Integrating Researcher Identifiers into University and Library Systems

Presented by:

Micah Altman, escience@mit.edu
Director of Research, MIT Libraries
Non-Resident Senior Fellow, Brookings Institution

Karen Smith-Yoshimura smithyok@oclc.org
Program Officer, OCLC Research
Related Work


Draft for community review and comment at:

http://www.oclc.org/research/activities/registering-researchers/progress.html
Collaborators & Co-Conspirators

Registering Researchers in Authority Files Task Group Members

- Micah Altman, MIT - ORCID Board member
- Michael Conlon, U. Florida – PI for VIVO
- Ana Lupe Cristan, Library of Congress – LC/NACO trainer
- Laura Dawson, Bowker – ISNI Board member
- Joanne Dunham, U. Leicester
- Amanda Hill, U. Manchester – UK Names Project
- Daniel Hook, Symplectic Limited
- Wolfram Horstmann, U. Oxford
- Andrew MacEwan, British Library – ISNI Board member
- Philip Schreur, Stanford – Program for Cooperative Cataloging
- Laura Smart, Caltech – LC/NACO contributor
- Melanie Wacker, Columbia – LC/NACO contributor
- Saskia Woutersen, U. Amsterdam

- Thom Hickey, OCLC Research – VIAF Council, ORCID Board
DISCLAIMER

These opinions are our own, they are not the opinions of MIT, Brookings, OCLC, any of the project funders, nor (with the exception of co-authored previously published work) my collaborators

Secondary disclaimer:

“It’s tough to make predictions, especially about the future!”

— Attributed to Woody Allen, Yogi Berra, Niels Bohr, Vint Cerf, Winston Churchill, Confucius, Disraeli (sic), Freeman Dyson, Cecil B. Demille, Albert Einstein, Enrico Fermi, Edgar R. Fiedler, Bob Fourer, Sam Goldwyn, Allan Lamport, Groucho Marx, Dan Quayle, George Bernard Shaw, Casey Stengel, Will Rogers, M. Taub, Mark Twain, Kerr L. White, etc.
Integrating Researcher Identifiers into University and Library Systems

News

Right-scaling Stewardship report provides multi-scale perspective on cooperative print management
25 March 2014
This report explores regional-scale cooperative print management from two perspectives: a local academic library print book collection and a consensual-scale collective print book resource. #printscale

New infokit details information about findings and outputs from Visitors and Residents project
24 March 2014
This infokit contains findings, outputs and video interviews of the co-principal investigators from two years of investigation of the collaborative longitudinal JISC-funded Visitors and Residents study.

OCLC Research identifies top ten alien abduction items in libraries in honor of Extraterrestrial Abduction Day
20 March 2014
This list of the top 10 most widely held items with the subject heading “alien abduction” was generated from the OCLC WorldCat database on 19 March based on the number of OCLC member libraries that hold the items. #oclc

More news »

Publications

Does Every Research Library Need a Digital Humanities Center?
by Jennifer Schaffner and Ricky Erway

Understanding the Collective Collection: Towards a System-wide Perspective on Library Print Collections

Follow Us

We Tweet

OCLC Research
Tweets from a list by OCLC

WebJunction
@WebJunction
2h
We’re looking for an intern! Please pass this along to any LIS students you know who would value the experience. webjunction.org/news/webjunc

Ricky Erway
@rickyerway
7m
Including 500 unknown file formats! @aigd: Dryad policy had to attempt to address 100s of different formats. #dap14

Ricky Erway
@rickyerway
18m
3:1 female to male ratio at #dap14. Theories? pic.twitter.com/2YHgwWdAS

Upcoming Events

Regional Print Management: Right-Scaling Solutions Symposium
27 March 2014 - 28 March 2014
Library managers, collection development officers, consortium administrators and others interested in shared print are invited and encouraged to attend this event co-sponsored by OCLC Research, the CIC and The Ohio State University Library, with support from OhioLINK. #regionalprint

