Check for updates

SOFTWARE TOOL ARTICLE

REVISED pubassistant.ch: consolidating publication profiles of researchers [version 2; peer review: 1 not approved]

Reto Gerber[1,2], Mark D. Robinson[1,2]

[1]Department of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland

v2 **First published:** 30 Sep 2021, **10**:989
https://doi.org/10.12688/f1000research.73493.1

Latest published: 20 Dec 2021, **10**:989
https://doi.org/10.12688/f1000research.73493.2

Abstract

Online accounts to keep track of scientific publications, such as Open Researcher and Contributor ID (ORCID) or Google Scholar, can be time consuming to maintain and synchronize. Furthermore, the open access status of publications is often not easily accessible, hindering potential opening of closed publications. To lessen the burden of managing personal profiles, we developed a R shiny app that allows publication lists from multiple platforms to be retrieved and consolidated, as well as interactive exploration and comparison of publication profiles. A live version can be found at pubassistant.ch.

Keywords

open access, publication profiles, R shiny

This article is included in the RPackage gateway.

This article is included in the Research on Research, Policy & Culture gateway.

Open Peer Review

Reviewer Status ✗

	Invited Reviewers
	1
version 2 (revision) 20 Dec 2021	
version 1 30 Sep 2021	✗ report

1. **Daniel W Hook**, Digital Science, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Mark D. Robinson (mark.robinson@mls.uzh.ch)

Author roles: Gerber R: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Robinson MD**: Conceptualization, Investigation, Methodology, Project Administration, Software, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: A financial contribution to the project was provided by Michael Schaepman. MDR acknowledges support from the University Research Priority Program Evolution in Action at the University of Zurich.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Gerber R and Robinson MD. **pubassistant.ch: consolidating publication profiles of researchers [version 2; peer review: 1 not approved]** F1000Research 2021, **10**:989 https://doi.org/10.12688/f1000research.73493.2

First published: 30 Sep 2021, **10**:989 https://doi.org/10.12688/f1000research.73493.1

REVISED **Amendments from Version 1**

Updates to the Introduction include: more theoretical background on the issue of name disambiguation was added, the important role of ORCID and DOI is better highlighted, imprecisely-worded statements (mainly about capabilities of Dimensions) were updated or removed, a description of similar tools was added, and all of this was put into perspective compared to our application.
Furthermore, the fact that web-scraping is done to retrieve publications from Google Scholar is highlighted and a more detailed description of the used fuzzy matching of publications is added.

Any further responses from the reviewers can be found at the end of the article

Introduction

Given the increasing number of both researchers and publications as well as publishing modes,[1,2] it becomes a challenge to identify and consolidate all publications from a single author. A few of the main issues are transliteration of names into roman alphabetic system, the non-uniqueness of names, differently written names (e.g., with or without middle initial) and changing affiliation over time. There are broadly speaking two approaches to solve this ambiguity: "unattended" and "attended". The "unattended" approach tries to automatically resolve ambiguity using additional existing metadata. The "attended" approach relies on human intervention in the form of unique identifiers that enable robust linkage of publications to authors, assuming researchers and their collaborators use them consistently. The most important, widely used and *de facto* standard identifier in many fields is the Open Researcher and Contributor ID (ORCID).[3] Other identifiers such as Google Scholar ID[4] or ResearcherID (Publons)[5] are also used, although they are not as broadly used as ORCID and persistence of identifiers is not always guaranteed. Having multiple identifiers on multiple platforms is not unusual and automatic publication detection and syncing between accounts is possible to some degree. However, automatic synchronization of accounts for different identifiers can be hindered by the fact that not all systems use the standardized DOI (Digital Object Identifier) as document identifier to match publications.

Although the two main standardized identifiers for authors (ORCID) and documents (DOI) are widely adopted, other identifiers are still used, making it often necessary to synchronize publication records on different platforms manually to obtain complete records. For instance, there is no simple one-click solution to synchronize publications between ORCID and Google Scholar. In Google Scholar, publications need to be searched and added manually (if they are not detected automatically) while in ORCID it is possible to input a citation file. A typical workflow to update ORCID based on Google Scholar would therefore be to first search (one by one) in Google Scholar all publications that are listed in ORCID and then add the missing ones. But since it is possible that publications listed in Google Scholar are not in ORCID, the reverse needs to be done to be sure the accounts are up to date. If more accounts need to be synced (e.g., Publons), the complexity and time needed increases accordingly. Although it is possible, and probably advisable, to link accounts for automatic updates (e.g., linking Publons with ORCID), this cannot be done under all circumstances and missing publications are still possible.

While some (commercial) services (such as Dimensions[6] or Web of Science[7]) provide extensive data mining to retrieve publication data, they often also rely on unique identifiers (such as ORCID in the case of Dimensions) for correct assignment. Furthermore, on many platforms that combine different sources, it is not easy to determine where the data originated (e.g., is a publication listed in ORCID or in Publons? or both?). In addition, data exploration and visualization is often restricted to citations over time (except costly commercial services). With the growing awareness, interest and mandates towards Open Science, open access (OA) status of articles can also be of interest. The same is true for preprints, which are not always taken into account despite becoming increasingly important in many research fields.[8,9]

Another inconvenience can be the existence of duplicated publications, which can stem either from the association of preprint and peer-reviewed publication or from revisions or different versions. In many cases, it is sensible to treat those closely linked publication as just one publication instead of multiple. If the required information to link publications is missing, automatic detection is not always possible and manual intervention is needed.

