Registering Researchers in Authority Files

Registering Researchers in Authority Files

#oclcresearch #rrafreport

Karen Smith-Yoshimura, OCLC Research

Micah Altman, Massachusetts Institute of Technology

Michael Conlon, University of Florida

Ana Lupe Cristán, Library of Congress

Laura Dawson, Bowker

Joanne Dunham, University of Leicester

Thom Hickey, OCLC Research

Daniel Hook, Symplectic Limited

Wolfram Horstmann, University of Oxford

Andrew MacEwan, British Library

Philip Schreur, Stanford University

Laura Smart, California Institute of Technology

Melanie Wacker, Columbia University

Saskia Woutersen, University of Amsterdam

OCLC Research

© 2014 OCLC Online Computer Library Center, Inc.
This work is licensed under a Creative Commons Attribution 3.0 Unported License.
http://creativecommons.org/licenses/by/3.0/

October 2014

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-488-4 (978-1-55653-488-1)
OCLC Control Number: 893635903

Please direct correspondence and feedback to:
Karen Smith-Yoshimura
Program Officer
karen_smith-yoshmura@oclc.org

Suggested citation:
Smith-Yoshimura, Karen; Micah Altman; Michael Conlon; Ana Lupe Cristán; Laura Dawson; Joanne Dunham; Thom Hickey; Daniel Hook; Wolfram Horstmann; Andrew MacEwan; Philip Schreur; Laura Smart; Melanie Wacker; and Saskia Woutersen. 2014. *Registering Researchers in Authority Files*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-a4.pdf.

Contents

Acknowledgments 5
Introduction 6
Recommendations 8
Recommended Practices for Researchers 10
Recommended Practices for Librarians/University Administrators 10
Recommended Practices for Funders 11
Recommended Practices for Identity Management Systems/Aggregators 11
Methodology 12
Challenges 13
Landscape Overview 14
Changes in the Field 19
Functional Requirements by Stakeholder 20
Emerging Trends 21
Opportunities 23
Supplementary Data 25
Image Attribution 26
Notes 26

Tables

Table 1. Comparison of name authorities and researcher ID systems 8
Table 2. Stakeholders and needs 9
Table 3. Scope and size of authority hubs and ID hubs profiled 16
Table 4. Scope and size of research information management and researcher profile systems profiled 17

Acknowledgments

The authors wish to thank Amanda Hill (University of Manchester) for her contributions to the use-case scenarios and profiles supplements to this report and to the following reviewers for their expertise and time: Linda Barnhart (University of California, San Diego), Christopher Brown (Jisc), Ricky Erway (OCLC Research), Janifer Gatenby (OCLC), Peter Gorman (University of Wisconsin-Madison), Stephen Hearn (University of Minnesota), Constance Malpas (OCLC Research), Merrilee Proffitt (OCLC Research), Verena Weigert (Jisc). Special thanks to JD Shipengrover (OCLC Research) for designing our Researcher ID Information Flow diagram. We'd also like to thank Sloan Galaxy Zoo whose *Mergers Author Poster* image appears on the cover of this report. (See the Image Attribution section at the end of this report for additional image details.)

Introduction

Registering researchers in some type of authority file or identifier (ID) system has become more compelling as both institutions and researchers recognize the need to compile their scholarly output. This need to uniquely identify researchers and correctly associate them with their scholarly output has given rise to bibliometrics and its extension, altmetrics—the attempt to measure the impact of a work as it is reflected by mentions in social networks and news media. University ranking tables rely in part on how often the works produced by the institutions' researchers are cited in professional and academic journals. Open Access mandates and funding reporting requirements represent other inducements to track an institution's scholarly output. These drivers present institutions with a challenge: how can they accurately measure and reflect the entire scholarly output of all their researchers? Information about a specific researcher may be represented in multiple databases, with only a subset interoperating with each other.

Scholarly output impacts the reputation and ranking of the institution. Three global university rankings of particular interest to research institutions—Times Higher Education World University Rankings, Academic Ranking of World Universities and QS Top Universities—all use citations as a factor in determining rankings.[1] Notes QS, "Citations . . . are the best understood and most widely accepted measure of research strength."[2] To better track researcher outputs, publishers have introduced a variety of ID schemes to support name disambiguation. Libraries have also made significant investment in disambiguating author names through name authority files. Authority files create a unique ID for a given entity represented by a text string constructed according to a specific rule set. Their focus is on the monographic literature; journal articles and other scholarly outputs are typically not represented.

Over the past two years the Registering Researchers in Authority Files Task Group has examined a range of researcher registration and profiling services. A number of approaches to providing authoritative researcher IDs have emerged, but they tend to be limited by discipline, affiliation, or publisher. The Task Group developed use cases and functional requirements for researcher ID management systems and then compared the functional requirements against a sample of currently available systems to identify gaps, challenges and opportunities. A key objective of this work was to understand how the various types of researcher ID schemes can be leveraged to support improved discoverability of research output for individual authors and institutions.

A registration file and an authority file may serve two distinct functions. A registration file strives to create a unique ID for a given entity. An authority file, while doing the same, may impose additional constraints such as formulating the text string associated with the entity according to specific rule sets (e.g., Resource Description and Access or RDA)[3].

