の生没年，号（ニックネーム）等情報の断片から容易に典拠 ID を辿ることができる．その典拠 ID は各書誌に紐づけられている．そのため，利用者が求めている著者の書誌だけを，検索キーワード入力時の表記揺れを考慮せずにピックアップすることができる．

NDLA で著者名を検索すると，図 7 のように画面に候補者がリストされるので，その中から利用者にとって関心のある著者名を選んでもらう．このとき，作成年を考慮してさらに候補者を絞り込むことができる．典拠 ID から引用可能である RSS 形式の書誌データは JSON 形式に自動的に変換される．この時，JSON に変換されたデータの各属性は，セマンティックが等しいので，データを組み合わせることができる．組み合わせたデータは Exhibit3.0 上で可視化した(図 5 ⑥から⑨)．

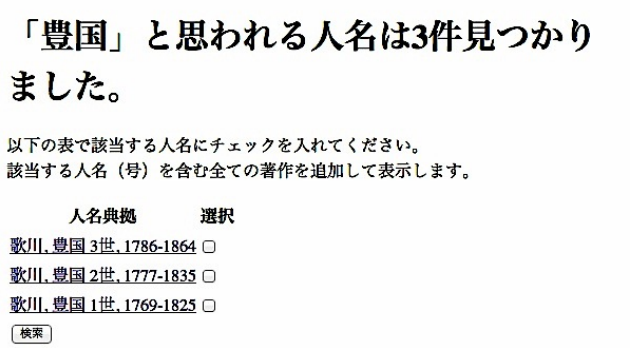

図 7 NDLA から「豊国」という呼び名を含む人名典拠を検索した結果

5. まとめ

デジタルアーカイブに収められた書誌目録データを利用者が効率的に活用できる仕組みを作るため，貴重書のためのオントロジー構築を行った．すなわち，現代の書籍目録に対応する Dublin Core などの汎用メタデータスキーマおよびアプリケーション・プロファイルの利用のみでは，貴重書の成立過程を詳細に分析する必要がある利用者にとっての概念表現が不足していた．そのため，文化財としての取り扱いが可能なオントロジーである，$FRBR_{OO}$ の利用および $FRBR_{OO}$ 内下位クラスの一部補填を提案することで，貴重書のためのオントロジーを構築した．また，構築したオントロジーに準じたメタデータスキーマを選択して，*ichiba* のデータを RDF ファイルに変換し，SPARQL サーバにストアして，エンドポイントを構築した．

次に，応用例として，構築したオントロジーに準じた属性による検索が可能で，サムネイルおよびタイムラインで SPARQL クエリ結果を可視化できる検索システムを構築した．さらに，検索システムに付加されたフォームを用い，NDLA にて著者の典拠を検索し，名寄せを行った．この機能により，著者の候補が画面上にリスト表示されるので，利用者が候補者を選ぶと，NDL は選んだ候補者の書誌をシステムに返す．その書誌データはシステム上で *ichiba* のデータとともに表示され，*ichiba* と NDL 書誌の比較がタイムライン上で可能になった．

今後は歴史・人文学の研究者や目録作成者へのインタビューを行いながら，貴重書目録における，より信頼度の高いデータモデリングと，このようなデータに対応する検索システムの改良を行う予定である．

参考文献

[山中 08] 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，総合研究大学院大学博士論文，2008.

[和中 04] 和中幹雄，古川肇，永田治樹（訳）: 書誌レコードの機能要件，日本図書館協会，2004.

[谷口 08] 谷口祥一: FRBR のその後:FRBR 目録規則? FRBR OPAC?，TP&D フォーラムシリーズ:整理技術・情報管理等研究論集，No.17, pp. 3-23, 2008.

[Tokita 12] Takuya Tokita, Maiko Kimura, Yosuke Miyata, Yukio Yokoyama, Shoichi Taniguchi, and Shuichi Ueda: Identifying Works of Japanese Classics for Construction of FRBRized OPACs, Cataloging & Classification Quarterly, Vol. 50, Issue 5-7, pp. 670-687, 2012.

[DCMI 04] Guidelines for Dublin Core Application Profiles, http://dublincore.org/documents/profile-guidelines/, 2004.

[CIDOC 11] Definition of the CIDOC Conceptual Reference Model, ICOM/CIDOC CRM Special Interest Group, http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf, 2011.

[FRBR 12] FRBR object-oriented definition and mapping to $FRBR_{ER}$ (Version 1.0.2), International Working Group on FRBR and CIDOC CRM Harmonisation, http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.2.pdf, 2012.

[Bœuf 12] Patrick Le Bœuf: Modeling Rare and Unique Documents: Using $FRBR_{OO}$/CIDOC CRM , Journal of Archival Organization, Vol. 10, Issue 2, pp. 96-106, 2012.

[NDL 11] 国立国会図書館ダブリンコアメタデータ記述（DC-NDL），http://www.ndl.go.jp/jp/aboutus/standards/meta.html, 2011.

[井上 07] 井上敏幸編集: 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.

[神崎 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，61(11), pp. 453-459, 2011.