様式 C-19

科学研究費補助金研究成果報告書

平成２３年　４月２５日現在

機関番号：１２１０２
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５００２２１
研究課題名（和文）　根拠の記録を伴う書誌データと記述メタデータの効率的作成法・高度活用法に関する研究
研究課題名（英文）　evidence recording in bibliographic records and descriptive metadata: its efficient creation and effective use
研究代表者
谷口　祥一　（TANIGUCHI SHOICHI）
筑波大学・大学院図書館情報メディア研究科・教授
研究者番号：５０２０７１８０

研究成果の概要（和文）：
メタデータの品質・信頼性の向上および相互運用性の実現等に向けて、メタデータとそれを構成するデータ項目（要素）値に必要な根拠を記録することを提案し、その有効かつ効率的な記録法と高度な活用法をシステム試作や評価実験を通して検証した。特に、図書等に対する書誌データ（書誌レコード）を対象とし、書誌同定（重複書誌レコードの自動判定）の性能向上、根拠記録の作成自動化、著作同定用根拠記録の作成と活用などを試みた。

研究成果の概要（英文）：
Recording evidence for data values, in addition to the values themselves, in metadata is proposed, with the aim of improving the quality, reliability and interoperability of such metadata. Some prototype systems were developed and also experiments with those systems were conducted to evaluate and validate the ways of recording evidence and utilizing it. Bibliographic records for books and other resources, in particular, were focused on, and (a) identifying records for the same resource (i.e., duplicate records) with recorded evidence, (b) creating recorded evidence in an automatic way, and (c) creating and utilizing recorded evidence for work identification were examined.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,600,000	480,000	2,080,000
２００９年度	1,000,000	300,000	1,300,000
２０１０年度	700,000	210,000	910,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：図書館情報学
科研費の分科・細目：情報学　・　図書館情報学情報学・人文社会情報学
キーワード：書誌データ、メタデータ、根拠記録、著作同定、書誌同定

１．研究開始当初の背景

図書館等における所蔵リソースに対する書誌データや、ネットワーク上の Web リソースに対する記述メタデータは、基本的に個々の定義されたデータ項目に対してその採用された値のみを記録している。図書館等が作成する書誌データは、現在でも高品質なメタデータの実例ではあるが、より広範な交換・共有等の観点からは、さらなる永続性と相互運用性を備えたメタデータとすることが求められている。そのために多様な方策が模索され検討されてきているが、現時点では課題が多数残されている。

メタデータの品質・信頼性の向上および相互運用性の実現に向けて、本研究代表者は、書誌データや記述メタデータにデータ作成の根拠を記録することを提案し、その有効性の検証を以前から試みてきた。本研究は、それらの成果を踏まえ、さらに展開を図ることを意図している。

他方、書誌データや記述メタデータにデータ作成の根拠を記録するという検討や検証は、これまでに取り組み例はない。例えば、(a)書誌データや記述メタデータの作成者向けの学習支援、データ作成処理自動化の研究、(b)同一リソースに対する書誌データやメタデータの書誌同定（重複データ同定）の研究などについて、これまでに相当数の蓄積があるとはいえ、根拠の記録とその活用との観点から研究を試みたものは皆無である。唯一、本研究と接点を有するのは、メタデータ自体の由来・出所（provenance）の記録化を図る研究であるが、本研究はそれら由来・出所を超えて広範な根拠を記録し活用することを当初から意図している。

２．研究の目的

メタデータの品質・信頼性の向上および相互運用性の実現等に向けて、メタデータとそれを構成するデータ項目（要素）値に必要な根拠を記録することを提案し、その有効かつ効率的な記録法と高度な活用法をシステム試作や評価実験を通して検証する。

根拠の記録から活用までの場面を想定した複数のシナリオを設け、それぞれに適した根拠の記録内容および方式の検討、根拠記録作業の自動化または支援システムの開発、記録された根拠の活用法の検討とそのシステム開発、そして有効性検証実験による評価などを行うことを目的としている。

