Title	エージェントと著作等に対する典拠コントロール支援用統合型典拠データベースの構築
Sub Title	Construction of an integrated authority database for supporting agent and work authority control
Author	谷口, 祥一(Taniguchi, Shōichi)
Publisher	
Publication year	2021
Jtitle	科学研究費補助金研究成果報告書 (2020.)
JaLC DOI	
Abstract	図書館目録における典拠コントロールの一層の充実をめざして、(a)個人や団体等というエージェントに対する国内の典拠データを仮想的に統合し、より包括的な典拠データとする方策を検討し検証した。また、(b)既存の書誌データから著作に関する事項を抽出し、包括的な著作典拠データを形成するために、抽出したデータが特定の著作か否かを機械学習の適用により判定する方式を試行した。(c)統合型の典拠データを適切に表現し管理できるメタデータスキーマの策定を意図して、既存の多様な語彙やスキーマの検討、複数モデルや語彙のマッピングとマージなどを実行した。 With the aim of further enhancing authority control in library catalogs, (a) we examined and verified ways to virtually integrate domestic authority data for agents such as persons and corporate bodies into more comprehensive authority data. In addition, (b) in order to extract information related to works from existing bibliographic data and form comprehensive authority data, we tried a method to judge whether the extracted data represents a given work by applying machine learning. (c) With the intention of developing a metadata schema that can appropriately represent and manage integrated authority data, we examined various existing metadata vocabularies and schemas for authority data, and also tried to the mapping and merge of multiple models and vocabularies.
Notes	研究種目：基盤研究 (C) (一般) 研究期間：2017～2020 課題番号：17K00452 研究分野：図書館情報学
Genre	Research Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KAKEN_17K00452seika

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ４ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００４５２

研究課題名（和文）エージェントと著作等に対する典拠コントロール支援用統合型典拠データベースの構築

研究課題名（英文）Construction of an integrated authority database for supporting agent and work authority control

研究代表者
谷口　祥一（Taniguchi, Shoichi）

慶應義塾大学・文学部（三田）・教授

研究者番号：５０２０７１８０

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：図書館目録における典拠コントロールの一層の充実をめざして、(a)個人や団体等というエージェントに対する国内の典拠データを仮想的に統合し、より包括的な典拠データとする方策を検討し検証した。また、(b)既存の書誌データから著作に関する事項を抽出し、包括的な著作典拠データを形成するために、抽出したデータが特定の著作か否かを機械学習の適用により判定する方式を試行した。(c)統合型の典拠データを適切に表現し管理できるメタデータスキーマの策定を意図して、既存の多様な語彙やスキーマの検討、複数モデルや語彙のマッピングとマージなどを実行した。

研究成果の学術的意義や社会的意義
図書館目録における典拠コントロールの一層の充実をめざして、わが国の実情に合致し、かつ必要性が高い事項に焦点を当て、有効な方策の提案とその検証を実施した。大きくは、(a)従来から実施されてきているが、国内において複数の典拠データが存在するエージェントの典拠データに関わる課題の解決と、(b)これまで殆ど実績がない著作の典拠コントロールの適用に伴う課題の解決に分けて、有効な方策の提案とその検証を実施した。これらによって、図書館目録が抱える大きな課題の解決に、研究レベルとはいえ、ある程度の見通しをつけることができたものと考える。

研究成果の概要（英文）：With the aim of further enhancing authority control in library catalogs, (a) we examined and verified ways to virtually integrate domestic authority data for agents such as persons and corporate bodies into more comprehensive authority data. In addition, (b) in order to extract information related to works from existing bibliographic data and form comprehensive authority data, we tried a method to judge whether the extracted data represents a given work by applying machine learning. (c) With the intention of developing a metadata schema that can appropriately represent and manage integrated authority data, we examined various existing metadata vocabularies and schemas for authority data, and also tried to the mapping and merge of multiple models and vocabularies.

