第 7 回 SPARC Japan セミナー2010

著者 ID の動向

ORCID プロトタイプシステムと 著者 ID 関連技術の動向

蔵川 圭
（国立情報学研究所学術コンテンツサービス研究開発センター特任准教授
／ ORCID Technical working group member ）

講演要旨

ORCID の活動は、いくつかのワーキンググループに分かれて行われている。ビジネスワーキンググループ、アウトリーチワーキンググループ、テクニカルワーキンググループである。そのうち、テクニカルワーキンググループは、 ORCID が NPO 法人として認可される 9 月ごろまで、サービスやシステムとして必要な要求の洗い出しを行いながら、α版のプロトタイプシステムを精力的に構築してきた。
本講演では、プロトタイプシステムの構築にあたってメンバーが議論してきたシステム要求や機能、その実現手段について紹介する。合わせて、ORCID 以外の著者 ID の関連技術とのかかわりについても触れる。

蔵川 圭

1994 年東京大学工学部精密機械工学科卒業。1996 年東京大学大学院工学系研究科修士課程修了。2001 年東京大学大学院工学系研究科博士課程修了。博士（工学）。1999 年より奈良先端科学技術大学院大学情報科学研究科助手を経て、2006 年より現在、国立情報学研究所学術コンテンツサービス研究開発センター特任准教授。専門は図書館情報学と設計学。図書館情報学分野のシステム研究開発に従事。学術コミュニティーに不可欠な情報サービスの開発を進めている。

Open Researcher and Contributor ID（ORCID）は、大きくは Business working group と Technical working group という二つに分かれます。私からはシステム的な観点から、Technical working group での議論を紹介したいと思います。

タイムラインとしては､2010 年の 1～2 月ごろから、どんなシステムを作るべきかという議論が始まりました。6 月末をめどにプロトタイプを作り、Business working group へのデモンストレーションも含めて、このシステムのテストを、ORCID が正式な組織になる 9 月ごろを目指して行ってきました。そのあたりの議論を紹介したいと思います。

実は本日使用するスライドは、Technical working group でディスカッションしてきた内容をまとめたサマリーレポートをベースに翻訳したものです。Open Researcher and Contributor ID の Researcher は研究者ですが、Contributor は貢献者と訳しています。

ORCID 誕生以前の二つのアプローチ

ORCID 誕生以前には、大きく分けて二つのアプローチがありました。一つは、計算機を使って論文書誌を対象に著者単位に分類してまとめる方式です。Scopus の Author Identifier やトムソン・ロイターの Distinct Author Identification System などがそうでした。しかし、一般的にプロダクションシステムとして必要な名寄せ精度は 100%近くと言われているので、これでは不十分でした。

そこで、論文書誌があって著者ごとにまとめるのではなく、著者をまず定めて、そこに論文書誌を集めるパターンで手動登録するというアプローチが出てきました。ここでは Researcher ID が研究者業績ショーケースとして機能する形です。これでいいではないかという話もあるのですが、実はこれだけでは不十分で、数が集まりません。それで、やはりインディペンデントなシステムとして、一つの出版社や一つプロプライエタリーに限定することなく、みんなでシステムを作りましょうということになりました。

ORCID identity system

ORCID でアイデンティティとして扱う情報は、貢献者自身の記述（名前や所属）と、貢献者とその出版物間の関係の記述です。登録の方法は、貢献者自身による登録と組織による登録のハイブリッド型です。研究者自身が自分で登録すればいいのですが、それだけでは集まりません。そこで呼び水として組織による登録も必要ですが、これだけでも十分ではありません。だからこの両方をする方法がいいということになっています。そもそも ORCID が始まる前に CrossRef で Contributor ID の議論があり、この議論をしっかりベースにして要求定義をしています。

CrossRef が集めた要求

CrossRef は、まずは「グローバルに一意に定められる貢献者 ID のユースケースは二つのカテゴリに大別される」、こんなことができるといいということを言っています。

一つ目の大きなカテゴリは「質問回答を含む知識発見のシナリオ」です。例えば、誰が文書 X を書いたか。ID Y の人が書いた、または査読した文書はどれか。ID Z の人はどの ID の人と関係するか。その関係はどんなものか（Z は Y と論文を共著した、Y の論文を Z が編集して査読した等）。ID Z のプロファイル情報は何か（所属機関、e-mail アドレス等）。こんなことが分かるといいということです。

もう一つは、「さまざまな状況における、ネットワーク上で貢献者自身を特定するシナリオ」です。これは、幾つか関連するシステムが外側にあって、その中である研究者を特定する条件があるということです。例えば原稿追跡システム（Manuscript Tracking System）があって、そこにシングルサインオンして、編集事務局、マーケティング部門、ロイヤルティ支払いシステムなどと連絡先情報を共有することができる。あとは、Table of Contents アラートや他の自動メール送信用の e-mail アドレスを自動更新する、あるいは査読者候補を自動選定するツール（同一の興味がある人の自動選定を含む）として使いたい。複数の出版社の Web サイトのユーザープロファイルと同期したり、その外部プロファイル間の ID 妥当性確認とその表明をしたりする。あるいは、研究者自身がその内容にカスタマイズや特権的なアクセスを保証する。このようなことがあるといいのではないかということです。

ORCID identity system の要求概要

こういう CrossRef が集めた要求を詳細化し、拡張する形で、さらに、CrossRef に関係する団体だけではなく、より広く機関リポジトリや助成機関、その他のステークホルダーと共に全体の要求を常に議論しながら整理しています。それは、エンドユーザーの要求、パートナーシステムの要求、コアシステムのキー属性および能力として整理されます。

これがその絵です（図 1）。エンドユーザーと、先ほどの原稿追跡システムのようなパートナーシステムと

のかかわりで、コアシステム（ORCID identity system）がどうであるべきか。エンドユーザーやパートナーシステムはコアシステムにどういうことができて、だからこのコアシステムにどういう機能があるといいのか。そういう話を延々としています。