Libraries and Research: Supporting Change/Changing Support
10 June 2014 - 12 June 2014
Exclusively for OCLC Research Library Partners, this gathering will focus on supporting the academic research process. Tweet: #forip
OCLC Research Mission

• To act as a *community resource* for shared Research and Development (R&D),
• To provide advanced development and technical support within OCLC itself, and
• To enhance OCLC's engagement with members and to mobilize the community around shared concerns.
Current Research Activities

- ArchiveGrid
- assignFAST
- Changes in Scholarly Communication
- Classify
- Cloud Library: Cloud-sourcing Shared Research Collections
- COBOAT
- Cookbook Finder
- Cyber Synergy: Seeking Sustainability through Collaboration between Virtual Reference and Social Q&A Sites
- Define Policy and Infrastructure Requirements for Building and Managing Shared Print Collections
- Demystifying Born Digital
- Digital Information Seeker Report
- Digital Visitors and Residents: What Motivates Engagement with the Digital Information Environment?
- Europeana Innovation Pilots
- FAST (Faceted Application of Subject Terminology)
- FAST Converter
- FictionFinder: A FRBR-based Prototype for Fiction in WorldCat
- info URI Registry
- Kindred Works
- Library Finder
- mapFAST
- mapFINDS Ohio
- MARC Usage in WorldCat
- Metadata Schema Transformation Services
- NACO Normalization Service
- Name Extraction
- OAICat
- OAICatMuseum 1.0
- OCLC Crosswalk Web Service Demo
- OCLC Linked Data Research
- OhioLINK Collection and Circulation Analysis
- PREMIS Maintenance Activity and Editorial Committee
- Preservation Health Check
- Print Management at "Mega-scale": a Regional Perspective on Print Book Collections in North America
- PURL
- **Registering Researchers in Authority Files**
- Right-scaling Stewardship: A System-wide Perspective on Print Books in the CIC
- Role of Libraries in Data Curation
- Rough and Ready: Typescript Finding Aid Conversion
- Scholars’ Contributions to VIAF
- Seeking Synchronicity: Evaluating Virtual Reference Services from User, Non-User and Librarian Perspectives
- SHARES Program
- Sharing and Aggregating Social Metadata
- Sharing Special Collections
- SRW/U
- Support for Research Workflows
- Survey of Special Collections and Archives in the UK and Ireland
- Survey of Special Collections and Archives in the US and Canada
- Terminology Services
- User-Centered Design of a Recommender System for a "Universal" Library Catalogue
- Virtual Research Environment (VRE) Study
- Work Records in WorldCat
- WorldCat Genres
- WorldCat Identities
- WorldCat Identities Network
- WorldCat Live!
Registering Researchers in Authority Files Task Group

How to make it easier for researchers and institutions to more accurately measure their scholarly output?

- Challenges to integrate author identification
- Approaches to reconcile data from multiple sources
- Models, workflows to register and maintain integrated researcher information
Registering Researchers in Authority Files

The Registering Researchers Task Group aims to create a concise report that summarizes the benefits and trade-offs of emerging approaches to the problem of incomplete national authority files.

Background

National name authority files are incomplete. Many researchers—tenured and non-tenured faculty and graduate students—are only partially represented in national name authority files. National name authority files cover poorly authors of journal articles and exclude researchers who do not publish but who create or contribute to data sets and other research activities. Uniquely identifying the academic authors of all publications, including journal articles, and researchers who do not publish, facilitates compiling individuals' scholarly output, especially as their affiliations change over time. The scholarly output is a factor in the reputation and ranking of the scholar's affiliated institution.

Impact

Our goal is to identify:

- the benefits, needs, and challenges for integrated author identification;
- approaches to effectively integrate multiple author identifier systems, and to reconcile information from multiple sources; and
- models, channels and workflows for registering and maintaining integrated author and researcher information.

The broader impact of this work will be to make it easier for researchers and institutions to more accurately measure their scholarly output.

Details

There are a number of emerging approaches to this issue, including authority databases, researcher identifier registries, and researcher profile networks. There is potential overlap across these approaches, and uncertainty in the library community regarding the challenges, benefits and trade-offs among each approach.
Preview

* Motivations *

* Landscape *

* Observations & Recommendations *
What’s the problem?

