We should be just a number, and we should embrace it

April 18, 2016

Abstract

Purpose - This viewpoint article supports the use of unique identifiers for the authors of scientific publications. This, we believe, aligns with the views of many others as it would solve the problem of author disambiguation. If every researcher had a unique identifier there would be significant opportunities to provide even more services. These extensions are proposed in this paper.
Design/methodology/approach - We discuss the bibliographic services that are currently available. This leads to a discussion of how these services could be developed and extended.
Findings - We suggest a number of ways that a unique identifier for scientific authors could support many other areas of importance to the scientific community. This will provide a much more robust system that provides a much richer, and more easily maintained, scientific environment.
Originality/value - The scientific community lags behind most other communities with regard to the way it identifies individuals. Even if the current vision for a unique identifier for authors were to become more widespread, there would still be many areas where the community could improve its operations. This viewpoint paper suggests some of these, along with a financial model that could underpin the functionality.
Keywords: Author disambiguation, Publishing, Unique Identifier

Paper Type Viewpoint

1 Introduction

Like many other industries, scientific publishing is going through a revolution, largely brought about by the ever increasing functionality of the Internet. The business model that underpins scientific publishing is seeing significant changes, primarily led by the move towards open access publishing (Collins, 2005, Ennas and Guardo, 2015, Masrek and Yaakub, 2015, Rodriguez, 2014).

Whilst the Internet has brought significant advantages, it has also led to new challenges. A significant challenge is author disambiguation (Chin et al., 2014, Ferreira et al., 2012, Kawashima and Tomizawa, 2015, Liu et al., 2015, Shoaib et al., 2015), which has only become an issue as access to papers through an electronic medium has become possible. Automatically processing author's names when attempting to collate all the papers for a given author using just the text version of their names is a challenging problem. Typical issues include authors changing names (e.g. through marriage), using different versions of an author's name (e.g. inclusion, or not, of their initials), affiliation with different institutions, the use (or not) of accented characters etc. Until recently, the only way to identify an author was by using the text used to write their name. In recent years, some tools have become available that enable a unique identifier to be assigned to a given author. If this had widespread take up, it would not only solve the author disambiguation problem but could also open up a range of other opportunities which are not currently easily available to the scientific community. That is, it would not only resolve the author disambiguation problem but also provide many *added value* benefits, which are not available at the current time.

In this viewpoint article we propose that all researchers should have a unique identifier that is used to identify themselves. We believe that the industry needs to agree on just one standard, and all publishers/researchers should be encouraged to adopt that standard. Indeed, we would like to get to a position where peer pressure, and normal working practices, ensure that researchers see the benefit of adopting a unique ID for their professional persona.

Having a unique identifier might also be a condition of publishing in the peer reviewed scientific literature and that this ID should be included in the work flows of scientific publishers. We note that earlier this year (2016), seven publishers announced that during 2016 they would require authors to use an ORCID ID if they wished to publish[1]. We support this initiative.

The ideas presented in this article would require the publishers to agree to some basic standards and to agree to work together, possibly through an independent intermediary. Moreover, it would require the academic community to agree to also work within the proposed structure, many of which already do, in order to get the benefits from the proposals in this paper.

The remainder of this paper is structured as follows. In the next section, we draw out the motivation behind having a single author identity, and highlight some of the challenges in reaching this goal. In Section 3 we discuss the current provision with regard to scientific publishing, drawing out the shortcomings with the current model (Section 4). In Section 5 we discuss the benefits of adopting a common standard. A proposed financial model is presented in Section 6 so that any system is sustainable. In Section 7 we state our proposals, and challenge the scientific community to adopt them. Finally, we present a closing discussion and conclude the paper in Sections 8 and 9.

[1] https://elifesciences.org/elife-news/publishers-require-orcid-identifiers-authors, last accessed 18 Apr 2016

2 Motivation

In almost every aspect of our lives, we have a unique identifier so that we can be be unambiguously identified. Whether that is a student/staff ID, a bank account number, a customer reference for a utility company, a customer number for on-line services etc., you are little more than a number inside an IT system. Indeed, one of the main principles that underpins any database is that each record should have a unique identifier. Whether you recognize it or not, every time you register for any type of service you will be assigned a unique identifier.