エンドユーザーの要求

まず、著者、貢献者、部門管理者、その他エンドユーザーが Web ベース UI を介して以下のことが可能であるべきだという議論がありました。

一つは、システムに筆頭著者を登録してプロファイルを作成し、続けてこの自己申告したプロファイルを編集および更新する（研究者が自分で登録する）ことです。

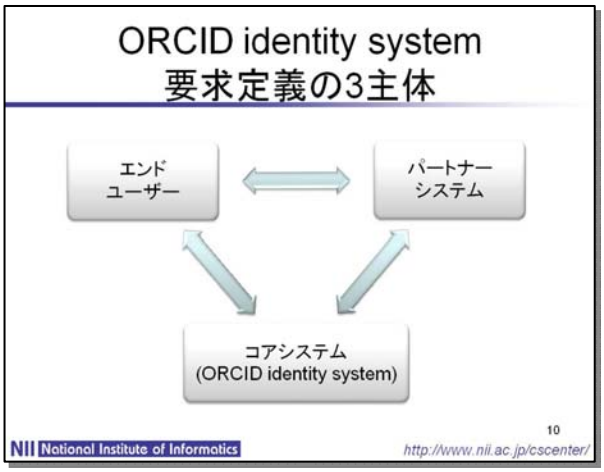
次に、第三者団体のシステム（パートナーシステム）で公開されているプロファイルで、その団体がデポジットしたプロファイルの登録を検索し、さらに幾つかのプロファイルに対して申告によって「自己名寄せ」をすることです。つまり、研究者のプロファイルが外の世界にもいろいろあって、ORCID のシステムにも登録し、それを研究者は検索できて、自分のものだということが分かったら手で名寄せして「これは私のものです」と宣言する。これを自己名寄せ（self disambiguation）と呼んでいます。

さらに、プロファイルに対して、例えばパブリック、プライベート、限定という 3 種類のプライバシーを設定できることです。

あとは、CrossRef の DOI で特定された寄稿した学術出版物を CrossRef の書誌データベースで検索、または手動で入力し、申告することです。

それから、CrossRef が主に出版物に DOI を付けているのに対して、DataCite サービスは、研究者が作ったデータに DOI をふるサービスをいろいろ実験的に模索している組織ですが、そこで特定される研究データセットを検索し、申告できるようにもした方がいいのではないかということが考えられます。

また、CrossRef に収められていないものもいろいろ



（図 1）ORCID identity system 要求定義の 3 主体

あります。例えば Online Computer Library Center（OCLC）のモノグラフと業界向け出版物、政府や NGO その他の出版した準学術的な著作、arXiv と機関リポジトリにある記事・報告・ワーキングペーパー、特許局の特許などです。それらも検索して申告できたらいいのではないか。また、Concept Web Alliance というセマンティック Web のエンティティを登録するシステムがあり、そのエンティティのトリプルまたはファクトを登録対象とできる、あるいはブログのネットワークのポストとして、Public Library of Science のブログ、Nature Network、Science Blogs、また Wikipedia の記事も登録できた方がいいのではないか。それから、登録したものに対して、誤っている、または不正なプロファイルと出版物申告を報告できる方がいい。また、他のユーザーのプロファイルを編集・消去したりアカウントを消したりできないと困るので、その本人だけではなく、特別な権限を与えて、登録したプロファイルに対してさまざまなバックオフィス的管理タスクをこなすようにしたらよいのではないか。

以上が、エンドユーザーの ORCID identity system に対する要求の概要です。

パートナーシステムの要求

次に、パートナーシステムは Web UI または API を介してどんなことが可能であるべきかという議論です。

まずは、大学やその他の組織がプロファイルデータと出版物申告をデポジットおよび取得して、その研究者にデポジットした情報をベースとして ORCID のプロファイルを登録し、簡単に穴埋めができるようにする。つまり、大学など外の組織が研究者の代わりにプロファイルデータをデポジットしますが、正式に登録するのは研究者なので、研究者はそのデポジットされたデータを眺めて、既にある程度情報が埋まっている状態で簡単に登録できるようにしたいということです。

それから、ジャーナル誌やその他の出版者が、そこで出版した著者の検証済みの出版物申告をデポジットする。「出版物申告」は英語では publication claims という言い方をしていますので、業績リストのような感じです。それをここでは「出版物申告」と訳しています。

さらに、学会、大学、その他の機関・組織が所属の申告を検証する。方々から出てくるプロファイルや publication claims で、自分のところに関係するものの申告を検証できるようにした方がいいということです。

また、ジャーナル誌の原稿追跡システムなどが、ID によって貢献者のプロファイル情報を照合し、限定公開されたプロファイルデータ（e-mail、電話番号等）をシステムで出版した貢献者から要求・取得し、このプロファイル情報との変更を記録してシステムに自動的に反映させる。Manuscript Tracking System などが ORCID の identity system に対して、ユーザーとのやり取り等の間で一番新しいものをそれぞれが発見して自動的に反映させていくというようなことです。

そして、ジャーナル誌の原稿追跡システムや機関ポジトリ、その他のシステムが、例えばパートナーサイトの登録・ログインプロセスの部分として ORCID ID の取得を促し、貢献者が中央の ORCID システムとやり取りを仲介する。つまり、ジャーナルに投稿したいときに、その人がもし ORCID の ID を持っていなかったら、ORCID の ID をまず作ってプロセスを進めてもらうということです。

さらに、エンドユーザーとパートナーシステムとのインタラクションを通して収集した情報にアクセスするためのさまざま知識発見のユースケースにおいて、すべてのステークホルダーが持つ一般的な関心事があります。重要なユースケースとして、次のようなことが含まれます。例えば、助成機関が助成したプロジェクトに関係のある ORCID ID と出版物を取得することには結構みんな興味がありますし、機関がそのファカルティの ORCID ID と出版物を取得することは、例えば研究評価のために使えるということをみんな言っています。