研究分野：図書館情報学

キーワード：典拠データ　典拠コントロール　エージェント　著作　メタデータ　情報組織化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

図書館目録のメタデータは、移行期にあり、RDA（Resource Description and Access；資源の記述とアクセス）の刊行、欧米の国立図書館等による RDA 採用などにより大きく変化の途上にあった。国内でも、限られた数の図書館が既に RDA の適用を開始していた。一方、日本目録規則（NCR）は当時の RDA に依拠した 2018 年版の刊行に向けて、改訂作業が進行中であった。また、国立情報学研究所が運用する共同分担目録システム NACSIS-CAT は 2020 年に向けて、軽量化・合理化の方向で検討を進めている段階にあった。このように、わが国の目録作成・管理の進展に向けて、一層の軽量化・合理化を図りつつも、同時に RDA や新 NCR を積極的に採用し、それらが力点を置いている典拠コントロールに一層の充実化を図ることが求められていた。これらゆえ、わが国のデータや状況に即した典拠コントロールの実現およびその支援の方策、支援システムの開発などが特に求められている段階にあった。

２．研究の目的

(1)図書館目録における典拠コントロールの一層の充実をめざして、個人や団体等というエージェントに対する国内の典拠データを仮想的に統合し、より包括的な典拠データとすること、併せて、国際的な典拠データ間のマッピングとして公開されている VIAF（バーチャル国際典拠ファイル）のマッピング結果について、その妥当性を検証することを目的とした。個人や団体に対する典拠コントロールは従来から実施されてきているが、典拠形アクセス・ポイントの国内図書館における多様性（構成要素と表記言語の多様性）の存在、さらなる典拠作業の効率化に向けて国外を含む外部機関により作成された典拠データの流用の可能性などを背景とした研究目的の設定である。

(2)著作や表現形に対する典拠コントロール作業を支援するために、既存の書誌データから著作および表現形に関する事項を抽出し、包括的な著作・表現形典拠データを形成すること、特に機械的な抽出後に同一著作または同一表現形の判定を効率的に実行できる方策を検討することを目的とした。わが国では一部を除いて、殆ど実績の蓄積がない著作や表現形に対する典拠コントロール作業を新たに開始するに当たって必要となる支援を意図した研究目的の設定である。

(3)統合型の典拠データを適切に表現し管理できるメタデータスキーマを策定することを研究目的とした。旧来からの MARC フォーマットが典拠データの記録や共有に現在でも用いられているが、その限界は明らかであり、書誌データと同様に新たなメタデータスキーマが求められていることを背景とした研究目的の設定である。

３．研究の方法

(1)国立国会図書館作成の著者名典拠データと NACSIS-CAT の著者名典拠データとを照合し、加えてそれら典拠データからリンクする書誌データ間の照合を加えた典拠データの照合を試行した。また、国際的な典拠データ間のマッピングとして公開されている VIAF のマッピング結果（VIAF クラスタ）を取得し、上記２つの典拠データセットと重ね合わせることにより、VIAF におけるマッピングの誤同定と同定漏れの可能性の高い部分を抽出し、その後、人手により確認した。

(2)著作同定を支援する目的の下、難度が高いとされる日本古典著作の同定に向けて、人手により著作を判定された書誌データ群に対して、複数の方式により著作の手がかりを特徴量として抽出し、複数の機械学習モデルを適用して、その性能を評価した。

(3)エージェントおよび著作、表現形などを表現し管理する適切な典拠データ用スキーマの策定に向けて、既提案のメタデータスキーマとメタデータ用の語彙（RDF クラスとプロパティ）を複数取り上げ、それらの特徴、共通点および相違点、他スキーマへの事後的な変換などの観点から検討した。加えて、典拠データに関する複数のメタデータモデルの対応づけ（マッピング）と併合（マージ）の試行、多様な選択肢を取り得る RDA や日本目録規則 2018 年版（NCR2018）のエレメント等の語彙に対して適切な RDF 定義を導くフレームワークの提示などを試行した。

