Author Accepted Manuscript. Published version available (with minor proof edits and corrections) as: Youtie, J., Carley, S., Porter, A.L., and Shapira, P. (2017), "Tracking researchers and their outputs: new insights from ORCIDs," *Scientometrics*. https://doi.org/10.1007/s11192-017-2473-0

Tracking researchers and their outputs: New insights from ORCIDs

Jan Youtie[a]

Stephen Carley[b]

Alan L. Porter[c]

Philip Shapira[d]

[a]Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta, Georgia 30308 USA, +1 404-894-6111 (phone), +1 404-894-8194 (fax), jan.youtie@innovate.gatech.edu (corresponding author)

[b]Search Technology, 6025 The Corners Pkwy Suite 202, Norcross, GA 30092, USA, +1 770-441-1457 (phone), +1 770-263-0802 (fax), stephen.carley@searchtech.com

[c] Search Technology, 6025 The Corners Pkwy Suite 202, Norcross, GA 30092, USA, +1 770-441-1457 (phone), +1 770-263-0802 (fax), aporter@searchtech.com; and School of Public Policy, Georgia Institute of Technology, 685 Cherry Street, Atlanta, GA 30332 – 0345 USA, +1 404-385-8577 (phone) +1 404-385-0504 (fax)

[d] Manchester Institute of Innovation Research, Alliance Manchester Business School, University of Manchester, Manchester M13 9PL, UK, +44 161 275 5921 (phone), +44 (0) 161 275 0923 (fax), pshapira@manchester.ac.uk; and School of Public Policy, Georgia Institute of Technology, 685 Cherry Street, Atlanta, GA 30332 - 0345 USA, +1 404-385-8577 (phone) +1 404-385-0504 (fax)

Abstract

The ability to identify scholarly authors is central to bibliometric analysis. Efforts to disambiguate author names using algorithms or national or societal registries become less effective with increases in the number of publications from China and other nations where shared and similar names are prevalent. This work analyzes the adoption and integration of an open source, cross-national identification system, the Open Researcher and Contributor ID system (ORCID), in Web of Science metadata. Results at the article level show greater adoption, to date, of the ORCID iD in Europe as compared with Asia and the US. Focusing analysis on individual highly cited researchers with the shared Chinese surname "Wang," results indicate wide scope for greater adoption of ORCID. The mechanisms for integrating ORCID iDs into articles also come into question in an analysis of co-authors of one particular highly cited researcher who have varying percentages of articles with ORCID iDs attached. These results suggest that systematic variations in adoption and integration of ORCID into publication metadata should be considered in any bibliometric analysis based on it.

Introduction

Author name disambiguation continues to be a thorny problem despite advances in algorithms and national and organizational programs to register scholars. The dramatic rise of scholarship from countries where similar names are frequent, such as China (Zhou and Leydesdorff, 2006) and South Korea, has contributed to difficulties in linking names with publications because of ambiguities and duplications occurring in anglicized versions of the names (Strotmann and Zhao, 2012). Aiming to address the problem of linking authors to their publications in this context, some publishers have developed unique author identifiers. One identification system that has been promoted for its open source and cross-national approach is ORCID, the Open Researcher and Contributor ID system. ORCID refers to the ORCID identifier as the "ORCID iD." ORCID has been in operation since October, 2012, and its adoption has the potential to give hope to bibliometric researchers (among others) who seek to conduct studies that require better connections between scholars and their publications.

The aim of this paper is to advance understanding of the usefulness of ORCID for bibliometric research. We use data from the Web of Science (WoS) Core Collection to understand where ORCID adoption is stronger and weaker. We could also have conducted the analyses with Scopus data (because ORCID also integrates with the Scopus Author ID) but we choose WoS for this analysis because of the ease of searching for the ORCID iD in the WoS search function. The analysis is conducted at two levels. First, we perform macro-level analyses, through searches of WoS aggregated to the country, organization, and journal levels. In this macro-level analysis, we assess the level of penetration of the ORCID iD, which is defined as the number of WoS publication records with at least one ORCID iD divided by the total number of

WoS publication records. Second, we focus on the thorny problem of name disambiguation of scholars with the common anglicized Chinese name of "Wang" by focusing on those who have received the "Highly Cited Researcher" designation based on Web of Science citations and on one of these Researchers who provided us with a verified list of publications. The results suggest that ORCID adoption is uneven at the country level: stronger in Europe and weaker in Asia, where the need for author identification is perhaps the greatest. This regional difference also filters down to the organization level, with research organizations in Europe generally having higher ORCID penetration at the article level than those in Asia or the US. Our review of highly cited researchers with the surname "Wang" found that most of these researchers did not have an ORCID iD. These results suggest that bibliometricians may use the ORCID identifier as one of many search tools, but should do so with care until its use has diffused more widely into the scholarly population, especially in Asia.

Background

There has been great interest in analyzing characteristics of individual researchers, such as their mobility across nations, institutions, topical areas, and time, using author data in publications. Being able to associate descriptive information for a set of research papers with their actual authors with high accuracy has proven challenging for relatively common names. Development of name disambiguation algorithms has progressed, but these have their own issues (for a review, see Smalheiser and Torvik, 2009). Machine learning technologies distinguish named entities and link them with certain word patterns in the text. Algorithms have been developed to relate individuals to citation patterns (Tang and Walsh, 2010, and US

Patent No. 8799237 for their work, licensed to Search Technology, Inc.). New developments in blocking schemes that put similar information (i.e., variations of a name) into the same category and calculate similarities and differences (Kelley, 1984) have made these processes more accurate and efficient (Mitra et al., 2005; Li et al., 2014). However, these automated approaches still have serious limitations in dealing with ethnic and geographic differences.

If the US had an official registry of unique author identifiers, that might preclude the need for these algorithms for US-based analyses. Such registries are used in some other countries, such as Brazil (Altman et al., 2014; http://lattes.cnpq.br). There are difficulties with creating such a registry, such as security, authentication, programming, the lack of standardization that individuals use when indicating their name in publications, and the attacks on freedom and liberty that such a registry implies in the context of cultural norms in the US (Garfield, 1969). Nevertheless, having a unique author identifier would greatly benefit research into the mobility of individual authors and their changing roles in networks, as careers unfold and following key events (e.g., overseas post-doc).

Amidst these author identification challenges has arisen the ORCID iD. Publicly launched in 2012, ORCID is "an open, non-profit, community-based effort to provide a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers" (*What is Orcid?,* 2016). ORCID allows individual researchers to create an ID. Use of ORCID iDs by authors is being promoted by publishers and associations including Crossref, Thomson Reuters, Elsevier, PLOS, Wiley, ACM, and Springer. Studies are beginning to appear that use ORCID data, for example, to track (albeit with recognition of limitations) the international movement and migration of scientists (Bohannon, 2017).