Indeed, this will happen when researchers create an account with a journal. The unique identifier is likely to be a name that the user chooses, or their email address. However, this leads to a two-fold problem. **1)** Each user may have different services (e.g. journal) that they use and **2)** They will have a different identifier on different services, with no way of being able to link them.

Finding ourselves in this position is not a surprise as the scientific publishing industry has evolved over hundreds of years, rather than being consciously designed. The rise of the Internet has highlighted the limitations of the current model.

Having said that, there is evidence that the industry is able to come together in order to improve the system. The Digital Object Identifier (DOI) (Chandrakar, 2006) system was introduced in 1997[2]. This system is a way of uniquely identifying a scientific paper. Importantly, it also provides a *persistent* URL so that once you know the DOI, you can access a paper using a URL of the form http://dx.doi.org/*DOI*. Where the paper is actually located is of no concern, as long as the underlying DOI system is updated when the owner relocates the paper. Since the DOI system was introduced, many (if not the majority) papers that were published prior to its introduction have been classified. It is even possible to locate Einstein's 1905 seminal paper on special relativity (Einstein, 1905) using its DOI (10.1002/andp.19053221004[3])

If we could get to a position where *every* academic had a unique identifier it would eliminate the problem of author disambiguation. There have been some notable attempts to do this (see Section 3), and these are still ongoing. However, in our view, without five key criteria being met, the aim of scientists having a unique identifier is still a long way off. These criteria are:

1. Every academic should be encouraged to have a unique identifier that is persistent throughout their academic career
2. There should be one standard, rather than many, often competing, standards
3. Publishers should build any system into their work flows
4. The data should be maintained by a not-for profit organization and not be held by any one publisher or company

[2]http://www.doi.org/doi_handbook/1_Introduction.html#1.2, last accessed 21 Mar 2016
[3]http://dx.doi.org/10.1002/andp.19053221004, last accessed 21 Mar 2016

5. There should be a sustainable financial model underpinning any initiative

Others have promoted the use of single author identification. Qiu (2008) outlines problems faced by Chinese authors in particular. It is interesting to see just one of the problems faced by this community. Qiu says:

"*Chinese researchers adopt a phonetic version of their names, converted through the pinyin romanization system, which uses the Latin alphabet to represent sounds from Chinese. This approach, however, is not bidirectionally unique. There are two Chinese surnames that can be 'spelt' as Wang, for instance. And the problem is compounded by the sheer number of Chinese researchers who have not just the same surname, but also the same initial. Searching the biomedical-literature database PubMed, curated by the US National Library of Medicine, for articles published by''Wang X' results in 8,904 entries, and this number rises almost daily.*"

But this is just one complexity and others are presented in the article.

3 Current Services

The publishing industry provide a range of services for the scientific community to maintain and monitor scientific articles. These are usually provided free of charge to the end user, recognizing that the service provider will have access to the (potentially valuable) data. These services enable researchers to record their papers in a central repository, enabling them to keep track of their own papers, as well as locating the papers of their colleagues. The system may also suggest papers that you may have authored, and other authors will suggest papers that you have co-authored. Social media services may also be incorporated, as well as access to limited analytics such as the number of citations, the number of downloads, h-indices etc. Some of the most well known providers are ResearchGate[4], ReseacherID[5] and Google Scholar[6].

There are even services which cross *sector* boundaries. The International Standard Name Identifier (ISNI)[7] is an ISO (27729) certified standard for "*identifying the millions of contributors to creative works and those active in their distribution, including researchers, inventors, writers, artists, visual creators, performers, producers, publishers, aggregators, and more.*" It aims to have "*a persistent unique identifying number in order to resolve the problem of name ambiguity.*" This is exactly in line with the aims of many of the other services provided in this paper. As far as the authors are aware ISNI has not seen strong adoption within the scientific community, even though one of the authors of this paper was surprised to learn that they do have an ISNI identifier.

Other services, such as Ringgold[8] aim to provide a unique identifier for institutions, for many of the reasons mentioned in this paper.