(more)
Then

Later

• By 1980, average number of authors in high-ranked medical journals was 4.5
• By 2000, average number of authors was 6.9

[Weeks et al. 2004]
Now
Now is More

Capturing Contributor Roles in Scholarly Publications
<table>
<thead>
<tr>
<th>Trend</th>
<th>Potential Authorship Issues</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase in number of coauthors</td>
<td>- ‘honorary’ authorship - ‘ghost’ authorship - disputes</td>
<td>- How to disambiguate author names? - How to communicate attribution in citation? - How to describe contributions to work? - How to evaluate and predict impact? - Who is responsible?</td>
</tr>
<tr>
<td>Shift from academic publishing in books to journals</td>
<td>- loss of sole-author-book as a evaluation measure</td>
<td>- How to integrate name authority and researcher identifier systems?</td>
</tr>
<tr>
<td>Decreasing granularity of publications</td>
<td>- persistence of “nano” publication vs. authorship</td>
<td>- How to document authorship over substructure of work?</td>
</tr>
<tr>
<td>Dynamic documents</td>
<td>- version misattribution</td>
<td>- How to document authorship over time?</td>
</tr>
<tr>
<td>Increasing diversity in citable scholarly outputs</td>
<td>- citation cannibalization, overcounting</td>
<td>- How to cite data, software, presentations(?) , blogs(?) , tweets(?)</td>
</tr>
</tbody>
</table>
Scholarly output impacts the reputation and ranking of the institution

We initially use *bibliometric analysis* to look at the top institutions, by publications and citation count for the past ten years…

Universities are ranked by several indicators of academic or research performance, including… highly cited researchers…

Citations… are the best understood and most widely accepted measure of research strength.
A scholar may be published under many forms of names

Also published as:
Avram Noam Chomsky
N. Chomsky

Works translated into 50 languages (WorldCat)

Noam Chomsky
Linguist

Journal articles
Same name, different people

Michael Conlon

Mike is Associate Director and Chief Operating Officer of the University of Florida Clinical and Translational Science Institute, and Director of Biomedical Informatics, UF College of Medicine. His responsibilities include development of academic biomedical informatics, expansion and integration of research and clinical information resources and strategic planning for academic health and university research. As principal investigator of the National Institutes of Health VIVO project, Mike led a team of 120 investigators at seven schools in the development, implementation and advancement of VIVO, an open source, semantic web application for research discovery. In his spare time, Mike enjoys his family and friends, Florida Gators football, and Python programming.

Michael Conlon QC is a practising barrister at Pump Court Tax Chambers, specialising in revenue law. He is President of the VAT Practitioners Group and of the Institute of Indirect Tax, of which he is also a Fellow. He is a Chartered Tax Adviser (Fellow) and a member of numerous other professional and learned bodies. He is a member of the Editorial Boards of Tax Journal and De Voi’s Indirect Tax Intelligence. He is a frequent writer and lecturer on taxation topics.


Human gut microbiota and future prebiotics

Author(s): Michael Conlon *, Anthony Bird and Claus Christoffersen
CSIRO Food Futures National Research Flagship, CSIRO Animal, Food and Health Sciences, Adelaide, SA 5000, Australia
Dr Michael Conlon, CSIRO Animal, Food and Health Sciences, PO Box 10041, Adelaide BC, SA 5000, Australia
The Landscape of Researcher Identification
One researcher may have many profiles or identifiers...