The literature on ORCID is oriented toward explaining the author identification problem, proposing ORCID as a possible solution, and describing how ORCID works. Many of the early articles were published by ORCID organizers. In 2011, an early article published in *Information Standards Quarterly* by a member of ORCID's board of directors and chair of its outreach group discussed the problem of author identification as stemming from fragmentation across multiple unconnected systems, including national level systems (such as Brazil's LATTES), field-level systems, and systems maintained by funding organizations (Fenner, 2011). ORCID is designed to extend beyond these boundaries through its open source design, but it requires coordination and management involving a large number of stakeholder organizations, consent of the scholar, trust in the system, and continuity for the length of time required to make the system work. Also in 2011, Fenner and colleagues published an article in *Serials* with other European members of the ORCID outreach group. The article described the history of ORCID, the initial size of its organizational membership as of 2010, its initial software platform based on the then-Thomson Reuters ResearcherID software, and its plans to accept individual scholar registrations in 2012. The cross-disciplinary and cross-national nature of ORCID are among the benefits mentioned, but the authors state that the value of ORCID would ultimately depend on attracting a "critical number" of users, including users claiming works at the time of manuscript submission, users with the ability to integrate their ORCID profile with other author identifying applications, and users actively identifying previously published works (Fenner et al., 2011). Laurel Haak, the Executive Director of ORCID, published an article with board members of the organization in *Learned Publishing* on the use case of ORCID for publishers (Haak et al., 2012a). The use case encompasses certification of authors, finding reviewers, tracking citations, and

integration with paper repositories such as Crossref. Another article published in 2012 in *Science* discussed standardization issues and efforts to address these in integrating ORCID with other publication databases (Haak et al., 2012b). Subsequent articles (Anstey, 2014; Thomas et al., 2015; Meadows, 2016) relate to the use of ORCID in automatically importing author information from multiple sources for medical researchers and facilitating the provision of library services in university and college settings. Butler (2012) offers, perhaps, the clearest justification for the need for ORCID by highlighting the rise of articles by Chinese (and Korean) authors and the difficulty of linking an article to common surnames such as Wang, with "Y. Wang" being the most common combination of first initial and last name at the time the article was written.

Not all views of identifier-based author disambiguation are universally supportive, however. Rosenkrantz de Lasson (2015) notes that ORCID (and other researcher identifier applications) require detailed active claiming of articles published prior to ORCID registration, which involves investing time into adding publications and keeping the list complete. In contrast, the author claims that Google Scholar automatically updates one's publication list (although our experience is that sometimes the author has to manually add or remove publications even in Google Scholar). The author can make corrections to the list and it can be manipulated, but nevertheless, the basic list is provided without requiring extensive time resources from the researcher. Likewise, a study at Texas A&M reported that the university pre-arranged for ORCID iDs for more than 10,000 graduate students, but only one-fifth of them actively claimed their IDs, suggesting that more is needed besides obtaining organizational participation and technically making these identifiers available (Clement, 2014).

In sum, these works highlight the promise of ORCID, as well as some of the challenges, but there have been few studies that examine its potential usefulness in bibliometric analysis of authorship patterns. This paper provides basic demographic information about the adoption of ORCID at the macro scale and with an individual example. At the macro scale, the paper addresses national differences in researcher adoption of ORCID and how such adoption has changed over time based on the presence of this information in WoS. It also examines how disciplinary fields, as proxied by journals, differ in the presence of ORCID iDs as well as changes in these differences over time. We also investigate organizational differences in ORCID adoption given that some organizations have announced mandating ORCID registration. At the individual scale, the paper focuses on a selected group of scholars with the last name "Wang" and trace their adoption of ORCID and other types of author identification. Foundational for this study is obtaining a basic understanding of the attributes of the individuals with ORCID iD's vis-à-vis those without, or put another way, what "sampling biases" would affect bibliometric analyses relying on ORCID iDs.

ORCID at the Macro Level

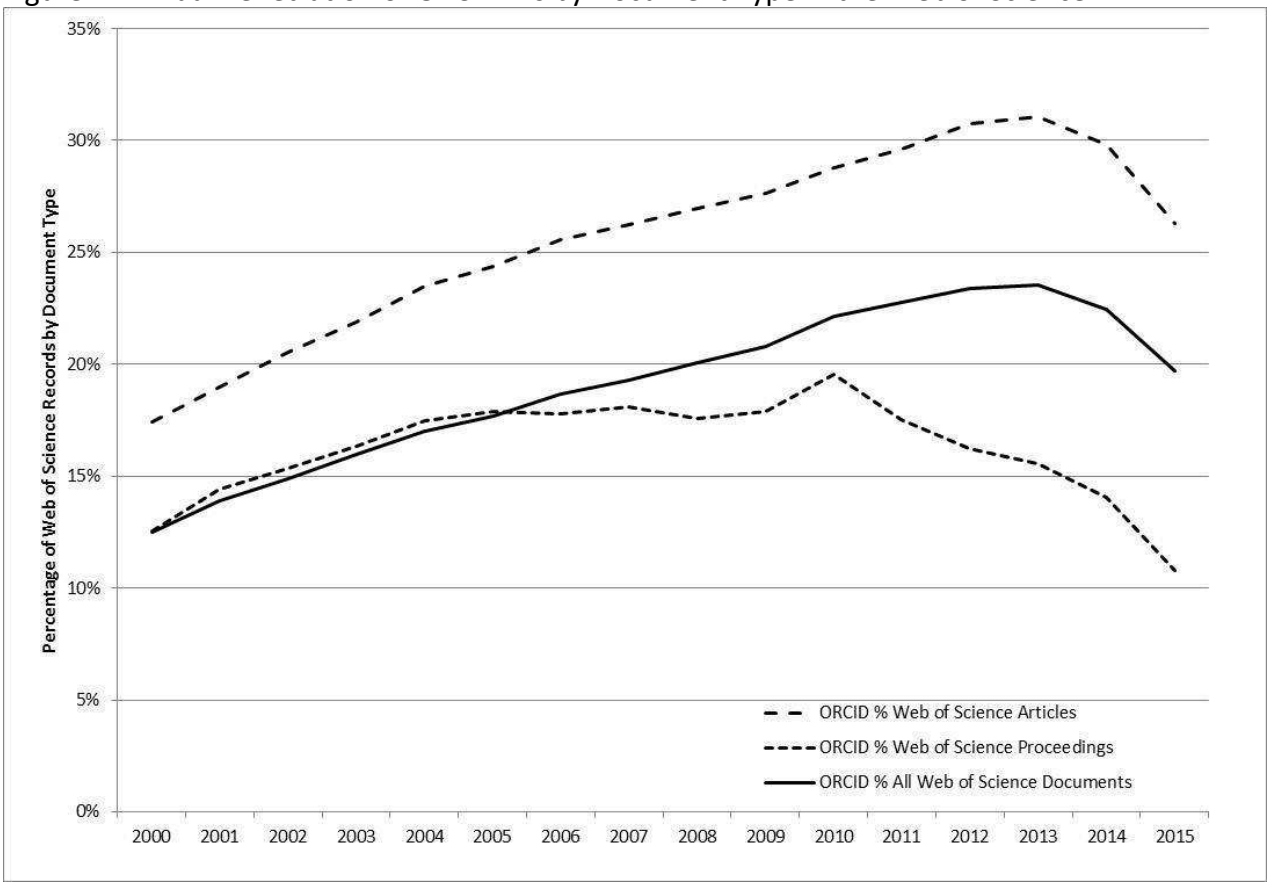
Adoption of ORCID has grown dramatically, from nearly 47,000 active ORCID iDs as of January 6, 2013 to more than 2.9 million as of January 6, 2017 (Number of ORCID iDs, 2017). This growth suggests that ORCID could be useful to bibliometric analyses. However, this growth is based on ORCID iD registrations; not all of these registrants are publishing researchers. The issue is the extent to which these registrations translate into publication metadata for bibliometric analysis and the growth trajectory of this integration of publications into metadata.