(1)書誌同定（重複書誌レコードの自動判定）の性能向上：書誌レコード同士を比較・照合し同一リソースに対するレコードであるのか判定する従来型の手法に加えて、根拠の記録を組み合わせた照合法・判定法を確立する。根拠記録を伴う書誌レコードと、根拠記録を持たない既存書誌レコードとの照合を想定し、特に精度の高い書誌同定（重複判定）に向けて複数の判定方式の適用を試みる。

(2)書誌レコード／記述メタデータ統合・更新処理の最適化：異なるデータ項目セットまたは記述方式に基づき作成された書誌レコード／記述メタデータの統合や更新において根拠を活用する方策を確立する。

(3)根拠記録の作成自動化：書誌レコード／記述メタデータへの根拠記録作業を自動化する、またはメタデータ作成者が当該作業を効率的に行えるよう支援するシステム機能を充実させる。

(4)著作同定用根拠記録の作成と活用：同一著作に属するリソース群の書誌レコードを同定する著作同定作業に、根拠記録の作成と活用を組み入れることにより、同定作業を精度高くかつ効率的に実施できる方策として確立する。

３．研究の方法

(1)書誌同定の性能向上：複数の根拠記録方式、判定方式により同一リソースに対する書誌レコード（重複書誌レコード）を自動判定するシステムを構築した。使用する根拠記録は、対象とするリソース（図書など）の情報源を OCR ソフトを用いてテキストデータ化した情報源コピー、それに書誌的カテゴリ（タイトル、責任表示等の区分）のマークアップを適用したもの、書誌レコードの個別データ項目値に処理ルール（タスクとアクション）を記録したものなど、複数方式を採用した。

評価実験には、所属大学の図書館が所蔵する欧文図書を使用し、対応する書誌レコードは当該図書館によるもの、米国議会図書館によるもの、その他を使用した。

また、(a)書誌レコード同士、つまり起点とするレコードと照合対象レコードとの照合機能、(b)照合対象レコードと根拠記録（起点レコードの根拠記録）との照合機能、(c)両者を組み合わせた照合機能の、それぞれに基づき書誌同定（重複判定）を行う機能を実装した。

(2)書誌レコード／記述メタデータ統合・更新処理の最適化：同一リソースに対する書誌レコード群とその根拠記録を多様性と整合性に配慮しつつ統合・更新すると同時に、根拠記録自体の統合による充実化を試みた。統合された根拠記録は、その後の書誌同定（重複

判定）に直接用いるとともに、多数の蓄積を形成することにより判定処理やルールの洗練化などに活用する。

(3)根拠記録の作成自動化：情報源コピーと、当該リソースに対応するデータ項目値（または既存の書誌レコード）を受け取り、情報源コピーにおいて該当する出現箇所を特定しマークアップを行い、併せて記録された値を得るまでの各タスクのアクションと入出力データを推定する機能をもつシステムを構築した。ここで「タスク」とは、データ項目の記述処理をその手順に沿って分割し定義したものを指す。データ項目値ごとにその取扱の詳細を規定した「目録規則」のルールは全般的に抽象度が高く、また場合分け（処理の分岐）が必ずしも明瞭かつ整合的とは限らないため、根拠記録に適したタスクそしてアクションを別途定義した。加えて、個々のアクションと目録規則中の個別ルールとの対応づけを人手で実施することにより、目録規則のルールからも参照できるようにした。

なお、システム構築においては、特にマークアップ処理や記述処理のアクション推定などの洗練化（処理ルールの優先度の調整など）に重点を置いた。

評価実験には、所属大学の図書館が所蔵する欧文図書を使用し、対応する書誌レコードは当該図書館によるもの、米国議会図書館によるもの、その他を使用した。

(4)著作同定用根拠記録の作成と活用：
①人手による著作同定用根拠記録の作成：人手による著作同定作業を、わが国の古典著作など特に同定処理の困難度が高い著作に対して実施した。JAPAN/MARC 書誌レコードを用いて、主要な古典著作ごとに包括的な検索を行い、ヒットした書誌レコード群をダウンロードし同定作業用ファイルとした。それらレコードに対して人手による判定を実施し、個々の書誌レコードに該当する著作のタイトルが出現する箇所を記録しマークアップする形式で判定結果を蓄積した。