(from an email signature block)

Profiles: Academia / Google Scholar / ISNI / Mendeley / Microsoft Academic / ORCID / ResearcherID / ResearchGate / Scopus / Slideshare / VIAF / Worldcat
## Stakeholders & needs

<table>
<thead>
<tr>
<th>Role</th>
<th>Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher</td>
<td>Disseminate research</td>
</tr>
<tr>
<td></td>
<td>Compile all output</td>
</tr>
<tr>
<td></td>
<td>Find collaborators</td>
</tr>
<tr>
<td></td>
<td>Ensure network presence correct</td>
</tr>
<tr>
<td>Funder</td>
<td>Track research outputs for grants</td>
</tr>
<tr>
<td>University administrator</td>
<td>Collate intellectual output of their researchers</td>
</tr>
<tr>
<td>Journalist</td>
<td>Retrieve all output of a specific researcher</td>
</tr>
<tr>
<td>Librarian</td>
<td>Uniquely identify each author</td>
</tr>
<tr>
<td>Identity management system</td>
<td>Associate metadata, output to researcher</td>
</tr>
<tr>
<td></td>
<td>Disambiguate names</td>
</tr>
<tr>
<td></td>
<td>Link researcher's multiple identifiers</td>
</tr>
<tr>
<td></td>
<td>Disseminate identifiers</td>
</tr>
<tr>
<td>Aggregator (includes publishers)</td>
<td>Associate metadata, output to researcher</td>
</tr>
<tr>
<td></td>
<td>Collate intellectual output of each researcher</td>
</tr>
<tr>
<td></td>
<td>Disambiguate names</td>
</tr>
<tr>
<td></td>
<td>Link researcher's multiple identifiers</td>
</tr>
<tr>
<td></td>
<td>Track history of researcher's affiliations</td>
</tr>
<tr>
<td></td>
<td>Track &amp; communicate updates</td>
</tr>
</tbody>
</table>
Some functional requirements

Librarian as a stakeholder

- Create consistent and robust metadata
- Associate metadata for a researcher’s output with the correct identifier
- Disambiguate similar results
- Merge entities that represent the same researcher and split entities that represent different researchers
More functional requirements

Researcher & university administrator as a stakeholder

- Link multiple identifiers a researcher might have to collate output
- Associate metadata with a researcher’s identifier that resolves to the researcher’s intellectual output.
- Verify a researcher/work related to a researcher is represented
- Register a researcher who does not yet have a persistent identifier

Funder and university administrator as a stakeholder

- Link metadata for a researcher’s output to grant funder’s data
Systems profiled (20)

Authority hubs:
- Digital Author Identifier (DAI)
- Lattes Platform
- LC/NACO Authority File
- Names Project
- Open Researcher and Contributor ID (ORCID)
- ResearcherID
- Virtual International Authority File (VIAF)

Current Research Information System (CRIS): Symplectic

Identifier hub: International Standard Name Identifier

National research portal: National Academic Research and Collaborations Information System (NARCIS)
Systems profiled (20)


Reference management: Mendeley

Research & collaboration hub: nanoHUB

Researcher profile systems:
  - Community of Scholars
  - Google Scholar
  - LinkedIn
  - SciENcv
  - VIVO

Subject author identifier system:

Subject repository: arXiv
## Partial overview: Authority & identifier hubs

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Author Identifier</td>
<td>Researchers in all Dutch CRIS &amp; library catalogs</td>
<td>66K</td>
</tr>
<tr>
<td>Lattes Platform</td>
<td>Brazilian researchers and research institutions</td>
<td>2M people, 4K inst.</td>
</tr>
<tr>
<td>ISNI</td>
<td>Data from libraries, open source resource files, commercial aggregators, rights management organizations. Includes performers, artists, producers, publishers</td>
<td>7M total; 720 K researchers</td>
</tr>
<tr>
<td>LC/NACO Authority File</td>
<td>Persons, organizations, conferences, place names, works</td>
<td>9M total; ? researchers</td>
</tr>
<tr>
<td>ORCID</td>
<td>Individual researchers plus data from CrossRef/Scopus, institutions, publishers</td>
<td>500K</td>
</tr>
<tr>
<td>ResearcherID</td>
<td>Researchers in any field, in any country</td>
<td>250K</td>
</tr>
<tr>
<td>VIAF</td>
<td>Library authority files for persons, organizations, conferences, place names, works</td>
<td>26M people; ? researchers</td>
</tr>
</tbody>
</table>
Some overlaps
Observations
Where is Everyone?