Library name authority files manage identities in library catalogs to enable users to easily and quickly find all the known works—individual monographic publications such as books, musical scores, sound recordings, etc.—associated with a given person. Authority records are created to control the names of authors or creators of works and names that are subjects of works, such as biographies, as they are recorded in library catalogs. Librarians record as many names by which an identity is known in an authority record together with other differentiating information. Traditionally a library authority record contains sufficient information to uniquely identify the name in a library's catalog. To that end, it will likely contain more differentiating information such as dates of birth, known affiliations, etc. in records for common names than in records for unusual names. The goal is to ensure that an author search on the catalog will retrieve all and only those works associated with that author. Since libraries usually share an authority file, such as the international Library of Congress/Name Authority Cooperative Program (LC/NACO)[4] Name Authority File, an authority record may have to uniquely identify a name in the context of hundreds of libraries' catalogs representing millions of works and millions of authors.

A name ID is meant to uniquely identify a person, who may share the same name with others or may have variant forms of the name appearing in publications. Traditional library practice is to select a "preferred name form" that is used in the library's catalog. This preferred name form may differ from one community and language to another. For example, "Confucius" is used in Anglo-American communities and "孔子" in Chinese, Japanese and Korean communities. The potential to link between different authority files which may have very different preferred forms for the same name has been demonstrated by the Virtual International Authority File (VIAF)[5], a compilation of name authority files from dozens of different countries and agencies. It uses the differentiating data contained in library authority records and the associated works they are linked to in library catalog records to cluster the authority records together and then assigns that cluster a unique ID.

Table 1 compares some key attributes of name authority files and researcher ID systems.

Table 1. Comparison of name authorities and researcher ID systems

Key Characteristics	Traditional Name Authorities	Researcher ID Systems
Primary Stakeholders	Libraries	Publishers, researchers, funders, libraries
Internal Standardization/Integration	Standardized and well integrated within libraries but new models are emerging	Fragmented. Some well-integrated communities of practice
Structural Organization	Primarily top-down, careful controlled entry from participating organizations	Varies: top down, bottom-up, middle out; often individual contributors
External Integration	Very limited: High barriers to entry, few simple API's	Varies, but more open. Some services offer simple open API's; integration with web 2.0 protocols (e.g. OpenId)
Works Covered	Primarily books and other works traditionally catalogued by libraries	Journal articles; grants; datasets
People Covered	Authors and people written about represented in library catalogs	Authors of research articles, fundees, members of research institutions—international
Key Function	Persistent and unambiguous ID with a preferred label for the community served	Persistent and unambiguous ID for an individual contributor

Recommendations

Our recommendations and the functional requirements for registering researchers vary across different stakeholder communities. From 18 use-case scenarios[6] we identified six primary stakeholders: researcher, funder, university administrator, librarian, identity management system and aggregator (including publishers). More than one stakeholder may share the same requirement (e.g., a librarian and an identity management system may both need to disambiguate names.) Stakeholder needs are summarized in table 2.

Table 2. Stakeholders and needs

Key Characteristics	Traditional Name Authorities
Researcher	Disseminate research
	Compile all publications and other scholarly output
	Find collaborators
	Ensure network presence is correct
	Retrieve others' scholarly output to track a given discipline
Funder	Track funded research outputs
University Administrator	Collate intellectual output of their researchers to fulfill funder or national mandates or internal reporting requirements
Librarian	Disambiguate names
Identity Management System	Associate metadata and output with researcher
	Disambiguate names
	Link researcher's multiple IDs
	Disseminate IDs
Aggregator (includes publishers)	Associate metadata and output with researcher
	Collate intellectual output of each researcher
	Disambiguate names
	Link researcher's multiple IDs
	Track history of researcher's affiliations
	Track and communicate updates

Adoption of researcher IDs has been rapid within scholarly publishing. Funders and commercial publishers have seen clear benefits in adopting researcher IDs. We believe that it is equally important for universities, research institutions and libraries to recognize that authors are not strings and need to have persistent IDs to link them to their scholarly output. It is likely that measuring and reporting this output will be more complex in an environment of multiple ID systems.

The criteria for selecting which of the various IDs to use will depend on the stakeholder. Among the factors to be considered is to select the ID system which attracts the "critical mass" representing one's peers. National or funding mandates may also influence one's choice. The earlier in the career an ID is used in association with a researcher's scholarly output, the less likely misattribution will occur—especially for common names—and the more likely the scholarly output associated with a researcher will be comprehensive and accurate. University administrators and librarians can also play an important role by disseminating information about their researchers. All stakeholders are encouraged to be as open as possible with their IDs and associated attributes (in the absence of data protections or other legal concerns) so that these IDs can be reused in other contexts and in other systems. Key recommendations for each stakeholder follow.

Recommended Practices for Researchers

The task group's review of the use cases, landscape and trends show increasing importance of modes of scholarly communication that go beyond books and even beyond articles to include many other diverse forms and formats.[7] Unique, persistent IDs are needed to increase the likelihood that all researchers' output is represented and that this output is attributed to the correct individual researcher and the researcher's institution. Accurate affiliation is critical to both accurately compile scholarly output in all formats and for others to form an assessment of the trustworthiness or authority of research output.

- Obtain a persistent ID before submitting any output. Ask your librarian or university administrator if you are unsure which IDs are most suitable or don't know how to get one.
- Disseminate your persistent IDs on all external communications—faculty profiles, email signature, professional networks, LinkedIn, or anywhere you communicate with your peers.
- Include the ISNI of your organization(s) and funders in the research output that you submit. Search www.isni.org. If your organization does not have an ISNI, it can request one through an ISNI Registration Agency.[8]
- Resolve errors in your metadata (affiliations, attributions, etc.) or if you're represented in the same system more than once. Consider reporting them to your librarian or university administrator if you cannot correct the errors yourself.

