

浮世絵を対象とした異種データベースの 多言語統合アクセス手法の提案

久山岳夫[†]

Batjargal Biligsaikhan[‡]

木村文則[†]

前田亮[†]

[†]立命館大学情報理工学部 [‡]立命館大学理工学研究科

本論文は、複数のデジタル化された美術館・博物館の情報を統合的なアクセスを可能にするプロジェクトの一環について述べたものである。本研究の目標は、日本の人文系の研究の補助となるシステム、すなわち異なる形式で記述されている複数の人文系のデータベースに対して統合的なアクセスを可能にするシステムの構築を行うことである。本論文では人文系のデータベースの一つとして浮世絵のデータベースを対象とした。浮世絵のデータは世界中のデータベースに散在しており、所在の把握が困難である。それらの浮世絵を一度に検索可能なシステムの構築を行うことは人文系の研究にとって役立つことが期待される。そこで本論文では、検索の効率化を図るために、浮世絵が所蔵されている複数の異種データベースに対して、Linked Dataによる多言語統合アクセスの手法の提案を行う。

An Approach for Integrated Multilingual Access to Diverse Ukiyo-e Digital Archives

Takeo Kuyama[†] Biligsaikhan Batjargal[‡] Fuminori Kimura[†] Akira Maeda[†]

[†]College of Information Science and Engineering, Ritsumeikan University

[‡]Graduate School of Science and Engineering, Ritsumeikan University

This paper provides a summary of our ongoing project for providing integrated access to multiple digital libraries, archives, and museums. The main goal of our project is to construct a federated access system for Japanese humanities digital archives. In this paper, we target Ukiyo-e digital archives. Ukiyo-e, Japanese traditional woodblock printing is known world-wide as one of the fine arts of the Edo period (1603–1868). Many museums and organizations in Japan as well as in Western countries hold numerous Ukiyo-e prints in their vast collections. As a result of worldwide digitization over the last decade, many cultural institutions including libraries, archives, and museums started to expose digitized images of Ukiyo-e prints on the Internet. It has become difficult to know which organization has what Ukiyo-e prints. Thus, federated searching of multiple Ukiyo-e digital archives all around the world is becoming crucial demand for humanities researchers of Japanese arts and culture. This paper proposes a method of integrated multilingual access to diverse Ukiyo-e digital archives, which improves search efficiency of existing federated search system by using Linked Data.

1. はじめに

浮世絵は江戸時代に成立し、演劇や古典文学から人々の日常生活や風物に至るまで、当時の様々な風俗を題材として描かれている。浮世絵を媒介とすることにより、後世の人々にも江戸時代の庶民の生活や風俗を垣間見ることが可能である。それゆえ、江戸時代の文化を理解するための資料として重要視されている。

一方、美術品としての価値は、もともと大衆文化の一部として描かれたいきさつから、それほど高かったわけではなかった。しかし、現代では浮世絵の芸術性の高さが見直され、美術品としての価値が高まっている。また海外においても、浮世絵は西洋の近代画にも影響を与えた

日本を代表とする美術品として注目されている。従来、浮世絵は美術館や大学の研究所などに所蔵されていることが多く、観賞するためには所蔵されている場所まで行く必要があった。しかし、近年では美術品や芸術作品などをデジタル化してデジタルアーカイブとして保存することが行われ始めている。浮世絵も同様であり、美術館や大学の研究所などにおいてデジタル化されている。また、これらのデジタル化されたアーカイブはインターネット上に公開することが可能であり、非常に多くのデジタルアーカイブが公開されるようになってきている。そこで多数のデジタルアーカイブから目的の浮世絵を効率よく探すことが重要となっている。そのための検索手法、例えばデータベースを横