ここまでは、エンドユーザーとパートナーシステムがコアシステムに対してどうあるべきかということをお話ししました。

コアシステムのキー属性および能力

今度は、コアシステムのキー属性および能力としてどのようなものがあるとよいかという議論の結果です。

まず、ORCID 識別子は、それ自体は容易に分からない、数字の、連続的でない文字列で、International Standard Name Identifier（ISNI）の命名スキームと互換性があるとよいということを言っていました。ISNI と ORCID の協調は、今のところ思案中です。

それから、そのスキームに積極的に参加した登録者は、そのコアシステムに対しては登録とともに ID が割り当てられ、自己申告プロファイルを作成します。研究者が自分で申告したときには自己申告といいますが、第三者（機関や出版社）から勝手に上がってくるプロファイルがあるので、これを第三者団体のプロファイルといって区別しています。

他の貢献者の識別子は、第三者団体のプロファイルをシードとして算出する第一プロファイルを自動生成するときに処理されます。しかし、これをどうするかはまだはっきりしていません。

ORCID のプロファイル自身は最小限のフィールドセット（e-mail、名前、所属など）を持つべきであり、後で拡張可能なようにします。そのプロファイルに入

れるときに、例えば所属機関を書き込む必要がありますが、そのような機関の識別子は既にあり、例えば商用ベースで機関の識別子を公開している Ringgold や、RePEc という経済学の論文を集めた論文検索サービスの派生物として出てくる ARIW のものを利用したらいいのではないか。あとは、プロファイルとして誕生日を入れるべきかどうかという議論もありました。

さらに、コアシステムは基本的なプロファイルマッチングができるようになる予定で、登録ユーザーの自己名寄せをサポートします。コアシステムに作り込まれるべき高性能な自動マッチングや重複解消の能力の程度を考えると、初期ベータのプロダクションシステムとしてどれだけ重要かは意見が分かれています。基本的には研究者が自分で名寄せし、あくまでもそれをサポートするようなシステムに構成します。一方で、方々からいろいろなプロファイルがやってきますから、そのマッチングの自動化もしなければいけませんが、この自動化は難しく、どれだけ作り込めるかは分からないということです。

また、方々からやってくるプロファイルの由来も追跡した方がいいでしょう。その際、プロファイル記録、プロファイル申告、そして出版物申告の情報源やその他属性が記述されたメタデータを捕獲することで、それを実現します。

そして、パートナーシステム（出版社、大学等）からバルクでデポジットされたレコードをバッチローディングするツールを提供します。

さらに、研究者自身が登録して変更したりするときにも、併せて認証・認可のためのセキュリティレイヤーを設けます。ユーザーはフロントエンドの UI にインタラクションする形で、原稿追跡システムとインタラクションするときにも認証しなければいけませんが、そのときには限定公開のプロファイルデータやオープンでないサービスに、例えば技術的には OAuth ベースのワークフローのために API を使って実現しようということを考えています。

あとは、いろいろなプロファイルが方々からやってきたときに、恐らくプロファイルの内容記述に対して矛盾が生じるでしょうから、その矛盾解決のメカニズムが必要だと言われています。

ここまでは、コアシステムに対して最初に特にすべきことなのですが、後のシステム構築サイクルで考慮すべき重要なこともあります。例えば、ORCID のプロファイルデータを代替フォーマット（Atom/RSS フィード、JSON、Linked Data）で出力して、軽量で Web2.0 スタイルの統合型のマッシュアップや、他の Web サイトに埋め込みできるようにするということを実現したらいいのではないか。既に arXiv.org の Author Identifier サービスもそうなっています。それから、ソーシャルネットワークが Web 上に一般的に普及しているので、そのような機能を実現するときには、例えば Universal Widget API や OpenSocial のプロトコルを準拠するなど、コミュニティーが考えている標準と互換性がある技術を使った方がいいのではないか。認証に関しては、フェデレーション認証（OpenID、Shibboleth 等）のサポート、かつ ORCID がアイデンティティプロバイダとして機能する可能性もあるのではないか。ただ、これらは議論としては上がっているのですが、確定的ではないというのが現状です。

ORCID identity system の要求詳細

identity system 自体は、主に次のような要素から成り立っています。

まず、最初にあるのが ORCID の ID（識別子）です。これに対して、研究者がどんな人か、どんな所属かというプロファイルがくっつきます。さらに ID に対して出版物申告（publication claims）がくっつくという構造になります。

こういう構造がある中で、さらにプロファイルはどのように生成して管理したらいいのか、また、そのプロファイルの内容に対するオーソリティや、その情報を書き換えたりするコントロールはどうしたらいいのか、プライバシーについてはどのような仕組みが必要か、方々から来るプロファイル同士のマッチングや重

複部分の解消はどうしたらいいのかということを議論します。出版物申告に対しても、どんな機能が必要か。また、これら全体の ORCID identity system を通してデータ公開をしてサービスを提供しますが、持続可能性との兼ね合いの中から、どういうところでお金が取れるか、逆に取れないか、取るべきではないかという議論もあります。こういったことを 1 カ月に 1 回ぐらい、電話で議論しました。

ORCID ID の構造

ID は、まずは意味的に不透明にするべきです。これはほとんど CrossRef の Contributor ID の議論を継承していて、Geoffrey Bilder のレポートに基づいています。まず、数字とするけれども、不連続とし、チェックサムを含める。理想的には ISNI 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 だと分かるようにしたいということがあります。ISNI と連携した方がいいのでしょうけれども、ISNI の ID をそのまま採用すれば、相互互換性をそのまま確保できる利点はありますが、ISNI の ID と ORCID の ID が同じだと混在してしまうので、全く別物にして対応表を作った方がいいのではないかという考え方もあります。その中では両極のシナリオがあります。a 案は、ORCID が ISNI の Registration Agency になる、つまり概念的には ISNI の方が上なので、その研究者版、学術出版コミュニケーション版としての Registration Agency になればいいのではないかというものです。b 案は、中間的なオーソリティ、例えば ISNI と連携している Virtual International Authority File（VIAF）を介して緩く ISNI とつながるという考え方です。これが議論されています。