Recommended Practices for Librarians/University Administrators

As a result of the task group's landscape review, we see that with the increasing adoption of researcher IDs by both publishers and funders, research institutions also need to implement unique IDs for their researchers. These IDs will facilitate aggregating and tracking researcher output to meet funder requirements and to contribute to institutional assessments. Unique, persistent IDs are crucial to validate that the output is indeed done by the individual researcher under the auspices of the institution.

- Assign persistent IDs to authors if they don't already have them. This includes authors of electronic dissertations in institutional repositories and papers or datasets uploaded to research websites.
- Retain traditional IDs (e.g., name authority file and VIAF IDs) as they are well supported.

- Ensure that the ISNI or other ID for your organization is accurate and well maintained, and promote its use.
- Integrate researchers' external IDs within library applications and services as appropriate.
- Advocate the benefits and reasons for researchers to register, use and diffuse their IDs.
- Find out from your identity management system provider or aggregator how to report errors.
- Provide guidance and training materials on why using persistent IDs is important, conveying good practices for how to get them, where to include them and how to report errors.

Recommended Practices for Funders

As a result of our landscape review, the task group notes the increasing pressure on funders to show the impact of the research they fund, and with that comes the requirement to track many types of research output. Given the myriad of new output types, unique, persistent IDs are key in making this type of tracking efficient, cost-effective and scalable. As the funding landscape becomes more diverse and complex, funders must disambiguate both funded researchers and those applying for funding. To meet stakeholder expectations, funders need to take advantage of the significant changes in the infrastructure, including the increased adoption of IDs by publishers, aggregators and research institutions.

- Encourage researchers who receive grants to obtain a persistent, standard name ID and use it systematically when research outputs are disseminated.
- Embed IDs in grant submission systems and workflows.
- Declare a persistent public ID for the organization administrating the funds.

Recommended Practices for Identity Management Systems/Aggregators

Our examination of use cases, tracing of information flow and mapping profile characteristics to functional requirements revealed a number of gaps and led the task group to recommend the following:

- Design your system so each data element is tagged with provenance information. This information is important for your system users to assess the

"trustworthiness" of the information displayed, especially when you have similar information from multiple sources. You'll need this information to pass on error reports or corrections.

- Establish maintenance mechanisms to:
    - Correct information about a researcher.
    - Merge identities representing the same person.
    - Split entities representing different researchers.
- Establish protocols to communicate changes and corrections to the original source
- Create framework to identify and handle privacy and rights issues. Be willing to share information for matching information between different systems even if the information is not displayed (such as birth dates, provided this does not violate legal requirements and agreements).
- Support batch searching and updating. Enable organizations to export thousands of names at a time to obtain IDs.
- Address interoperability of standards for both formats and data elements.
- Include the IDs used in other systems.
- Link researcher IDs to the institutions or agencies they are affiliated with.

Methodology

The OCLC Research Registering Researchers in Authority Files Task Group comprises specialists from the Netherlands, United Kingdom and the United States with different perspectives: researcher, librarian, publisher, ORCID or ISNI Board member, LC/NACO contributors, Program for Cooperative Cataloging, VIAF, VIVO, a Research Information Management System (Symplectic) and a national researcher system. The group communicated through conference calls, email and shared documents and ideas using a web-based project management and collaboration tool.

For this work, we defined researcher as "anyone who produces or is in the process of producing scholarly output." By "registering" we mean assigning a persistent ID—a unique reference to a single individual that does not change over time or between systems—usually with some attributes such as discipline, institutional affiliation, or titles of the individual's publications.

The group developed 18 use-case scenarios around six types of actors. The group derived a set of functional requirements from the use-case scenarios. We augmented a list of 100 research networking/ID management systems compiled by Dr. Michael Conlon and selected 20 systems to profile based on two criteria:

1. The system had to have significant uptake or "mind-share" by researchers.
2. The system had to represent researchers with a persistent, unique and publicly accessible URI.

We wanted a representative sample of different researcher information/ID systems. For some system types we profiled only one system to represent a category. We mapped each profile to the functional requirements and identified functional requirements not met and overlaps. We diagramed the relationships among the systems. We solicited external feedback on our use-case scenarios, research networking systems characteristics profiles, functional requirements, and a researcher ID information flow diagram as "work in progress" in July 2013.

From this research, we identified changes in the field, emerging trends, opportunities and recommendations targeted for each stakeholder. This report summarizes the results of our research; all our research is available as supplementary datasets listed at the end of this document.

Challenges

We identified several key challenges associated with disambiguating researcher identities:

- A scholar may be published under many names variations. Abbreviated given names used in journal articles are generally absent from national authority files and authors that publish only in journal articles may not be represented in authority files at all. Middle names or initials may be used inconsistently. If a scholar's work is translated, the transliteration of the scholar's name in non-Latin scripts— such as Arabic, Chinese, Cyrillic, Hebrew, Japanese katakana, Korean hangul, or any of the Indic scripts— makes it difficult to rely on text string matching to determine if two authors represent the same person or not.
- Multiple people can share the same name, requiring additional attributes or metadata to distinguish them, such as discipline or research topics, institutional affiliations, or links to publications. This is particularly true for Chinese names, where 87% of the population in China shares just 100 family names (compared to the United States where 90% of the population uses 151,671 family names). 270 million Chinese have the family name of Li, Wang or Zhang, not including all the overseas Chinese.[9]

- Some researchers already have multiple profiles or IDs, which may not be linked. A researcher may have profiles or IDs in systems such as Academia.edu, Google Scholar, ISNI (International Standard Name ID), Mendeley, Microsoft Academic, ORCID (Open Researcher and Contributor ID), ReseachGate, Scopus, VIAF (Virtual International Authority File), and VIVO as well as be represented in the institution's research information management system. The scholar's web presence may thus be fragmented. Sometimes scholars deliberately maintain distinct identities when publishing in different subject areas or writing under pseudonyms. Privacy control is an additional layer of complexity to consider when developing mechanisms for associating IDs.
- Information related to a researcher or the researcher's scholarly output that is updated in one system may not be reflected in other systems that include the researcher's work.
- Standards among different ID systems for both formats and data elements are often not interoperable.