To understand adoption of the ORCID iD and integration into publication metadata, our analysis focuses on integration of the ORCID iD into WoS. The macro level analysis examines the extent to which any WoS publication record has at least one ORCID iD. To this end, we perform a simple WoS search using the wildcard search term "0000*" in the Author Identifier field. The weakness of this macro-level search is that it is at the publication record level and not the author level. Thus, we can readily discern the number of articles with *at least one* ORCID iD, but we cannot handily distinguish the penetration of ORCID iDs among the authors of multi-authored papers at the macro level. We will address this issue through micro-level explorations of an individual author's verified publication record, but for now, we take this macro-level method as providing an indicator of publication record adoption rather than author-level adoption.

The results of this macro-level search indicate that 19% of all WoS documents published in the 2000-2016 period have at least one associated ORCID iD. The penetration rate for all document types is at 12% in 2000, peaking at 24% in 2013, and declining to 20% in 2015 (with records likely incomplete for 2016) (Figure 1). If this rate is examined only for the set of *articles* in the WoS, penetration is 17% in 2000 and peaks at 31% in 2013 before declining to 20% in 2015. In contrast, penetration of ORCID iDs into proceedings papers is much smaller, peaking at just under 20% in 2010. This decline after 2013 is unexpected given the rapid growth of ORCID iD registrations. One explanation is that, prior to 2015, WoS obtained ORCID iDs from the WoS ResearcherID system, when an author signed up for ResearcherID and linked his or her ORCID iD to the ResearcherID. In November of 2015, WoS began obtaining ORCID iD information

directly from a database feed from ORCID. This source switchover could account for the drop-off in ORCID iDs reported in WoS after 2013.

Figure 1. Annual Penetration of ORCID iDs by Document Type in the Web of Science



Source: Web of Science, January 31, 2017 based on 5,129,893 articles, 802,632 proceedings, and 19,913,828 total document record counts in the Web of Science based on searches of the publication year, article and proceedings paper document types, using "0000*" to represent the ORCID iD

The first set of factors that could affect this appearance of the ORCID iD in WoS records potentially has to do with broader policy efforts of ORCID to encourage adoption at the national level. Table 1 suggests that country level differences exist based on the share of article records with ORCID iDs and that these differences have changed from 2013 to 2015. We examine the top 20 countries in article production in WoS, and define ORCID penetration as having at least one author based in this country. The caveat with this approach is that we are using full

Table 1. Top Article Publishing Countries and their ORCID iD Penetration in the Web of Science: 2013 and 2015

Country	Articles		With ORCID iDs	
	2013	2015	2013	2015
USA	394,003	414,897	29.4%	25.2%
China	218,440	284,643	21.0%	17.0%
UK	115,769	124,423	45.4%	42.1%
Germany	102,228	109,147	42.4%	38.8%
Japan	78,663	78,141	26.9%	24.8%
France	70,935	75,118	43.2%	38.6%
Canada	64,342	69,971	31.8%	27.0%
Italy	61,228	68,177	68.7%	61.6%
Spain	55,439	61,873	73.2%	60.7%
Australia	54,765	65,637	54.4%	47.0%
India	53,642	68,860	25.6%	19.9%
South Korea	51,528	58,011	24.6%	20.9%
Brazil	39,170	47,338	51.1%	40.5%
Netherlands	37,482	39,795	48.6%	42.9%
Russia	29,539	37,678	42.7%	38.7%
Taiwan	28,036	26,686	27.9%	25.1%
Switzerland	27,259	29,882	50.0%	46.7%
Turkey	26,581	31,451	25.0%	21.6%
Iran	26,087	31,723	30.5%	24.6%
Sweden	24,794	27,504	53.6%	47.9%
All Articles	1,484,889	1,647,765	31.0%	26.3%

Source: Total Web of Science article records: 1,484,889 in 2013 and 1,647,765 in 2015. Web of Science article records with ORCID iDs: 460,889 in 2013; 432,992 in 2015 based on searches of the publication year and "0000*" to represent at least one ORCID iD in an article record, performed on January 31, 2017.

counting in case of multiple authors from different countries who each have their own ORCID iD. Within this caveat, seven countries/territories have less than 30% of articles with authors based at organizations in these countries in 2013 associated with at least one ORCID iD: the USA, China, Japan, India, South Korea, Taiwan, and Turkey. These percentages are slightly lower in 2015. In contrast, five countries have ORCID iDs in 50% of more of the WoS indexed articles

whose authors are based there: Italy, Spain, Australia, Brazil, and Switzerland. It is interesting that the low ORCID iD group, with the exception of the USA, is comprised of Asian countries, while three of the five countries in the high ORCID iD penetration group are European.

Country differences in ORCID adoption may well stem from the ORCID organization's membership fee-based agreements with institutions and nations. The Portuguese national funding agency for science, research and technology (Fundação para a Ciência e a Tecnologia) required in 2014 that its researchers register for an ORCID iD and connect through Scopus to be eligible to apply for research grants. This requirement led to 14,000 scholars registering for ORCID iDs in the three weeks after the announcement. In 2015, Italy, through its Ministry of Education, similarly required researchers to register for an ORCID iD and connect through Scopus. Nearly, 60,000 researchers registered for IDs in the two months after the announcement, such that 80% of Italian researchers have so registered. Other membership increases have resulted from consortia initiatives. Six of Denmark's universities and a library consortium adopted ORCID in 2014 to enable researcher tracking across institutions. Taiwan also adopted this type of consortium platform in 2016. Pilot projects have been carried out in the UK, Australia, and Germany. As of 2016, the UK consortium had nearly 80 member universities and seven research councils, the latter of which have integrated the ORCID iD into their grant application systems. Australia's consortium over the same time period has 40 universities and two research funding agencies. There also are platforms in China, through the Chinese Academy of Sciences, and in the US, through regional consortia of universities and libraries, although neither of these consortia is explicitly reinforced by a national research funding agency (Table 2).