Many tools, both commercial and free, exist that allow to explore certain aspects of bibliographies. Furthermore, many of the available tools are not made for individual authors but rather operate on the department, institutions or even country level. Some of the existing commercial tools include Elements (from the company Symplectic[10]) and Dimensions, both of which are mainly intended for institutional use; but, especially Dimensions also offers functionalities for authors to explore their bibliographies.

Commercial as well as institutional tools provide valuable improvements to the quality of bibliographies, especially in non-STEM subjects where accurate representations of scholars is often more difficult.

Some of the free and/or open tools include Vivo[11] (Institutional level, creates ontologies for representing scholarship), Profiles Research Networking[12] (Institutional level, help to discover collaborators), ReCiter[13] (Institutional level, find publications of authors in PubMed), ImpactStory[14] (author level, impact and open access status of publications from ORCID).

Furthermore, tools mainly intended for monitoring open science include the open science monitoring of the European commission[15] (country-level), the German open access monitor (institution-level) and OpenAIRE (Open Access Infrastructure for Research in Europe), which provides dashboards (country- or institution-level).

In our case, we took inspiration from the Swiss National Science Foundation's Open Access Check,[16] which allows Swiss researchers to reflect on their publishing practices and encourages various forms of OA, including green OA; importantly, such resources rely on the source databases being up to date in the first place.

To facilitate overview and synchronization of publication records, we provide a web-based application that allows publications for an author to be retrieved from different sources, combines entries, checks for duplicates and downloads citations to easily update records across platforms. Furthermore, the open access status of each publication is provided, which can help to select publications that could be "greened" (i.e., depositing documents in institutional repositories). Taken together, this allows researchers to organize their public publication profiles and to interactively explore the accuracy of records across the various entry points.

Methods

The workflow is as follows: The user needs to first specify the unique identifiers of the researcher of interest for at least one of ORCID, Google Scholar and Publons. Additionally, a search query for Pubmed can be generated. Furthermore, the