Landscape Overview

Researcher IDs are used in a variety of systems at different scales—institutional, disciplinary, national, international and webscale. The total number of researchers worldwide was estimated to be about nine million in 2012.[10] No one system includes all of them, and it can be difficult to determine how many active, professional researchers are represented. The current researcher ID information flow represents a complex ecosystem, as illustrated in figure 1.

The 20 systems we profiled that met our selection criteria fell into 10 categories or types:

- **Authority hubs**, providing a centralized location of authority records for multiple institutions (6): Digital Author ID (DAI) in The Netherlands; Lattes Platform in Brazil; LC/NACO Authority File; Names Project in the United Kingdom; ResearcherID; Virtual International Authority File (VIAF)
- **Research information management systems** (sometimes known as CRIS in the context of European systems), storing and managing data about research conducted at an institution and integrating it with data from external sources: Symplectic
- **ID hubs**, providing a centralized registry of IDs (2): International Standard Name ID (ISNI); Open Researcher and Contributor ID (ORCID)

- **National research portals**, providing access to all research data stored in a nation's network of repositories: National Academic Research and Collaborations Information System (NARCIS) in The Netherlands
- **Online encyclopedias**, providing information divided into articles which includes references to the works by scholars and articles about individual scholars: Wikipedia
- **Reference management systems**, helping scholars organize their research, collaborate with others, and discover the latest research: Mendeley
- **Research and collaboration hubs**, providing a centralized portal where scholars in a particular discipline can work together: nanoHUB
- **Researcher profile systems**, facilitating professional networking among scholars (5): Community of Scholars; Google Scholar; LinkedIn; SciENcv; VIVO
- **Subject author ID systems**, linking registered scholars with the records about the works they have written: AuthorClaim
- **Subject repositories**, facilitating scholarly exchanges within a discipline-based centralized repository in a particular field: arXiv

Each profile includes: URL for the site; year started; purpose; description; scope; sources; content; size; who it is used by; public functions; restricted functions; which other research network or ID systems it interoperates with; overlaps with other systems; whether it supports linked data; access methods; metadata schema; licenses; fees; responsible agency; and references.[11] Some of the information was not available publicly, and where possible we interviewed the system's developers to provide as complete a profile as possible.

Comparisons were more meaningful where we had profiled more than one system within a category. For example, the comparison of just the scope and size of authority hubs and ID hubs illustrates a key difference between those that focus only on researchers and those that are very broad, with millions of people represented, in which the number of professional, active researchers cannot be ascertained. The Lattes Platform is the largest of the researcher-only databases profiled; it includes Brazilian researchers in all disciplines plus those outside Brazil who work with them.

A link to the spreadsheet compiling all the characteristics of all the research networking systems profiles is included in the list of supplementary data.

Table 3. Scope and size of authority hubs and ID hubs profiled

Hub	Scope	Size (as of Aug 2014)
Digital Author ID	Researchers in all Dutch research information management systems & library catalogs	• 66K researchers
Lattes Platform	Brazilian researchers and research institutions	• 2M researchers • 4K institutions
ISNI	Data from libraries, open source resource files, commercial aggregators, rights management organizations. Includes performers, artists, producers, publishers	• 8M names total • 830K researchers*
LC/NACO Authority File	Persons, organizations, conferences, place names, works	• 9M names total • ? researchers
ORCID	Individual researchers plus data from CrossRef/Scopus, institutions, publishers	• Over 800K registrants
ResearcherID	Researchers in any field, in any country	• 350K researchers
VIAF	Library authority files for persons, organizations, conferences, place names, works	• 30M people • ? researchers

* These numbers are for *assigned* ISNIs; total number of names in the ISNI database is 18.7 million of which 1.8 million are researchers.

These authority and ID hubs overlap to varying degrees. For example, ISNI is a VIAF contributor, and has loaded and matched records from the Dutch DAI. LC/NACO Authority File records occasionally include ISNI or ORCID IDs and are also contributed to VIAF. There are some key differences in approaches. For example, ISNI consolidates data from multiple databases while ORCID receives only the data from individual or institutional registrations.

Systems characterized as research information management systems and researcher profile systems similarly include those that focus only on researchers, while those like Google Scholar, ResearchGate and LinkedIn serve millions of people with an unknown percentage representing professional, active researchers.