併行して、著作同定の基準の明確化（ガイドラインの策定）を図った。FRBR（「書誌レコードの機能要件」）における著作の定義に依拠しつつ、わが国の古典著作に合致した詳細な基準を順次定めた。また、同定作業を支援するツール群を試作した。

こうして同定箇所を記録したレコード群から、個々の著作ごとに同定用の根拠記録を集約した著作典拠レコードに準じるもの、および同定ルール等の機械的生成を試みた。
②機械的照合による著作同定用根拠記録の作成：同一著作に属する書誌レコード群を機械的に同定する手法を、国立国会図書館の JAPAN/MARC レコード、さらには同館が運営する総合目録「ゆにかねっと」のレコード（DC-NDL 形式）に適用した。なお、後者については同一リソースに対応するレコード群を同定する書誌同定を併行して実施した。

個々の書誌レコードから著作同定（および書誌同定）に必要な項目値を抽出した後、正規化を行い、照合する方式とした。レコードの形式の相違（JAPAN/MARC レコードと DC-NDL 形式レコードの相違）に依拠して、それぞれ適切な方式を設定した。さらに、照合用のデータベースを内部的に保持する方式と、それらを用いず、直接同定用のキーの一致を照合する方式とをそれぞれ採用した。可能な複数の項目値の組み合わせによって同定処理を実施し、いずれの組み合わせによる同定性能が優れているのかを、別途人手により形成した正解集合を用いて性能評価した。また、性能が優れていた方式により同定されたレコード群から、著作同定用の根拠記録の集約に該当するものを機械的に生成した。

４．研究成果
(1)書誌同定の性能向上：実験の結果、テキストデータ化した情報源コピーを用いた同定処理、さらにはそれに書誌的カテゴリのマークアップを行った情報源コピーを用いた同定処理は、書誌レコードの個々のデータ項目（書誌要素）のレベルでは同定性能の多少の上昇をもたらすが、それらを組み合わせたレコードのレベルでは顕著な性能向上を示すに至らなかった。具体的には、今回の実験集合においては、単一または少数の書誌要素（たとえば出版年など）のみで高い性能値が得られており、それを大きく向上させることはかなり困難であった。

今後は、書誌要素ごとに異なるスコア、さらには一致の程度に応じたスコアを適用し、閾値を超えたときに一致と判定する方式などの適用が必要となろう。

(2)書誌レコード／記述メタデータ統合・更新処理の最適化：同一リソースに対する書誌レコード群の根拠記録を統合化したが、今回の実験は事例数が少なく、かつ類似した構成をもつレコード群であったことなどにより、その後の書誌同定（重複判定）への活用に際して明確な性能向上は確認できなかった。

(3)根拠記録の作成自動化：構築したシステムの機能である、情報源コピーにおいて該当する出現箇所を特定しマークアップを行い、併せて記録された値を得るまでの各タスクのアクションと入出力データを推定する機能については、それぞれ一定程度の精度が得られた。同時に、実用レベルの処理とするには、

いくつか問題が残されていることが判明した。
なお、作成された根拠記録は人手による確認と修正を経ることにより、書誌レコードそして情報源コピーと合わせて蓄積していくことにより、(a)書誌レコード作成者によるレコード作成作業の理解支援、すなわち記録された根拠をそれぞれのデータ項目値の記述作業の内容説明として参照することで理解を図る、あるいは(b)書誌同定処理システムの性能向上に向けた大規模な正解集合の形成に通じる。

