機関横断型文献情報 Wiki による著者情報対応付けの試み

日向野達郎† 増田 英孝†† 山田 剛一†† 清田 陽司††† 中川 裕志††††

† 東京電機大学大学院 未来科学研究科 情報メディア学専攻 〒 120–8551 東京都足立区千住旭町 5
†† 東京電機大学 未来科学部 情報メディア学科 〒 120–8551 東京都足立区千住旭町 5
††† 株式会社ネクスト 技術基盤本部 リッテル研究所 〒 108–0075 東京都港区港南 2–3–13 品川フロントビル
†††† 東京大学 情報基盤センター 学術情報研究部門 〒 113–0033 東京都文京区本郷 7–3–1
E-mail: †higano@cdl.im.dendai.ac.jp, ††{masuda,yamada}@im.dendai.ac.jp, †††kiyota@littel.co.jp,
††††n3@dl.itc.u-tokyo.ac.jp

あらまし 論文や書籍を集積している機関が Web 上で文献検索サイトと呼ばれるポータルサイトを提供している．文献検索サイトはそれぞれ独立した機関が提供しているので各々のメタデータの対応付けがなされていない．そのため，網羅的に文献を探しているユーザにとっては，複数のサイトで検索を繰り返す必要があり，手間がかかってしまう．そこで本研究では MediaWiki を用いることによって，複数の文献検索サイトのメタデータを容易に対応付けることを可能にする枠組みを提案している．現在は検証のため文献検索サイト内の著者情報を対象としてメタデータを対応付ける試みを行なっている．

キーワード メタデータの統合，MediaWiki，名寄せ，情報修正

Metadata Mapping of the Same Author by Cross-agency Bibliographic Information System using MediaWiki

Tatsuro HIGANO†, Hidetaka MASUDA††, Koichi YAMADA††, Yoji KIYOTA†††, and Hiroshi NAKAGAWA††††

† Tokyo Denki University
5 Senjuasahicho, Adachi, Tokyo, 120–8551 Japan
†† Tokyo Denki University
5 Senjuasahicho, Adachi, Tokyo, 120–8551 Japan
††† NEXT Co,.Ltd
2–3–13 Konan, Minato, Tokyo, 108–0075 Japan
†††† University of Tokyo
7–3–1 Hongo, Bunkyo, Tokyo, 113–0033 Japan
E-mail: †higano@cdl.im.dendai.ac.jp, ††{masuda,yamada}@im.dendai.ac.jp, †††kiyota@littel.co.jp,
††††n3@dl.itc.u-tokyo.ac.jp

1. はじめに

現在 Web 上には，「CiNii」[1] や「J-GLOBAL」[2] 等の，文献検索サイトと呼ばれるポータルサイトが存在する．これらのサイトは論文や書籍等の文献をを集積している機関によって運営され，過去から最新の研究成果や，関連研究の調査の際に非常に便利なサービスとして多くの研究者に利用されている．しかし，それぞれ別々の機関によってメタデータが管理されているため，複数のサイト上の文献を横断的に検索することができない．そのため，網羅的に文献を探しているユーザにとっては，複数のサイトで検索を繰り返す必要があり，非常に手間がかかってしまうというのが現状である．

本研究では，機関の枠を超えて文献情報を横断的に検索することを可能にするサービスの開発を目的としている．このためには各機関がメタデータに対して，それぞれ割り当てている固有の識別番号 (論文 ID，著者 ID 等) を互いに対応付ける必

要がある．そこで MediaWiki を用いることによって，複数の文献検索サイトのメタデータを容易に対応付けるための枠組みを提案している．本研究では，主に人物情報を対象として，MediaWiki 上で対応付けが可能であるかどうかを検証する．具体的には，各文献検索サイトから機械的にメタデータを収集し，MediaWiki に自動的に登録するシステムを試作し，登録されたデータを人手で名寄せする作業を行い，MediaWiki の仕組みが有効であることを示す．