Table 4. Scope and size of research information management and researcher profile systems profiled

Profile System	Scope	Size (as of Aug 2014)
Symplectic Elements	Institutional data sources plus article & citation, reference databases, bibliographies. author names, IDs, affiliations, bibliography, grants, professional activities, educational history, employment history, teaching activities, etc.	• Over 100K researchers
Community of Scholars	Researcher-created profiles plus more than 70 article and citation databases	• Over 3M profiles
Google Scholar	Authors of publications in commercial, institutional, and web sources indexed by Google Scholar	• Millions • Unknown "verified" research profiles
LinkedIn	International user base; profiles created by individuals and organizations	• Over 300M members • ? researchers
SciENcv	National Institutes of Health pilot to link researchers to their grants and output; data from federal and non-federal sources; embedding ORCIDS	• 100K researchers
VIVO	Sources: funding agencies, institutions, scholar self-reports, open source and publisher data. Profiles include all scholarly output and research, teaching, service activities.	• Over 100K researchers

Research information management systems like Symplectic pull information from a variety of bibliographic and bibliometric databases such as ArXiv, CiNii, CrossRef, DBLP, figshare, Google Books, Mendeley, ORCID, PubMed, RePEc, SciVal, Scopus and Web of Science. Community of Scholars includes content from over 70 ProQuest and CSA proprietary databases and other certain verified publications such as ABI/INFORM, ERIC and PubMed; information is also pulled from the scholar's personal and institutional websites, if it exists. SciENcv's beta release provides a shared, voluntary researcher profile system for individuals who receive or are associated with research investments from federal agencies. By embedding ORCID IDs in the grant application workflow, SciENcv can help funders and administrators better track research and understand the impact of the National Institutes of Health's funding.

The same information about a specific researcher may be represented in multiple databases, and only a subset interoperates with each other. Figure 1 depicts the flow of information describing researchers and identified researcher outputs such as publications among classes of actors and systems. Each category of actor is depicted with an icon and a caption. An arrow going from one entity to another is used to indicate that information flows from one entity to the other.

Specific entities (Crossref, ISNI, VIAF, ORCID) are indicated with bold text; all other entities represent a category of entity within the general types—public aggregator, internal aggregator or public view. The presence or absence of arrows is based on analyzing the detailed research networking systems characteristics profiles[12]. Arrows are shown in the diagram wherever there is at least one pair of profiled entities of those categories transmitting information.

Figure 1: Researcher ID information flow[13]

The researcher ID information flow diagram reveals several patterns:

1. Information flow is quite complex overall, especially between public aggregators and private aggregators.

2. Most of the information flow is one way—there are few automatic or systematic channels for corrections or annotations at later stages to flow back up to public aggregator sources, or to the actors providing them. ISNI—which focuses on managing links between member ID systems and databases—is an exception. It tracks all sources, and corrections are fed back to the source as they are made. The interoperability flows between ISNI and ISNI registration agencies, and between ISNI and VIAF, are indicated in the diagram by bidirectional arrows.

3. Many actors are contributing to multiple public views of Researcher ID information.
4. Information provided by the same class of actor may flow through multiple (possible concurrent) paths to internal aggregator systems and public views—suggesting the potential for duplication and inconsistency.
5. Individual categories of public views seldom represent more than one type of internal aggregator.

A key question is how corrections or updates can be communicated between systems. Researchers are frustrated when they see errors in their profile, works incorrectly assigned to them or works missing. Even if the information is corrected in the local instance, it often is not reflected in the aggregated databases or hubs.

Changes in the Field

Since the task group started its work in September 2012, we have seen a number of changes in the field; both in the organizations involved with researcher IDs and the increased adoption of researcher IDs. Some of the key changes include:

- The Names Project ended in July 2013. Its deliverables are available for re-use and the results of the project live on in the form of ISNI IDs for UK researchers.[14]
- Elsevier acquired Atira (a Research Information Management System) and Mendeley.
- Thomson Reuters acquired AVEDAS, the supplier of CONVERIS, another Research Information Management System.
- Elsevier and University College London launched the "UCL Big Data Institute" (building on Elsevier's acquisition of Mendeley), a new innovation lab to " tackle the challenges researchers face as they seek to forecast trends, synthesize information from thousands of research papers, and show the potential societal impact of their research so it will be eligible for funding."[15]
- ISNI continues to increase the number of researchers represented, now including identities from 12 research-based sources as of January 2014: American Musicological Society, AuthorClaim, British Library Theses, Digital Author IDs (DAI, Dutch), Jisc Names (UK), Modern Languages Association, OCLC Theses, ORCID and DataCite Interoperability Network (ODIN), Proquest Theses, RePec, Scholar Universe and Electronic Tables of Contents (ZETOC).[16]
- ISNI officially submitted a "notice of inquiry" to the United States Copyright Office to make ISNI part of the author's process of obtaining copyright.[17]

- More research sites are using or experimenting with ORCID to identify their researchers, including PubMed, nanoHUB and SciENcv.
- Wikipedia has been adding ISNIs, ORCIDs and VIAFs via an "authority control template"[18]; ISNIs and VIAFs have been added to Wikidata (structured data shared by Wikimedia projects including Wikipedia, Wikivoyage and Wikisource).[19]
- The Library of Congress' Bibliographic Framework Initiative has proposed linking local authority files to external authority sources as its model for the future.[20]
- VIVO and Harvard's Profiles Research Networking Software (using VIVO format) has become more widespread, with more than 100 institutions in the United States plus universities in Australia and China.[21]
- Google added Thomson Reuters Web of Science citation links to Google Scholar results pages for Web of Science subscribers.[22]
- ÜberWizard launched the "ÜberWizard for ORCID" to help researchers easily add their grants from multiple funders into their ORCID record.[23]

Functional Requirements by Stakeholder

Working from the set of stakeholder needs, the Task Group identified corresponding functional requirements for research networking systems supporting different communities (The full set is in the Functional Requirements supplementary data spreadsheet.[24])

For researchers and university administrators

- Link multiple IDs a researcher might have to collate output.
- Associate metadata with a researcher's ID that resolves to the researcher's intellectual output.
- Verify that a researcher and his or her scholarly output is correctly represented.
- Register a researcher who does not yet have a persistent ID.