[4]https://www.researchgate.net/, last accessed 21 Mar 2016
[5]http://www.researcherid.com/, last accessed 21 Mar 2016
[6]https://scholar.google.com/, last accessed 21 Mar 2016
[7]http://isni.org/, last accessed 21 Mar 2016
[8]http://www.ringgold.com/, last accessed 21 Mar 2016

As far as the authors of this paper are aware, the service that has the highest level of integration with publisher work flows, albeit only partially, is ORCID. On some publishers web sites, you can register your ORCID ID and this will then be associated with any papers that you write.

There has been some work on promoting the wider use of ORCID. Thomas et al. (2015) discusses how ORCID can be used, presenting case studies from the Modern Languages Association and Texas A&M University. The paper is enthusiastic about the wider take up and usage of ORCID, but also notes several issues about wider uptake.

A useful historical reference to ORCID is available from Anon (2009) and Butler (2012), with another good overview of ORCID being provided by (Haak et al., 2012).

There are some other services starting to appear which enable different aspects of an academic's work to be registered. Publons[9] is one such example, which enables academics to store details of their peer review activities.

Institutions increasingly want to capture data about their staff, in addition to details on their publications. There are providers that provide this functionality, such as Pure[10], Worktribe[11] and Converis[12]. These platforms are aimed at institutions who wish to collect and analyse data on the research activities of their staff. The data that can be stored includes information about funding, conference attendance, peer review activity etc.

Individuals may also have access to the data but this assumes that the institution is using the platform. If an individual moves institutions then access to the data may be lost or may have to be re-entered into the system used by the new institution.

As far as we are aware, there is no personal, free platform that enables an individual to store all their academic record in a way that is easily accessible to both the individual concerned and to the wider community. At least there is no single system that provides this functionality using a single unique ID, and which is integrated into work flows.

4 Key Issues

The key issues that we believe are present in the current *system* are as follows:

1. There are various systems (as outlined in Section 3), meaning that:
   (a) Academics have to register themselves on different systems
   (b) Academics have to maintain different systems
   (c) Each system will have a different way of being updated
   (d) There is little, or no, interaction between the various systems

[9] https://publons.com/, last accessed 21 Mar 2016
[10] http://www.elsevier.com/solutions/pure, last accessed 21 Mar 2016
[11] http://www.worktribe.com/, last accessed 21 Mar 2016
[12] http://converis.thomsonreuters.com/, last accessed 21 Mar 2016

(e) As a result of having different systems, researchers will have a number of different *unique* identifiers

(f) When trying to track down other research, you may have to look at a number of systems to find out the information you need

(g) Many of the systems are fundamentally commercially motivated meaning that some collected data may not be available to anybody that wishes to access it

2. Current systems, at least those that are publically available, can only be utilized for tracking publications. They generally do not provide other functionality which might be possible by leveraging on a single unique identifier for every researcher

3. The current systems compete with each other and, while there are benefits to them in collecting data, advertising etc., it does not provide the scientific community with the services that it requires

5 Potential Benefits

If we were able to arrive at position where every member of the academic community had a unique identifier, that was aligned with one provider, it would provide the community with significant value. The most obvious benefit is that the author disambiguation problem would disappear. Other benefits could include (but certainly not be limited to) the following.

1. **Common pool of reviewers**: At the present time, every publisher (at best) or journal maintains its own pool of reviewers. This is not only wasteful in terms of scientists having to maintain numerous accounts but many potential reviewers will not be immediately available, unless they have previously published, or reviewed, for that journal/publisher. If there was a single pool of reviewers this would have the benefit that scientists would only have to maintain a single system but every journal, and potentially conference, would have access to many more reviewers than they currently do.

2. **Access to reviews undertaken**: Many scientists like to keep a record of reviews they have carried out, or at least the journals they have reviewed for. This is especially important for early career researchers who present this information as part of their growing international esteem. If this data was stored as part of a scientists unique ID, then this data would be easily accessible.

3. **Conference Attendance**: Having a central repository would enable conference organisers to register their delegates via their unique ID. This data would then be available to recall who attended a given conference, and what conferences have been attended by a given scientist.