プロファイルの生成と管理

ID を定義した後に、プロファイルをくっつけていきます。プロファイルの生成と管理には二とおりあり、一つが組織表明型の ID 申告です。ORCID メンバー（組織）によってデポジットされるもので、第三者団体プロファイル（third party profile）と呼んだりします。もう一つが個人の登録で生成される個人表明型の ID 申告で、第一 ORCID プロファイル（primary ORCID profile）といいます。明らかに個人が最初に決めたのだから、それをプライマリーにしてしまうという言い方です。この二つには、登録の主体が組織か個人かという違いがあります。

プロファイルを管理するシステムの配置として、中央集権型にしてみんなが登録するのかいいのか、いろいろなシステムが分散的に置かれている形がいいのかと考えると、中央集権型で進めた方がいいという話になります。中央集権型だと、よりシンプルで、より経費がかからず、ORCID データを定型で長期保存させること（学術の記録としては重要な用件）を確保でき、他の重要な問題（データのライセンシング、再利用等）にも対処できるというように、いろいろ楽だからです。

しかし、実は混合型の意見もあります。ユーザーが外部のパートナーシステムにあるレコードを指定すると、プロファイルデータが中央の ORCID システムに取得・コピーされるという、いわゆるパートナーシステムを分散的なもう一つのデータソースとして、そこから取ってくるという方法もあるということです。

あとは、ORCID ID にプロファイルを関連付けるのですが、それには、直接登録ユーザーが自己申告したプロファイルをくっつけるというパターンと、いろいろな出版社や機関リポジトリなどからプロファイルがやってくるので、そこから一つのプライマリーの算出プロファイルを疑似的に作って ORCID ID を付与するというパターンの二つがあります。

直接登録の場合は簡単で、ORCID ID はその新規作成の第一プロファイルに関連付けられます。ユーザーが後に第三者団体のプロファイルを申告して自分の第一プロファイルと関連付けると、ORCID ID が組織申告と自己申告の ID を集約したものをその人と特定します。これをハイブリッド・アイデンティティ・モデルと言っています。一方、算出プロファイルから ORCID ID を付与する場合、登録したがらないユーザ

ーで、一つまたはそれ以上の第三者団体のプロファイルが登録された個人に ORCID ID を付与していきます。実は、このようにしながらも網羅していくことは、ORCID 成功のための重要なミッションだということを確認しました。

コアシステムが、アクティブな研究者（後で自己申告するかもしれない）、引退した研究者、休眠中の研究者、いずれのためにもバルクでデポジットされたレコードから算出第一プロファイル（computed primary profile）の生成を支援することが大事です。算出プロファイルの生成と管理はコアシステム内の自動バッチマッチングの能力に依存するので、この問題の解決は難しく、最初のプロトタイプはアクティブな研究者の自己登録にフォーカスすべきだろうとみんな思っています。

プロファイルのオーソリティとコントロール

プロファイルのオーソリティをつけるためにはどうすればよいか、そのコントロールはどうすればよいかについては、二つの意見に分かれました。一つは図書館の視点（組織の視点）、もう一つは個人の視点（アイデンティティ・プライバシーの視点）です。

図書館の視点というのは、ボードメンバーである MIT の MacKenzie Smith の意見なのですが、機関がバルクでデポジットした後、研究者をプッシュする必要があるということです。図書館や他の管理スタッフが、ファカルティの ID 作成にバルクデポジットすることが期待されていますが、この場合、図書館がアップしても、多くの研究者が「さらにもう一つのプロファイル」のためにサインアップすらしないことは予想できるので、サインアップしろと積極的にプッシュするアプローチが必要です。これは、機関リポジトリでもいろいろと情報を入れろ、入れろと言ったのに説得がうまくいかなかったという経験に基づいています。

さらに、機関が管理することが実は情報の質を高めることだと実感しています。ですから、機関の管理は個人データの質を高めるために重要であり、機関が情報の管理をずっと行ってきていて、これからもそうであることが期待されています。ただ、こうしたモデルだと、機関を通して ORCID に登録された研究者は第一プロファイルにひも付けられ、機関が提供したレコードと矛盾するようにそれを編集することは許されなくなります。機関にオーソリティを付け過ぎると研究者は編集できなくなることがデメリットです。

次に、アイデンティティ・プライバシーの視点です。研究者自身が ORCID ID とプロファイルを管理すると、図書館の視点にあった機関による情報の間違いを見つけにくくするということを回避できますし、本当は ORCID の ID は研究者自身の手元にあるべきで、研究者コミュニティーに受け入れられるためにはそうあることが重要なので、そう考えましょうということです。

また、アイデンティティやプライバシーには十分に配慮すべきだということが確認されました。従って、自己表明（研究者が自ら表明するもの）と組織の表明の申告は両方保持し、矛盾したら「矛盾していますよ」とシステムにフラグを付けるなどするハイブリッドモデルも提案されました。

事例からオーソリティコントロールを考える

実は香港大学（Hong Kong University：HKU）が既に Researcher ID と統合して実験をしています。これは Researcher ID のアプローチそのままなのですが、HKU のレコードと書誌データを最初にバッチアップロードしてプロファイルを作ります。後に HKU のファカルティに e-mail が届きます。そうしたら、研究者自身は登録および新しく作られた ID を申告するように促され、そこから自分でデータをメンテナンスし始めます。もちろんやめることもできます。

ここから、さらに二つの拡張モデルが考えられました。一つは、申告されていない組織管理の ORCID プロファイルに、はっきりマークした方がいいのではないかというものです。研究者が自分で yes とは言っていないけれども組織がアップロードしたというものに

はっきりマークすることによって、メンテナンスされ、管理されたプロファイルと混ざってしまうことを避けることができます。

もう一つは、代理メカニズムを立てるということです。別のユーザーに自分のデータの編集許可を与えて、例えば学部や図書館の管理者が学部全体の申告した第一プロファイルを編集できるようにする、あるいは、プロファイルの無効化や、死去したり引退したりした人のプロファイルをそれ以上更新しないようにロックできるようにした方がいいのではないかということも挙がりました。