４．研究成果

(1)個人や団体等というエージェントに対する国内の典拠データを仮想的に統合し、より包括的な典拠データとすることに関して、以下の成果を得た。

①国立国会図書館作成の著者名典拠データと NACSIS-CAT 著者名典拠データの照合、およびリンクする書誌データ間の照合を加えた典拠データの照合を試行した。それぞれの典拠データの特性、言語・文字種の区分や表記と読みの分離など、データの最小構成要素を適切に分節化した上

での照合法としている。それぞれの典拠データ集合に含まれる名称について、総名称数、そのうち重複がない名称と重複がある名称を算出し、参照形を加えたときには、重複となる名称が増加すること、さらに読みと異体字処理を加えることにより重複となる部分が大幅に増大することを確認している。次に 2 つの典拠データ集合における名称の重複出現状況を調査し、名称総数（異なり数）、一方のみに出現する名称、両集合に対応する名称が 1 つずつ含まれる（1 対 1 対応）名称、1 対多または多対多の対応となる名称群を得た。さらに、参照形を加えたり、異体字処理を適用した場合には、分布が多少とも変動することを確認した。

②併せて、国際的な典拠データ間のマッピングとして公開されている VIAF のマッピング結果（VIAF クラスタ）について、その妥当性検証を目的に、日本名の典拠形アクセス・ポイントをもつ個人を対象に、効率的な検証方法の提案とその試行を行った。すなわち、国立国会図書館と NACSIS-CAT の典拠データおよび書誌データを用いて、誤同定や同定漏れの可能性が高い部分を効率的に特定し、特定された部分のみ人手による検証を行う方法を提案した。誤同定の可能性が高い部分として、単一 VIAF クラスタ内で、①同一機関作成の典拠データが複数属するもの、②両機関の典拠データの名称、参照形、名称カナ読みのいずれも一致しないものを機械的に特定し、それらに誤同定が含まれていることを確認した。同様に、同定漏れの可能性がある部分として、名称が一致した部分（1 対 1、1 対多、または多対多で一致）について、VIAF によるマッピング結果を重ね合わせ、異なるクラスタとされた部分を特定した上で、それらがリンクしている書誌データ同士の機械的照合を実行し合致するものを見つけることを試みた。書誌データの機械的照合は一定程度の性能のみ期待でき、完全さを求めることはできないが、今回の試行では、典拠データにリンクしている書誌データ同士の照合に限定しているため、照合回数は一定数内に抑えることができている。これらの結果、同定漏れの事例を多数検出することができた。

(2)既存の書誌データから著作に関する情報を抽出し、包括的な著作典拠データを形成することを意図して、特に難度が高いとされる日本古典著作の著作同定について機械学習の適用を試みた。

①実験 1 は、人手により判定された書誌データ群からタイトルと読み、責任表示と著者標目など著作判定に関わる項目群から値を抽出し特徴量とし、設定済みの個別の著作を予測させる多クラス分類問題として実施した（データ数 22 万件）。そこでは、FRBR 研究会が以前に人手により判定した書誌データ群のうち、当該著作に属する書誌データが 10 件以上ある古典著作 89 と、これらのいずれにも含まれないものからランダムに抽出して「非該当」クラスのデータとして整備した。この結果、「非該当」クラスを含めて分類対象クラス数 90、総データ数 22 万件を実験対象データとした。それらのデータに対して、記号の除去などの表記の正規化を加え、機械学習実験用の特徴量（属性値）とした。特徴量の抽出方式については複数設定し、著作同定性能への影響を実験により検証した。大きくは特徴量の抽出方式を 3 つに区分し、最小限のものから順次情報量が増加する仕様とした。機械学習モデルとして比較的計算量が少なく、かつ性能が安定しているといわれるロジスティック回帰、線形 SVM、ランダムフォレストなどを採用し、一定程度以上の性能値が得られている。併せて、個別の著作によって、その同定性能に幅があることも確認できた。