4. **Analysis**: Having data in a central repository would make analysis a lot easier and more meaningful (subject to privacy settings). For example, at the present time, we do not really know when an author started publishing so it is difficult to compare one author with another with regard to things like average number of papers published since they were awarded their PhD, average number of papers published in each year of their career etc. If we had access to this data, it would be a lot easier to carry out analysis for different authors, different institutions etc.

5. **Produce CVs:** If an academic's data was maintained in a central repository, indexed by a unique ID, the database could be accessed in order to provide data for their CV

6. **API:** If an API (Application Programming Interface) was developed, it would enable access, by a variety of agencies (funding agencies, individuals, institutions etc.), to the underlying data This would enable easy maintenance of web sites, assist in promotion and job applications, assist in funding decisions etc.

7. **Program Committee Membership:** Conferences could register their program committee members with a central service. This could then provide details about people, conferences etc.

8. **DOI integration:** If the DOI system, could be integrated with the author ID system, this would provide benefits that are simply not possible at the moment. It would be easy to produce a list of papers for each of your co-authors, it would be easy to produce a list of authors who are publishing in similar areas, who have published with authors within n degrees of freedom of a given author etc.

9. **Author permalink:** In same way that every paper has a permalink, if every academic had a unique ID, the they could also be associated with a permalink, so knowing just knowing the ID for an academic would enable you to access all their (public) information.

10. **Governments:** Many governments would like to track research that they fund and/or is carried out in their country. If the meta-data associated with each author included their affiliation/country, this would enable a suitable API to return various data, rather than embarking on a major data collection exercise.

11. **Funding Agencies:** Funding agencies, for example Research Councils UK[13] have an interest in tracking the research undertaken on their behalf. The UK has recently initiated a program[14] where all their funded research outputs are reported. Researchers are asked to store an ORCID ID as part of their profile though it is not clear how integrated this is into the work flow.

[13] http://www.rcuk.ac.uk/

[14] https://www.researchfish.com/, last accessed 13 Nov 2015

12. **Research Assessment Exercises:** Many countries regularly carry out research assessments of their universities and research institutions. The UK, for example, has the Research Excellence Framework (REF)[15] and Malaysia has the Malaysia Research Assessment (MyRA). These assessment instruments could benefit if every research had a unique identifier, along with suitable meta-data.

13. **Single Login:** Perhaps the biggest benefit is that every researcher would have a single login to maintain their data. This not only saves time, but also ensures data consistency. For example, if a researcher marked themselves as being unavailable for reviewing, this information would be available to all journals. If they decided that their keywords were incorrect, then a single change would be available to all journals.

14. **Research Data:** If the data were made available to the scientific community then it would enable many research projects to be undertaken. For example, what universities carry out most reviews, do some researchers mark themselves as permanently unavailable to carry out reviews, are there are difference across different disciplines etc. The data held would be a rich repository for research projects, which can only be imagined at the moment.

15. **Financial Leverage:** Although we would expect the data to be freely available to those that wished to access it, there could be opportunities to provide consultancy services for interested parties who wished to have the data analysed by suitably qualified analysts.

6 Financial Sustainability

If the proposals in this paper (see Section 7) were adopted there would inevitably be questions about how it would be funded. There are some obvious options such as asking the publishers, asking one or more professional bodies or user groups or trying to fund it through advertising and sponsorship. None of these are ideal due to potential conflict of interests and the reliance on ongoing support, which may not be forthcoming. The proposals would need the support of publishers, professional bodies and societies, but that should, in our view, stop short of financial support.

One possible financial model, based on the concept that everybody who publishes has to pay, by an annual micro-payment. This payment could be as low as five USD per annum. There could also be different rates for developing countries, institutional subscriptions etc.

A conservative estimate[16] places the number of active researchers at fifteen million worldwide. If each of those paid five USD this would generate an annual income of USD 75M. This may seem a lot, but the start up costs would

[15] http://www.ref.ac.uk/, last accessed 21 Mar 2016

[16] http://www.richardprice.io/post/12855561694/the-number-of-academics-and-graduate-students-in, last accessed 21 Mar 2016, data drawn from (Anon, 2010)

be substantial and there would be on-going costs for staff, premises, technical resources (e.g. servers), development costs, sales and marketing, raising awareness, ensuring data integrity etc.