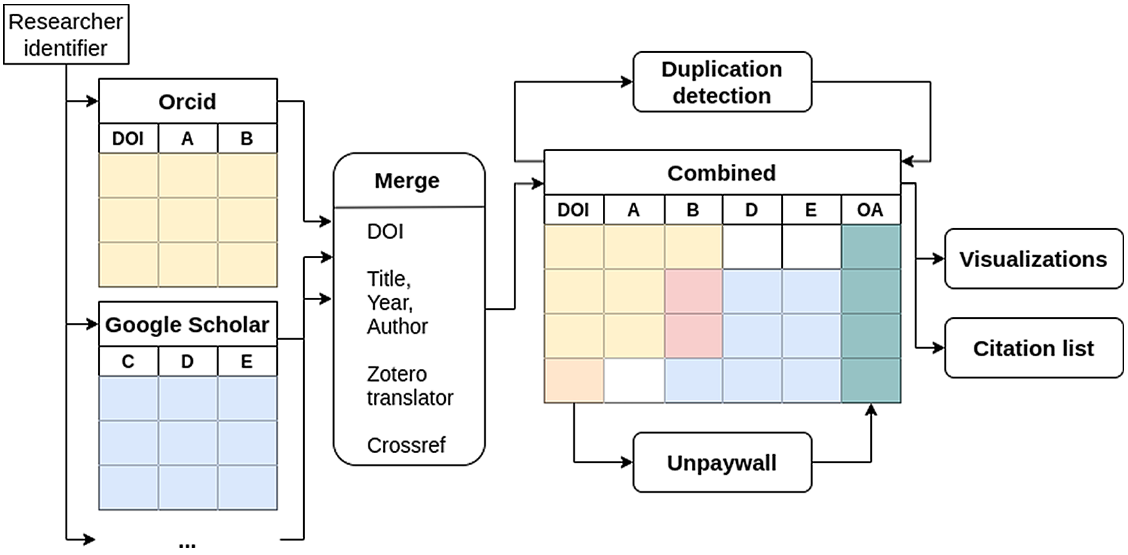
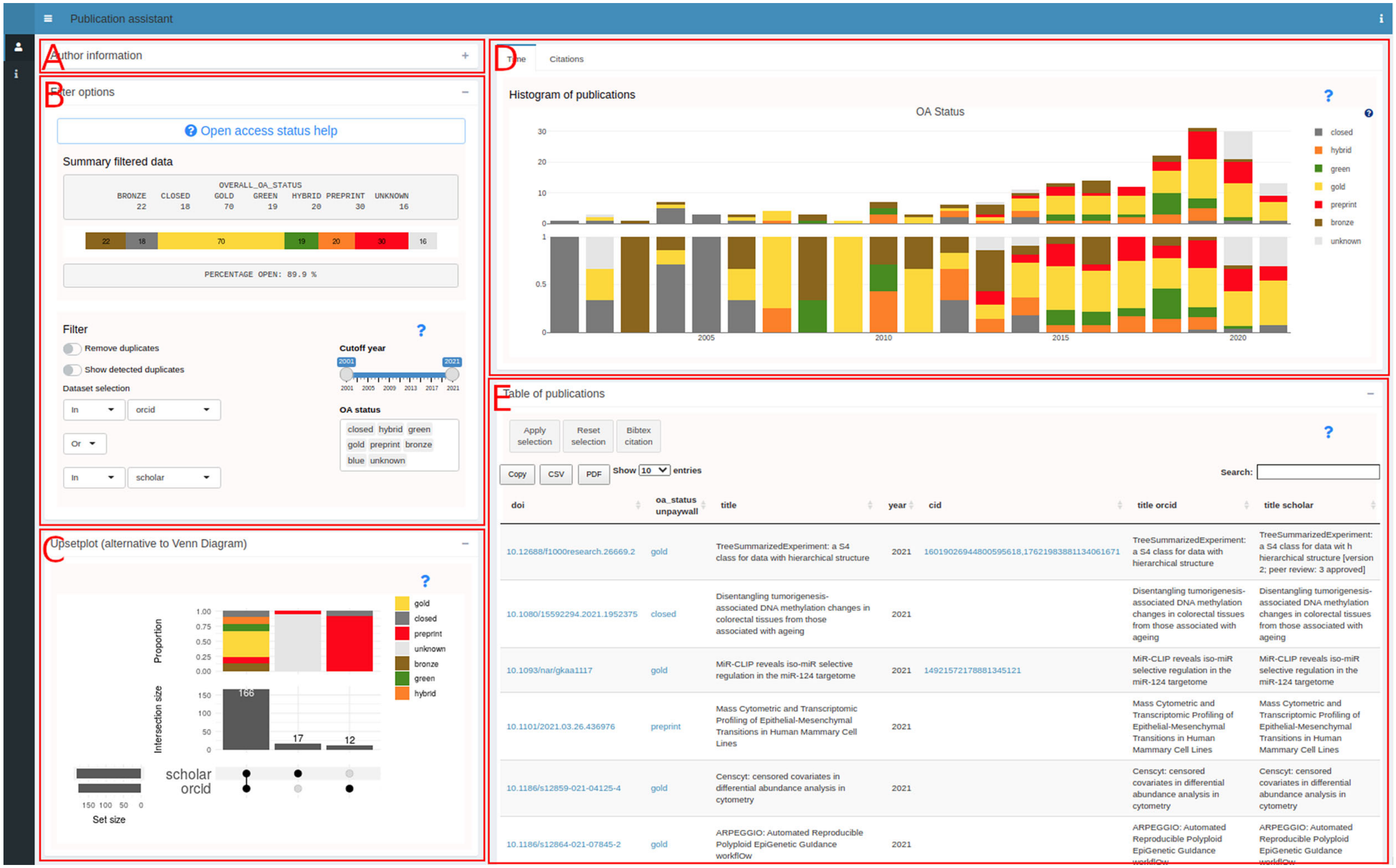


Figure 1. Overview of the data processing. The identifiers given by the user are used to obtain the data from each platform independently. The data is then merged and the open access status (column OA) is obtained using the Digital Object Identifier (DOI). Furthermore duplicates are detected by comparing the titles of the publications.

Table 1. Open access (OA) definition used by Unpaywall.

OA status	Open accessible	Description
Gold	Yes	Published in open-access journal
Green	Yes	Publication in free repository
Hybrid	Yes	Open licence
Bronze	Yes	No open licence
Closed	No	



Summary filtered data

```
                        OVERALL_OA_STATUS
BRONZE  CLOSED   GOLD   GREEN  HYBRID PREPRINT  UNKNOWN
    22      18     70      19      20       30       16
```

```
PERCENTAGE OPEN: 89.9 %
```

Filter

○ Remove duplicates

○ Show detected duplicates

Dataset selection

In ▾ orcid ▾

Or ▾

In ▾ scholar ▾

Cutoff year: 2001 – 2021

OA status: closed hybrid green gold preprint bronze blue unknown

Apply selection | Reset selection | Bibtex citation

Copy | CSV | PDF | Show 10 entries | Search:

doi	oa_status unpaywall	title	year	cid	title orcid	title scholar
10.12688/f1000research.26669.2	gold	TreeSummarizedExperiment: a S4 class for data with hierarchical structure	2021	16019026944800595618,17621983881134061671	TreeSummarizedExperiment: a S4 class for data with hierarchical structure	TreeSummarizedExperiment: a S4 class for data wit h hierarchical structure [version 2; peer review: 3 approved]
10.1080/15592294.2021.1952375	closed	Disentangling tumorigenesis-associated DNA methylation changes in colorectal tissues from those associated with ageing	2021		Disentangling tumorigenesis-associated DNA methylation changes in colorectal tissues from those associated with ageing	Disentangling tumorigenesis-associated DNA methylation changes in colorectal tissues from those associated with ageing
10.1093/nar/gkaa1117	gold	MiR-CLIP reveals iso-miR selective regulation in the miR-124 targetome	2021	14921572178881345121	MiR-CLIP reveals iso-miR selective regulation in the miR-124 targetome	MiR-CLIP reveals iso-miR selective regulation in the miR-124 targetome
10.1101/2021.03.26.436976	preprint	Mass Cytometric and Transcriptomic Profiling of Epithelial-Mesenchymal Transitions in Human Mammary Cell Lines	2021		Mass Cytometric and Transcriptomic Profiling of Epithelial-Mesenchymal Transitions in Human Mammary Cell Lines	Mass Cytometric and Transcriptomic Profiling of Epithelial-Mesenchymal Transitions in Human Mammary Cell Lines
10.1186/s12859-021-04125-4	gold	Censcyt: censored covariates in differential abundance analysis in cytometry	2021		Censcyt: censored covariates in differential abundance analysis in cytometry	Censcyt: censored covariates in differential abundance analysis in cytometry
10.1186/s12864-021-07845-2	gold	ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw	2021		ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw	ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw

Figure 2. pubassistant.ch panel overview. After entering identifiers in panel A, successful retrieval and merging, panels B-E appear. Panel B is the main panel for filtering. Visualizations are in panel C (upsetplot), D (histogram) and E (table).

option to search for bibliometrics, obtained from the NIH Open Citation Collection using iCite,[17] can be selected. After confirmation, publications are retrieved from the specified sources and combined into a table based on the DOI (see Figure 1) or, in case of publications from Google Scholar, based on (fuzzy) matching of titles and/or metadata retrieval from Zotero (Zotero translator, i.e., web scraping)[18] or Crossref (i.e., query the available metadata to obtain a DOI).[19] Since the set of considered publications stem from the same author, matching of publications is solely based on the title of the publication, by calculating the pairwise relative Levenshtein distances (Levenshtein distance divided by the maximum possible Levenshtein distance, i.e., number of characters) between titles and setting a threshold of 0.1 below which publications are assumed to be the same. No formal validation of this approach was done, but manual checking of a large number cases showed good matching in most cases. After joining the publications list, the open access status of each publication with a DOI is retrieved using Unpaywall,[20] who provide a publicly accessible database containing open access information for publications. The definitions of the different open access status that Unpaywall uses is provided in Table 1. Additionally, preprints are defined as having OA status "green" in Unpaywall with the attribute "version" equal to "submittedVersion". A database snapshot of Unpaywall can be downloaded https://unpaywall.org/products/snapshot.

After this step, interactive exploration of the publications is possible. Various options to filter the data according to OA status, year and source (ORCID, Google Scholar, etc.) are available with the possibility to remove or show duplicates (detected using fuzzy matching of titles, similar to matching of publications). Several metrics, tables and plots are available for exploration of the data. Examples include a upset plot that shows how many publications are associated with each identifier, a histogram of the number of publications per year colored by open access status, and a table listing the individual publications. After exploration, specific subsets can be generated using the filtering options, which are then imposed on the visualizations and tables presented. In all cases, relevant snapshots of the citation information can be obtained in the form of a downloadable file.

Another possible application is the integration of local databases, such as university repositories. For example, the Zurich Open Research Archive (ZORA),[21] developed and maintained by the Main Library at the University of Zurich, has been integrated in an alternative version of the app that allows local entries to be compared with public profiles, allowing synchronization of publication profiles with local repositories.

Implementation

The application is written in R (Version 4.1.0)[22] and shiny (Version 1.6.0),[23] see *Software availability*. As a back-end database, PostgreSQL is used to store a local copy of Unpaywall (and ZORA). Such a local database for Unpaywall is not strictly needed, but a large speedup of the retrieval of the open access status is achieved compared to access over the Unpaywall API. Furthermore, since only a fraction of the data from Unpaywall is used (only the DOI, the open access status and two additional columns for preprint identification) the actual table, containing open access status, is comparably small with a size of about 6 GB (compared to more than 165 GB of the complete version). Unpaywall does daily updates that can be downloaded and are used to update the local database to keep it in sync with the online version. The DOIs for publications listed in Google Scholar are obtained by either matches to publications from other sources, metadata retrieval using the Zotero translator service or a Crossref query.

Various R packages that facilitate retrieval of publications from a specific resource such as https://docs.ropensci.org/rorcid (ORCID),[24] https://github.com/jkeirstead/scholar (Google Scholar)[25] or https://docs.ropensci.org/rentrez (Pubmed)[26] have been included.

Operation

The app is containerized using Docker (Version 19.03.13, dockerfiles and docker-compose file are provided in *Software availability*). Multiple, interacting containers are deployed using docker-compose, the two most important are a container running the R shiny application and another running PostgreSQL. Furthermore, the Zotero translator service is run in a separate container. As already stated, the PostgreSQL service is not strictly needed, but substantially increases retrieval speed of the OA status.

Use case

Figure 2 shows a use case for an author where the ORCID (0000-0002-3048-5518) and Google Scholar ID (XPfrRQEAAAAJ) were given as an input (collapsed panel in Figure 2A). Panel B provides a summary of the publication list and options to filter by dataset, by year and by OA status. Additionally, the possibility to remove duplicates or only show duplicates is available. The other panels contain visualizations including an upsetplot[27] (C), a histogram (D) and a table (E). The table can be further filtered by selecting rows allowing to create specific citation lists that can be created based on the rows in the table. The contents of the table can be copied to the clipboard or downloaded in CSV format.

