CA1740

動向レビュー

著者の名寄せと研究者識別子 ORCID

1. はじめに

学術研究成果の多くは論文として出版され公表される。論文は、すでに存在する論文を引用しながら、それが表す知識の体系を位置づける。そのような知識の体系を構成することに、誰が貢献したか、どのような組織が貢献したかがわかるように、内容とともに著者の名前や所属組織名が明記される。助成機関に対して謝辞を加えることも多い。ある研究者がどのくらい知識の体系化に貢献したかを測ってみたいとき、その研究者の論文を並べてみればよい。それがいわゆる業績リストである。著者本人の申告だけでなく、より客観性を帯びた形でリスト化されればより正確な評価が可能となるであろう。今では、論文や業績リストがWeb上に公開されるようになり、瞬時にそのような情報を得ることが可能となった。出版者の論文検索システム、機関リポジトリ、出版者や機関の研究者ディレクトリなどから直接、または大手の検索サービスを介して取得可能である。

このとき、名前の表記だけで論文などの研究成果を分類すると困ったことが起きる。ある論文に書いてある著者名と別の論文に書いてある著者名は同じ表記であるが、同姓同名の別の人物かもしれないということである。これが英語論文に明記されるローマ字による表記となれば、漢字に比べて同姓同名の割合はもっと増える。また、表記が異なるが同一人物であることもある。論文に明記される名前は、結婚などを機に姓を変え、別の理由で名まで変化することがある。論文の指定する表記方法の違いから、名前の表記揺れもある。

客観性があり正確な業績リストを作成するためには、このような名前の問題を解決して、研究者ごとに研究成果をリスト化する必要がある[(1)(2)(3)]。名前の問題を解決して同一性を判断することを「名寄せ」（Name Disambiguation）という。

本稿で取り上げるORCID（Open Researcher and Contributor ID）は、学術情報流通の世界を対象として様々なステークホルダーが集まってこのような名寄せの問題に取り組む国際的な組織である。IDとは識別子のことであり、ORCIDでは研究者だけでなく貢献者[(4)]に付与される。筆者はORCIDのテクニカルワーキンググループのメンバーであり、名寄せのためのシステム構築の議論に参加してきた。以降では、先ずORCIDができるまでの名寄せの取り組みについて概観する。続けて、ORCIDについて、組織や掲げられた原則、IDシステム、外部識別子との関係、パートナーシステムとの関係について述べる。そして、その他関連する識別子を紹介し、最後にまとめを行う。

2. これまでの名寄せの取り組み

図書館の目録のように閉じたデータベースの中では、名前の問題に対処するために、著者ごとに英数字記号の識別子を付与して区別する著者名典拠を構成してきた。国立国会図書館が提供する全国書誌であるJAPAN/MARCの2008年7月5日付けの典拠ファイルを解析したところ、著者名として個人名681,924件のレコードが登録されており、そのうち漢字圏の東洋人を抜粋すると572,638件が登録されていた[(5)]。漢字の姓名部分を文字列比較してみたところ73,138件のレコードに同一の表記を持つ別のレコードの存在が認められた。ざっと1割を超えている。

学術論文のデータベースにおいては、主に二つのアプローチがとられてきた。計算機による方法と人手による方法である。計算機を用いた方法では、論文書誌集合に対し機械学習をベースとしたクラスタリングの技術を用いて著者ごとに分類する。ある論文書誌に明記された著者と別の論文書誌に明記された著者が同一著者であることを様々な素性を対象として確率的に判定していく。素性とは書誌に記述された姓名表記や所属、共著関係、分野、キーワードなどで与えられる。商業出版者のデータベースはこの種の方法で、独自のアルゴリズムを開発して自らのサービスに実装している。たとえば、トムソン・ロイター社の文献データベースWeb of ScienceにはDistinct Author Identification System[(6)]が実装され、エルゼビア社の文献データベースScopusにはScopus Author Identifiers[(7)]が実装されている。しかしながら、実用レベルに必要だといわれている100パーセントに近い精度には達していない。もう一つの方法の人手による方法では、研究者自身がIDを登録し、自らの業績リストを構築していく。たとえば、2008年1月にスタートしたトムソン・ロイター社の研究者ディレクトリResearcherID[(8)]がある。商業出版者のサービスとタイアップした、研究者自らが自身をID登録するサービスは、既存の研究者ディレクトリにはなく画期的である。しかしながら、名寄せをするのに十分なほどの登録数は得られていない。