Adopting this policy would have the added advantage that the system would have the details of every active researcher, along with a history of previously active researchers, which would provide added value to the data that is stored, and the analysis that can be performed.

7 Proposal

If the ideas in this paper are to become a reality, we would encourage adoption of the following proposals.

1. The publishing and scientific community should agree to adopt a single standard to uniquely identify members of the scientific community.
2. Every academic should have a unique identifier. This should not be reused when they are no longer active (either through retirement or death).
3. Organisations (e.g. universities, funding agencies, government agencies, publishers etc.) should also have an identifier, which would provide a way to more easily access researchers of interest to them.
4. Every academic should be strongly encouraged to use the system.
5. An international committee should be established to oversee the author ID initiative, and seek to develop it for other benefits for the community.
6. There should be a privacy *policy* developed as part of the system. This would enable individual researchers to define what data can be accessed both via the web portal and the API.
7. The affiliation IDs of the authors should form part of their data, so that institutions can access all researchers who work, or have worked, at a given university, research institute etc.
8. The affiliations of the authors should form part of their data, so that institutions can access all researchers who work, or have worked, at a given university, research institute etc.
9. Publishers should update their work flows so that the author ID is included (and possibly required) and that it is attached to every scientific paper that is submitted. This should include conferences, books, monographs etc.
10. Work that has already been published should be updated, in the same way that DOIs are now attributed to the vast majority of papers.

11. Search facilities on publishing web sites should enable users to search by author ID.

12. Every author, as well as their name, should also include their author ID as part of the title information on a paper. We would suggest that this is placed in parenthesis after the author's name.

13. As well authors ID's appearing on papers, it should appear on other key areas such as editorial boards, program committees, external examining roles, funding agency committees etc.

14. A suitable API should be developed so that anybody can access the data in an automated way.

15. The initiative should be managed by a not-for-profit organization to preclude any conflict of interest and to ensure the sustainability of the project

We acknowledge that the proposal presented above is a shift change in the way the scientific community operates. There are many hurdles that would need to be overcome and many more discussions that would need to take place, but we hope that these proposals provide the catalyst for change.

8 Discussion

The scientific publishing industry is changing, perhaps faster than it has ever changed before. A lot of the change is being financially driven as publishers, and other providers, seeking to adapt to the new way of working and establish new business models so that they remain competitive in the market. This can most obviously been seen in the transition from *traditional* publishing to open access publishing (Kendall et al., 2015b) and in new services that are now being offered (Kendall et al., 2015a).

There is also a battle being fought to take control of information, such as publication details, authors, social media etc.

Various providers have entered the market in recent years and this is leading to different products which researchers have to register for, and then keep up to date.

Moreover, these systems are not providing everything that could be useful to a researcher, and the scientific community, in general, are not being fully consulted.

In this paper, we have argued for a **single**, independent system that provides every researcher with one identifier, enabling maintenance to be a lot easier, author disambiguation to be a problem of the past and for many additional services to be provided that would be of great benefit to researchers and their institutions.

Finally, the system should form part of the work flows of publishers, conference organizers etc. so that all the data is captured, stored in one place and can be interrogated easily.

9 Conclusion

In almost every aspect of our lives, we are governed by having a unique identifier that discriminates us from everybody else in the system. If this were not the case, then any system would be unworkable and would result in chaos. Yet, the scientific community works within this chaotic system and as the connected, global age has developed, it has shown up its inadequacies. We can no longer rely on name and/or institution to differentiate ourselves. We must move towards each academic having a single identifier. Moreover, we suggest that the community should utilize a single identifier system, rather than having competing systems where, at best, data has to be entered into several places and, at worst, the systems are not integrated or commercial pressures come into play. Therefore, any system that supports this important area of the academic community should be run by a not-for profit organization, and there should be one system. Of course, there is a need for a sustainable financially model, but that is different to developing a business model that has ultimate responsibility to its shareholders.

If the scientific community does not resolve this now, we may get to the point of no return.