プライバシーモデル

次に考えたのは、プライバシーのモデルとしてプロファイルの属性は最大限にした方がいいのか、最小限にした方がいいのかという二極の選択です。最大限にすると、名寄せ目的にはより情報量が多いので使い勝手がいいですし、情報量が多いということは、ORCID を自分のホームページのように使ってくる人も出てくるでしょう。しかし、最大限のプロファイルモデルは最小限のプロファイルモデルよりも実現が難しく、識別目的以上に個人情報を集めることは EU プライバシー法を侵害するかもしれないので、学術出版物で典型的に使われている名寄せに必要なデータ要素だけをパブリック・ファクトとして捕捉することに専念しましょうということに落ち着いています。

では、最小限のプロファイルとは言っても、そこに誕生日は入れるべきかどうか。誕生日はプライバシーに関する項目ですが、名寄せのためには、年を除いて日だけであっても重要です。誕生日を必須にしてシステム外には公開しないという手もあるでしょうが、アイデンティティ・プライバシーに反して名寄せに必要である以上に個人情報を集めると、ORCID が mini-Facebook と受け取られる危険性があるので、ここは名寄せという目的に注力しようという意見がありました。

それから、プロファイルのフィールドはどの程度見えるようにするべきか。名前も含めた幾つかのフィールドのセットは常に見えた方がいいのではないかという意見がある一方で、ID 以外のフィールドは全部隠せた方がいいのではないかという意見もありました。しかし、全部隠すと、取りあえず作っておいて後は放置するという「“放っておく” 第二の自己のプロファイル（throw away alter ego profile）」を作ってしまうことになるだけでなく、疑似的に匿名となるようなモードでの操作を可能にしてしまうことになる。全然知らない著者を作り出して、それをうまくコントロールしてやろうというやからもいるので、名前を隠すのはやめた方がいいのではないかといった議論がありました。

プロファイルマッチングと重複解消

プロファイルが出版社や図書館などいろいろなところから集まってくると、本人が登録したものと併せて必ず重複するので、それをうまく解消することが必要です。そのために、コアシステムには 2 レベルのマッチング能力が必要で、最初に自分で登録したときと、全くシステム側からプロファイルがやってきたときの両方のシナリオを考えています。

自分で登録するときのプロファイルの申告ステップは比較的簡単です。まず、ユーザーは新しく作成するプロファイルにマッチする可能性のあるデポジットされたプロファイルのリストを提示され、これらプロファイルの受け入れ、または却下をします。このような自己名寄せには、フィールドベースのマッチングよりは、あまり精緻ではない、取りあえず候補として挙げてくれるぐらいの緩いマッチングが必要です。ここでは、条件つきで受け入れたり、修正してから受け入れたりする仕組みがあった方がいいのではないかという意見がありました。

一方で、個人がまずデポジットしない場合、同一人を示す異なる ID システムにおけるデポジットされたレコードをつなげる働きをします。いろいろなところから来たものに対して、まずはプロファイルをマッチングさせて、この人だということを言う、つまり単一

ID を作成します。これは、重複した ORCID プロファイルを作成することを可能な限り避けるために行います。コアシステムが名前の衝突検知や回避のメカニズムを持たない限り、重複は必ず起こり得て、重複した第一プロファイルがあることが後で判明したとき、それらを結合できる必要があります。いろいろなところからやってきていたり、重複していたりして、くっつけたりはがしたりといういろいろな操作をしなければいけません。

こういういろいろなプロファイルのマッチングと重複解消能力を考えていますが、そもそも、まだどうしたらいいかが分からない未解決問題が二つ残っています。一つは、算出プロファイルは最初のプロダクションシステムで優先させるか、後で実現することにするか、もう一つは重複解消能力をどの程度拡張させるかという問題です。各プロファイルが既に他のシステムで名寄せされてきた後に、その範囲でプロファイルを受け入れて、別のところから来るプロファイルに対して「これとこれは同じ人のプロファイルです」ということに限定すべきか、そうではなくて、やってくるプロファイルの中に重複がたくさんあるときに、それすべてに対して完全なバッチ名寄せをするべきか。どの範囲ですべきかということは、まだ考え中です。

マッチングや重複解消の技術については、実はいろいろなプロジェクトや企業があるので、ORCID がこの手の問題に深入りするのはどうなのか、アウトソーシングしてバッチ名寄せすることでコストを下げるべきだという意見もあります。その一方で、これはとても重要な問題なので、コアシステムは外に頼らずに自分ですべきではないかという意見もあります。

出版物申告

今度は、publication claims（出版物申告）に関してどんなものが必要かということです。2 タイプの重要な出版物関連の ID の表明があります。一つは著者自身が「私がこれを書きました」と申告するもの、もう一つは出版社が検証して、「うちの出版物には誰々がどんな論文 X を書きました」と申告するもので、これは CrossRef が以前に行った要求作業に基づいた ID の取り方です。RePEc の論文検索システムで Author Claim という似たようなことが既に行われていたので、それに発想を得てワイヤーフレームを作成したりしました。

出版物申告を ORCID のシステムに集めると、次のようなことが可能になります。まず、決定すべきさまざまな付加的、二次的な表明ができ、例えば、ある論文の共著者による代理申告を検証することができますし、矛盾とエラーの自動検知ができます。同じ論文を同じ名前で重複して申告する人がいますので、これらの著者のプロファイルを確認して名寄せのきっかけとすることが必要です。

今後、どのようなタイプの申告が最も利便性があるか、中央のフロントエンド UI でこのような情報をどう表示すればいいかを決定する必要があります。最初のプロダクションシステムでは、もちろん第一申告表明データを、関連する由来のメタデータを初めから付けて蓄積できる必要があります。由来についてはちゃんと最初からやっておかなければいけないということです。