Table 2. Chronology of ORCID Institutional and Consortia Adoption (Selected)

Year	ORCID Adoption
2013	• ORCID added national-level membership fee model

2014	• Portugal's Fundação para a Ciência e a Tecnologia required ORCID researcher registration
	• Denmark consortium of universities and libraries encouraged ORCID registration and integration with its research monitoring system
	• The Sloan Foundation funded projects at seven US universities, a Canadian-led research network, and a US-based society to develop tools to integrate ORCID into existing systems
	• Organizations requiring ORCID include:
	  ○ Autism Speaks (US)
	  ○ Qatar National Research Fund
	  ○ Swedish Research Council

2015	• Australia consortium of 40 universities, National Health and Medical Research Council, Australian Research Council encourages ORCID researcher registration
	• UK consortium of 77 universities and seven research councils requiring ORCID researcher registration
	• Organizations requiring ORCID
	  ○ National Institute for Health Research, UK
	  ○ Wellcome Trust, UK
	  ○ National Research Foundation, South Africa

2016	• Center for Science Ltd – IT Center, Finland encourages ORCID researcher registration
	• German Research Foundation (DFG), Helmholz, and two libraries encourage ORCID researcher registration
	• New Zealand ORCID Consortium (ministries and funding organizations) encourages ORCID researcher registration
	• Three US regional library associations (Greater Western Library Alliance, NorthEast Research Libraries, Big Ten Library Alliance) for access to ORCID information
	• Organizations requiring ORCID
	  ○ Austrian Science Fund (FWF)
	  ○ Department of Transportation, US (ORCID required for publications)

Source: Compiled from https://orcid.org/ and interview with ORCID executive director, October 31, 2016

Differences between distributions of WoS article counts overall and those with ORCID iDs are less apparent at the organizational level than the national level (Table 3). The focus of this analysis is on the WoS "organization-enhanced" field, specifically nineteen of the top-ranked organizations by WoS article counts in 2013, and the penetration of ORCID iDs in articles

Table 3. Top Article Publishing Organizations and their ORCID iD Penetration in the Web of Science: 2013 and 2015

Organization-Enhanced	Articles		With ORCID iDS	
	2013	2015	2013	2015
University of California System	37,266	38,287	38.1%	33.3%
Chinese Academy of Sciences	31,353	37,125	31.7%	27.2%
French National Center for Scientific Research (CNRS)	29,675	31,209	51.4%	46.6%
University of London	19,609	21,304	49.9%	45.5%
Harvard University	19,471	20,311	40.7%	34.1%
Russian Academy of Sciences	16,409	19,032	45.3%	41.8%
US Department of Energy	13,105	12,943	56.6%	54.1%
Pennsylvania Commonwealth System of Higher Education	12,739	13,085	38.3%	33.7%
State University System of Florida	12,431	13,541	32.3%	27.5%
University of Toronto	10,629	11,312	37.3%	31.5%
University of North Carolina	10,404	10,843	32.2%	26.2%
Max Planck Society	10,313	10,217	62.5%	58.4%
National Research Council of Spain (CSIC)	9,590	9,389	87.0%	79.4%
University of Oxford	9,056	9,757	52.6%	50.1%
University of Michigan System	8,918	9,373	39.4%	35.3%
University College London	8,878	9,911	58.2%	52.0%
French National Institute of Health and Medical Research (Inserm)	8,726	9,067	52.6%	47.7%
University of São Paulo	8,285	9,197	65.9%	58.5%
US National Institutes of Health (NIH)	8,214	7,735	48.1%	42.5%
All Articles	1,484,889	1,647,765	31.0%	26.3%

Source: See Table 1. Organizations-enhanced reported.

which have authors in these organizations. The lowest penetration rate was in articles authored by scholars at the Chinese Academy of Sciences (just below 32% in 2013), followed by the University of North Carolina and the State University System of Florida (both just above 32% in 2013). Three organizations had 60% or more of their articles associated with an ORCID iD: Max Planck Society, Consejo Superior de Investigaciones Cientificas (CISC), and Universidade de Sao Paulo. Indeed nearly all of the articles with a CISC author (87% in 2013, 79% in 2015) are

associated with an ORCID iD. These organizations likely adopted explicit incentives or requirements for their researchers to obtain ORCID registration and link this identifier to their publication records.

Second, there are increasing efforts at the publisher level to incentivize use of ORCID. ORCID announced that several publishers were requiring that authors have ORCID iDs. It was further noted that 3,000 journals collect ORCID iDs from authors during the manuscript submission process (Haak, 2016). At the time of writing of this paper, we are unable to capture the effects of the publishers requiring ORCID iDs in 2016, although we are able to observe associations through the manuscript submission process based on an analysis of journals with the largest number of WoS articles (Table 4). The top 20 article publishing journals in 2013, with articles also shown for these journals in 2015, reflect the WoS orientation toward physics and chemistry. Still, within the journals in these fields, we are able to discern differences. The highest percentage of articles with ORCID iDs in 2013 are in *Physical Review Letters* (60%), *Physical Review B* (59%), and *Journal of the American Chemical Society* (56%). The former two are American Physical Society (APS) publications and the latter is an American Chemical Society (ACS) publication. The lowest percentages are in *Journal of Alloys and Compounds* (33%), *Physical Review D* (39%), and *Journal of Biological Chemistry* (just under 40%). The first is an Elsevier publication, the second also an APS publication, and the third a publication of the American Society for Biochemistry and Molecular Biology (ASBMB). Although ASBMB's ORCID requirements are unknown, APS, ACS, and Elsevier collect ORCID iDs during the manuscript submission process. There does not seem to be a specific pattern as far as publisher policies

and ORCID iD penetration as yet at the journal level. To reiterate, penetration means "at least one ORCID iD" for a given paper; it does not mean the paper has an ID for each of its authors.