同様に主要な学術出版者を横断的に網羅したサービスとして、非営利組織である出版者国際リンキング連盟（Publishers International Linking Association, Inc.：PILA）の運営するCrossRef[(9)]（CA1521参照）がある。CrossRefは、論文などの学術コンテンツに

ID を付与して、ID と Web 上の URL とを結びつける仕組みを提供してきたが、同様な方法で学術コンテンツの作者に ID を付与する方法を考案するため、Contributor ID プロジェクト[10]を進めていた。

このような背景の中で、2009 年 11 月 9 日、研究者の識別子に関心のあるいくつかの主要なステークホルダーが集まって、名前識別子サミット（The Name Identifier Summit）が開かれた[11]。チェアは、トムソン・ロイター社のコチャルコ（David Kochalko）とネイチャー出版グループのラトナー（Howard Ratner）であった。これが本稿で紹介する ORCID 発足のための最初の会議である。

3. ORCID

3.1. 設立趣旨

ORCID の設立趣旨は公式ホームページ上に掲げられている[12]。原文を翻訳すると以下の通りである。

「ORCID は、学術コミュニケーションにおける著者 / 貢献者の名前の曖昧性の問題を解決することを目的とし、個々の研究者に対する固有の識別子の中央レジストリと、ORCID と現存する他の著者 ID スキームとの間のオープンで透過的なリンクメカニズムを構築することによって実現する。これらの識別子及び識別子間の関係は研究者のアウトプットにリンクすることが可能であり、科学的発見プロセスを拡大させ、研究コミュニティにおける研究助成や協働の効率性を改善する。」

3.2. 組織

ORCID は正式な組織となる前から活動を開始し、2010 年 8 月に米国デラウェア州の非営利組織となり、そのことが同年 9 月 7 日にプレスリリースされた[13]。組織発足時のボードは、出版者、学会、財団、大学、研究所など多種多様な組織からのメンバーで構成されている。国立情報学研究所もその一組織である。

その後、10 月 8 日にボードメンバー内の選挙によって[14]、ネイチャー出版グループのラトナーがボードの代表に、トムソン・ロイター社のコチャルコが会計、ハーバード大学のブランド（Amy Brand）が秘書に選出された。そのほか、ウェルカム財団のアレン（Liz Allen）、ACM のラウス（Bernard Rous）、ワイリー・ブラックウェル社のバン・ディック（Craig Van Dyck）の 3 名がエグゼクティブコミッティに選出された。

3.3. 参加組織

ORCID への参加は組織単位となっている。2010 年 11 月 16 日の CrossRef の会議での公表スライドによると[15]、144 の参加組織があり、組織の形態で分類すると学術機関 47、出版者 28、企業 19、学会 15、政府 11、NPO 17、その他 7 という内訳になっている。学術研究に関係する様々なステークホルダーで構成されているが、大学と出版者が多い。

また、地理的には、米国 70、英国 30、ドイツ 8、オーストラリア 6、日本 3、イタリア 3、インド 3、スペイン 2、中国 2、カナダ 2 で、1 組織の参加の国は、トルコ、スイス、スウェーデン、韓国、シンガポール、セルビア、オランダ、イスラエル、ギリシャ、フランス、エジプト、コロンビア、ブラジル、ベルギー、オーストリアとなっている。米国と英国が圧倒的多数であり、アジアからの参加は少数である。

