Open Researcher & Contributor ID (ORCID): Solving the Name Ambiguity Problem

The Open Researcher & Contributor ID (ORCID) initiative was started in November 2009 to solve the author name ambiguity problem in scholarly communication. Author name ambiguity means that the author name can't be used to reliably identify all scholarly authors, thus making it impossible to unanimously associate scholarly works with their authors. The inadequacy of author names as unique identifiers becomes obvious with common names such as Smith in the United States or Chen in China. Computer algorithms can disambiguate most authors using location, date, subject area, and authorship information but will still be wrong 5–10 percent of the time.

Author name ambiguity creates problems for everyone involved in scholarly communication. For universities and colleges, it means that there is no easy way to identify the publications and other scholarly works of their faculty and students—information they need for their institutional repositories, for expert discovery, for research assessment, and for other reporting purposes. Universities and colleges have several options to obtain this information: they can use library and administrative staff to collect this information, they can ask their researchers to regularly report their publications and other scholarly works, and they can obtain this information from a commercial service. Most institutions use a mix of these strategies, but a lot of information will simply be unavailable because it would require too large an effort. This means institutional repositories that are incomplete, researchers who are unaware of potential collaborators in their institution, and scholarly activities that go unnoticed.

The name ambiguity problem can be solved only by issuing unique identifiers for authors. Several author identifier initiatives have been started in the past ten years, and some of them are widely used in a particular discipline or geographic region—for example RePEc in economics or LATTEST in Brazil. Other author identifiers are global in scope but have not gained wide support in the scholarly community because they are controlled by a single commercial entity or because they lack the funding to build a sustainable service. And some author identifiers are generated from disambiguation algorithms, potentially creating multiple identifiers for the same person. Thus, since none of the available author identifier services looked like the solution to the name ambiguity problem, the ORCID initiative was started in late 2009 and formed as a nonprofit organization in August 2010.

The ORCID service intends to start issuing unique author identifiers in the late second or third quarter of 2012. The initial focus will be on active researchers, and they will be able to create, edit, and maintain an ORCID ID free of charge. ORCID identifiers will be 16 digit numbers, segmented into four-digit groups and including a checksum. They will be expressed as HTTP URI (such as http://orcid.org/0137-1963-7688-2319). ORCID identifiers contain no semantic information, such as the year the identifier was minted or the country of origin, and they are issued out of sequence. The ORCID service not only will issue unique author identifiers but also will enable linking to existing author identifier services. This is a core feature of the ORCID service, and many of the organizations behind these author identifier services (e.g., ResearcherID, Scopus Author ID, RePEC, INSPIRE) have been part of the ORCID initiative from day 1.

ORCID is not simply an author identifier registry. A unique author identifier is of limited value unless the service also helps with authorship claims of scholarly works. Although the initial focus will be on articles in scholarly journals, ORCID intends to help in collecting claims about all relevant scholarly contributions, from research datasets to grants. To achieve this functionality, ORCID is closely working with CrossRef, DataCite, and other providers of persistent identifiers for scholarly objects. It is anticipated that an open and universally accepted unique author identifier provided by ORCID will enable additional services provided by academic institutions, nonprofit organizations, and commercial entities. One example is the better attribution of data curation and other scholarly contributions that today often go unnoticed. ORCID will also facilitate the development of new scholarly metrics by linking authors, their publications and other works, and the relevant metrics (citations, usage data, etc.).

Early during the initiative, various organizations created alpha prototype environments to explore federated system interoperability, investigating the requirements for a compre-
hensive system that supports the needs of authors, colleges/universities, funding organizations, publishers, and other organizations involved in the ORCID initiative. The ORCID initiative then decided to build the service based on ResearcherID source code, which was licensed from Thomson Reuters in August 2011. Consistent with ORCID’s principles, ORCID was allowed to license the derivative work under the MIT open source license. While the ResearcherID code provided a strong headstart on creating a viable ORCID system, certain underlying systems needed to be changed to support the expanded requirements identified in the alpha exploration. This work started in the third quarter of 2011. Development APIs were openly released in December 2011 so that ORCID’s partnering organizations could start planning to interface with the system. The initial phase of development work (phase 1.0) was completed in April 2012 and focused on expanding the self-claim functionality, adding OAuth federated authorization and administrative functions and providing RESTful service interfaces. Specific features that were added include institutional seeding of profiles, delegated management of profiles, profile exchange into grant/manuscript submission systems, fine-grained control of privacy settings at the claim level (public/protected/private), ORCID identifier resolution, and metadata search.