For funders and university administrators

- Link metadata for a researcher's and institution's output to grant funder's data.

For librarians

- Create consistent and robust metadata through manual or automatic means.
- Associate metadata for a researcher's output with the correct ID.

- Disambiguate similar results.
- Merge entities that represent the same researcher and split entities that represent different researchers.

For aggregators and identity management systems

- Make pre-existing authoritative institutional IDs in systems (such as email, Single-Sign-On, Finance or Human Resources) interoperable with external systems such as ISNI, ORCID or VIAF.
- Link a researcher's multiple IDs.
- Determine whether an aggregated identity record represents a single identity.
- Affiliate a researcher with multiple departments, institutions, disciplines.
- Track history of a researcher's various affiliations over time.
- Merge entities representing the same researcher & split entities representing different researchers.
- Communicate information to other systems.
- Tag each data element with its source or provenance.
- Support batch searching and updating.
- Support Unicode to record researchers' names and citations in other languages and writing systems.

None of the systems profiled met all of the functional requirements. Please refer to the supplementary dataset [25] which maps each of the 20 profiled systems to the full set of functional requirements to determine which type of system might be most appropriate for your needs.

Emerging Trends

This field is changing so quickly that it is hard to tell whether a couple of examples represent isolated occurrences or indications of an emerging trend. The task group has identified the following as broad emerging trends:

- The need for persistent IDs for researchers has become widespread and there is increasing use of both ISNI and ORCID IDs to disambiguate names. More broadly, Wikipedia, search engines such as Google and the open web community have been investing efforts into disambiguating names. More researchers have multiple IDs in multiple systems.

- Open data or public access mandates are increasing, with the call that publicly-funded research be accessible to all. This will also increase the demand for researchers to have—and use—persistent IDs

- Registration files are being used more than authority files to identify researchers. Recognizing that there is no one central authority file, the Program for Cooperative Cataloging is considering changes to include references to researcher ID systems in the LC/NACO Authority File.[26] National programs have emerged to register all their researchers, such as the Dutch Digital Author Identification (DAI) system and the Lattes Platform in Brazil. The Portuguese national funder FCT requires ORCIDs for its national evaluation system[27]; all Dutch researchers with DAIs also have ISNIs.[28]

- Universities are assigning IDs to researchers. Some approaches which we have noted:

    - Assigning ORCIDs to authors when submitting electronic dissertations in institutional repositories. (Harvard)[29]

    - Automatically generating preliminary authority records from publisher files (Harvard pilot).[30]

    - Assigning ISNI IDs to all university researchers (LaTrobe)[31]

    - Streamlining ORCID implementations and developing models for effective adoption—Jisc and ARMA (Association of Research Managers and Administrators) launched a pilot with eight UK higher education institutions.[32]

- Publishers are among the early adopters of ORCIDs (ACM, Elsevier, PLOS, Springer, Taylor & Francis, Thomson Reuters, Wiley) and ISNI (Bowker and Ringgold are ISNI registration agencies; MacMillan's Digital Science is an ISNI member). Publishers have started to mark up their websites in schema.org, allowing more linking between library and non-library domains.[33]

- More machine processing is surfacing the need for further standardization of both formats and data elements in the metadata about researchers. Interoperability between systems is increasing:

    - ISNIs may be automatically added to LC/NACO authority records.

    - ISNI and VIAF have established interoperability procedures.

    - ORCID and ISNI are coordinating their services. ORCID now includes organization IDs to be cross-walked with their ISNI organization IDs.[34] The ORCID and ISNI boards recently signed a Memorandum of Understanding

defining forms of interoperation, investigating synergies and differences between their systems, and how to share or link IDs. ORCID has released a beta lookup system to search and retrieve ISNI IDs while inside ORCID.[35]

- Academic open source environments have started to integrate researchers' IDs into their platforms, such as the ORCID Adoption and Integration (A&I) Program.[36]
- More research information management systems are integrating with VIVO or other profiling systems. Two widely used profile systems are VIVO and Harvard's Profiles Research Networking Software.[37]

- More metadata formats can represent researchers and their output, including: MARC, XML, MODS, MADS, ZThes, Common European Research Information Format (CERIF), VIVO RDF. Most if not all research information management systems in the United Kingdom are CERIF-compliant. More organizations and universities are joining the Consortia Advancing Standards in Research Administration Information (CASRAI) to jointly develop a common data dictionary and advance best practices for data exchange and reuse.[38]

Opportunities

The desire to compile and present to the Internet community the complete and accurate output of a scholar or an institution offers new opportunities. Use of IDs rather than text strings to represent a researcher could enable a more comprehensive compilation of the researcher's work. New researcher ID systems are including many authors who have not been represented in authority files. The widespread use of persistent IDs could provide the means to link and integrate all scholarly output, including those in non-traditional forms such as datasets, e-prints, presentations, survey results and social media like blogs.

The changing landscape opens the opportunity to reconsider the role of authorities in an ecosystem of registered IDs. Universities may wish to consider automatically registering researchers and assigning persistent IDs beginning with the university's local faculty files. This is an opportunity to shift to a more "dynamic" way of establishing authority headings. Rather than waiting until sufficient information is available to create a national authority heading, librarians could generate a stub with an ID that others could augment or match and merge with other metadata.

Batch-querying ID databases would allow more linking of a researcher's multiple IDs. We believe it's unlikely that there will ever be one, comprehensive source for all researchers, so the ability to communicate information among systems becomes crucial.