Table 4. Top Article Publishing Journals and their ORCID iD Penetration in the Web of Science: 2013 and 2015

Source Title	Articles		With ORCID iD	
	2013	2015	2013	2015
PLOS ONE	31,233	27,871	45.2%	34.1%
Applied Physics Letters	5,363	3,437	50.9%	51.9%
Physical Review B	4,774	4,892	59.3%	53.7%
Journal of Applied Physics	4,342	3,267	43.1%	46.1%
Proceedings of the National Academy of Sciences of the United States of America	3,902	3,281	54.0%	47.9%
Physical Review Letters	3,557	2,502	60.9%	56.6%
Optics Express	3,287	3,321	43.8%	31.2%
Physical Review D	3,230	3,372	38.8%	31.3%
Journal of Biological Chemistry	3,218	2,463	39.8%	42.9%
RSC Advances	3,145	12,591	42.7%	29.5%
Chemical Communications	3,135	3,837	52.4%	46.1%
Journal of Physical Chemistry C	3,113	3,257	56.6%	50.6%
Astrophysical Journal	2,889	3,006	48.7%	47.8%
Journal of the American Chemical Society	2,840	2,379	56.7%	53.5%
Physical Review A	2,786	2,545	47.3%	40.4%
Journal of Chemical Physics	2,722	2,463	46.0%	53.9%
Monthly Notices of the Royal Astronomical Society	2,686	3,096	50.1%	46.2%
Physical Review E	2,502	2,467	43.1%	38.2%
Scientific Reports	2,484	10,643	56.8%	42.2%
Journal of Alloys and Compounds	2,318	3,321	33.4%	28.1%
All Articles	1,484,889	1,647,765	31.0%	26.3%

Source: See Table 1. Source title reported.

Third, these national, institutional and journal level differences in policy-level factors in adoption of ORCID iDs are further magnified by the process for acquiring an ORCID iD at the individual researcher level. For an individual researcher to obtain an ORCID iD, the process involves two activities: (1) registering for the ID, and (2) linking the ID to publications. Obtaining

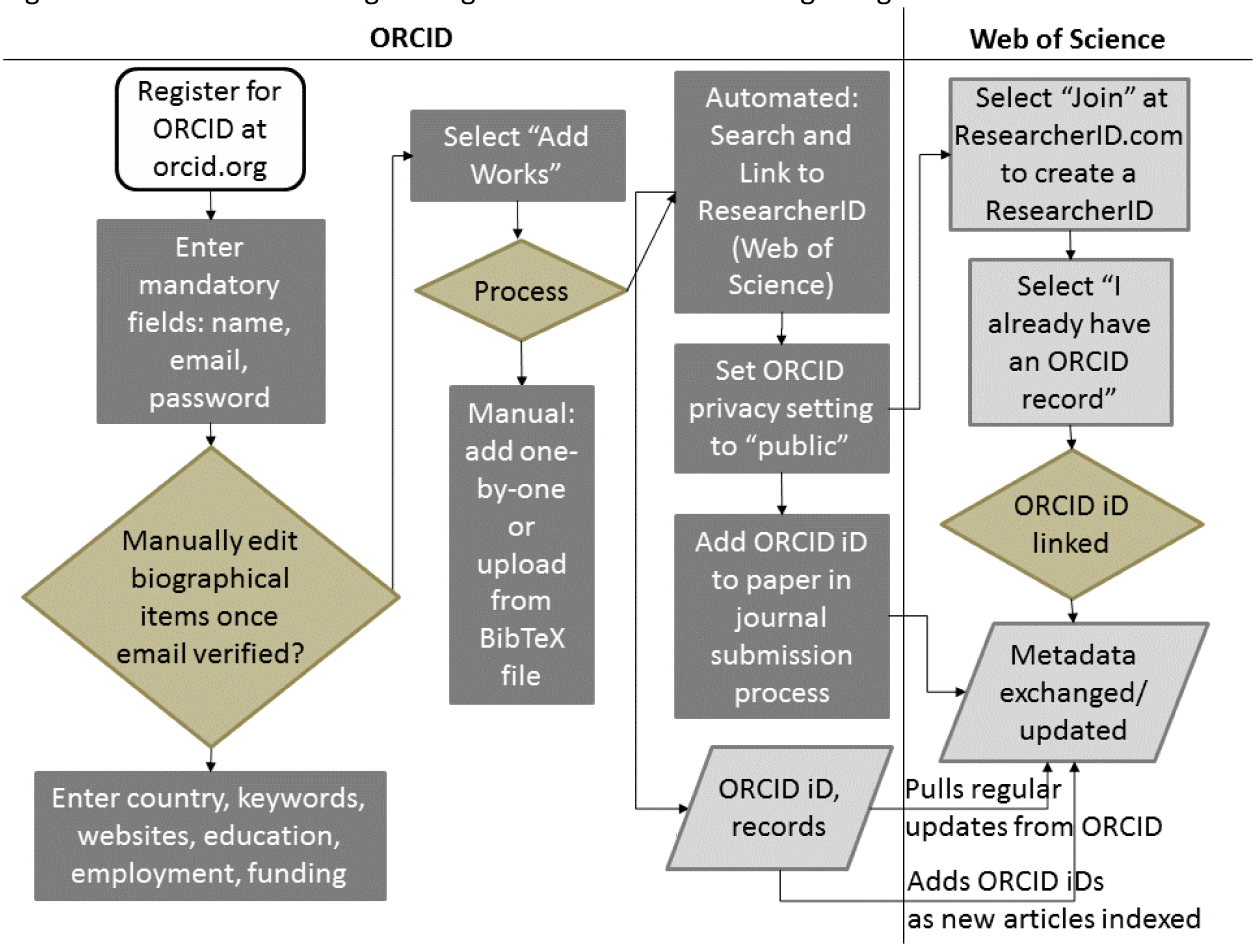
an ORCID requires minimal information: a name (first and last, as well as name in publications and other names the scholar goes by or has gone by in the past), email, and password. The initial registration is carried out by the researcher. The scholar can also add personal information such as country, keywords, and websites; education and employment information, which may be listed in drop-down menus; funding information, which can be added using a software wizard; and scholarly works. Scholarly works can be added manually one at a time, imported from a BibTeX file, or perhaps most commonly obtained through linkages with indexes such as the Web of Science (through its ResearcherID author identifier code), Scopus (through its Scopus Author ID author identifier code), Crossref Metadata Search, and national and regional online library tools where available (Figure 2). Despite these processes, duplicate IDs exist, although ORCID uses methods to prevent them through the registration process and through encouraging scholars to merge duplicate records.

There are multiple paths to connect a researcher's ORCID iD and publication record in WoS following ORCID registration. The most common route is to submit the ORCID iD to the journal along with a manuscript. A second route allows the scholar to go manually to WoS or Scopus (or other indexes), obtain ResearcherIDs or Scopus Author IDs, confirm any publications associated with the scholar's name, and authorize Scopus or WoS access to the ORCID account In either case, ORCID registrants may not realize that after they obtain their ORCID iD, they then have to enter it as part of the journal submission process or associate it with their WoS or Scopus IDs. The ORCID registrant may, instead of selecting WoS or Scopus, allow automatic association with Crossref; if this option is selected, then only publications in Crossref will be indexed in WoS (or Scopus). Any of these manual associations with publication indexes also

allow for links to the ORCID iD through the paper submission process. Assuming the process is followed in such as manner as to enable attachment of the ORCID iD to the researcher, the next step involves transmission of the ORCID information into the Web of Science. This ORCID iD information would then be integrated into the WoS metadata and attached to the author's publications (Gulpers, 2016). WoS has future plans to simplify these steps.