References

Anon. 2009. Credit where credit is due. *Nature* **462** 825. doi:10.1038/462825a.

Anon. 2010. National science board science and engineering indicators. Arlington, VA: National Science Foundation. (NSB 10-01).

Butler, D. 2012. Scientists: your number is up. *Nature* **485**(7400) 564. doi:10.1038/485564a.

Chandrakar, R. 2006. Digital object identifier system: an overview. *Electronic Library* **24**(4) 445–452. doi:10.1108/02640470610689151.

Chin, W-S., Y. Zhuang, Y-C. Juan, F. Wu, H-Y. Tung, T. Yu, J-P. Wang, C-X. Chang, C-P. Yang, W-C. Chang, K-H. Huang, T-M. K, S-W. Lin, Y-S. Lin, Y-C. Lu, Y-C. Su, C-K. Wei, T-C. Yin, C-L. Li, T-W. Lin, C-H. Tsai, S-D. Lin, H-T. Lin, C-J. Lin. 2014. Effective string processing and matching for author disambiguation. *Journal of Machine Learning Research* **15** 3037–3064.

Collins, J. 2005. The future of academic publishing: What is open access? *Journal of the American College of Radiology* **2**(4) 321 – 326. doi:10.1016/j.jacr.2004.07.018.

Einstein, A. 1905. Zur elektrodynamik bewegter kärper. *Annalen der Physik* **322**(10) 891–921. doi:10.1002/andp.19053221004.

Ennas, G., M. C. Di Guardo. 2015. Features of top-rated gold open access journals: An analysis of the scopus database. *Journal of Informetrics* **9**(1) 79 – 89. doi:10.1016/j.joi.2014.11.007.

Ferreira, A.A., M.A. Gonçalves, J.M. Almeida, A.H.F. Laender, A. Veloso. 2012. A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Information Sciences* **206** 42–62. doi: http://dx.doi.org/10.1016/j.ins.2012.04.022.

Haak, L. L., M. Fenner, L. Paglione, E. Pentz, H. Ratner. 2012. ORCID: a system to uniquely identify researchers. *Learned Publishing* **25**(4) 259–264. doi:10.1087/20120404.

Kawashima, H., H. Tomizawa. 2015. Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics* **103**(3) 1061–1071. doi:10.1007/s11192-015-1580-z.

Kendall, G., A. Yee., B. McCollum. 2015a. Is there a role for publication consultants and how should their contribution be recognized? *Science and Engineering Ethics* In pressdoi:10.1007/s11948-015-9710-9.

Kendall, G., A. Yee., B. McCollum. 2015b. The scientific publishing revolution and the challenges it presents. *Learned Publishing* (under review).

Liu, Y., W. Li, Z. Huang, Q. Fang. 2015. A Fast Method Based on Multiple Clustering for Name Disambiguation in Bibliographic Citations. *Journal of the Association for Information Science and Technology* **66**(3) 634–644. doi: 10.1002/asi.23183.

Masrek, M. N., M. S. Yaakub. 2015. Intention to publish in open access journal: The case of multimedia university malaysia. *Procedia - Social and Behavioral Sciences* **174**(0) 3420 – 3427. doi:10.1016/j.sbspro.2015.01.1013. International Conference on New Horizons in Education, {INTE} 2014, 25-27 June 2014, Paris, France.

Qiu, J. 2008. Scientific publishing: Identity crisis. *Nature* **451** 766–767. doi: 10.1038/451766a.

Rodriguez, J. E. 2014. Awareness and attitudes about open access publishing: A glance at generational differences. *The Journal of Academic Librarianship* **40**(6) 604 – 610. doi:10.1016/j.acalib.2014.07.013.

Shoaib, M., A. Daud, M.S.H Khiyal. 2015. Improving similarity measures for publications with special focus on author name disambiguation. *Arabian Journal for Science and Engineering* **40**(6) 1591–1605.

Thomas, Wm. J., B. Chen, G. Clement. 2015. ORCID identifiers: Planned and potential uses by associations, publishers, and librarians. *The Serials Librarian* **68**(1–4) 332–341. doi:10.1080/0361526X.2015.1017713.