2. 文献検索サイトの統合

国立情報学研究所の「CiNii」や，科学技術振興機構の「J-GLOBAL」等の文献検索サイトが様々な機関から提供されているが，横断検索等のサービスの統合はなされていない．例として「CiNii」の著者情報のページから他機関のサービスである「J-GLOBAL」へのリンクというものが存在しているが，あくまで「J-GLOBAL」でその著者の名前を検索した結果のページへのリンクであり，直接「J-GLOBAL」の著者情報ページへリンクされているわけではない．これは国立情報学研究所と科学技術振興機構がそれぞれ所有している情報に対して，それぞれが独自に割り当てている識別番号（論文 ID や著者 ID 等）が互いに対応付けられていないために起こる問題である．お互いの機関との対応付けを行おうとしても、それぞれが情報の管理に独自の規格 (メタデータフォーマット) を利用しているため共通のメタデータフォーマットを作成するためには人手や時間等の多大なコストを各組織が支払わなければならないため，対応付け行うことは難しい．

さらに著者データベースでは，著者の所属の変更や結婚などの理由による姓名の変更によって発生する重複レコードや，同姓同名の複数の人物のレコードを機械的に判別し修正することは難しい [3]．そこで人手による修正が必要になってくるがここでもいくつかの問題がある．一例として「CiNii」では，重複レコードが存在していることに気づいたユーザによる「同一人物の報告」という機能が存在する．しかし，ユーザによる報告の後，機関の人間が確認し報告が正しければ修正が行われるというように，間違いの発見から修正までに時間がかかってしまう．また他のサイトにおいては，著者情報の修正機能があっても，自分の情報のみであり，自分でない他の著者の情報に間違いを発見しても修正することができないので，著者本人が間違いに気づくまで情報が間違ったままとなってしまうのが現状である．

そこで本研究では，各機関を間接的につなぎ機関の情報を横断的に検索するサービスの開発を目的とする．さらにユーザの手で情報の修正を即座に行える仕組みを取り入れるということが本研究の特徴的な点である．　この目的のために，それぞれの機関のメタデータに割り当てられている ID に着目し，同一の情報同士の ID を各サイトから収集し，対応付けを行うことでサイトを統合するシステムを MediaWiki を利用して構築する．図 1 に ID 統合の概念図を示す．

図 1 に示すように，従来の文献検索サイトは一方向のハイパーリンク等のつながりが存在するのみであったが，本システムは各機関の ID を対応付けることで間接的にサイトをつなぐことを可能にしている．また，MediaWiki を利用することでユーザが情報に間違いを発見した場合，そのユーザの手で即座に修正をすることも可能になる．本研究では，機関の所有するメタデータの内，まず手始めに著者情報を対象として MediaWiki 上で対応付けを行う．

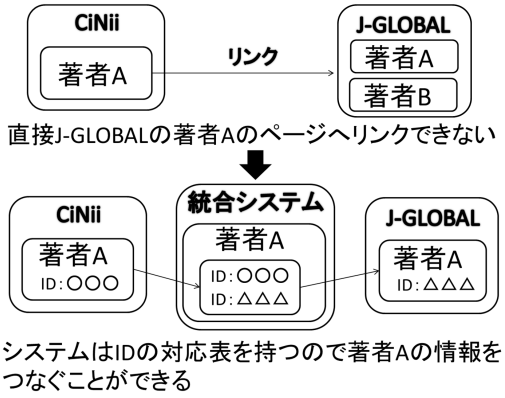

図 1　ID 統合の概念図

3. 関 連 研 究

様々な機関の所有するデータをつなぐ研究として Linked Data と呼ばれる取り組みが近年行われ始めている [4] [5]．Linked Data の実例として美術館や博物館の所有するデータを集めて関係づけた LODACMuseum [6] というサービスが公開されている．美術館や博物館が持つ情報に加え，Wikipedia の情報とも関連付けられており，様々な側面から情報を得ることができるサービスとなっている．

