2I1-R-4-7

機関横断型文献情報Wikiによるコミュニティベースのメタデータ対応付けの試み

Attempt mapping of community-based metadata by Cross-agency Bibliographic Information System using MediaWiki

日向野達郎[*1] 増田英孝[*1] 山田剛一[*1] 清田陽司[*2] 中川裕志[*3]
Tatsuro Higano Hidetaka Masuda Koichi Yamada Yoji Kiyota Hiroshi Nakagawa

[*1]東京電機大学大学院 Tokyo Denki University
[*2]株式会社ネクスト NEXT Co.,Ltd
[*3]東京大学 University of Tokyo

Bibliographic databases are provided as web services from various agencies. Since entities (e.g., persons, books, and papers) are not linked across other databases, users need to repeat searches at each database. This paper propose a method for mapping entities across databases using MediaWiki.

1. はじめに

現在 Web 上には,「CiNii」[1] や「J-GLOBAL」[2] 等の, Web 上で閲覧することのできる論文や書籍等の文献検索サイトと呼ばれるサービスが存在している．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ことを可能にする枠組みを提案している．本論文では，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した結果を報告する．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を行うというように，間違いの発見から修正まで時間がかかってしまうという．他のサイトにおいては，著者情報の修正を行うことができても自分の情報のみであり，他の著者情報の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いるのが現状である．

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．

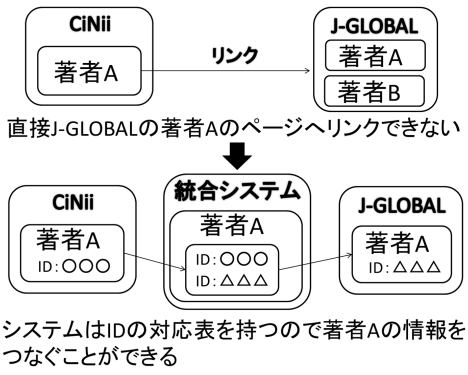

図 1: ID 統合の概念図

連絡先: 日向野達郎, 未来科学研究科 情報メディア学専攻, 東京都足立区千住旭町 5, higano@cdl.im.dendai.ac.jp

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，主に著者情報を対象としてまず手始めに MediaWiki 上で対応付けを行う．

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 LODACMusiam[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，実際には同じ実体に対してそれぞれの機関が ID を独自に割り当てているのが現状である．本研究では実体に対して統一的な ID を割り当てることを目的とせず，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．

また著者のリンクリゾルバを作成する「研究者リゾルバー」[7] というサービスが国立情報学研究所により提供されていて，様々なサイトにある研究者の情報を集約して，アクセスしやすくしている．ここでは科学技術振興機構の研究開発支援総合ディレクトリ「ReaD」[8] へのリンクという機能があり，国立情報学研究所の科学研究費補助金データベース「KAKEN」[9] における ID の対応関係を用いて直接リンクすることができる [10]．しかしユーザが著者の情報に誤りを発見した場合には修正する手段がなく，機関による修正を待つか，機関へ直接問い合わせるという方法しかない．本システムでは，ユーザの手によって情報を修正することができるという点が相違点である．

文献情報を横断的に検索することのできるサービスとして「Google Scholar」[11] がある．Web 上に存在する様々なデータベースを横断的に検索することができる非常に有用なサービスである．しかしある著者の執筆した論文を検索する場合，「Google Scholar」では著者名をクエリとして単純に文字列一致した検索結果を表示しているため，検索結果が本当に求めている著者の論文なのか，または同姓同名の別の人物の論文なのかの判断はユーザに委ねられているという問題点がある．図 2 に「Google Scholar」で「田中一郎」を検索した結果の画面を示す．図 2 に示すように，一番上の論文の著者の「田中一郎」と二番目の論文の「田中一郎」が別の人物なのか，または同一人物なのかは，ユーザに著者の研究分野等の事前知識がなければ一見して分からない．その点本提案手法を利用したシステムは，機関ごとの著者の対応付けがなされるので，検索結果はユーザが求めている著者の論文であるという保証がなされる．

4. 機関横断型文献情報 Wiki の構築

4.1 MediaWiki とは

MediaWiki とは，オンライン百科事典である Wikipedia にも利用されているウィキソフトウェアであり，OS である Linux，Web サーバである Apache HTTP Server，データベースである MySQL，スクリプト言語である PHP 環境で動作が保証されている．


馬鈴薯天狗巣病の虫媒伝染に関する研究
…， 塩田弘行， 関山英吉， 田中一郎… - 北海道大學農學部 …, 1955 - eprints.lib.hokudai.ac.jp
… Issue Date 1955-10-31 URL http://hdl.handle.net/2115/11593 Rights Type bulletin Additional Information Page 2. 馬鈴薯天狗巣病の虫媒伝染に関する研究 福士貞吉・四方英四郎骨 塩田弘行・関山：英吉附 田中一郎・大島信行・西尾美明山 …
全 3 バージョン

単一杆体の応答と光刺激量との関係について
…， 雅規， 田中一郎 - 東京女子医科大学雑誌, 1977 - ir.twmu.ac.jp
… http://ir.twmu.ac.jp/dspace/ Title 単一杆体の応答と光刺激量との関係について Author(s) 田内, 雅規; 田中, 一郎 Journal 東京女子医科大学雑誌, 47(7):837-838, 1977 URL http://hdl.handle.net/10470/3224 Page 2. 67 例と，死をまぬがれて生存した例とについて …
全 2 バージョン

