SOFTWARE TOOL ARTICLE

REVISED pubassistant.ch: consolidating publication profiles of researchers [version 3; peer review: 3 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

v3 **First published:** 30 Sep 2021, **10**:989
https://doi.org/10.12688/f1000research.73493.1
Second version: 20 Dec 2021, **10**:989
https://doi.org/10.12688/f1000research.73493.2
Latest published: 13 Apr 2022, **10**:989
https://doi.org/10.12688/f1000research.73493.3

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

Approval Status ✓ ✓ ✓

	1	2	3
version 3 (revision) 13 Apr 2022	✓ view	✓ view	
version 2 (revision) 20 Dec 2021	? view	✗ view	✓ view
version 1 30 Sep 2021	✗ view		

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

2. **Paul Albert**, Weill Cornell Medical College, New York, USA
Curtis L. Cole, Weill Cornell Medical College, New York, USA

3. **Griffin M Weber**, Harvard Medical School, Boston, USA
Beth Israel Deaconess Medical Center, Boston, USA

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 3; peer review: 3 approved]** F1000Research 2022, **10**:989 https://doi.org/10.12688/f1000research.73493.3

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

REVISED **Amendments from Version 2**

Several points in the introduction were made more specific and the description of rationale of this work has been extended. A clearer description of the use case and the target user has been added. Furthermore, a small evaluation on the accuracy of matching publications was conducted.

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. Updating data in ORCID is possible using a variety of methods, such as through CRIS (current research information) systems, as auto updates between Crossref and ORCID, linking to Dimensions, among others.

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 combine bibliographies and bibliometrics for a wide variety of use cases such as evaluation, compliance (e.g. OA), grant writing, literature review and keeping professional web profiles updated. Furthermore, many of the available tools are not made for individual authors but rather operate on the department, institutions or even country level. Although those tools allow research institutions to curate research profiles to meet some of the aforementioned use cases, they are usually not designed for individuals to curate their own individual profiles outside an institutional context. 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. It is worth noting that our tool is not meant in any way for evaluation of researchers and that initiatives such as DORA[17] and the Leiden Manifesto[18] represent important considerations toward responsible research evaluation.

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. In other words, pubassistant.ch is intended for researchers who want to cleanup their publication profiles across multiple platforms and are interested in the open access status of their publications. One specific use case would be to find publications where the ORCID was not included and therefore is not listed in the online ORCID profile, but for example found and listed by Google Scholar.

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 option to search for bibliometrics, obtained from the NIH Open Citation Collection using iCite,[19] 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

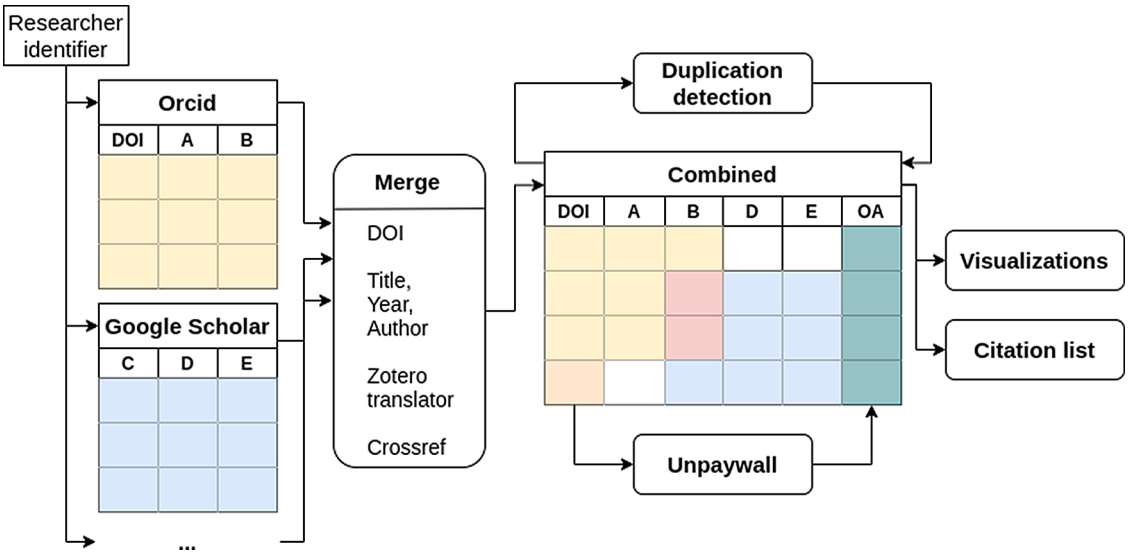

Figure 1. Overview of the data processing. The identifiers given by the user are used to obtain the data from each platform independently. The data is then merged and the open access status (column OA) is obtained using the Digital Object Identifier (DOI). Furthermore duplicates are detected by comparing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

from Zotero (Zotero translator, i.e., web scraping)[20] or Crossref (i.e., query the available metadata to obtain a DOI).[21] 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. The accuracy of detecting duplicates using a small test dataset (n=6929) of reasearch articles with an associated preprint on bioRxiv[22] was 0.72. No other 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,[23] 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),[24] 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)[25] and shiny (Version 1.6.0),[26] 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),[27] https://github.com/jkeirstead/scholar (Google Scholar)[28] or https://docs.ropensci.org/rentrez (Pubmed)[29] 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

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).

separate container. As already stated, the PostgreSQL service is not strictly needed, but substantially increases retrieval speed of the OA status.