Linked Data の研究ではひとつの実体に対してひとつ ID があることが前提であるが，現状ではひとつの実体に対して様々な機関がそれぞれの ID を割り当てているという文献検索サイトの現状と同様の問題を抱えている．Linked Open Data の研究の中に，実体に対して統一的な ID を用意するというアプローチがある [7]．本研究ではそれぞれの ID 同士を対応関係で結ぶことでつなぐというアプローチをとる．

4. 機関横断型文献情報 Wiki の構築

ここでは，MediaWiki の概要と，MediaWiki 上でどのように複数の文献検索サイトのメタデータを対応付けていくかについて述べる．

4.1 MediaWiki とは

MeidaWiki の基本的な特徴として，HTML よりも簡単な構文規則でページを記述することができるというものがある．MediaWiki のアカウントを取得すれば，誰でも自由に編集が可能になるので，ユーザが間違いを発見した場合，そのユーザの手によって即座に修正することができる．全ての編集には履歴が残るので，誤った編集をしてしまった場合でも過去の編集履歴をたどることで編集前の状態へ簡単に戻すこともできる．

他にも「bot」と呼ばれる自動編集プログラムがあらかじめ用意されており，大量の編集を機械的に行うことができる．

4.2　メタデータの収集

本システムでは，各サイトの所有するメタデータを，サイト内の人物情報ページをスクレイピングすることで機械的に収集する．各サイトから収集するメタデータの一覧を表 1 に示す．

表 1　各サイトから収集するメタデータの一覧表

サイト名	収集するメタデータ					
CiNii	名前	所属	ID	-	所蔵論文	-
J-GLOBAL	名前	所属	ID	研究分野	所蔵論文	HP アドレス
研究者リゾルバー	名前	-	ID	研究分野	-	キーワード

表 1 に示すように，各サイトから，メタデータに割り当てられている「ID」と，「名前」，「所属機関」といった人物に関する基本的な情報に加え，「研究分野」や，「論文の一覧」等のように人物を特定するために参考となる情報を収集する．今回は検証のために東京電機大学の教員計 335 名の人物情報を対象としている．

4.3　Wiki への登録

収集したメタデータの Wiki への登録は編集 bot により機械的に行う．ページタイトルはそのサイトにおける ID とする．登録された人物情報の例を図 2 に示す．図 2 は著者「増田英孝」

図 2　登録されたメタデータページの例

の「J-GLOBAL」でのメタデータページである．「J-GLOBAL」における「増田英孝」の ID は「200901009424739052」なのでページタイトルは「J-GLOBAL:200901009424739052」となる．ページ内の項目は，サイトから収集したその人物の基本情報を記載する．このようにしてページを作成していき，同名の著者のページをその人物名をページタイトルとした人物名ページに一覧としてまとめる．図 3 に人物名ページの例を示す．

図 3　作成された人物名ページの例

この例の場合，「J-GLOBAL」に一件，「CiNii」に二件レコードが存在するので，図 3 に示す通り「増田英孝の人物名ページ」には三つのメタデータページがまとめられることになる．

4.4　各サイトのメタデータの対応付け

各サイトから収集したメタデータが Wiki に登録された段階では，各サイト間の対応付けがなされていないので，同一著者のページを対応付けるという作業を行う必要がある．現状ではこの作業は人手で行なっている．図 4 に対応付けのイメージ図を示す．