出版物申告に関しては、さらに詳細化する話も出てきました。出版物申告に対してタイムスタンプや申告元、方法などを含める方がいい、それから、著者そのものと伝統的な出版物、論文だけではなく、研究データセットを含めて ORCID は広く対象をとらえたい。また、武田先生が言っていましたが、ビッグサイエンスでは、170 機関、2000 著者という例もあります。そういうときには第一、第二、ラストオーサーなどというもので貢献度を測ることには全く意味がなくなるので、貢献度に対しての記述も必要ではないか。最近のジャーナルではそんなことが書け、後でメタデータで上がってくるので、それも考慮したらいいのではないかという意見があります。しかし一方で、それは ORCID の仕事ではないからやめた方がいいのではないかという意見もあります。

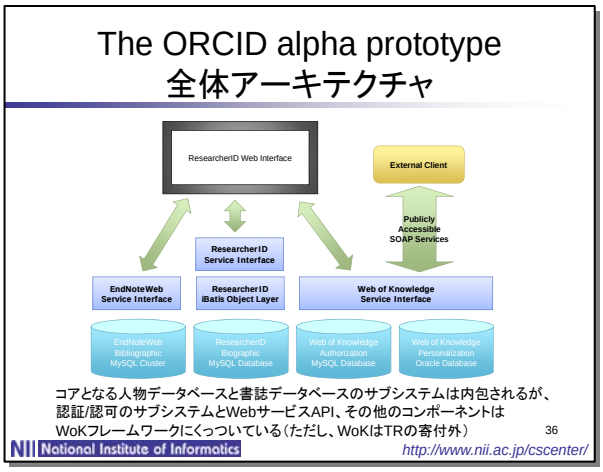
データ公開とサービスの在り方、vs.持続可能性

あとは、データ公開とサービスの在り方、そして、どうやってお金を取るのかという持続可能性の話です。持続可能であるにはお金を取らなければいけないのですが、広くみんなに使ってもらいたいので、基本はオープンにしたいです。オープンということに関しては既に考えられている概念があり、Open Knowledge Foundation の Open Knowledge Definition などの原則に沿って外に出した方がいいという意見があります。それから、オープンデータの原則を適用すると、ある程度は無料なのですが、持続可能であるためには収入が必要で、アクセスするときに完全な無料とはなり得ないから、どこからかお金を取りたいわけです。このあたりは ORCID Principles にも明示されるようになっています。

Technical working group として合意に達したのは、貢献者の参加は多い方がいいということや、組織の参加については、基本的にはクリティカルマスに達するための呼び水としてどんどん参加させているわけだから、最初のうちは課金すべきではないといったことです。しかし、データアクセスと再利用に関しては、例えば長期にわたる継続的な成功と幅広い適用を促すために、ID とプロファイル情報は Web サービスを介して、軽量で非商用であるならば無料でアクセスさせるべきであって、商用のときは、無料のために保障される品質のサービスとともに何か別個に対応したらいいのではないか。バルクデータダンプは Creative Commons の CC0 で権利放棄しますが、バルクでデータをアップロードするときには制限をかけて課金した方がいいのではないかという話です。

The ORCID alpha prototype —全体アーキテクチャ

ここからは ORCID alpha prototype のお話です。これまでの要求定義の議論から、取りあえず Business working group のメンバーでプロトタイプとして作りました。これはトムソン・ロイターの Researcher ID をベースにすることで迅速なプロトタイプが実現したもので、2010 年 3 月から 9 月ぐらいまでの話です。



（図 2）The ORCID alpha prototype 全体アーキテクチャ

構成としては、Researcher ID がベースになっています。Researcher ID 自身は EndNoteWeb の書誌の部分と、Researcher ID 本体が持っているバイオグラフィックのプロファイルの部分、認証系は Web of Knowledge の部分を使っており、外との話は全部認証のインターフェース込みの Web of Knowledge のインターフェースを介してパートナーシステムと情報をやり取りしていたという構成です（図 2）。これをベースに使いました。

The ORCID alpha prototype —ユーザー入力データ項目

そのときにどんなプロファイルの属性を提起したかというと、基本的には Researcher ID のプロファイルがベースになっていて、それを並べている感じです。ORCID 番号、名前（姓、名、ミドル）、別名、e-mail アドレス、固定 URL（研究者自身の ORCID の URL）、役割、主題、キーワード、記述（その人の説明）、ユーザー定義の URL、プライバシーの設定、機関名、サブ組織、サブ組織アドレス、サブ組織の役割、所属機関名、所属機関のサブ組織、所属機関の開始日、所属機関の役割、過去の所属機関情報、個人設定として出す・出さない、オプトイン・オプトアウトという機能も付

けています。

The ORCID alpha prototype—主機能

alpha 版としてのプロトタイプでは、容易な登録プロセスを採用しています。例えば自分が登録したいときに、パートナーシステムからデータをあらかじめ埋めておいて、研究者自身はそれに OK ボタンを押すだけでいいということです。ユーザーコントロールのプライバシー設定としては、出す・出さない、また英語だけではなくてローカル言語についても UTF-8 をサポートしています。

検索機能としては、姓・名、機関、キーワード、ORCID ナンバーで公開プロファイルを検索できます。機関とキーワードのときにはオートサジェストがあり、幾つか出た候補から選んでブラウジングしていく形です。

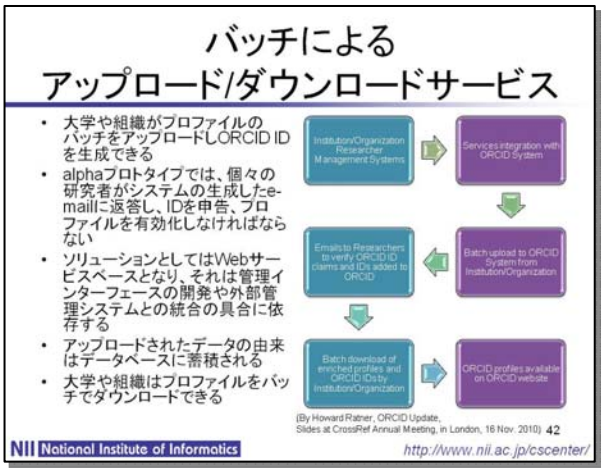
出版物の申告に関しては、CrossRef の DOI で検索できます。ORCID のパートナーシステムと統合する形でバルクでダウンロードしたりアップロードしたりするものを作りました。