Use case

As mentioned above, a key use case is the comparison of publications listed in different public profiles and the detection of possible duplicated entries that allows the respective profiles to be updated. 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[30] (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. An useful addition could be the integration of the API from OpenAlex for citation information and publication metadata.[31]

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. We thank the reviewers for valuable suggestions.

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. DORA, San Francisco Declaration on Research Assessment.
Reference Source

18. Hicks D, Wouters P, Waltman L, *et al.*: **Bibliometrics: The Leiden Manifesto for research metrics.** *Nature.* 2015/04/01.
Publisher Full Text

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

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

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

22. Fu DY, Hughey JJ: **Releasing a preprint is associated with more attention and citations for the peer-reviewed article.** *eLife.* December 2019.
Publisher Full Text

23. 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**

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

25. 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

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

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

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

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

30. 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**

31. OpenAlex: **The open catalog to the global research system.**
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 3

Reviewer Report 04 May 2022

https://doi.org/10.5256/f1000research.126603.r134743

✔ **Paul Albert**
Wood Library, Weill Cornell Medical College, New York, NY, USA
Curtis L. Cole
Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, USA

The authors did address several of our concerns and clearly disagree with much of our feedback. Some of the improvements they made are not reflected in the figures (e.g. Fig 2) so with little effort they could make more improvements. But we agree with the other reviewer's observation that the code itself might be useful and should be shared.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: We have both collaborated on publication tracking and disambiguation systems, such as VIVO and Reciter.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 14 April 2022

https://doi.org/10.5256/f1000research.126603.r134741

✔ **Daniel W Hook**
Digital Science, London, UK

I thank the authors for making further updates in response to my comments. In light of their

diligent changes, I am pleased to recommend this article to others.

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 confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 17 February 2022

https://doi.org/10.5256/f1000research.80816.r120551

Griffin M Weber
[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

pubassistant.ch is a website that helps users to manage the list of publications that they have authored. Publications can be imported from several sources, including ORCID and Google Scholar. Deduplication is handled by matching on DOI or similar titles (using Levenshtein distance). The software automatically imports metrics on the publications, such as their open access status and number of citations. Various visualizations show summary information about the publications. The publication list can be exported in several formats.

There are other software programs that provide similar functionality: reference management, open access reporting, and bibliometric analyses. Though, as the authors correctly point out, many of these are products that institutions acquire and make available only to their community. These include Dimensions, Web of Science, VIVO, and Profiles RNS. However, OpenVIVO and Mendeley allow anyone to sign up. Also, users of ORCID can import publications from Scopus.

The individual components of pubassistant.ch are not novel: the authors use public APIs and well known data sources, and their deduplication process uses a common algorithm. However, when put together, the website presents the data in a nice way. I liked the timelines and UpSet plot showing the different open access levels of the publications. It was easy to interpret the visualizations and understand how to use their settings.

Unfortunately, some features of the website are slow due to limits on the APIs the software calls; and, I encountered some bugs, especially with the Citations tab either not loading or producing an error. The reliance on web-scraping Google Scholar is problematic. It should be noted that publications from Google Scholar can be exported in BibTeX format and then imported into ORCID, which is alternative way of merging Google Scholar and ORCID publication lists, though pubassistant.ch

I'm not sure who the users of this website will be. Investigators managing their ORCID and Google Scholar profiles might not like having to go to a third website. Funding agencies, librarians, and institution administrators would probably want the open access and citation metrics, but having to enter IDs manually for each author one at a time and waiting for the results could be too time consuming.

I think the biggest impact of pubassistant.ch could be the code itself. The website is a great demonstration of how to (1) query the APIs some frequently used bibliometric tools, (2) integrate the data that are returned, and (3) visualize the results in some interesting ways. I can imagine this code being extended by other developers or even embedded in other software platforms in the future.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

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?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bibliometric analysis, social network analysis, research networking systems

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Apr 2022

Reto Gerber, University of Zurich, Zurich, Switzerland

Unfortunately, some features of the website are slow due to limits on the APIs the software calls; and, I encountered some bugs, especially with the Citations tab either not loading or producing an error.