断した検索などが求められている。また、浮世絵は明治時代に大量に海外に流出しており、欧米の多数の美術館や博物館に所蔵されている。これら海外に散逸した浮世絵の所在の把握は難しいのが現状である。それらの異なる言語で記述されたデータを検索するには言語を横断した検索技術が求められる。また、浮世絵に関するデータベースも世界の各所において公開されている。このような状況において、出来る限り全ての浮世絵を対象に検索を行いたいという要求にこたえるには、複数のデータベースを横断した検索の技術の向上が求められている。

そこで本論文では、国内外に散在する複数の浮世絵データベースの検索の効率化を図るために、異種の複数データベースに対して多言語統合アクセスを行う手法を提案する。ここで統合アクセスとは、単なる複数データベースの横断検索にとどまらず、検索結果に関連する他のデータベースの情報を「芋蔓式」にたどってアクセスすることを含めたものと定義する。この実現のために、本研究では Linked Data の技術を用いる。

2. 関連研究

嘉村ら[1]は、Lod.ac プロジェクトの一環である LODAC Museum において、主に日本の美術館・博物館情報の統合を行っている。この LODAC Museum では日本全国の美術館・博物館を Linked Data として共有して統合できる仕組みを作っている。本研究の嘉村らの手法との相違点は、あらかじめ各データベースの全データを収集せず、検索結果から動的にリンクを作成する点である。これによりあらかじめリンクを生成していないデータベースやサービスに対してリンクを作成することができる。他の相違点として、日本の美術館・博物館情報だけではなく海外のデータベースも対象としている点、関連項目のリンクを生成する際に典拠データを用いている点がある。

原ら[3]は地域研究資料の知的関連付けや定量分析を実現するための情報学的手法の研究開発を行っている。原らの研究の資源共有化システムでは、多くのメタデータに共通なデータ項目からなる「共有化メタデータ」を定義し、その上で、各データベースのデータ項目から意味的に対応する共有化メタデータのデータ項目へのマッピングを行い、「共有化メタデータベース」の構築を行っている。また山本ら[4]は実在する博物館の資料情報等の共有・活用をするために、インターネット上の情報資源の発見や記述に用いられる Dublin Core Metadata Element Set と博物館コアメタデータとの対応関係について議論している。本研究とこれらの研究との相違点は、

検索結果の個々のデータに着目して他のデータベースへのリンクの生成を行っている点である。

3. 提案手法

我々が提案するシステムの全体の概要図を図1に示す。全体的な流れとしては、まずユーザが浮世絵横断検索システム（後述）を用いて検索をする(1)。その検索結果(2)の中からユーザがある一つのデータを選択する(3)。人名リンク生成モジュールは選択されたデータに記述されているある一人の人名に着目し、NDL Authorities と LC Authorities （後述）にて典拠データの取得を行う(4)(5)。取得したデータをもとに関連するデータを検索(6)しそのリンクを生成する。最後にリンク群をユーザに出力する。