A third route involves WoS pulling regular updates from ORCID to update WoS metadata records. A fourth route entails WoS including ORCID iDs as it indexes new articles. Thus, there are varied ways, some automated and some manual, for ORCID iDs to get into a WoS record.

Figure 2. Process Flow of Registering for an ORCID iD and Integrating with Web of Science*



*ORCID used to allow registration by "trusted parties" but that is no longer the case.
Source: ResearcherID & ORCID Integration, http://wokinfo.com/researcherid/integration/

ORCID at the Micro Level

This analysis focuses on a set of individual researchers with the common last name of Wang who are designated as "Highly Cited Researchers" based on a citation analysis of the Web of Science (http://hcr.stateofinnovation.com/). We focus on those highly cited researchers with the last name "Wang," under the assumption that an identification number would be most useful if it could identify the most recognized researchers with a very common anglicized Chinese name that is difficult to disentangle using name disambiguation algorithms. We

searched the Highly Cited Researcher website for all scholars with the last name Wang in November, 2016. Doubtless these lists change and new information appears over time, but this is the time period we selected. The list at that time included 20 entries (Table 5).

Table 5. ResearcherID and ORCID iDs among Highly Cited Researchers with the Last Name of Wang

First Name	Last Name	Organization	ResearcherID	ORCID iD
Erkang	Wang	Chinese Acad Sci, China	Not available	Not available
Guiling	Wang	Harbin Engn Univ, China	Not available	Not available
Haijiang	Wang	Natl Res Council Canada, Canada	Not available	Not available
Hailiang	Wang	Yale Univ, USA	Not available	Not available
Jian	Wang	BGI Shenzhen, China	Not available	Not available
JinRong	Wang	Guizhou Univ, China	Not available	Not available
Joseph	Wang	Univ Calif San Diego, USA	Available	Not available
Jun	Wang	BGI Shenzhen, China	Available	Available
Jun	Wang		Unknown	Unknown
Meng	Wang	Hefei Univ Technol, China	Not available	Not available
Nan-Lin	Wang	Peking Univ, China	Not available	Not available
Shaobin	Wang	Curtin Univ, Australia	Not available	Not available
Thomas J	Wang	Vanderbilt Univ, USA	Not available	Not available
X L	Wang	Zhejiang Univ, China	Not available	Not available
Xiangke	Wang	King Abdulaziz Univ, Saudi Arabia; N China Elect Power Univ, China	Available	Available
Xinchen	Wang	King Abdulaziz Univ, Saudi Arabia; Fuzhou Univ, China	Not available	Not available
Yanli	Wang	Natl Inst Hlth (NIH), USA	Available	Not available
Yi-Hong	Wang	Univ Texas MD Anderson Canc Ctr, USA	Not available	Not available
Zhong Lin	Wang	Georgia Inst Technol, USA	Available	Not available
Zidong	Wang	Brunel Univ, UK	Available	Available

Source: http://hcr.stateofinnovation.com/, accessed November 17, 2016.

The list presents these authors with their first and last name and the name of their current organization. If we wanted to confirm these authors' publications, we might find ORCID or another identifier, such as ResearcherID, given that we are working with a list drawn from WoS. Of these 20 scholars, we were able to locate ResearcherIDs for only six of the scholars (by searching for their ResearcherIDs), and able to locate ORCID iDs for only three of them (by searching in the ORCID registry). The list of Highly Cited Researchers includes a name disambiguation problem; "Jun Wang" is associated with two entries. Luckily one of the two is linked to an organization, "BGI Shenzhen, China." We can go into the author's ORCID iD and obtain more information to be able to link the author to his publication record. A web searche confirms that Dr. Jun Wang is the founder of Beijing Genomics Institute (now BGI). When we go to ORCID and enter the name "Jun Wang," ORCID produces 193 entries of scholars with that name. No public information is available for 131 of these 193 "Jun Wang" entries. However, the remainder do have employment, educational, and other information, especially ORCID iDs, as well as ResearcherIDs and Scopus Author IDs. ORCID specifies three entries associated with the name "Jun Wang" at "BGI Shenzhen, China." Each of these three entries has a different ResearcherID and a different ORCID iD. The ResearcherIDs associated with "Jun Wang" at BGI Shenzhen, China in the ORCID repository as of November 17, 2016 are A-7261-2013, B-9503-2016, C-8434-2016; the ORCID iDs are 0000-0002-2113-5874, 0000-0002-8540-8931, 0000-0002-1422-3331. This situation (which could be a duplication of the same researcher, reference to more than one researchers with the same name at the same institution, or both) makes it difficult to confirm the link between publication and author using ORCID, even for a highly cited researcher.

To further test the ability to link publications with one of the highly cited authors on the list in Table 6, we conducted interviews in December 2016 and January 2017 with Professor Zhong Lin (ZL) Wang at the campus at Georgia Institute of Technology. In the interviews, we asked ZL Wang to carefully review the list of articles that we obtained from producing a simple WoS search using ZL Wang's first name, last name, and organization. We compared this verified list (search #1) with a basic search using last name and initials (search #2), a full name search with variations of spellings of ZL Wang's first and middle names (search #3), the above search plus the organization (search #4), a search of records based on ZL Wang's ResearcherID (search #5), and a search based on his ORCID iD (search #6). ZL Wang has been affiliated with Georgia Institute of Technology since 1995 (and a member of the Chinese Academy of Sciences since 2009), so using his name plus his organizational affiliation makes sense as a search strategy. ZL Wang did not have an ORCID iD at the time we conducted an initial interview with him, but he subsequently registered for an ORCID iD and connected it with his WoS publication record before the second interview. Table 6 reports the counts of articles that result from these searches (shown in the column labeled "WOS Search Input"), the number of ZL Wang's verified publications that are "missing" from the counts, and the number of false positives. False positives are articles with an author named ZL Wang, but these articles are not on ZL Wang's verified list of publications and therefore are likely written by someone with a similar name.