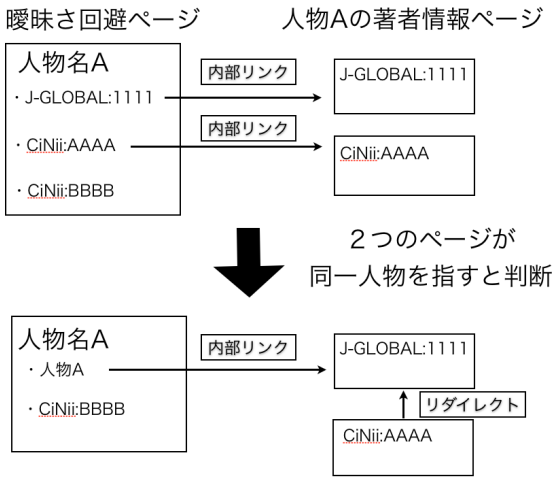

図 4　対応付けのイメージ

図 4 のように，編集者は，人物名ページにまとめられた著者情報ページ内の，論文の一覧や，所属，研究分野，研究キーワード等の情報をもとに同一人物であるか否かの判断を行い，同一人物であった場合にはページ同士をリダイレクト関係にする．リダイレクト先は「J-GLOBAL」のメタデータページとする．理由としては「J-GLOABAL」の人物情報ページは同姓同名の区別がなされレコードの重複が存在しない．そのため他のサイトのページから「J-GLOBAL」のメタデータページにリダイレクトさせることで，一人の人物に対して一つのレコードという関係が可能になるためである．リダイレクトをさせるためにはリダイレクト元のページをリダイレクトページに変更する必要がある．ページの編集画面に「#redirect[リダイレクト先のページタイトル]」という記述を加えるだけでそのページをリダイレクトページに変更することができる．このリダイレクト先のページに対して「リダイレクトしているページのタイトルの一覧」を取り出す．タイトルは各機関における ID なので，このページとページ（ID と ID）のリダイレクト関係が ID の対応表として機能する．

このような対応付けを行なっていくことで最終的に人物名ページは，同名の異なる人物の区別をするための曖昧さ回避ページとして機能する．図 5 に曖昧さ回避ページとして機能している様子を示す．同姓同名の人物が存在しなかった場合，人物名ページも J-GLOBAL のメタデータページへのリダイレクトページにする．こうすることで名前で検索すると直接 J-GLAOBAL のメタデータページへリンクすることができる．

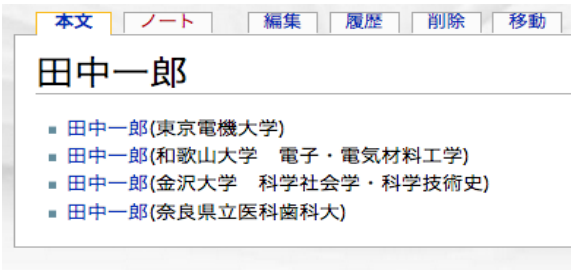

図 5　曖昧さ回避ページの例

4.5　リアルタイムでの編集状況の提示

Wiki の編集は全て履歴が残る．この編集履歴を利用し RSS 等の形式で確認することができるようにすることでどの著者情報ページが編集されたかリアルタイムに知ることができるようになる．また著者情報ページ同士の対応付けも上記のようにリダイレクトの関係に編集することで行なっているので，機関ごとの著者情報の対応付けの結果も同様に知ることができる．著者情報の重複を解決したいと考えている機関側がひとつの参考情報として活用することも期待できる．