Discussion

Our method relies on the DOI to retrieve the OA status, which is a limitation in domains where DOIs are not used. The DOI is also used to unambiguously match publications. If no DOI is present, the titles of the publications are used for matching, which can lead to ambiguity. Even if a publication has an assigned DOI, but it is missing in the data, it becomes difficult or time-consuming to retrieve the missing information with services such as the Zotero translator or Crossref.

Because of the non commercial nature of this application, some additional limits present themselves. Most notably, our application requires freely-available APIs, or in the case of Google Scholar web-scraping (contravening the terms of use of Google Scholar), for retrieving the open publication data from their respective sources. For the two main sources considered so far (ORCID and Google Scholar), no restrictions have been noticed, while for others rate limits in the number of requests are quite restrictive (e.g., for Publons). Other APIs not currently included in our application (e.g., from Dimensions or Mendeley) could be added in the future.

Data availability

No data are associated with this article.

Software availability

Software available from: https://pubassistant.ch/

Source code available from: https://github.com/markrobinsonuzh/os_monitor

Archived source code at time of publication: https://doi.org/10.5281/zenodo.5509626

License: MIT

Acknowledgements

We thank Izaskun Mallona for help with hosting the application and various helpful suggestions. We thank various members of the Statistical Bioinformatics Group at University of Zurich for feedback.

References

1. UNESCO: **Science Report.** 2021.
Reference Source

2. Bornmann L, Mutz R: **Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references.** *J. Assoc. Inf. Sci. Technol.* 2015; **66**(11): 2215–2222. 2330-1643.
Publisher Full Text | Reference Source | Reference Source

3. Haak LL, Fenner M, Paglione L, *et al.*: *ORCID: a system to uniquely identify researchers.* Learned Publishing; 2012; **25**(4): 259–264. 1741-4857.
Publisher Full Text | Reference Source | Reference Source

4. Google Scholar.
Reference Source

5. Publons.
Reference Source

6. Hook DW, Porter SJ, Herzog C: **Dimensions: Building Context for Search and Evaluation.** *Front. Res. Met. Analy.* 23, August 2018; **3**: 2504-0537.
Publisher Full Text | Reference Source

7. World's largest publisher-neutral citation index and research intelligence platform.
Reference Source

8. Vale RD: **Accelerating scientific publication in biology.** *Proc. Natl. Acad. Sci.* National Academy of Sciences Section: Perspective; November 2015; **112**(44): 13439–13446. 0027-8424, 1091-6490.
Publisher Full Text | Reference Source

9. Johansson MA, Reich NG, Meyers LA, *et al.*: **Preprints: An underutilized mechanism to accelerate outbreak science.** *PLoS Med.*PublicLibrary of Science; April 2018; **15**(4): e1002549. 1549-1676.
Publisher Full Text

10. Open Access.
Reference Source

11. Conlon M, Woods A, Triggs G, *et al.*: **VIVO: a system for research discovery.** *J. Open Source Softw.* 2019; **4**: 1182.
Publisher Full Text

12. *Deployments · wcmc-its/ReCiter*: **(kein Datum). Abgerufen am 15. 11 2021 von ReCiter: an enterprise open source author disambiguation system for academic institutions.**
Reference Source

13. *Impactstory: Discover the online impact of your research*: **(kein Datum). Abgerufen am 15. 11 2021 von.**
Reference Source

14. *Profiles Research Networking Software*: **(kein Datum). Abgerufen am 15. 11 2021 von.**
Reference Source

15. Trends for open access to publications.
Reference Source

16. SNSF Open Access Check.
Reference Source

17. ICite, B: **Ian Hutchins, and George Santangelo. iCite Database Snapshots (NIH Open Citation Collection).** The NIH Figshare Archive; 2019.
Publisher Full Text | Reference Source

18. Zotero Translation Server: July 2021. original-date: 2018-06-11T11:28:53Z.
Reference Source

19. Lammey R: **CrossRef developments and initiatives: an update on services for the scholarly publishing community from CrossRef.** page 6.

20. Piwowar H, Priem J, Larivière V, *et al.*: **The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles.**

PeerJ. February 2018; **6**: e4375. 2167-8359.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Welcome to Zurich Open Repository and Archive - Zurich Open Repository and Archive.
Reference Source

22. R R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2011; 3-900051-07-0. 16000706.
Publisher Full Text

23. Chang W, Cheng J, Allaire JJ, *et al.*: **shiny: Web Application Framework for R.** 2020.
Reference Source

24. Scott Chamberlain: **rorcid: Interface to the 'Orcid.org' API.** 2021.

25. Keirstead J: **scholar: analyse citation data from Google Scholar.** 2016.
Reference Source

26. Winter DJ: **rentrez: an R package for the NCBI eUtils API.** *The R Journal.* 2017; **9**(2): 520–526.
Publisher Full Text

27. Conway JR, Lex A, Gehlenborg N: **UpSetR: an R package for the visualization of intersecting sets and their properties.** *Bioinformatics.* September 2017; **33**(18): 2938–2940. 1367-4811.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✗

Version 1

Reviewer Report 26 October 2021

https://doi.org/10.5256/f1000research.77148.r96152

✗ **Daniel W Hook**

Digital Science, London, UK

The authors have created a free, open source piece of software to bring researcher and publication records together from different data sources. They detail their motivations and methodology in this paper.