Table 6. Comparison of Zhong Lin Wang's list of publications from a simple search, ResearcherID, and verification

WoS articles, 2009-2016	Count	Missing	False Positives	WOS Search Input
Confirmed by ZL Wang	704	0	0	NA
Simple search for ZL Wang's name	4,614	1	3,894	AUTHOR: (Wang, ZL) AND YEAR PUBLISHED: (2009-2016)
More complex search for ZL Wang's name	836	13	145	AU=(Wang, Zhong Lin OR Wang, Zhong-Lin OR Wang, Zhong L OR Wang, Zhonglin OR Wang, "Zhong Lin (Z L") AND PY=(2009-2016)
More complex search for ZL Wang's name along with an Organization Enhanced Search for Georgia Tech	658	46	0	AU=(Wang, Zhong Lin OR Wang, Zhong-Lin OR Wang, Zhong L OR Wang, Zhonglin OR Wang, "Zhong Lin (Z L") AND PY=(2009-2016) AND [Organization Enhanced Search for Georgia Tech*]
Using ZL Wang's ResearcherID	608	101	0	AUTHOR IDENTIFIERS: (E-2176-2011) AND YEAR PUBLISHED: (2009-2016)
Using ZL Wang's ORCID iD	615	96	0	AUTHOR IDENTIFIERS: (0000-0002-5530-0380) AND YEAR PUBLISHED: (2009-2016)

Source: Author search of Web of Science on February 9, 2017. Confirmation took place in December, 2016 and January, 2017.

The results of these searches show that the simple search on ZL Wang's name yields too many false positives; nearly 4,000 false positives using a simple name search and 145 false positives using a more complex search of the first name. Adding the organization to the full-name-plus-variations search returned more than 93% of the verified articles with no false positives. The ORCID iD and ResearcherID searches also produced no false positives, while returning 87% and 86% of ZL Wang's articles respectively. These results suggest that the ORCID iD is useful as an author search strategy, offering excellent precision, i.e., capturing truly relevant records with limited noise or unrelated records, and reasonable recall, i.e., finding and capturing the largest number of relevant records. However, in the case of ZL Wang, a search based on variations of his full name and organization offers even better recall at the same level

of precision. We note that there is no other researcher with the name of ZL Wang at Georgia Tech, hence the organization specification becomes a useful qualifier in this case.

To understand the penetration of ORCID iDs in multi-authored articles, we examined the most frequent co-authors of ZL Wang who had registered for ORCID iDs. Of the 69 scholars who have co-authored more than 10 articles with ZL Wang in the 2009 to 2016 time period, one-third (22 scholars) had ORCID iDs. Exploring the extent of multi-author coverage in WoS article metadata, we total the amount of each of these co-authors' articles with ZL Wang that have an ORCID iD and divide this number by the total number of co-authored articles over the 2009 to 2016 time period. ORCID iD coverage for these 22 co- authors ranged from 50% to 100% (Table 7). Six of these authors had all of their ORCID iDs (i.e., 100% coverage) listed alongside ZL Wang's ID in these articles. Five had 90%-99% coverage of their ORCID iDs in these co-authored articles while four had fewer than 70% of their ORCID iDs listed. This result suggests that in the case of ZL Wang, who pays attention to article tracking (according to our interviews), most of his co-authors also have their ORCID iDs listed in his articles, although a few co-authors have articles with him that are missing their ORCID iDs. We are unsure of the reasons underlying the variable appearance of co-author ID in the metadata. However, these overall higher adoption percentages seem to suggest a relationship between author identifier adoption and the extent to which senior co-authors make an effort to keep an accurate record of their research articles.

Table 7. Most frequent ZL Wang Co-authors with ORCID iDs: 2009-2016

ZL Wang co-authors with ORCID iDs		Co-authored articles including co-author's ORCID iD	
Name	Co-Authored articles with ZL Wang	Number	Percent
Pan, Cao Feng	51	47	92.2%
Niu, Simiao	49	44	89.8%
Lin, Long	61	41	67.2%
Wu, Wenzhuo	40	39	97.5%
Jing, Qingshen	38	36	94.7%
Zhou, Yusheng	45	35	77.8%
Lin, Zong-Hong	33	29	87.9%
Chen, Jun	50	25	50.0%
Liu, Ying	37	24	64.9%
Wen, Xiaonan	29	23	79.3%
Zhou, Jun	23	23	100.0%
Fan, Fengru	23	19	82.6%
Song, Jinhui	19	19	100.0%
Xie, Yannan	18	18	100.0%
Xu, Sheng	17	17	100.0%
Zi, Yunlong	20	16	80.0%
Wen, Zhen	19	13	68.4%
Yang, Qing	14	13	92.9%
Hong, Jung-Il	12	12	100.0%
Yang, Rusen	14	12	85.7%
Dong, Lin	12	11	91.7%
Hu, Bin	11	11	100.0%

Source: As for Table 6 (Verified list of ZL Wang's articles).

Conclusions

The ORCID iD offers great promise for tracking researchers across national, organizational, and publication boundaries. The open source construction of the registration application along with outreach efforts has led to a substantial growth of researcher registration for the ORCID iD. This growth makes using the ID appealing for bibliometric analyses that require a good method for author tracking, yet little is known about where ORCID iD take-up has been more or less prevalent. If ORCID iD usage is non-random, then these

systematic variations need to be identified and assessed so that they can be accounted for in bibliometric analyses. This study has contributed to this task by highlighting some of the systematic differences in ORCID iD usage.

The study has conducted analyses at article (or macro-level) and at the author (or micro-level). These two levels were examined to capture both types of variation in ORCID usage. We observed that article-level penetration likely overstates ORCID usage, because only one of the authors need to have an ORCID iD for the article to count as having an ORCID iD. Nonetheless, this is a convenient method for examining broad systematic patterns at the country, organization, and journal level, while micro-level analysis is useful to understand further variations in ORCID adoption around a particular author's research network.

We acknowledge that the term "ORCID usage" may be misleading at the author level. Some authors may feel that they have registered for ORCID and thus their identifier should automatically integrate their papers in publication indexes such as the WoS so long as they use their ID when they publish new papers. The process is not straightforward for registering for the ORCID iD and then having it continuously integrated with indexed publications. Moreover, even the manuscript submission process, in our experience, sometimes encourages usage of the ORCID iD, albeit our results did not uncover systematic differences in take-up of the ORCID iD at the level of WoS journal articles. However, the process does not always allow or require the entry of multiple ORCID iDs for each of the co-authors, although multiple ID capture by journal publishers may become easier in the future. We speculate that these registration and integration factors may play a part in why ORCID penetration of articles in WoS peaked in the 2012-2013 period and has subsequently declined. However, a larger factor in this unexpected

trend may be the switch in 2015 from obtaining ORCID iD information for WoS from the ResearcherID record (if authors had added their ORCID iD to this record) to obtaining the information directly from a data feed from ORCID. Still, it is difficult to disentangle the effects of multiple possible causes. These include whether there are issues with journal requests for the ID from authors, whether the researcher is providing the ID, whether the journal passes the ID along in metadata to the indexes and this properly updates the author's ORCID record (Haak et al., 2012b), or whether the data feed from ORCID to WoS has gaps.