Integrating Researcher Identifiers into University and Library Systems
Where are researchers?

Researchers

- DAI
- ISNI
- ORCID
- LC/NACO (?)
- VIAF (?)
- LinkedIn (?)
- Unlisted (?)

W.A.G.'s
# Changing Scholarly Landscape

## Books vs. Journals

<table>
<thead>
<tr>
<th></th>
<th>Books</th>
<th>Journals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read by</strong></td>
<td>All disciplines</td>
<td>Most Disciplines</td>
</tr>
<tr>
<td><strong>Tenure &amp; promotion driver</strong></td>
<td>Humanities, some Social Sciences (e.g. Political Science)</td>
<td>Some Humanities, Social Sciences, Life Sciences, Physical Sciences</td>
</tr>
<tr>
<td><strong>Robust system of name and subject</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Robust system of citation tracking</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Robust system of full-text indexing</strong></td>
<td>No</td>
<td>Yes, although fragmented</td>
</tr>
</tbody>
</table>
## Researcher Identifier ≠ Name Authorities

<table>
<thead>
<tr>
<th>(Example)</th>
<th>Traditional Name Authorities</th>
<th>Researcher Identifier Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Stakeholders</td>
<td>Research Libraries</td>
<td>Publishers, Researchers, Funders, Libraries</td>
</tr>
<tr>
<td>Internal standardization/integration</td>
<td>Standardized and well integrated within libraries</td>
<td>Fragmented. Some well-integrated communities of practice.</td>
</tr>
<tr>
<td>Organization</td>
<td>Primarily top-down, careful controlled entry from participating organizations</td>
<td>Varies: top down, bottom-up, middle out; often individual contributors</td>
</tr>
<tr>
<td>External integration</td>
<td>Very limited: High barriers to entry, few simple API’s</td>
<td>Varies, but more open. Some services offer simple open API’s; integration with web 2.0 protocols (e.g. OpenId)</td>
</tr>
<tr>
<td>Works Covered</td>
<td>Primarily Books &amp; other works traditionally catalogued by libraries</td>
<td>Journal articles; Grants; Datasets</td>
</tr>
<tr>
<td>People covered</td>
<td>Authors of works published in or about a single country*</td>
<td>Authors of research articles, fundees, members of research institutions – international;</td>
</tr>
<tr>
<td>Key record criterion</td>
<td>Complete record of work.</td>
<td>Persistent and unambiguous identifier for an individual contributor</td>
</tr>
</tbody>
</table>
Complex Environment

Types of Systems

- **Authority hubs**: providing a centralized location of records for multiple institutions
- **Current Research Information System (CRIS)**: stores and manages data about research conducted at an institution and integrates it with data from external sources:
- **Identifier hubs**: providing a centralized registry of identifiers
- **National research portal**: providing access to all research data stored in a nation’s network of repositories
- **Online encyclopedia**: a compendium of information divided into articles which includes references to the works by scholars
- **Reference management**: a system to help scholars organize their research, collaborate with others, and discover the latest research
- **Research and collaboration hub**: a centralized portal where scholars in a particular discipline can work together
- **Researcher profile systems**: networks that facilitate professional networking among scholars
- **Subject author identifier system**: a registration service to link scholars with the records about the works they have written
- **Subject repository**: a discipline-based centralized repository to facilitate scholarly exchanged in the fields covered

Roles

- **Systems overlap**
  - Google Scholar combines reference management, profiles, and ids
- **Systems can have both producers and consumers relationships with each other**
- **Institutional members/maintainers overlap systems, but do not necessarily coordinate**
- **How disputed information is resolved is often unclear**
How are differences in data models, provenance maintained?

How do corrections, annotations, and conflicting assertions on public profile presentation propagate back?