While phase 1.0 completed the creation of a robust, internally consistent system, work with community organizations identified several optimizations that would significantly improve the roll-out of ORCID services. This development (phase 1.1) is completing in the second quarter of 2012 and includes everything else needed to open up the system for public use. Following completion of phase 1.1, ORCID will proceed with an initial roll-out of the system with three areas of focus, all related to self-claim functionality:

- Allowing researchers to claim their profiles in an open environment that transcends geographic and national boundaries, discipline, and institutional constraints
- Allowing researchers to delegate control of the ongoing management of their profile to their institution
- Providing an interoperable platform for federated exchange of profile information with systems supplied by publishers, grant managers, research assessment tools, and other organizations in the scholarly community

The initial focus on self-claims by active researchers is a strategic decision, since engaging active researchers in the ecosystem of organizations and systems they use today is essential for the widespread adoption of the ORCID service. ORCID recognizes that an ideal future solution will include functionality to allow for completely granular and flexible assertions about researchers and their scholarly endeavors from a much wider set of sources. Managing and deriving value from this partially interlocking network of assertions presents a much more challenging problem than the initial phase and will be addressed in phase 2 of the ORCID service. Given the importance of reputation and transparency in the scholarly community, assertions from authors about their claimed publications can be assumed to have a high accuracy. As ORCID is presented with additional assertions about researchers from many different types of organizations that have collected varying subsets of information, establishing confidence about the totality of works by an individual becomes exponentially more complex. Add to that retrospective challenges—such as print-to-electronic transcription errors, partial data elements, subject- or geographic-specific subsets, and a host of other incomplete or “dirty data” problems—and the potential for inaccuracy increases. But many of the ideas emerging in the linked data community can help structure and inform solutions to these problems.

The ORCID initiative aims to make the ORCID author identifier an essential part of the scholarly infrastructure, similar to digital object identifiers (DOIs) and other persistent identifiers for scholarly objects. To facilitate widespread adoption, participation in ORCID is open to any organization that has an interest in scholarly communication. More than 310 organizations are already participating. ORCID is governed by representatives from a broad cross-section of stakeholders, the majority of whom are not-for-profit. ORCID is also committed to open data and open source. ORCID will make all profile data contributed or claimed by researchers available for free under the CC0 waiver, and all software developed by ORCID will be publicly released under an open-source software license.

Another factor critical for widespread adoption is the perception that the ORCID organization is sustainable and can guarantee the longevity of the service. ORCID has always made it clear that building and maintaining a global author identifier service requires money. Although the organization has received several grants, donations, and loans, ORCID will need to start raising membership fees in 2013. The size of these fees will depend on the type and size of the organization; small academic institutions will be expected to pay much lower membership fees than are large commercial publishers. In addition, membership fees could decrease over time as more organizations join the initiative.

Everyone involved in ORCID is excited about the imminent launch of the service. We hope that a few years from now ORCID identifiers for authors will have become as commonplace as persistent identifiers for scholarly works are today.

**By BRIAN WILSON and MARTIN FENNER**

**Brian Wilson** (brian.wilson@thomsonreuters.com) is Head of Architecture for the Intellectual Property & Science business of Thomson Reuters and chairs the ORCID Technical Working Group. **Martin Fenner** (fenner.martin@mh-hannover.de) is a Clinical Fellow at Hannover Medical School in Hannover, Germany, and chairs the ORCID Outreach Working Group.

© 2012 Brian Wilson and Martin Fenner. The text of this article is licensed under the Creative Commons Attribution 3.0 License (http://creativecommons.org/licenses/by/3.0/).