We think there is a huge opportunity for third-party reconciliation or resolution services to provide linking among different ID systems rather than have each organization attempting to construct such a linking system on its own. Coordinated with researcher ID systems, authority files and the catalogs with which they are associated represent a rich resource of curated data that can support data linking and semantic web applications. These connections can improve sharing of research outputs, collaboration and new research discoveries.

Supplementary Data

This report summarizes the results of the research conducted by the OCLC Research Registering Researchers in Authority Files Task Group in 2012-2014. Details of this research are in the following supplementary data sets:

Supplement A: Use Cases – The 18 use-case scenarios the task group developed around different stakeholders: researcher, funder, university administrator, librarian, identity management system and aggregator (including publishers). http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-supplement-a.pdf.

Supplement B: Research Networking Systems Characteristics Profiles – The 20 research networking systems the task group characterized. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-supplement-b.pdf.

Supplement C: Excel Workbook – This Excel file consists of four worksheets that contain data considered or derived by the OCLC Research Registering Researchers in Authority Files Task Group. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-supplement-c.xlxs.

Worksheet 1: Research Networking Systems (candidates considered) **–** This worksheet contains the list of 100 research networking and ID systems the task group considered, with the ones selected for profiling highlighted. The Registering Researchers in Authority Files Task Group augmented a list originally provided by task group member Dr. Michael Conlon.

Worksheet 2: Functional Requirements (derived from 18 use-case scenarios) **–** This worksheet contains the 48 functional requirements the Registering Researchers in Authority Files Task Group derived from the use-case scenarios and their associated stakeholders.

Worksheet 3: Profiles (compilation of research networking systems characteristics profiles—overview) **–** This worksheet contains all 20 research networking systems' profiles for easy comparison.

Worksheet 4: Systems Mapped to FRs –This worksheet contains each of the 20 research networking systems profiled and mapped to the functional requirements.

Image Attribution

Cover:
Galaxy Zoo. 2007. *Mergers Author Poster*. Licensed under CC-BY-NC-ND 2.0. http://zooniverse-resources.s3.amazonaws.com/advent-calendar-2010/MergersPoster_5000.jpg.

Notes

1. Times Higher Education World University Rankings: http://www.timeshighereducation.co.uk/world-university-rankings/
   Academic Ranking of World Universities: http://www.shanghairanking.com/
   QS Top Universities: http://www.topuniversities.com/
2. See QS (Quacquarelli Symonds Limited). 2013. "Citations per Faculty." *Intelligence Unit.* http://www.iu.qs.com/university-rankings/rankings-indicators/methodology-citations-per-faculty/. Citations representing "research influence" is listed as part of the Times Higher Education's World University Rankings 2013-2014 methodology (captured as the author saw it on 8 September 2014 at http://oc.lc/the2013-2014methodology); Academic Ranking of World Universities' methodology includes highly cited researchers and articles indexed in major citation indices (captured as the author saw it on 8 September 2014 at http://oc.lc/arw2014methodology).
3. http://www.rda-jsc.org/rda.html
4. http://www.loc.gov/aba/pcc/naco/index.html.
5. http://viaf.org.
6. See OCLC Research Registering Researchers in Authority Files Task Group. 2014. *Supplement A: Use Cases*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-supplement-a.pdf.
7. For a conceptual framework of the nature and scope of the evolving scholarly record driving the evolution of scholarship, see: Lavoie, Brian F., Eric Childress, Ricky Erway, Ixchel M. Faniel, Constance Malpas, Jennifer Schaffner, and Titia van der Werf. 2014. *The Evolving Scholarly Record*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-evolving-scholarly-record-2014.pdf.
8. ISNI Registration Agencies are listed at: http://isni.org/do-you-have-an-isni.
9. Fish, Isaac Stone. 2013. "Why Do So Many Chinese People Share the Same Name?" *Passport* (blog). Foreign Policy. 26 April. http://blog.foreignpolicy.com/posts/2013/04/26/why_do_so_many_chinese_people_share_the_same_name.
10. The nine million researchers worldwide figure is extrapolated from the World Bank's 2013 figure of 1,271 researchers (full-time equivalents) per million inhabitants (https://web.archive.org/web/20140225192253/http://wdi.worldbank.org/table/5.13) and a 2012 world population of 7.046 billion

(https://web.archive.org/web/20140225193314/http://wdi.worldbank.org/table/2.1). The World Bank's definition of researcher: "Professionals engaged in the conception or creation of new knowledge, products, processes, methods, or systems and in the management of the projects concerned. Postgraduate PhD students (ISCED97 level 6) engaged in R&D are included."