心筋活動電位の基礎としてのナトリウム– カリウム説
田中一郎 - 東京女子医科大学雑誌, 1961 - ir.twmu.ac.jp
○幹事会 '日時昭和36年1月9日（月）午後4時 場所 東京女子医大図書館 会議室 「議題 ユ．雑誌1月2月を合併号とする、 1．例会の件，症例検討会は新宿区医師会々員に連絡して聴講させる。 ．L 幹事補充は織姫教授と決定した6． ○例会 （第104回） …
全 2 バージョン

染料中間体アミノフェノール 3 異性体の中毒作用について
田中一郎 - 東京女子医科大学雑誌, 1985 - ir.twmu.ac.jp
… Knowledge Database. http://ir.twmu.ac.jp/dspace/ Title 染料中間体アミノフェノール3異性体の中毒作用について Author(s) 田中, 一郎 Journal 東京女子医科大学雑誌, 55(1):81-82, 1985 URL http://hdl.handle.net/10470/10127 Page 2. 結論 …

図 2: Google Scholar で田中一郎を検索した結果

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，誰でも自由に情報の修正，追加が可能になるので，ユーザが間違いを発見した場合，そのユーザが即座に情報を修正することができる．ページの編集履歴が残るので，誤った編集をしてしまった場合でも簡単に元の状態へ戻すこともできる．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．人手と機械の 2 つの方法による柔軟な編集ができるという点が MediaWiki の利点である．

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 3 に示す．図 3 は著者「増田英孝」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

図 3: 登録されたメタデータページの例

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 4 に人物名ページの例を示す．

図 4: 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 4 に示す通り「増田英孝の人物名ページ」には三つの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 5 に対応付けのイメージ図を示す．

図 5 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，

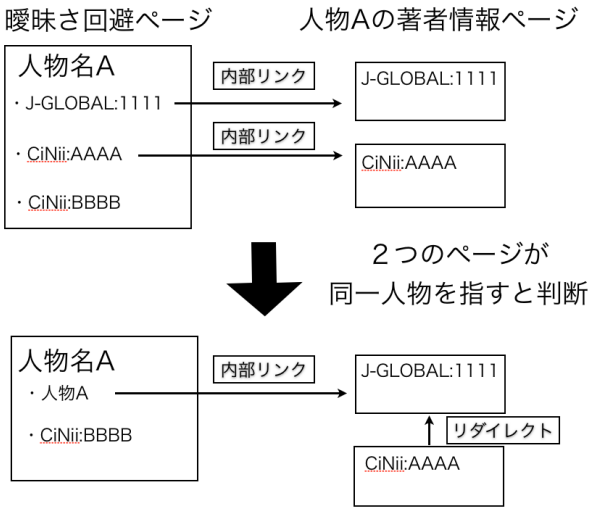

図 5: 対応付けのイメージ

こ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 6 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

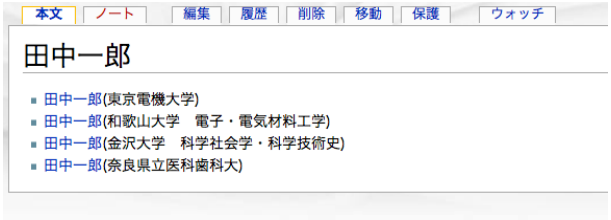

図 6: 曖昧さ回避ページの例

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．

5. 考察

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がその時点で即座に編集することができる．しかし，HTML よりも比較的簡単な構文規則で記述できるとはいえ，それまで Wiki の記法を知らない人にとって，初めから記法を覚えるということは，編集をためらう要因となってしまうことが考えられる．編集される情報の信頼性を向上させるという観点からも，多くの人に編集に参加してもらうということが必要となる．そこで Wiki の記法を知らなくても編集ができるようにするためのインタフェースを現在作成中である．

また同じく 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即座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．

図 7 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．

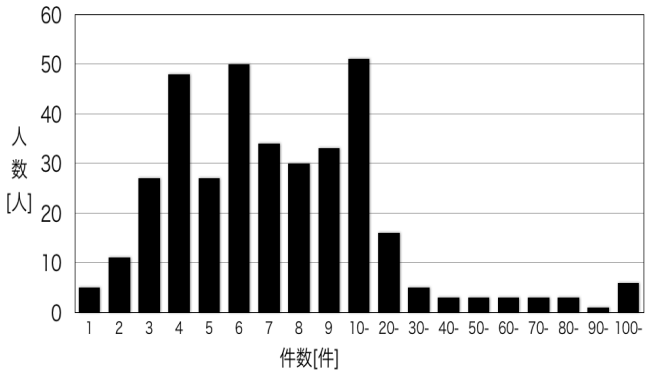

図 7: 人物名と件数

図 7 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．そこで所属機関等を参考にしてある程度の段階まで機械的に対応付けを行う仕組みを作る必要があると考えられる．

このように人間の判断による正確な対応付けに，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．

今後の課題としては考察でも述べた通り，一点目にシステム内の情報の信頼性の確保という面からより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できるようにするためのインタフェースを作成することが挙げられる．二点目に所属機関情報等を参考にしてある程度の段階まで機械的に対応付けを行う仕組みを取り入れることが挙げられる．

将来的には今回作成したシステムを利用して，機関の枠を超えて情報を収集し，ユーザに提供するシステムの開発を目指す．

参考文献

[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] 研究者リゾルバー, http://rns.nii.ac.jp/

[8] ReaD, http://read.jst.go.jp/

[9] KAKEN, http://kaken.nii.ac.jp/

[10] 研究者リゾルバー ヘルプ, http://rns.nii.ac.jp/html/help.html

[11] Google Scholar，http://scholar.google.co.jp/