バッチによるアップロード／ダウンロードサービス

バッチでアップロード・ダウンロードするというのは、先ほど説明した Researcher ID とほとんど同じパターンで、大学がプロファイルをバッチでアップロードし、個々の研究者がその結果を e-mail で受けて ID を申告し、プロファイルを有効化します。Web インターフェースのサービスベースの兼ね合いによって複雑にできることが決まりますが、それによってバッチでアップロードします。アップロードされたデータの由来はデータベースに蓄積され、大学組織は逆にプロファイルをバッチでダウンロードできるような形になっています（図 3）。

The ORCID alpha prototype—登録等の流れ

これは Researcher ID のインターフェースにとてもよく似ていますが、ORCID の identity system の alpha prototype です（図 4）。登録はここで「Register」と押します。



（図 3）バッチによるアップロード/ダウンロードサービス

（図 4）The ORCID alpha prototype 登録

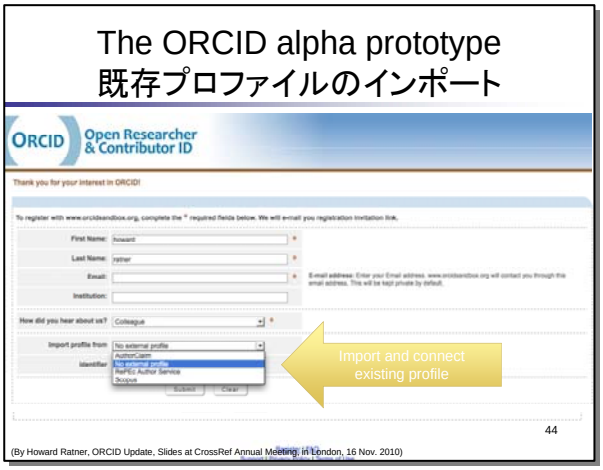
自分がまず登録したとき、プロファイルの候補が出てきて「あなたはどこのプロファイルを投入しますか。AuthorClaim ですか、外部なしですか、RePEc ですか、Scopus ですか」と聞かれて、そこでインポートします（図 5）。

その後、登録した e-mail アドレスで簡単にログインできます。そうすると著者のページが出てきます（図 6）。ORCID ID、名前、e-mail、URL などの属性があって、下の方にパブリケーションが出ます。Researcher ID を見たことがある人は同じでないかと思われるかもしれませんが、少しだけ違って、ちょっと ORCID らしくなっています。

出版物申告は、CrossRef から書誌を持ってきて DOI 検索ができるようになっています。パブリケーションクレームをしたいときに、例えば名前を入れると、その名前に関する CrossRef からの書誌の結果がばらばらと出てきます。自分のものならチェックして Add ボタンを押します（図 7～9）。検索は、名前、ファーストネーム、ラストネームなどでできます。Result として結果が出てきて、名前をクリックすると、その人のページにいきます。もちろんキーワードでもブラウジングできるので、そのキーワードに関係する人が検索結果として出てくることもあります（図 10）。

The ORCID alpha prototype —原稿追跡システムとの連携

原稿追跡システムとの連携ですが、これは Nature publishing group の Manuscript Tracking System の一つのサンプル画面のような感じで、自分の投稿をするときにこのようにできます（図 11）。ここで自分の名前を打ちますが、ORCID ID の「探します」というところにチェックを入れて、例えば名前を入れると、関係する ORCID の人が検索できます。ここから自分の ORCID ID を選ぶと、ORCID ID が埋め込められるというシステムを alpha prototype で作りました。



（図 5）The ORCID alpha prototype 既存プロファイルのインポート

（図 7）The ORCID alpha prototype 出版物申告

（図 6）The ORCID alpha prototype 基本的なプロファイル

（図 8）The ORCID alpha prototype CrossRef の検索

The ORCID beta production system

また、production system に向けて、beta 版を作ろうとしています。これはまだ決まっていないのですが、Researcher ID をベースに進めるのではないかという話が出ていました。ソフトウエアはオープンソースです。

未解決の問題

それでも未解決の問題はたくさんあります。プライバシーのモデルについては、最初のプロファイルにどこまで含めるかという問題があります。

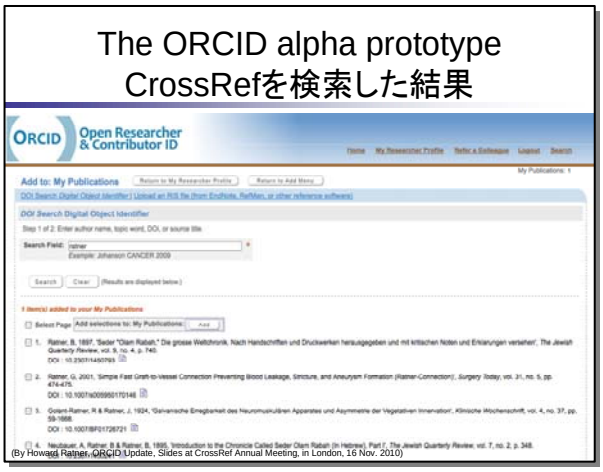
由来のデータモデルについては、機関によってもプッシュされない、プロファイルが登録されない、活動しないけれども論文をたくさん書いていた研究者たちについては、どのように ORCID ID を作成すればいいのか。勝手に ORCID が著者の第一プロファイルを作ってしまえばいいのか。これはまだ決まっていません。