11. See OCLC Research Registering Researchers in Authority Files Task Group. 2014. *Supplement B: Research Networking Systems Characteristics Profiles*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-supplement-b.pdf.
12. See OCLC Research Registering Researchers in Authority Files Task Group. 2014. *Supplement C: Excel Workbook*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-supplement-c.xlsx.
13. This image is available as a PowerPoint slide with notes. See: OCLC Research Registering Researchers in Authority Files Task Group. 2014. *Researcher ID Information Flow*. PowerPoint slide. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014-rid-info-flow-slide.pptx
14. Names Project deliverables are documented here: http://namesproject.wordpress.com/.
15. Kisjes, Iris. 2014. "University College London and Elsevier launch UCL Big Data Institute." Editors' Update. Elsevier. 6 January. http://editorsupdate.elsevier.com/short-communications/university-college-london-and-elsevier-launch-ucl-big-data-institute/.
16. For a current list of ISNI's research-based sources, see: http://www.isni.org/content/data-contributors.
17. ISNI International Agency. 2013. "The ISNI International Agency Recommends the use of the International Standard Name Identifier (ISNI) within the Copyright Office's Systems." Response to Library of Congress "Notice of Inquiry Docket 2013-2" dated 20 May. http://www.isni.org/filedepot_download/58/336.
18. The authority control template used in English-language Wikipedia at https://en.wikipedia.org/wiki/Template:Authority_control has been adapted by other Wikipedia sites but not all.
19. See, for example, the "Statements" on the Wikidata page for Noam Chomsky at http://www.wikidata.org/wiki/Q9049. (Captured as the author saw it on 23 September 2014 at http://oc.lc/chomsky).
20. BIBFRAME.ORG. 2013. "On BIBFRAME Authority." 15 August 2013 http://bibframe.org/documentation/bibframe-authority/.
21. See http://profiles.catalyst.harvard.edu/?pg=community and https://wiki.duraspace.org/display/VIVO/The+VIVO+Wiki. (Captured as the author saw them on 8 September 2014 at http://oc.lc/prns and http://oc.lc/vivo respectively.)
22. Cornell University Library. 2013. "Web of Science & Google Scholar Announce New Partnership" In Albert R. Mann Library *News*. Posted 16 December 2013. http://mannlib.cornell.edu/news/web-science-google-scholar-announce-new-partnership.
23. Über Research. 2014. "ÜberWizard for ORCID Launched—Supporting Researchers in Adding Grants from Various Funders to their ORCID Records with a Free and Open Tool." Posted 3 March 2014. http://www.uberresearch.com/uberwizard-for-orcid-launched-supporting-researchers-in-adding-grants-from-various-funders-to-their-orcid-records-with-a-free-and-open-tool/.
24. Refer to "Worksheet 1. Functional Requirements" in OCLC Research Registering Researchers in Authority Files Task Group. 2014. *Supplement C: Excel Workbook*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014d.xlxs.
25. Refer to "Worksheet 4. Systems Mapped to FRs" in OCLC Research Registering Researchers in Authority Files Task Group. 2014. *Supplement C: Excel Workbook*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-registering-researchers-2014d.xlxs.

26. Library of Congress. 2014. *Descriptive Cataloging Manual: Z1: Name and Series Authority Records.* Field 024: Other Standard Number. Washington, D.C.: Cataloging Distribution Service, Library of Congress. http://www.loc.gov/catdir/cpso/dcmz1.pdf.
27. De Castro, Pablo. 2014. "Building Pioneering Functionality around ORCID Integration: FCT and Portugal" *GrandIR Blog.* Posted 18 February (00:25). http://grandirblog.blogspot.com/2014/02/building-pioneering-functionality.html.
28. SURF (Collaborative organization for ICT in Dutch higher education and research) defines DAIs at http://www.surf.nl/en/themes/research/research-information/digital-author-identifier-dai/digital-author-identifier-dai.html. DAI is listed among ISNI's data contributors at http://www.isni.org/content/data-contributors.
29. "ETDs @Harvard" http://www.hsph.harvard.edu/registrar/etds-harvard/.
30. Christine Fernsebner Eslao gave a presentation on Harvard's "OAQ: Identifying Authors through Publisher-Collected Metadata" at the American Libraries Association annual meeting 3 July 2014: http://ala14.ala.org/files/ala14/Eslao_oaq%2020140628_1.pdf.
31. "Peak Global Researcher Identifier Group Welcomes LaTrobe University as a Member and Data Contributor." *LDawson's blog*. http://isni.org/content/peak-global-researcher-identifier-group-welcomes-la-trobe-university-member-and-data-contrib.
Sadler, Roderick and Simon Huggard. 2014. "International Standard Name Identifier (ISNI) Identifiers for One University's Researchers: What, Why, How." Paper presented at eResearch Australiasia Conference, Melbourne, Australia, October 2014. http://eresearchau.files.wordpress.com/2014/07/eresau2014_submission_34.pdf.
32. Jisc-ARMA ORCID pilot project is described at: http://www.jisc.ac.uk/fundingopportunities/funding_calls/2014/03/orcid.aspx.
The eight higher education institutions selected are listed at: http://orcidpilot.jiscinvolve.org/wp/hei-based-projects/.
33. See for example the source code for Elsevier's website, where the schema.org tag for "CreativeWork" is used: view-source:http://www.journals.elsevier.com/accounting-organizations-and-society.
34. http://support.orcid.org/knowledgebase/articles/276884.
35. MacEwin, Andrew, and Laure Haak. 2014. "ISNI and ORCID sign Memo of Understanding." *LDawson's blog*. ISNI. January. http://www.isni.org/content/isni-and-orcid-sign-memo-understanding.
36. https://orcid.org/content/adoption-and-integration-program.
37. From VIVO Project Director Layne Johnson correspondence 21 August 2014: "Some VIVO leaders estimate that there are at least 150 installations worldwide." For Profiles RNS, see: http://profiles.catalyst.harvard.edu/?pg=community.
38. The list of international research organizations collaborating with CASRAI has expanded beyond Canada to include Avedas, Jisc, EuroCRIS, Symplectic and VIVO. http://casrai.org/program/organizations.

For more information about our work registering researchers in authority files, please visit:
www.oclc.org/research/activities/registering-researchers.html

OCLC Research

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 1-55653-486-8
978-1-55653-486-7
1410/215261, OCLC