3.4. ORCID の原則

ORCID の運営指針となる原則（Principles）がビジネスワーキンググループによって議論され、2010 年 10 月に公開、12 月 8 日に公式ホームページに掲載された[16]。原則は 10 項目からなっており、これに基づいてビジネスモデルやシステムの機能が決定される。

原則では、まず、ORCID が著者と貢献者を信頼して特定できるようにすることによって、学術コミュニケーションにおける、固定の、明確な、曖昧でないレコードの作成を支援することを宣言し、学術分野、地理、国籍、機関の境界を超えた、オープンで透明性のある組織であることを明示している。

そして、研究者は ORCID のサービスを介して自由に ID とプロファイルを登録することが可能であり、その際プライバシーには十分に配慮することとしている。研究者のプロファイルデータは、プライバシー設定後、クリエイティブコモンズが CC0 と定義する権利放棄[17]の形で公開される。研究者のデータに対する権利について議論を積み重ねた結果、ORCID から公開するデータについて権利放棄を明示することになった経緯は強調しておきたい。

また、ORCID の開発したソフトウェアはオープンソースイニシアチブのオープンソース[18]として公にリリースされることとした。オープンソースとして公開することを決めたことは、ボランティアベースによる開発コミュニティを構成することでソフトウェア開発コストを削減したい思いがある。

ORCID のビジネスモデルは、組織が非営利でありながらも持続可能であるための必要最低限の収入を得ることを目的としている。そのためのシステムの API は有料と無料の双方によって構成されることを明示している。

最後に、組織内部の構成が非営利であり、活動内容について最大限に透明性を確保することを謳っている。

3.5. IDシステムの要求

ORCIDのコアシステムとなるIDシステムに関する議論は、2010年2月から9月ごろまでの間、テクニカルワーキンググループによって行われた。どのようなシステムであるべきか、システム要求が議論され、アルファ版のプロトタイプが構築された[19]。その後、2011年の1月にはプロダクションシステムのベータ版構築に向けて議論が進んでいる。

IDシステムにおいて、アイデンティティとして扱う基本的な情報は、

- 著者/貢献者自身の記述
- 著者/貢献者とその出版物間の関係の記述

の2種類である。「著者/貢献者自身の記述」は名前や所属などを含み、研究者のプロファイルである。「著者/貢献者とその出版物間の関係の記述」とは、研究の業績とする論文や記事、書籍、データなどを含むリストであり、出版物申告（Publication Claims）と呼ぶ。これらがORCIDのIDと紐づけられることになる。

プロファイルと出版物申告の登録は、著者/貢献者と組織の双方が行うハイブリッド型による方式が提案されている。著者/貢献者自身によるだけでは情報が集まりにくいので、組織がまとめて情報を登録することによって呼び水とするわけである。

システム要求の議論は、CrossRefがこれまでContributor IDとして議論してきた内容を拡張している。ここでは、エンドユーザー、パートナーシステム、コアシステムの3つの主体が登場する。エンドユーザーは、著者、貢献者、部門管理者、その他の様々な人である。パートナーシステムは、出版者の原稿追跡システム（Manuscript Tracking System）や研究者ディレクトリ、論文検索システムなど関連するシステムである。コアシステムは、ORCIDのIDシステムそのものである。エンドユーザーは、コアシステムに対して、プロファイルや出版物申告を登録して作成し、編集、更新する。大学や研究機関、出版者などの組織は、パートナーシステムからコアシステムに対し、プロファイルや出版物申告をまとめて登録する。コアシステムでは、複数のプロファイルを集めて、マッチングや重複解消をして著者/貢献者の主プロファイルを自動で作成したり、著者/貢献者自らが手動で名寄せするのを支援したりする。そのほか、システムと人、システム間のやり取りを示す個々のユースケースがテクニカルワーキンググループで議論され想定される技術が列挙されたが、ここでは紙面の関係で触れないことにする。