- The authors begin their article by motivating the development of their software based on the difficulty of consolidating all the publications from a single author resulting from:
   - The non-uniqueness of names of authors
   - The proliferation of unique identifiers that aim to solve this problem (e.g. ResearcherID/Publons, ORCID *et al.*)It is reasonable to claim that we, as a community, have not yet realised a comprehensive solution to these problems.

  However, the authors choose to take a peculiarly western-centric view of this issue. The greatest challenges of name disambiguation are typically found when authors who might not natively use a roman alphabetic system are forced to transliterate their names when publishing in the western-centred publishing system. This fact goes unacknowledged in this paper but would seem to be at the heart of the name disambiguation issue.

  There are several approaches to name disambiguation but there are broadly two “schools”: attended and unattended. “Attended” is where humans interact with the disambiguation and “unattended” is where humans have no role. Attended disambiguation is the focus for identity management systems like ORCID - incentives need to be aligned for authors and others to participate in this and significant work goes into understanding the motivations and concerns of researchers in order to make these types of system successful. Unattended disambiguation makes use of the data that are available in the ecosystem (including the outputs of the attended disambiguation approach) to create a calculated output. We feel that the authors should give some of this background in their paper for context.

  In this context, it is important to acknowledge that only one identifier system has been successful in engaging the broad global academic community and that is ORCID. Other solutions are self-acknowledged proprietary solutions that aim to solve this problem in limited contexts (Researcherid.com/Publons to improve the Web of Science data (attended disambiguation); ScopusID - unattended disambiguation approach). Dimensions explicitly leverage ORCID data in an unattended person disambiguation approach.[1]

- The authors claim that there is no standardization of unique identifiers for authors and documents.

  This is a very strong claim when viewed at an international ecosystem level. The majority of funders, publishers, institutions, scholarly societies and government agencies involved with research from around the world acknowledge DOIs as the key identifier for a research paper and ORCID as the principal identifier of researchers.

  Some complexity lies in the authority that issues and maintains an article DOI (Crossref or DataCite in most cases but also, for example, J-Stage in Japan).

  ORCID may currently be a "de facto standard" but the authors' statement to this effect underplays the central role played by ORCID in the scholarly infrastructure community and the extent to which ORCID is the standard with which the majority of the commercial and open infrastructure providers engage. Google Scholar provides no API and makes no claim of persistence of identifiers; authors have limited control of their profile and of privacy. Researcherid.com (the progenitor of Publons/ResearcherID) is a founding member, supporter and participant of ORCID. We suggest that the authors should ensure that acknowledgement be made of ORCID's suitability and level of adoption in academia. If the authors wish to note that ORCID is not the only standard, it would be appropriate to mention parallel efforts such as researchmap.jp in Japan and note that some countries remain reluctant to adopt ORCID as their principal identifier at this time.

- The authors claim that it is not easy to determine the provenance of data in Dimensions has derived its data from ORCID or Publons. Arguably this is not the case. Dimensions clearly states that algorithmic methods are used with the input of ORCID data.[1] Publons/Researchid.com data is a proprietary source that could not be used in Dimensions without explicit acknowledgement of its use.
- The authors further claim that no information is given about the completeness of these sources (e.g. Dimensions/Web of Science/Scopus), yet significant academic work has been undertaken to understand and benchmark coverage and completeness of these sources (see reference 2 for example).[2]

- We agree with the authors that Open Science is indeed critical. This is why Dimensions, Web of Science, Scopus and others contain information on Open Access statuses of articles, provided by Unpaywall (https://unpaywall.org/integrations). Dimensions also contains full listings of preprint articles from many preprint servers including ArXiv.org, BioArXiv, the Center for Open Science, PeerJ and so on. As such we find the authors' comment that these things are "not taken into account" to be misleading.

- The authors then appear to contradict themselves by claiming that preprints and publications cause duplicate entries in these systems. In Dimensions, where the current reviewer has the most experience, publications from preprints are linked to final publications where this information is available, see for example

https://app.dimensions.ai/details/publication/pub.1118864658?and_facet_researcher=ur.01123321343.51
. We agree that manual intervention is often needed in these cases.

- The authors claim that a free tool does not exist to explore metadata brought together from multiple sources, however, this does not acknowledge the rich lineage of free tools written to serve institutional use cases in author disambiguation and metadata aggregation including: VIVO (https://duraspace.org/vivo/) and Catalyst Profiles (http://profiles.catalyst.harvard.edu), both funded by the NIH; ImpactStory (e.g. https://profiles.impactstory.org/u/0000-0001-6728-7745/publications); and, ReCiter (https://github.com/wcmc-its/ReCiter). The authors should endeavour to situate their work in the context of this prior work.