Issues in the integration of the ORCID iD into publication indexes such as WoS and Scopus) constitutes an important topic for future research. Such research should take into account two elements. The first would examine the extent to which individual ORCID iD registrants link to their WoS (or Scopus) publication record. The second would focus on the integration of ORCID iD information with WoS (or Scopus) metadata. Such research would go far in explaining the anomalies in the presence or absence of ORCID in these indexes.

Notwithstanding these usage factors and anomalies, we do observe systematic differences at the country and (filtering down to) the organization level. European countries, especially among those in the top publishing frequency segment such as Italy and Spain, have very high penetration of ORCID associated with their article records, while penetration is much lower in Asian countries and even the US. Yet these Asian countries with fast growing publication activity, such as China and Korea, pose the greatest need for ORCID usage because of difficulties in using name disambiguation algorithms on anglicized versions of their names. We would expect the most highly cited researchers from these Asian countries at least to have ORCID iDs, but that was not the case for the set of researchers with the last name of "Wang" in

the Highly Cited Researcher listing based on WoS citations. It would seem that if these types of prominent researchers could be encouraged to register for ORCID iDs and enable integration with their indexed articles, then their network of co-authors might follow. This relationship between leading researchers and their co-author networks in adoption of ORCID constitutes another topic for future research. In any case, it could signal a new round of efforts to encourage adoption of ORCID outside the Western research enterprise. Indeed, issues in the adoption of ORCID in Asian countries and/or organizations represent another promising area of future research related to ORCID usage.

In response to these differences, we recommend that researchers register for the ORCID iD, use it when they submit a manuscript, and grant permission for record updates. We recommend that organizations encourage this effort by not solely relying on technical solutions, as Clement (2014) notes. Organizations can assist with the process through providing marketing and training into how to register for the ORCID iD and use it in article submissions and updating.

At present, we recommend to bibliometricians and others seeking to track author publication (or award) patterns that careful use of ORCID is warranted as an aid alongside other tracking methods. Gaps in coverage warn against relying solely on ORCID iD to collect researcher data. Even as a handy way to sample, the ORCID iD varies in coverage at the national, organizational, disciplinary, and individual level, so caution is in order, requiring attention to possible biases as we have uncovered in this paper. The ORCID iD is currently strongest when it is used as a component to support a comprehensive search strategy, as illustrated example discussed in this paper. While the situation may improve in the future, at

present any searches that rely on ORCIDs should undertake a sensitivity analysis to assess coverage issues and report and recognize limitations in subsequent bibliometric analyses.

Acknowledgements

The authors thank Laurel Haak and ZL Wang for their assistance with this study. We also thank Joshua Brown and Adèniké Deane-Pratt from ORCID, and Patricia Brennan, Helen Muth, and Joe Barton from Clarivate Analytics for their help with interpreting the findings. This study was undertaken with support from the US National Science Foundation under Award 1645237 (EAGER: Using the ORCID and Emergence Scoring to Study Frontier Researchers). Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the sponsors or individuals who provided assistance.

References

Altman, M., Conlon, M., Cristan, A. L., Dawson, L., Dunham, J., Hickey, T., & Smart, L. (2014). *Registering researchers in authority files*. OCLC Research.

Anstey, A. (2014). How can we be certain who authors really are? Why ORCID is important to the British Journal of Dermatology. *British Journal of Dermatology*, *171*(4), 679-680.

Bohannon, J. (2017). Vast set of public CVs reveals the world's most migratory scientists. *Science*, *356* (6339), doi:10.1126/science.aal1189.

Butler, D. (2012). Scientists: Your number is up: ORCID scheme will give researchers unique identifiers to improve tracking of publications. *Nature*, *485*(7400), 564-565.

Clement, G. (2014). ORCID-opoly, Where High-touch Meets High-Tech: Learning and Outreach efforts in support of ORCID Integration at Texas A&M. ORCID Outreach Conference, May 22-24, 2014. Chicago, IL.

Fenner, M. (2011). ORCID: unique identifiers for authors and contributors. *Information Standards Quarterly*, *23*(3), 10-13.

Fenner, M., Gómez, C. G., & Thorisson, G. (2011). Collective action for the Open researcher & contributor ID (Orcid). *Serials*, *24*(3).

Garfield, E. (1969). British quest for uniqueness versus American egocentrism. Nature, 223, 763.

Gulpers, J. (2016, February 15). Creating your ORCID. Retrieved from https://www.eur.nl/fileadmin/ASSETS/UB/Training Support/e-learning/researchimpacts/Handout Orcid.pdf.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012a). ORCID: a system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259-264.

Haak, L. L., Baker, D., Ginther, D. K., Gordon, G. J., Probus, M. A., Kannankutty, N., & Weinberg, B. A. (2012b). Standards and infrastructure for innovation data exchange. *Science*, *338*(6104), 196-197.

Haak, L (2016, January). Publishers start requiring ORCID iDs. Retrieved from https://orcid.org/blog/2016/01/07/publishers-start-requiring-orcid-ids.

Kelley, R. P. (1984). Blocking Considerations for Record Linkage under Conditions of Uncertainty. In Proceedings of the Social Statistics Section, 602–605.

Li, G. C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., et al. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy*, *43*(6), 941-955.

Meadows, A. (2016). Everything you ever wanted know about ORCID... but were afraid to ask. *College & Research Libraries News*, *77*(1), 23-30.

Mitra, P., Kang, J., Lee, D., & On, B. W. (2005). Comparative study of name disambiguation problem using a scalable blocking-based framework. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on* (pp. 344-353). IEEE.

Number of ORCID iDs (2017, February 17). Retrieved from http://support.orcid.org/knowledgebase/articles/150557-number-of-orcid-ids

ResearcherID & ORCID Integration (2017, January). Retrieved from http://wokinfo.com/researcherid/integration/.

Rosenkrantz de Lasson, J. (2015, February 15). Why ORCID and ResearcherID When We Have Google Scholar? [Blog post]. Retrieved from http://www.jakobrdl.dk/blog/2015/02/why-orcid-and-researcherid-when-we-have-google-scholar.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation.*Annual review of information science and technology*, *43*(1), 1-43.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis?. *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833.

Tang, L., and Walsh, J.P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps, *Scientometrics* DOI 10.1007/s11192-010-0196-6.

Thomas, W. J., Chen, B., & Clement, G. (2015). ORCID Identifiers: Planned and Potential Uses by Associations, Publishers, and Librarians. *The Serials Librarian*, *68*(1-4), 332-341.

What is Orcid? (2016, March). Retrieved from http://orcid.org/content/about-orcid.

Zhou, P., & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research policy*, *35*(1), 83-104.