3.6. ORCIDのIDと外部識別子

IDシステムに利用するORCID IDの表現方法は様々に議論された[20]。プライバシーの問題や外部識別子との連携が念頭に置かれて要求が整理された。その結果、IDの要件は以下の通りである[21]。

- 表記に意味を持たせない
- 数字とするが、不連続とし、チェックサムを含める。理想的には国際標準名前識別子“ISNI”（International Standard Name Identifier）と互換性があるようにする
- 人間が覚えられなくてもよいが、書けて、ORCID IDだとわかるようにする

ここに互換性が取り上げられたISNI[22]は、現在ドラフト段階のISO規格27729であり、メディアコンテンツ産業に従事する団体に使われることを想定された、ORCIDより対象が広い範囲のクリエータ識別子の規格である。ISNIのIDは16ケタの数字で、最後はチェックサムとなっている。ISNIのIDは、商用のIDシステムの上に展開されるオープンレイヤーとして、IDシステム同士が必要最低限の情報を交換してID間の対応を付けるブリッジ識別子（Bridge Identifier）として機能する。その結果、ISNIは外部のIDとマッチングの結果を保持するので、例えば、ORCIDやバーチャル国際典拠ファイルVIAF[23]（CA1521参照）とも識別子同士の対応をつけることができる。ORCIDのテクニカルワーキンググループの議論では、IDをISNIと同一にするという提案がある一方で、まったく同じだと区別がつかなくなることを懸念し、いっそのことORCIDがVIAFとIDマッピングをするだけにとどめ、ISNIとはVIAFを通してゆるく連携する可能性もあわせて提案されている。

3.7. パートナーシステムとの連携

ORCIDのIDシステムは、様々なパートナーシステムと連携する。連携の在り方は様々なシナリオとして考えられているが、最も重要なものはパートナーシステムとORCIDのIDシステムがプロファイル交換を行うことである。研究者のIDやプロファイルを独自に保持して、すでに利用されている、たとえば次のようなシステムと連携する。トムソン・ロイター社のResearcherID、エルゼビア社のScopus、国立衛生研究所（NIH）の助成を受けて開発し全米で利用される予定の研究者ディレクトリVIVO[24]、高エネルギー分野の論文を対象とした論文検索システム

INSPIRE[25]、経済学分野の論文を対象とした論文検索システム RePEc[26]、ProQuest 社の研究者ディレクトリ Author Resolver[27]、NIH の運営する医学生物系論文検索システム PubMed[28]である。このようにすでに権威があり、利用頻度が高くユーザー数の多い既存のシステムと連携することは、より信頼性高くすべての研究者を網羅することを可能とする。

4. その他関連する研究者の識別子

研究者の識別子という観点からすると、ORCID 以外にも取り上げるべき活動は多く存在する。たとえば、オランダの SURF 財団の行った DAI（Digital Author Identifier）[29]はオランダの研究者に研究者番号を割り振っている。数物系のプレプリントサーバー ArXiv も Author Identifiers[30]をオプトインの方式で導入している。英国の情報システム合同委員会（Joint Information Systems Committee：JISC）の助成を受けた Names Project[31]は、機関リポジトリの典拠を目指して、研究者の ID を英国図書館の ZETOC 書誌から研究者をクラスタリングして自動で ID を構築している。国立情報学研究所では、機関リポジトリの典拠となることを目的の一つとした、研究者リゾルバー[32]を構築している。これは科学研究費補助金データベース KAKEN[33]をベースにして研究者に ID を付与している。

これらの研究者識別子に関するシステムも ORCID のパートナーシステムとなることが可能であり、ID の登録とプロファイル交換の可能性がある。さらに、Web 上に公開され利用されることを前提としていることから、今後はこれらの識別子同士が Linked Data の技術をベースに互いに同一人物を関係付けることによって連携することも予想される。

5. まとめ

本稿では、学術に対する貢献度を正確に明示するためには論文などの研究成果の著者や貢献者を識別することが重要であることを示し、その歴史的展開から ORCID の活動へつながっていったことを述べた。そして、ORCID の組織について、ORCID の活動で議論されていることの概要を述べた。あわせて、別の研究者識別子を取り上げ、関係性にも触れた。