- Symplectic Elements is, as the authors state, a commercial system focused on institutional use cases that meets this need. However, the authors fail to acknowledge that outside STEM subjects, coverage of research outputs becomes more challenging. The strength of commercial tools is often that, as they must meet institutional needs, work is put into diversifying coverage of different output types beyond the journals and conference proceedings that STEM subjects favour. While composing a faithful representation of a STEM academic and their output can be challenging from disparate data sources, doing the same for a non-STEM academic can be an order of magnitude more difficult. We believe that it is also fair to acknowledge that institutional engagement is not unimportant to improving the overall data quality in the bibliographic ecosystem.

- The authors also appear to be unaware that Dimensions does bring records together from multiple sources including author information and that this is transparently documented in scholarly articles written by the Dimensions team,[1] as well as in system and API documentation (https://docs.dimensions.ai/dsl/). Open access information and much more (citation information, funding information) and is available in the version of Dimensions that is available for free for personal use.

- The authors note that Dimensions, Mendeley and others do not offer a public open API.

  While Dimensions does offer a free end-user tool, it does not offer a full open public API. However, Dimensions does offer a free metrics API for non-commercial purposes (https://www.dimensions.ai/dimensions-apis/) and Mendeley continues to offer a free API (albeit with recent changes to some of the functionality around search functionality - https://dev.mendeley.com/).

  In addition, given the use of Google Scholar by the authors in their tool, I think that it is important to note explicitly that Google Scholar does not offer an API. Indeed, from the software source code deposited by the authors, it appears that data is scraped from the Google Scholar website contravening the terms of use of Google Scholar. This fact should be explicitly noted in the paper.

- The paper contains no critical analysis on the accuracy or success of the algorithmic approaches taken by the authors as regards fuzzy matching of papers and authors between sources.

- The paper would also be improved by including a flow diagram and description of matching approach taken by the authors.

References
1. Hook D, Porter S, Herzog C: Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics*. 2018; **3**. Publisher Full Text
2. Visser M, van Eck N, Waltman L: Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*. 2021; **2** (1): 20-41 Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
No

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
No

Competing Interests: Daniel Hook is the CEO of Digital Science, the owner of Altmetric, Dimensions, Figshare, IFI Claims, ReadCube and Symplectic. He is also a co-founder of Symplectic and a Board Member (and Treasurer) of ORCID.

Reviewer Expertise: Open Research, Bibliometrics, Sociology of Research, Theoretical Physics (Quantum Statistical Mechanics, PT-Symmetric Quantum Mechanics).

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 10 Dec 2021
Reto Gerber, University of Zurich, Zurich, Switzerland

- *However, the authors choose to take a peculiarly western-centric view of this issue. The greatest challenges of name disambiguation are typically found when authors who might not natively use a roman alphabetic system are forced to transliterate their names when* *publishing in the western-centred publishing system. This fact goes unacknowledged in this paper but would seem to be at the heart of the name disambiguation issue.*

- **Response:** Missing issue of ambiguous transliteration of names into roman alphabetic system has been added to the Introduction.

- *There are several approaches to name disambiguation but there are broadly two "schools": attended and unattended. "Attended" is where humans interact with the disambiguation and "unattended" is where humans have no role. Attended disambiguation is the focus for identity management systems like ORCID - incentives need to be aligned for authors and others to participate in this and significant work goes into understanding the motivations and concerns of researchers in order to make these types of system successful. Unattended disambiguation makes use of the data that are available in the ecosystem (including the outputs of the attended disambiguation approach) to create a calculated output. We feel that the authors should give some of this background in their paper for context. In this context, it is important to acknowledge that only one identifier system has been successful in engaging the broad global academic community and that is ORCID. Other solutions are self-acknowledged proprietary solutions that aim to solve this problem in limited contexts (Researcherid.com/Publons to improve the Web of Science data (attended disambiguation); ScopusID - unattended disambiguation approach). Dimensions explicitly leverage ORCID data in an unattended person disambiguation approach.*[1]
- **Response:** A short description of the unattended vs attended approaches of name disambiguation were added to the Introduction.

- *The authors claim that there is no standardization of unique identifiers for authors and documents.This is a very strong claim when viewed at an international ecosystem level. The majority of funders, publishers, institutions, scholarly societies and government agencies involved with research from around the world acknowledge DOIs as the key identifier for a research paper and ORCID as the principal identifier of researchers.*
- **Response:** We rewrote this statement to say that various other identifiers exist beyond the two most important ones, ORCID and DOI.

- *Some complexity lies in the authority that issues and maintains an article DOI (Crossref or DataCite in most cases but also, for example, J-Stage in Japan).*
  ORCID may currently be a "de facto standard" but the authors' statement to this effect underplays the central role played by ORCID in the scholarly infrastructure community and the extent to which ORCID is the standard with which the majority of the commercial and open infrastructure providers engage. Google Scholar provides no API and makes no claim of persistence of identifiers; authors have limited control of their profile and of privacy. Researcherid.com (the progenitor of Publons/ResearcherID) is a founding member, supporter and participant of ORCID. We suggest that the authors should ensure that acknowledgement be made of ORCID's suitability and level of adoption in academia. If the authors wish to note that ORCID is not the only standard, it would be appropriate to mention parallel efforts such as researchmap.jp in Japan and note that some countries remain reluctant to adopt ORCID as their principal identifier at this time.
- **Response:** We have rephrased parts of the text to highlight the important role of ORCID.