5.　考　　察

図 6 に人物名に対する重複レコードの件数の関係をグラフに示す．

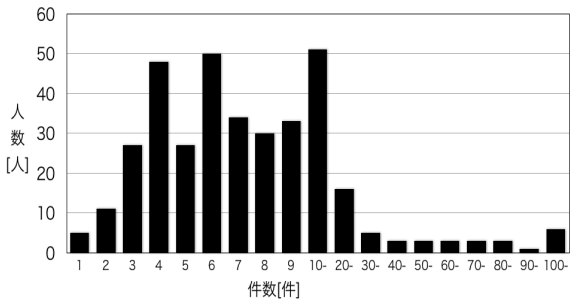

図 6　人物名と件数

図 6 は各サイトからメタデータを収集し，Wiki に追加する際に人物名一つに対して同姓同名の別の人物を含む重複レコードが何件存在するかを示している．縦軸は人数，横軸は件数を表し，「10-」は 10 件台 (10 件以上 20 件未満) を表している．このグラフから東京電機大学の教員 335 人中 238 人の人物には重複レコードが 1 件から 9 件存在する．これはすなわち Wiki に登録された段階では約 70 %の人物名ページに 1 件から 9 件のメタデータページがまとめられることを意味している．この範囲であればそれぞれのページを参照し対応付ける作業を人手で行うことは十分可能であるが，10 から 19 件の重複レコードを持つ人物は 51 人、20 から 29 件の重複レコードを持つ人物は 16 人と重複レコードが数多く存在する人物も多い．特に 100 件以上の重複レコードをもつ人物も 6 人存在し，ここまで膨大な重複レコードを人手で対応付けていくことは現実的ではない．そこで重複レコードの件数を減らすことを目的として機械的な大規模編集を行う仕組みを作る必要があると考えられる．

さらに既存の文献検索サイトでは，ユーザが誤りのある情報を発見しても機関がその情報を修正するまでに時間がかかってしまうという問題があった．その点本システムでは MediaWiki を利用しているので，間違いに気づいたユーザの手によって即座に編集することができる．しかし，本システムと同様に MediaWiki を利用したサービスであるオンライン百科事典「Wikipedia」でも問題視されている通り，サービス内の情報の信頼性という点では，多くの課題がある．アカウントさえあれば誰でも自由に編集できるという MediaWiki の特徴から，悪意ある編集者が容易にでたらめな情報を追加するという可能性がある．しかしでたらめな編集が行われていることに他のユーザが気づけば，編集履歴を参照することで簡単に元の状態に戻すことが可能である．そのため多くのユーザが利用してくれるようになれば，情報の修正も多く行われるようになるため，ある程度の情報の信頼性は確保できるものと考えられる．

このように人間の判断による精度の高い対応付けに，機械的な大規模編集を組み合わせることのできる環境である MediaWiki は今回の目的に適していると考えられる．

6.　おわりに

既存の文献検索サイトは，それぞれが別々の機関により提供されているので，メタデータの対応付けがなされておらず，横断的に文献を探しているユーザはそれぞれのサイトで検索を繰り返す必要があった．そこで我々は，各機関が情報に対して割り当てている固有 ID の対応付けを行い各サイトの情報を間接的につなぐための仕組みである機関横断型の文献情報統合システムを MediaWiki を利用することで構築した．このシステムを ID の対応表として利用することで，機関を横断して情報を収集するということが可能になる．

また情報の信頼性の確保や，人手による精度の高い対応付けのためには，より多くのユーザに編集に参加してもらう必要がある．

今後の課題としては考察でも述べた通り，重複レコードを減らすために，所属情報に加えて，過去に執筆した論文の情報等を参考にして機械的に対応付けを行う仕組みを作成すること．また多くのユーザに編集に参加してもらうために Wiki の記法を知らなくても編集に参加できる仕組み等を取り入れる必要があると考えられる．

文　　献

[1] CiNii, http://ci.nii.ac.jp/
[2] J-GLOBAL, http://jglobal.jst.go.jp/
[3] 相澤　彰子 他, レコード同定問題に関する研究の課題と現状, 電子情報通信学会論文誌 J88-D-I, No. 3, pp. 576–589 (2005).
[4] 武田　英明, Web におけるアイデンティティとセマンティックスの表現と利用, 人工知能学会誌 Vol. 24, No. 4, pp. 512–518 (2009).
[5] 神埼　正英, リンクするデータ、未来へのリンク, 第 19 回 Web インテリジェンスとインタラクション研究会, http://www.kanzaki.com/works/2011/pub/0307wi2.html
[6] LODACMuseum, http://lod.ac/
[7] 神崎 正英, 連携するデータ、リンクするデータ, デジタルアーカイブフォーラム研究会, http://www.kanzaki.com/works/2007/pub/1129keio.html