(4)著作同定用根拠記録の作成と活用：
①人手による著作同定作業を、わが国の古典著作などに実施し、同定作業結果を着実に蓄積している。個々の書誌レコードに記録された同定結果から機械的に著作のレコードを生成し、それを根拠と位置づけ活用した。具体的には、新規または未判定の書誌レコードに対する著作同定処理において、先の根拠記録としての著作レコードを参照することで、同定処理の精度向上または作業効率向上に寄与できることを確認した。同時に、同定結果の分析・集計から、単純な文字列の照合に基づく機械的な同定処理では十分なレベルの同定が極めて困難であることを明らかにした。併せて、著作同定の基準自体も根拠の記録と位置づけられることを確認した。
今後は、(a)さらに広範な著作に対する同定作業、(b)複数人によるダブルチェックなど判定結果の妥当性の保証、(c)JAPAN/MARC 書誌レコード以外の書誌レコードに対する同定作業（追加の手間を最小限に抑えた同定作業方式の確立）が求められる。
②著作同定作業が極めて困難な古典著作などを除いては、機械的な著作同定は実用的な性能を示すことが判明した。また、採用するデータ項目の範囲、適用する文字列正規化の範囲などによって性能値が多少変化することも確認した。全体的には、採用する項目値の範囲を広くし、広範な文字列正規化を適用するほど、性能が上昇する結果となった。すなわち、全般的に誤同定が発生する可能性は低く、いかに同定漏れを抑えるかが性能を決定することが分かった。
また、機械的な同定では性能上の限界があり、前記の人手による同定作業結果を流用し、機械的同定結果を上書きして訂正する方式が有効に適用できることが分かった。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

①谷口祥一．メタデータの現在：最近のトピック，ダブリンコア，そしてセマンティックWeb．情報の科学と技術．査読無．Vol.60，No.12，2010，p.482-488．

②谷口祥一，鴇田拓哉．書誌情報とメタデータ：理論，ツールのわが国における展開．図書館界．査読無．Vol.61，No.5，2010，p.572-580．

③谷口祥一．FRBR OPAC 構築に向けた著作の機械的同定法の検証：JAPAN/MARC 書誌レコードによる実験．Library and Information Science．査読有．No.61，2009，p.119-151．

④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．査読無．No.17，2008，p.3-23．

⑤谷口祥一．Google 時代の目録教育・メタデータ教育．情報の科学と技術．査読無．Vol.58，No.9，2008，p.454-459．

〔学会発表〕（計４件）
①Shoichi Taniguchi，Keita Tsuji，Fuyuki Yoshikane．LIS research in Japan．Conference 2010: Research and Education of Library and Information Science in China，Korea and Japan．2010 年 9 月 10 日．筑波大学（茨城県）

②谷口祥一，上田修一，横山幸雄，鴇田拓哉，向當麻衣子，宮田洋輔．OPAC の FRBR 化を目指した人手による著作同定作業：FRBR 研究会の取り組み．日本図書館情報学会．2010 年 5 月 29 日．同志社大学（京都府）

③松本聖，谷口祥一．NDC Finder：自由語からの主題検索機能を提供する図書館 OPAC 検索支援システム．日本図書館情報学会．2010 年 5 月 29 日．同志社大学（京都府）

④谷口祥一．FRBR OPAC 構築に向けた著作の機械的同定法の検証：JAPAN/MARC 書誌レコードによる実験．日本図書館情報学会．2009 年 5 月 23 日．駿河台大学（埼玉県）

〔図書〕（計１件）
①谷口祥一．勉誠出版．メタデータの「現在」：情報組織化の新たな展開．2010，154p．

〔その他〕
ホームページ等
http://www.slis.tsukuba.ac.jp/~taniguch/

６．研究組織
(1)研究代表者
谷口　祥一　（TANIGUCHI SHOICHI）
筑波大学・大学院図書館情報メディア研究科・教授
研究者番号：５０２０７１８０

(2)研究分担者
なし

(3)連携研究者
なし