- *The authors claim that it is not easy to determine the provenance of data in Dimensions has derived its data from ORCID or Publons. Arguably this is not the case. Dimensions clearly states that algorithmic methods are used with the input of ORCID data.*[1] *Publons/Researchid.com data is a proprietary source that could not be used in Dimensions without explicit acknowledgement of its use.*
- **Response:** We have removed this imprecisely-worded statement.

- *The authors further claim that no information is given about the completeness of these sources (e.g. Dimensions/Web of Science/Scopus), yet significant academic work has been undertaken to understand and benchmark coverage and completeness of these sources (see reference 2 for example).*[2]
- **Response:** We have now removed claim about completeness of these sources since in our case, we are more interested in author-level completeness, e.g., are publications listed in Google scholar missing in ORCID for a specific author?

- *We agree with the authors that Open Science is indeed critical. This is why Dimensions, Web of Science, Scopus and others contain information on Open Access statuses of articles, provided by Unpaywall (https://unpaywall.org/integrations). Dimensions also contains full listings of preprint articles from many preprint servers including ArXiv.org, BioArXiv, the Center for Open Science, PeerJ and so on. As such we find the authors' comment that these things are "not taken into account" to be misleading.*
- **Response:** We removed the imprecisely-worded claim about Dimensions' open access status and preprints.

- *The authors then appear to contradict themselves by claiming that preprints and publications cause duplicate entries in these systems. In Dimensions, where the current reviewer has the most experience, publications from preprints are linked to final publications where this information is available, see for example https://app.dimensions.ai/details/publication/pub.1118864658?and_facet_researcher=ur.01123321343.51. We agree that manual intervention is often needed in these cases.*
- **Response:** This contradiction was removed by acknowledging the listing of preprints on other platforms, such as Dimensions.

- *The authors claim that a free tool does not exist to explore metadata brought together from multiple sources, however, this does not acknowledge the rich lineage of free tools written to serve institutional use cases in author disambiguation and metadata aggregation including: VIVO (https://duraspace.org/vivo/) and Catalyst Profiles (http://profiles.catalyst.harvard.edu), both funded by the NIH; ImpactStory (e.g. https://profiles.impactstory.org/u/0000-0001-6728-7745/publications); and, ReCiter (https://github.com/wcmc-its/ReCiter). The authors should endeavour to situate their work in the context of this prior work.*
- **Response:** We have added the above mentioned tools to the Introduction text and now describe our application in the context of those tools.

- *Symplectic Elements is, as the authors state, a commercial system focused on institutional use cases that meets this need. However, the authors fail to acknowledge that outside*

STEM subjects, coverage of research outputs becomes more challenging. The strength of commercial tools is often that, as they must meet institutional needs, work is put into diversifying coverage of different output types beyond the journals and conference proceedings that STEM subjects favour. While composing a faithful representation of a STEM academic and their output can be challenging from disparate data sources, doing the same for a non-STEM academic can be an order of magnitude more difficult. We believe that it is also fair to acknowledge that institutional engagement is not unimportant to improving the overall data quality in the bibliographic ecosystem.

- **Response:** We have now added Introduction text, highlighting that commercial systems can make a big impact, especially in institutional cases.

- *The authors also appear to be unaware that Dimensions does bring records together from multiple sources including author information and that this is transparently documented in scholarly articles written by the Dimensions team,[1] as well as in system and API documentation (https://docs.dimensions.ai/dsl/). Open access information and much more (citation information, funding information) and is available in the version of Dimensions that is available for free for personal use.*
- **Response:** We are aware of this functionality within Dimensions, but it is not open for use beyond personal use and most of the data that we are interested in is already in the public domain.

- *The authors note that Dimensions, Mendeley and others do not offer a public open API. While Dimensions does offer a free end-user tool, it does not offer a full open public API. However, Dimensions does offer a free metrics API for non-commercial purposes (https://www.dimensions.ai/dimensions-apis/) and Mendeley continues to offer a free API (albeit with recent changes to some of the functionality around search functionality - https://dev.mendeley.com/).*
- **Response:** We now removed the imprecisely-worded statement about closed APIs of Dimensions and Mendeley.

- *In addition, given the use of Google Scholar by the authors in their tool, I think that it is important to note explicitly that Google Scholar does not offer an API. Indeed, from the software source code deposited by the authors, it appears that data is scraped from the Google Scholar website contravening the terms of use of Google Scholar. This fact should be explicitly noted in the paper.*
- **Response:** We have now added a statement that web scraping is done to retrieve information from Google Scholar and that this may contravene the terms of use of Google Scholar.

- *The paper contains no critical analysis on the accuracy or success of the algorithmic approaches taken by the authors as regards fuzzy matching of papers and authors between sources.*
- **Response:** A description of the fuzzy matching mechanism that was applied was added together with a statement about accuracy of matching.

- *The paper would also be improved by including a flow diagram and description of matching approach taken by the authors.*

- ○ **Response:** The description of the approach for matching of publications has been expanded.

Competing Interests: No competing interests were disclosed.