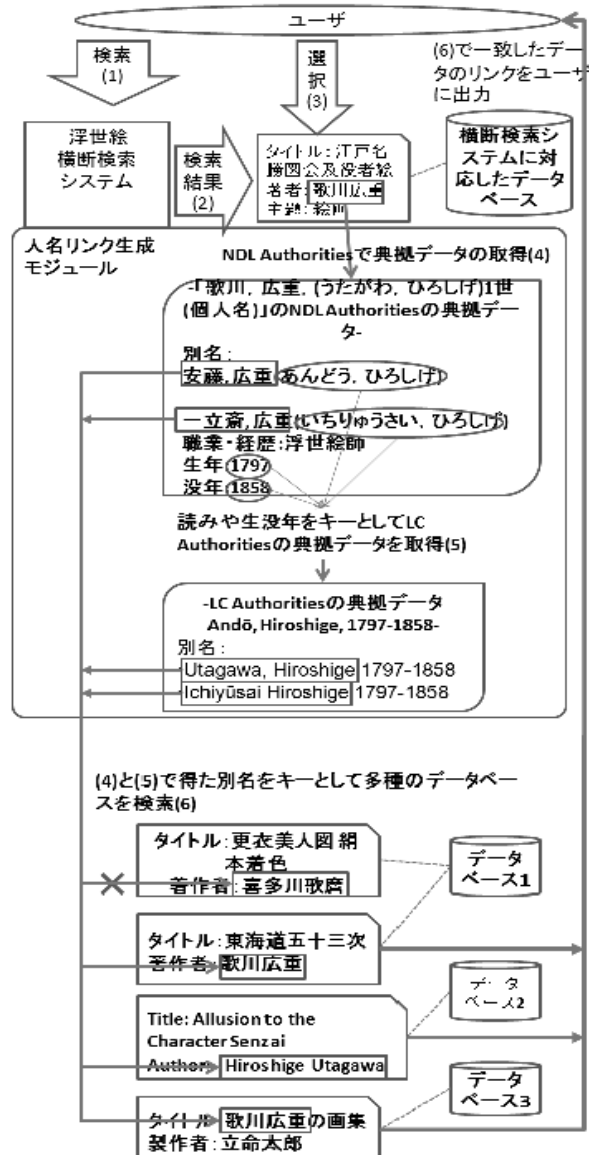


図1：提案するシステムの概要図

3.1 浮世絵横断検索システム

本研究では、我々がこれまでに構築している浮世絵横断検索システム[2]での検索結果から多言語統合アクセスを行う。このシステムでは各データベースに対して、SRU(Search Retrieve via URL)またはスクレイピングを用いて検索を行う。SRUは横断検索用のプロトコルであり、検索要求情報を含んだURLをサーバ側に送ることで、検索結果をXML形式で返すものである。スクレイピングとは、Webページから必要な情報を抜き出す処理のことであり、本システムでは検索結果のページからメタデータを抽出するために用いる。SRUは容易に各データを取得できるが、各データベースを提供するサービスがSRUによる検索に対応している必要がある。スクレイピングではデータベースを提供するサービスがAPI等を提供していなくても各データを取得できるが、サービス毎にスクレイピングの手法を変えなければならない。本システムで現在対象としているデータベースは、アメリカ議会図書館¹、大英博物館²、ボストン美術館³、ヴィクトリア&アルバート博物館⁴、立命館アトリサーチセンター⁵、国立国会図書館デジタルアーカイブポータル⁶、国立国会図書館サーチ⁷である。その内SRUでの検索を行っているのは国立国会図書館デジタルアーカイブポータル、国立国会図書館サーチ、アメリカ議会図書館であり、他のデータベースに対してはスクレイピングを用いている。システムの処理の概要を図2に、システムでの検索結果の一例を図3に示す。

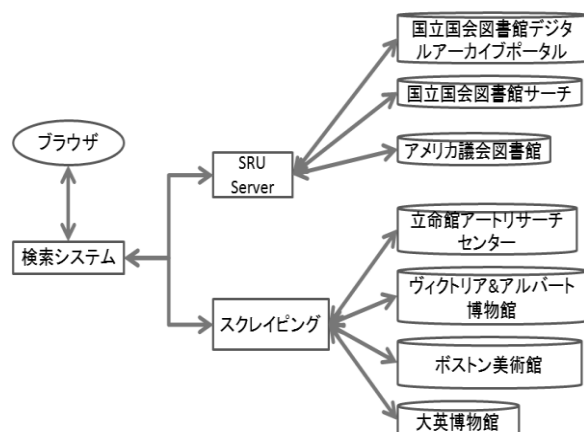


図2: 浮世絵横断検索システムの処理概要図

¹ <http://www.loc.gov/index.html>

² <http://www.britishmuseum.org/>

³ <http://www.mfa.org/>

⁴ <http://www.vam.ac.uk/>

⁵ <http://www.arc.ritsumei.ac.jp/>

⁶ <http://porta.ndl.go.jp/>

⁷ <http://iss.ndl.go.jp/>



図3: 浮世絵横断検索システムにおける検索結果の例
(‘utagawa hirosighe’で検索)