②実験 2 は、2 つの書誌データが同一著作を表すかを予測する 2 クラス分類問題として構成し、その性能値を確認した（データ数 376 万件）。前述の実験 1 による機械学習適用方式は、正解データが存在する事前設定の著作についてのみ予測が行われ、事前に設定されていない著作については、「非該当」との予測が行えるのみである。つまり、新たな（未知の）個別著作を同定する機能はない。そこで、2 つの書誌データの組み合わせ（ペア）が同一著作を表すのか、それとも異なる著作を表すのかを機械学習により予測する 2 クラス分類問題とした。書誌データのペアにおいて、タイトルなど、13 項目についてその値の一致・不一致を表したものを特徴量として採用した。機械学習モデルとして、ロジスティック回帰、ランダムフォレストなどを採用した。最終的には書誌データペアを形成しているレコード ID から書誌データ単位でのグループ化と集計を別途行い、その性能値を複数の著作について求めた。その結果は、全体的に精度が低く、再現率が高くなる傾向が見られた。これは複数の著作のデータが混合して大きなグループを形成した結果であり、こうしたグループ（2 部グラフ）を適切に分割する方法の検討が残されている。

(3)エージェントおよび著作、表現形などを表現し管理する適切なメタデータスキーマ、かつ統合型の典拠データを適切に表現し管理できるメタデータスキーマを策定することに関して、以下の成果を得た。

①既提案のメタデータスキーマとメタデータ用の語彙を複数取り上げ、それらの特徴、共通点および相違点、他スキーマへの事後的な変換などの観点から検討した。取り上げたスキーマ等は、(a)RDA に対応して RDA Registry に登録されている RDF クラスとプロパティ、(b)米国議会図書館が主導するプロジェクトにおいて提案された最新版の BIBFRAME 2.0、さらには(c)IFLA Library Reference Model(IFLA LRM)を RDF による表現に変換したモデルおよび語彙、(d)FRBR

等と CIDOC CRM との統合モデルである FRBRoo などとした。たとえば、RDA の適用を想定したメタデータ用のスキーマの場合、候補となる MARC21、RDA 語彙を用いたスキーマ、BIBFRAME、および MARC21 から他の 2 つへの事後的な変換という 5 つの方式それぞれに対して、設定した複数の観点から検討し、それぞれの利点と問題点などを明らかにした。これに伴い、今後の採用が有力視されている BIBFRAME について、RDA に依拠したメタデータの適切な受け皿となるのか、あるいは RDA を超えて情報資源に対する多様な記述メタデータの受け皿となりうるのかを検討し、その問題点や課題等を明らかにした。

②メタデータスキーマの基盤となる概念モデルレベルの検討として IFLA LRM と BIBFRAME 2.0 を取り上げ、前者の IFLA LRM を実体関連モデルから RDF による表現に変換した上で、先ずクラス間、プロパティ間のマッピングを検討し、両者のモデルおよび語彙の共通点と相違点を確認した。さらに、両者のマージ（併合）を検討し、主な相違点ごとにマージ処理における選択肢が存在すること、および選択肢によっては先行する他の選択肢群における選択結果に依存するもの、さらには選択肢間で衝突するものや冗長さを生み出すものがあることを示し、併せてマージ処理結果を示した。

③RDF データには、個別のリソース（インスタンス）に対するメタデータ記述と、そうしたメタデータ記述に用いる語彙（クラスとプロパティ）自体を定義したものがあり、後者はオントロジーとも呼ばれる。RDF と RDFS が規定するクラスとプロパティのみを用いて定義されているオントロジーを対象として想定し、複数の異なるオントロジー間のマッピンとマージの形式化を試みた。異なるオントロジーにおいてクラス間のマッピングが成立する要件として、意味範囲の包含関係にかかわる条件を 2 つ明示し、またプロパティ間のマッピングが成立する要件については、意味範囲、定義域、値域にかかわる条件 4 つを明示した。次に、2 つの異なるオントロジーから第 3 のオントロジーを生成するマージについて、その処理のレベルを 3 つに分け、それぞれにおけるクラスとプロパティのマージ処理の要件と取り得る選択肢を提示した。なお、クラスの意味範囲の包含関係は通常機械的には判断できないため、ここでは人手による判定を前提としている。