ORCID はプロダクションシステムのリリースに向けて活動中である。まだ検討すべき事項は多く残っており、アクティブなメンバーによって議論が積み上げられている。組織として持続可能なビジネスの在り方やシステムの使われ方を議論し、研究者や貢献者へのプロモーションを行っている。ORCID は活動に賛同するメンバー組織を募集中であり、メンバーが積極的にワーキンググループに参加することが望まれている。

（国立情報学研究所：蔵川（くらかわ）　圭（けい））

（1）Enserink, Martin. Are you ready to become a number?. Science. 2009, 323（5922）, p. 1662-1664.
（2）Credit where credit is due. Nature. 2009, 462（7275）, p. 852.
（3）Hellman, Eric. “Authors are Not People: ORCID and the Challenges of Name Disambiguation”. Go To Hellman. 2010-05-04. http://go-to-hellman.blogspot.com/2010/05/authors-are-not-people-orcid-and.html,（accessed 2011-01-14）.
（4）“Contributor” の訳語として、ここでは「貢献者」とした。ダブリンコアにおける同一表記の要素の訳語として「寄与者」が使われることがあるが、意味としては同じである。
（5）蔵川圭ほか . “ 研究者リゾルバー α の同姓同名推定モデルと実データによる分析 ”. 2009 年度新領域融合プロジェクト研究による研究会「大規模データ・リンケージ，データマイニングと統計手法」. 2009-10-08/09, 国立情報学研究所 . 2009, p. 65-74.
（6）“Distinct Author Identification System”. Thomson Reuters. http://science.thomsonreuters.com/support/faq/wok3new/dais/,（accessed 2011-01-14）.
（7）“Author Identifier”. Sciverse. http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier/,（accessed 2011-01-14）.
（8）ResearcherID.com. http://www.researcherid.com/,（accessed 2011-01-14）.
（9）crossref.org. http://www.crossref.org/,（accessed 2011-01-14）.
（10）Fenner, Martin. “Interview with Geoffrey Bilder”. Nature.com Blogs. 2009-02-17. http://blogs.nature.com/mfenner/2009/02/17/interview-with-geoffrey-bilder,（accessed 2011-01-14）.
（11）“Research Stakeholders Announce Collaboration among Broad Cross-Section of Community to Resolve Name Ambiguity in Scholarly Research”. ORCID. 2009-12-01. http://www.orcid.org/sites/default/files/ORCID_Announcement.pdf,（accessed 2011-01-14）.
（12）“Mission Statement”. ORCID. http://www.orcid.org/mission-statement,（accessed 2011-01-14）.
（13）“Organization Launched to Solve the Name Ambiguity Problem in Scholarly Research”. ORCID. 2010-09-07. http://www.orcid.org/sites/default/files/ORCIDInc-Press.pdf,（accessed 2011-01-14）.
（14）“ORCID Board Meeting October 8, 2010”. ORCID. http://orcid.org/sites/default/files/ORCIDBoardOct10_0.pdf,（accessed 2011-01-14）.
（15）Ratner, Howard. “ORCID Update, CrossRef Members meeting 16 November 2010”. http://www.slideshare.net/CrossRef/orcid-update-2010-annual-meeting,（accessed 2011-01-14）.
（16）“ORCID Principles”. ORCID. http://www.orcid.org/principles,（accessed 2011-01-14）.
（17）“CC0 1.0 Universal（CC0 1.0）Public Domain Dedication”. Creative Commons. http://creativecommons.org/publicdomain/zero/1.0/,（accessed 2011-01-14）.
（18）“Open Source Definition（Annotated）, Version 1.9”. Open Source Initiative. http://www.opensource.org/osd.html,（accessed 2011-01-14）.
（19）2010 年 11 月 11 日にワーキンググループメンバーにメールで配布された報告資料による。
（20）ワーキンググループの報告資料による。
（21）2010 年 11 月 11 日にワーキンググループメンバーにメールで配布された報告資料による。
（22）International Standard Name Identifier. http://www.isni.org/,（accessed 2011-01-14）.
（23）VIAF, Virtual International Authority File. http://viaf.org/,（accessed 2011-01-14）.
（24）VIVO. http://www.vivoweb.org/,（accessed 2011-01-14）.
（25）INSPIRE, beta. http://inspirebeta.net/,（accessed 2011-01-14）.
（26）RePEc. http://repec.org/,（accessed 2011-01-14）.