3.2 典拠データと Linked Data の利用

本研究では異種データベース間の多言語統合アクセスを実現するために Linked Data を用いる。Linked Data とは個々の情報源に対してグローバルな ID として URI を与え、個々の情報源の関係性を記述した任意の種類のリックを付与したデータの集合で構成される。Linked Data は RDF 形式で記述される。従来は個々のデータベースに個別にアクセスするしか方法がなかったが、Linked Data を用いることで、異なるデータベース間で関連する情報の間にリンクを張ることができ、複数のデータベースを跨いだアクセスが可能になる。

統合アクセスを可能にするための RDF データを得るために、国立国会図書館典拠データ検索・提供サービス(Web NDL Authorities)を用いる。Web NDL Authorities は国立国会図書館が維持管理する典拠データを一元的に検索・提供するサービスで、RDF 形式でデータを取得でき、SPARQL(SPARQL Protocol and RDF Query Language)による検索にも対応している。典拠データとは、標目を統制するために、標目の別表記や同義語、標目として選んだ根拠を示す資料等の情報をまとめて記録したものである。標目とは、異なる表記方法や同義語がある場合は一つにまとめ、同じ名称で異なるものを指す場合はそれぞれ区別して検索ができるようにしたものという。図4にその例を示す。国立国会図書館の典拠データには個人名、家族名、団体名、地名、統一タイトル、普通件名が含まれている。本システムではこのうち人名と普通件名を用いる。

著者の表記	標目
シェークスピア	→ Shakespeare, William, 1564-1616
シェイクスピア	
シェークスピア	
シェイクスピア, ウィリアム	
沙士比阿	

図 4：典拠データの例

浮世絵の分野でどの程度この典拠データが対応しているかを調べるために、原色浮世絵百科大事典[5]と浮世絵事典[6]から絵師の名前の一覧を取得し、Web NDL Authorities にて検索を行った。絵師のデータは、表記ができない漢字を含む名前は無視し、歌川広重 1 世 等の世代を含むものは「○世」の部分を削除した状態で重複しているものは一つの名前としている。絵師の名前の数は 1212 件でそのうち典拠データに含まれていた数は 310 件で約 26%であった。また同様に、Wikipedia の浮世絵師のリスト¹から検索した結果、668 件中 284 件で、約 57%が典拠データに含まれていた。この結果から、典拠データは浮世絵師に対しては必ずしも網羅性が高いとは言えないが、典拠データは既に Linked Data に対応しているという大きな利点があり、典拠データに含まれない人名については、浮世絵に関する辞典など他の情報源によって補完が可能であると考えている。

3.3 Linked Data を用いた統合アクセス

図 1 の(4)に示す通り、人名リンク生成モジュールはまずユーザが選択したデータから Web NDL Authorities の SPARQL 検索を用いて検索結果に関する典拠データを取得する。次に取得した典拠データの別名の項目の値を用いて、別のデータへアクセスすることが可能になる。着目する属性によって得られる結果が変わるが、今回は人名(絵師)に着目した統合アクセスを行う。これにより著者に関連するデータ群に対して統合的なアクセスが可能になる。図 5 にアクセスの一例を示す。この例では、「葛飾北斎」という検索結果から「葛飾北斎」の典拠データを取得し、その中の別名の項目から「前北斎」、「時太郎可候」、「勝川春朗」の情報源へのリンクを出力している。また「歌川広重 1 世」「歌川広重 2 世」のようにキーを含む典拠が複数検出された場合、ユーザが選択したデータの別項目を利用することにより、それらを区別することを検討してい