- **Some of the less stable components of the applications have been deactivated.**

The reliance on web-scraping Google Scholar is problematic. It should be noted that publications from Google Scholar can be exported in BibTeX format and then imported into ORCID, which is alternative way of merging Google Scholar and ORCID publication lists, though pubassistant.ch

I'm not sure who the users of this website will be. Investigators managing their ORCID and Google Scholar profiles might not like having to go to a third website. Funding agencies, librarians, and institution administrators would probably want the open access and citation metrics, but having to enter IDs manually for each author one at a time and waiting for the results could be too time consuming.

- **The target user could be a researcher who wants to align their multiple public profiles in a systematic and semi-automatic way that also gives information about timelines and open access. The manuscript has been updated to make it clearer who the target user is.**

Competing Interests: No competing interests were disclosed.

Reviewer Report 31 January 2022

https://doi.org/10.5256/f1000research.80816.r120552

Paul Albert
Wood Library, Weill Cornell Medical College, New York, NY, USA

Curtis L. Cole
Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, USA

This paper describes pubassistant.ch, a freely available website using open-source code. Pubassistant.ch is designed to perform the following functions:

- Allows users to manually import data from the following services: Google Scholar, Publons, PubMed, and ORCID.
- Disambiguate (deduplicate) articles.
- Provide the open access status of each article.

- Allow users to download citation citation data as Bibtex for updating these profiles.

We applaud the initiative to design and deploy a tool meant to improve researchers' work lives.

We agree with the first reviewer's careful comments and appreciate that many of the concerns were addressed in this revision. However, one key problem remains, the unclear perspective and motivation for the tool. A key question for any software tool is whether it addresses an unmet need. It's not quite clear what this tool could uniquely perform. Is it that Pubassistant allows users to easily sync publication lists between profile systems? Maybe the tool provides readily available statistics on open access status of a given author's output. “Who exactly is the target user?” I don’t think the author would agree this tool is designed for unaffiliated researchers.

While this revision does now mention many other contributions in this area, it remains unclear what niche pubassistant.ch is filling within this larger space. Also, as noted by the first reviewer, Vivo should be written VIVO.

While any software, especially open source, is often used in ways not imagined by the original author, without a key use case the success or failure of the program to meet its goal cannot be evaluated. This points to one of the key problems with the paper which is the lack of any evaluation other than a high level subjective sense that it works.

In our opinion, the functionality in pubassistant.ch is not ready for deployment calling into question even that subjective assessment. It has several bugs and could benefit from user feedback and iteration. For example:

- This reviewer could not perform a PubMed search until clicking on the "Example PubMed Query" button.

- The "Time" tab displays two histograms. The top graph would appear to be a raw count and the bottom graph is a percentage by year. But this distinction is not explained by the site either via a text-based description or using any labels on the axes. In this reviewer's estimation, only one graph is warranted.

- The units on the "Citations" graph are sometimes unnecessarily unconventional. One year is "2,016.5". Years do not typically include commas or periods which makes the display confusing.

- The x-axis dropdown in the "Citations" tab shows up behind the "Copy", "CSV", and "Excel" buttons.

- After attempting to click on the "Download Bibtex citation" button, this user got a "503 Service Unavailable" error. This did not allow this reviewer to verify whether the tool can allow users to populate publication lists into a tool like Google Scholar.

- The integration with Sci-Hub has a certain illicit appeal. But when clicking on the "Get Full text links" button, nothing happens.

- The "Apply selection" button is confusing. One assumes this applies to the filter in the left hand column. Wouldn't this button be more appropriate to include immediately below the

filters?

- When this reviewer clicks on "Remove duplicates" and then "Apply selection," it removes all data from the table including non-duplicates.

- In the "Table of publications" section, the label "cid" is used. What does this label mean? One assumes it's a unique ID that Google uses. A more descriptive label might be more appropriate.

- The method for checking for duplicates is to look for duplicate article titles within the same author's namespace.

- As designed, the article disambiguation can run into problems when there are no DOIs. This is a concern in cases where publication records are imported from Google Scholar as Pubassistant.ch appears to only provide article title and Google Scholar's local identifier. The authors indicate that they use fuzzy matching on article title with an author's namespace to perform deduplication. In the context of a research article, such an assertion should be tested and validated. Here's an example where such an approach would fail. All the following articles were written by Robert S. Brown, a faculty member at Weill Cornell Medicine, and all of them have article titles that are non-unique within Robert Brown's namespace: See here.

This is an interesting work that has the potential to be a real contribution, but some additional development is probably warranted before publication and distribution.

Is the rationale for developing the new software tool clearly explained?

No

Is the description of the software tool technically sound?

Yes

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: Paul Albert and Curtis Cole are both active in the VIVO consortium and key personnel in the development of ReCiter. Both are open source tools in this space.