(27) “Author Resolver”. RefWorks-COS. http://www.refworks-cos.com/authorresolver/,（accessed 2011-01-14）.
(28) “PubMed”. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/pubmed,（accessed 2011-01-14）.
(29) “Digital Author Identifier（DAI）”. SURF Foundation. http://www.surffoundation.nl/en/themas/openonderzoek/infrastructuur/Pages/digitalauthoridentifierdai.aspx,（accessed 2011-01-14）.
(30) “Author Identifiers”. ArXiv.org. http://arxiv.org/help/author_identifiers,（accessed 2011-01-14）.
(31) “Names Project”. Mimas. http://names.mimas.ac.uk/,（accessed 2011-01-14）.
(32) “研究者リゾルバー”. 国立情報学研究所. http://rns.nii.ac.jp/,（参照 2011-01-14）.
(33) “科学研究費補助金データベース KAKEN”. 国立情報学研究所. http://kaken.nii.ac.jp/,（参照 2011-01-14）.

CA1741

動向レビュー

人文学研究と電子アーカイブ

1. 電子アーカイブプロジェクト

1.1. 人文学研究と資料アクセス

14世紀英国の物語詩『農夫ピアズの夢』（“Piers Plowman”）には3つの稿と十指にあまる写本がある。各バージョンを合わせると60以上の基礎資料が存在し、手稿のページ数は1万にのぼる(1)。

手稿や異稿の研究は人文学に欠かせないが、原資料に直接アクセスしたり、各地に分散する異稿を調べて回ったりするのは容易ではない。そこでこうした資料を整理し、注釈や関連情報とともに提供する学術版、批判校訂版、あるいはファクシミリ版（たとえば「『農夫ピアズの夢』コンコーダンス」や手稿B.15.17ケンブリッジ版）が重要な役割を果たすことになる。

しかし印刷物では、物理的、経済的な制約から盛り込める情報が限定され、検索や相互参照などの活用に限界がある。そこで学術資料をデジタル化する試みが重ねられ、さらにインターネットの発達とともに、「『農夫ピアズの夢』電子アーカイブ」(2)のようなデジタル化されたアーカイブに発展し、資料へのアクセス性が飛躍的に向上してきた（電子アーカイブはデジタル・アーカイブとも呼ばれる。これらの場合のアーカイブは、資料の集成だけでなく、一般に学術版としての研究成果も盛り込んだサイトの意味で使われる(3)）。

1.2. 資料デジタル化の役割

資料のデジタル化は、膨大な資料へのアクセス性を改善するばかりではない。マッギャン（Jerome McGan）は、紙ベースのテキストを電子形態に変換すると原資料の見方が大きく変わることを指摘し、それは「自然現象研究に対する数学的アプローチが理論的視点のレベルを高度化するのと同じように、電子ツールが批判的抽象度のレベルを引き上げる」からだと述べている(4)。

研究ツールとしてデジタルテキストを利用するためには、手稿や印刷物などの形の原資料から文字を転写し、さらにそのテキストがどのような構造になっているか（ページ構成、章節構造など）を何らかの方法で明示しなければならない。またアーカイブされた資料を検索利用し、共有するためには、メタデータを適切に付与することも重要である。

テキストのデジタル化に関しては、転写の方法、文字コードなど、問題となる点は多々あるが、ここ