る。例として図 6 では作品の作成年から典拠の特定を行っている。



図 5：統合アクセスの例

また Web NDL Authorities が提供するデータの一部には LCSH(アメリカ議会図書館件名標目表)における対応語が存在する場合があるため、図 7 に示す通り日本語と英語の言語横断アクセスへの利用が可能である。ここでは国立国会図書館の典拠データにある「浮世絵」の同義語である「江戸絵」、「錦絵」、「Ukiyoe」、さらに LCSH における「Ukiyoe」の同義語である「Hashirae」、「Pillar Prints」、「Ukiyo-e」等のデータ取得が可能となる。これらのデータを用いることにより、浮世絵に関する知識がなくても、各データにおいて表記が違うがデータの内容が類似しているデータにアクセスが可能になる。

¹ <http://ja.wikipedia.org/wiki/浮世絵師一覧>

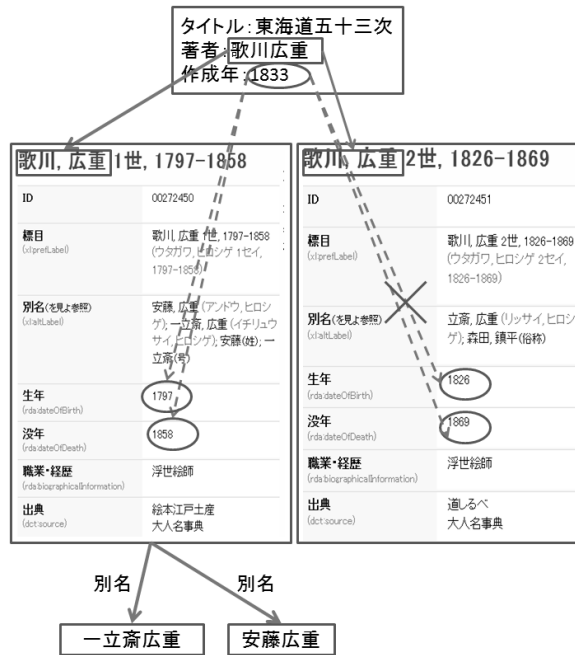


図 6: 典拠データの区別の例

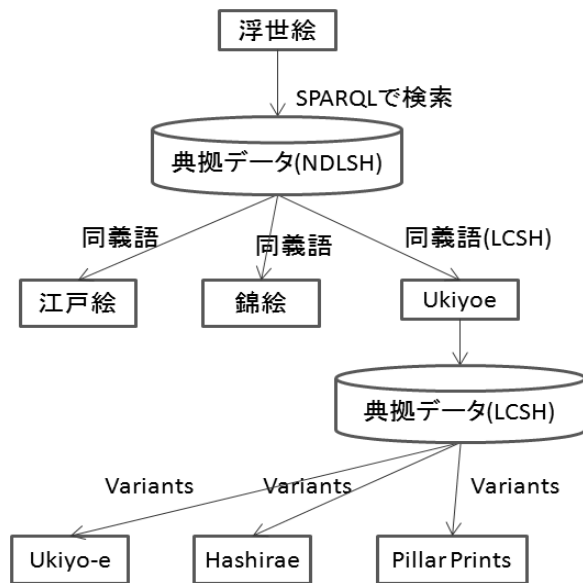


図 7: 多言語アクセスの例

ただしこれは NDL Authorities の典拠データに LCSH へのリンクがある場合に可能な事である。本研究では NDL Authorities の LCSH へのリンクがない場合でも可能な限りリンクを自動で作成するようにしている。図 8 にその例を示す。ユーザが選択したデータの著者の属性値である「歌川広重」を Web NDL Authorities で検索する。そこで得られたデータである「安藤広重」「一立斎広重」をまず検索クエリとする。図 5 に示すように典拠データには読みが表記されているので、読みをローマ字表記に変換することで英

語での検索にも対応できる。また、読みをローマ字変換した文字列や漢字で表記された人名をキーとして LCSH 内を検索することにより、LC の典拠データも利用できる。