プロファイルマッチングについては、計算機が出したプロファイルに対して重複を最小限に抑えるというプロファイルマッチングおよび重複解消の機能は必要ですが、beta 版のプロダクションシステムにすら必要なのか、それとも最初のシステムは自己登録の支援に注力した方がいいのか。最終的には、コアシステムに組み込まれるバッチマッチングと自動名寄せの能力はどの程度あるべきか。これもまだ決まっていません。

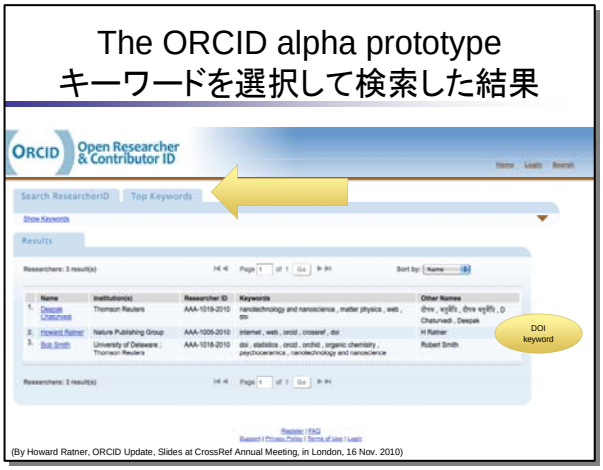
プロファイルデータのコントロールについては、研究者自身や組織が提供したプロファイルに矛盾があるときにどちらを優先するのか。それともスーパーバイザー的な仲裁機能を何かシステムに組み込むべきかどうか。回答には、さらに議論が必要です。

さらなる検討事項

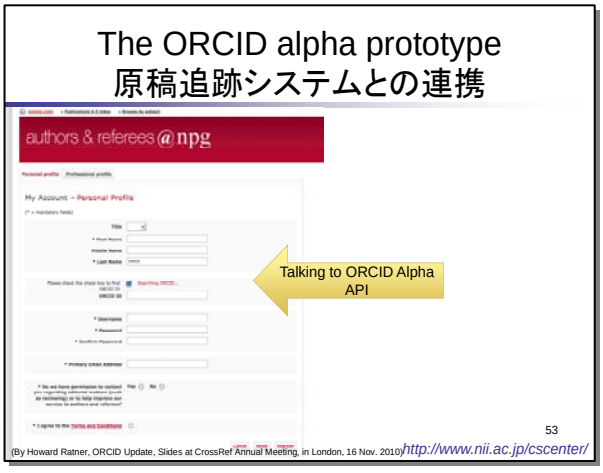
より遠い話になりますが、さらなる検討事項もあります。研究者の所属のための、利用可能な機関の識別子を作らなければいけないのではないか。また、機関の管轄するシステムとインタラクションし、その API を介していくつかの機能を実行することをユーザーに



（図 9）The ORCID alpha prototype CrossRef を検索した結果



（図 10）The ORCID alpha prototype キーワードを選択して検索した結果



（図 11）The ORCID alpha prototype 原稿追跡システムとの連携

許すかどうか。データの所有権として、ORCID から集めたものが出ていくときには creative commons の CC0 で出しますが、第三者団体のプロファイルから勝手にコピーしたデータを所有して自己申告プロファイルとしていいのかどうか。そして、貢献者に外部の Web ページや API からプロファイル情報をプルして ORCID に入れることを研究者自身に許すのかどうかも検討しなければいけません。

今日の参考資料は、「A summary report on ORCID core system requirement and current status of development」です。これは Technical working group のメンバーになると見られます。また、Howard Ratner が CrossRef の Members Meeting のときに上げたのが、先ほどのスクリーンダンプのものです。それから、Geoffrey Bilder が書いた、ORCID の identifier についての議論など、この辺は ORCID の Google sites に組織として参加すると見ることができるので、後追いすることが可能です。

●**林**　どうもありがとうございました。かなり詳しい議論の経緯やシステムの要件に資するための議論を紹介いただきましたが、ご質問はございますか。

●**谷藤**　物質・材料研究機構の谷藤です。この ORCID は、何万人、あるいは何百万人くらいの研究者の人をイメージしておられるのでしょうか。

●**蔵川**　数は聞いていないですが、イメージとしては全員です。しかし VIAF との兼ね合いなどから、まずできるところからと考えています。いろいろな議論を聞いた中では、出版社は生きている人を相手にしているので、生きている人からまず登録を促します。たくさん論文を書いていても、死んだ人やもういない人は後回しになりますが、後々には登録されるでしょう。

●**谷藤**　もちろん生きている、アクティブな研究者の話をしているのですが、その人という ID の概念は、もちろん個人の研究者の方もおられるものの、平均的には所属する機関があって、そこと結び付いて e-mail アドレスなども生きた情報になっていると思うのです。ただ、どうしてそのプロファイルを機関の方に戻すという逆のベクトルが、最初のプライマリーの段階の議論にないのでしょうか。

全部吸い取って ORCID で名寄せをして、パブリケーション重複の排除をしてきれいにしていこう、その確認は研究者個人がしようという、個人と ORCID システムの関係で成功していくのだと思うのです。もちろんそれを否定はしていないのですが、そのときに、人ではなく、その人が所属している機関と照会する、参照するという仕組みがプライマリーな段階の設計にないのはなぜかという質問です。

●**蔵川**　ないというよりは、基本的には個人が自分のプロファイルを登録するのですが、それを補足・支援する形で組織がバックアップします。まず、登録には多分それがあります。それは最初にはしようとしています。しかし、本人が登録することが、より優先順位が高いのです。ですから、本人の登録があって、それをバックアップする組織があって、それをダウンロードした先で自分の目的のために使います。例えば出版社には、自分のところに登録した研究者の評価やプロファイルを作成してあげるなど、いろいろなサービスがあり、それに使う。あるいは、大学が自分のファカルティの評価に使う。助成機関は、どんな結果が研究成果として出てきたのかをトラッキングするために使う。そういう目的がまた次に来るというイメージです。