④適切な典拠データ用スキーマの策定に向けて、RDA や NCR2018 の語彙に対する適切な RDF 定義を導く１つの方法論としてフレームワークを提示した。現在公開されている RDA 語彙および公開に向けて準備がなされていた NCR2018 の語彙を対象に、それを理解し、適切な RDF 定義を導く方法論であるフレームワークを提示した。このフレームワークは、(a)概念モデル（FRBR または IFLA LRM）を適用した段階に対応する定義、(b)属性および関連のエレメントの「記録の方法」の指示に依拠した定義、(c)関連のエレメントをクラス化して再構成した定義、(d)上記以外の構造表現力を導入した定義という４層から構成した。(c)は膨大な数のプロパティを抱える RDA 語彙に特に有効であり、(d)はエレメントの下でのグループ化と名称等の値の構造的表現を実現するために必要となる。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 谷口祥一	４．巻 67
２．論文標題 日本目録規則2018年版の語彙をRDFによって定義する：フレームワークアプローチ	５．発行年 2021年
３．雑誌名 日本図書館情報学会誌	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shoichi Taniguchi	４．巻 56
２．論文標題 Mapping and Merging of IFLA Library Reference Model and BIBFRAME 2.0	５．発行年 2018年
３．雑誌名 Cataloging & Classification Quarterly	６．最初と最後の頁 427～454
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/01639374.2018.1501457	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 橋詰秋子，谷口祥一	４．巻 70
２．論文標題 書誌情報とメタデータ：理論，ツールの2010年代のわが国における展開	５．発行年 2018年
３．雑誌名 図書館界	６．最初と最後の頁 305～314
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Taniguchi Shoichi	４．巻 56
２．論文標題 Is BIBFRAME 2.0 a Suitable Schema for Exchanging and Sharing Diverse Descriptive Metadata about Bibliographic Resources?	５．発行年 2017年
３．雑誌名 Cataloging & Classification Quarterly	６．最初と最後の頁 40～61
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/01639374.2017.1382643	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 谷口祥一	4．巻 67
2．論文標題 概念モデル構築を中心としたデータベース設計とメタデータ設計	5．発行年 2017年
3．雑誌名 情報の科学と技術	6．最初と最後の頁 442～447
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 谷口祥一
2．発表標題 NCR2018とRDAによる語彙のRDF定義とメタデータスキーマ
3．学会等名 2020年度日本図書館情報学会春季研究集会
4．発表年 2020年

1．発表者名 谷口祥一
2．発表標題 VIAFによる典拠レコードマッピングは適切か：日本名個人名を対象とした検証方法の提案
3．学会等名 2019年度日本図書館情報学会春季研究集会
4．発表年 2019年

1．発表者名 谷口祥一
2．発表標題 書誌レコードに対する著作同定に機械学習を適用する試み：日本古典著作の事例
3．学会等名 第67回日本図書館情報学会研究大会
4．発表年 2019年

１．発表者名
谷口祥一

２．発表標題
IFLA Library Reference ModelとBIBFRAME 2.0の統合：マッピングからマージへ

３．学会等名
2018年度日本図書館情報学会春季研究集会

４．発表年
2018年

１．発表者名
谷口祥一

２．発表標題
RDFオントロジーのマッピングとマージの形式化

３．学会等名
2018年度三田図書館・情報学会研究大会

４．発表年
2018年

１．発表者名
谷口祥一

２．発表標題
RDAとNCR2018にとって適切なメタデータスキーマとは何か

３．学会等名
第66回日本図書館情報学会研究大会

４．発表年
2018年

１．発表者名
谷口祥一

２．発表標題
多様な記述メタデータの交換・共有用スキーマとしてBIBFRAME 2.0は適切か

３．学会等名
2017年度日本図書館情報学会春季研究集会

４．発表年
2017年

１．発表者名
谷口祥一

２．発表標題
BIBFRAME 2.0の概要と問題点：米国議会図書館の本気度を改めて問う

３．学会等名
日本図書館研究会情報組織化研究グループ月例会

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

谷口祥一ホームページ
http://user.keio.ac.jp/~taniguchi/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関