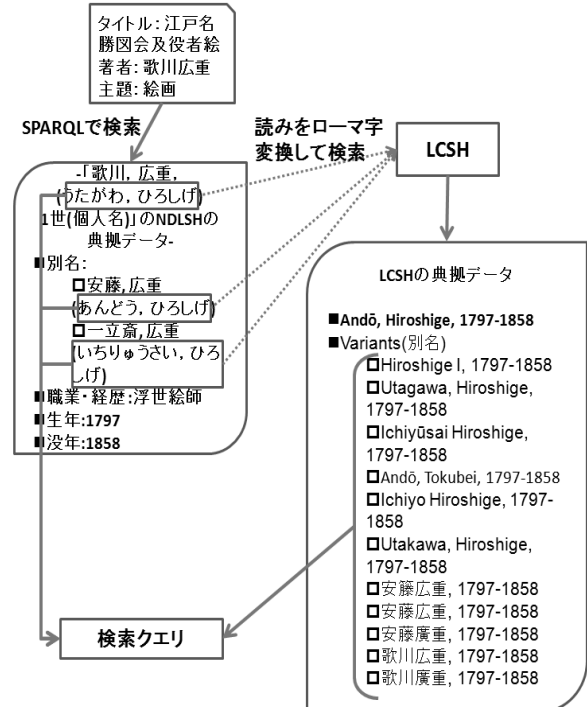


図 8: NDL Authorities から LCSH へのリンクの自動生成の例

最後に図 1 の(4)(5)で得られた別名をキーとして横断検索システムが対応していないデータベースを含む多種のデータベースに対して検索を行い、一致したデータへのリンク群をユーザに出力する。

4. おわりに

本論文では、複数の異なるフォーマットを持つデータベース間における多言語統合アクセス手法の一つとして Linked Data を用いた手法を提案した。これにより、ユーザが容易に関連付けられたデータにアクセスすることができ、単なる検索の効率化にとどまらず、ユーザの新たな知識の獲得の手助けになる可能性があると考えている。

本研究の今後の課題として、主に二つ挙げられる。一つは 3.2 節で述べたように浮世絵分野での典拠データの数が少ないことである。この点は浮世絵に関する辞典等のデータから機械的に情報を抽出・構造化したものを別途用意することで改善を図る予定である。二つ目に個々のデータへの URI の付与が考えられる。本研究では URI を動的に自動付与することを予定しているが、自動付与によって付与された URI が情報

源を一意に識別できるようにすることが課題である。

謝辞

本研究の一部は文部科学省グローバル COE プログラム「日本文化デジタル・ヒューマニティーズ拠点」、文部科学省私立大学戦略的研究基盤形成支援事業「芸術・文化分野の資料デジタル化と活用を軸とした研究資源共有化研究」、文部科学省科学研究費補助金若手研究(B)「言語・時代・文化横断型の情報アクセスに関する研究」(研究代表者:前田亮, 課題番号:21700271)の支援を受けている。

参考文献

- [1] 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: Linked Dataによる芸術・文化情報統合の試み, 人工知能学会研究会資料, SIG-SWO-A1101-04, 2011
- [2] Biligsaikhan Batjargal, Fuminori Kimura, Akira Maeda: Metadata-related Challenges for Realizing Federated Searching System for Japanese Humanities Databases, In Proceedings of the 11th International Conference on Dublin Core and Metadata Applications 2011 (DC-2011), pp.80-85, 2011
- [3] 原正一郎, 関野樹, 久保正敏, 柴山守: 地域研究支援のための時空間情報処理ツールの構築. 人文科学とコンピュータシンポジウム (じんもんこん 2009) 論文集. pp.163-170, 2009
- [4] 山本泰則, 安達文夫: 博物館資料情報統合検索のためのコアメタデータ, 人文科学とコンピュータシンポジウム (じんもんこん 2009) 論文集. pp.287-291, 2009
- [5] 檜崎宗重, 岡畏三郎: 原色浮世絵大百科事典, 大修館書店, 1981
- [6] 吉田暎二: 浮世絵事典, 画文堂, 1965