Overlap among members of group actor types?
Some Emerging trends:

- Widespread recognition that persistent identifiers for researchers are needed
- Registration services rather than authority files as a solution for researcher identification
- Interoperability between systems is increasing:
  - ISNI & VIAF interoperability.
  - ORCID and ISNI coordination
  - CRIS system integration with ORCID, ISNI, VIVO

- Adoption trends ...
Adoption Trends: Publishers -- Early Adopters

- Widespread integration of ORCIDs with manuscript subscription systems
- MacMillan integrating ISNIs in Digital Science family of systems
- Integration of ORCIDs with CrossRef platform for DOI indexing and interlinking services
Adoption Trends: Funders -- National Adoption & Beyond

- FCT, the Portuguese national funder, requires ORCIDS for their national evaluation system
- DAI, the Netherlands national funder, has created ISNIs for each researcher
- SNSF, the Swiss national funder, has created ISNIs for each researcher
- Wellcome Trust has integrated ORCIDs into grant submission and evaluation.
- NIH integrated ORCIDs into the inter-agency biosketch platform SciENcv.
- U.S. D.O.E. integrated ORCIDs into grant submission
Adoption trends:

Increasing number of Universities assigning identifiers to researchers

- Assigning ORCIDs to authors when submitting electronic dissertations in institutional repositories
- Pilot to automatically generate preliminary authority records from publisher files (Harvard U. press, one other)
- Assigning ISNI identifiers to their researchers.
- Assigning local identifiers to researchers who don’t have one.
- Using UUIDs (Universally Unique identifiers) to map to other identifiers like ORCID.
(But wait, there’s more! ...)

- New universities now systematically creating ORCID’s for their researchers:

  - *Boston University (medical school), Chalmers Institute of Technology, Hong Kong University, La Universidad de Zaragoza, National Taiwan University, Texas A&M, and University of Michigan* (in progress)

- *University efforts to integrate ORCID’s and software Infrastructure*
  - BU is integrating ORCID into the *Profiles* research profile platform (used by 30+ institutions)
  - *Cornell University,* is integrating ORCID into the VIVO open source researcher profiling system, used by over a hundred institutions
  - *University of Notre Dame,* is integrating ORCID identifiers into their Institutional Repository and into the *Hydra* open source IR tool.
  - *Purdue University,* is integrating ORCID identifiers into their data management platform, and the HUBzero open source platform for supporting research collaboration.
  - Texas A&M University is integrating ORCID identifiers into the open source Vireo electronic theses and dissertations (ETD) workflow
  - University of Colorado, is integrating ORCID identifiers into their Faculty information System (FIS).
  - University of Missouri, is integrating ORCID identifiers into DSpace.
Recommendations for Universities
Prepare to Engage

• Adoption of researcher identifiers has been rapid within scholarly publishing
• Funders see clear benefits, and are engaged
• **It is time for many universities to transition from watchful waiting to and engagement**

Starting to engage...

• Develop outreach and educational materials for researchers, stakeholders
• Future-proof systems:
  – Authors are not a string
  – Identifiers are multi-valued, with multiple authorities
• Demand more than PDF’s ...
  – Many publishers are already associating each article with:
    – Multi-valued author list
    – Identifiers – author, funder, institution
    – Contribution/COI statements
• Prepare for more complex measurement & reporting of usage
Choosing Identifiers

• Broad Researcher Identifiers: Compare ORCID & ISNI
  – National mandates
  – Capabilities
  – Usage patterns

• Retain traditional identifiers: VIAF, NACO
  – Well supported in library systems
  – Primarily describe authors of books and similar works

• Be aware of community identifiers for local integration (e.g. ArXiV)
Manage Risks

- Environment is evolving
  - Funder mandates and policies are incomplete
  - No dominant business model
  - Incomplete adoption, no single comprehensive data source
  - Integration between classic and new name authority is lacking

- Researchers ...
  - will not drive change alone;
  - are sensitive to who controls their profile, and how information can be “corrected”;

- Incentive mechanisms, well-timed nudges, setting norms with junior scholars, and establishing information feedback loops are critical.


• VIAF Presentations and Publications [http://oclc.org/research/activities/viaf.html](http://oclc.org/research/activities/viaf.html)

We welcome feedback on the draft!

http://www.oclc.org/research/activities/registering-researchers/progress.html