Reviewer Expertise: We have both collaborated on publication tracking and disambiguation

systems, such as VIVO and Reciter.

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

Author Response 11 Apr 2022

Reto Gerber, University of Zurich, Zurich, Switzerland

We agree with the first reviewer's careful comments and appreciate that many of the concerns were addressed in this revision. However, one key problem remains, the unclear perspective and motivation for the tool. A key question for any software tool is whether it addresses an unmet need. It's not quite clear what this tool could uniquely perform. Is it that Pubassistant allows users to easily sync publication lists between profile systems? Maybe the tool provides readily available statistics on open access status of a given author's output. "Who exactly is the target user?" I don't think the author would agree this tool is designed for unaffiliated researchers.

- **The target user could be a researcher who wants to align their multiple public profiles in a systematic and semi-automatic way that also gives information about timelines and open access. The manuscript has been updated to make it clearer who the target user is.**

While this revision does now mention many other contributions in this area, it remains unclear what niche pubassistant.ch is filling within this larger space. Also, as noted by the first reviewer, Vivo should be written VIVO.

While any software, especially open source, is often used in ways not imagined by the original author, without a key use case the success or failure of the program to meet its goal cannot be evaluated. This points to one of the key problems with the paper which is the lack of any evaluation other than a high level subjective sense that it works.

- **Pubassistant.ch relies heavily on other tools to interact with the various APIs (R packages, such as rorcid), retrieve open access information (Unpaywall), and retrieval of DOI from metadata (Crossref, Zotero translator). We assume that those heavily used tools work as intended. The main part of the applications that need evaluation in our opinion is the matching of publications. A note on the accuracy of detecting duplicates using a small dataset consisting of publications with associated preprints was added.**

In our opinion, the functionality in pubassistant.ch is not ready for deployment calling into question even that subjective assessment. It has several bugs and could benefit from user feedback and iteration. For example:

- *This reviewer could not perform a PubMed search until clicking on the "Example PubMed Query" button.*
  - **Added placeholder text to indicate that the button on the right has to be pressed first.**

- *The "Time" tab displays two histograms. The top graph would appear to be a raw count and the bottom graph is a percentage by year. But this distinction is not explained by the*

site either via a text-based description or using any labels on the axes. In this reviewer's estimation, only one graph is warranted.
- **The missing axis labels have been added.**

- *The units on the "Citations" graph are sometimes unnecessarily unconventional. One year is "2,016.5". Years do not typically include commas or periods which makes the display confusing.*
  - **The experimental "Citations" tab has been deactivated.**

- *The x-axis dropdown in the "Citations" tab shows up behind the "Copy", "CSV", and "Excel" buttons.*
  - **The experimental "Citations" tab has been deactivated.**

- *After attempting to click on the "Download Bibtex citation" button, this user got a "503 Service Unavailable" error. This did not allow this reviewer to verify whether the tool can allow users to populate publication lists into a tool like Google Scholar.*
  - **This bug has been fixed.**

- *The integration with Sci-Hub has a certain illicit appeal. But when clicking on the "Get Full text links" button, nothing happens.*
  - **This feature has been deactivated.**

- *The "Apply selection" button is confusing. One assumes this applies to the filter in the left hand column. Wouldn't this button be more appropriate to include immediately below the filters?*
  - **The "Apply selection" button does filter entries selected in the list below. The name of the button has been changed and additional hints were added to make this clearer.**

- *When this reviewer clicks on "Remove duplicates" and then "Apply selection," it removes all data from the table including non-duplicates.*
  - **This behavior is as intended if no rows in the table have been selected. The same will happen without pressing "Remove duplicates" or with pressing "Show detected duplicates".**

- *In the "Table of publications" section, the label "cid" is used. What does this label mean? One assumes it's a unique ID that Google uses. A more descriptive label might be more appropriate.*
  - **"Cid" is an identifier for publications used by Google. This label has been extended to make it clearer.**

As designed, the article disambiguation can run into problems when there are no DOIs. This is a concern in cases where publication records are imported from Google Scholar as Pubassistant.ch appears to only provide article title and Google Scholar's local identifier. The authors indicate that they use fuzzy matching on article title with an author's namespace to perform deduplication. In the context of a research article, such an assertion should be tested and validated. Here's an example where such an approach would fail. All the following articles were

written by Robert S. Brown, a faculty member at Weill Cornell Medicine, and all of them have article titles that are non-unique within Robert Brown's namespace: See here.

- **It is true that the finding of duplicates with only the title will not always work and with the available data (not all available data is shown in the table), such as coauthors, publication date, open access status, etc., there are certainly improvements possible. The duplication identification performed in Pubassistant.ch is mainly intended to assist in finding duplicates and manual checking of results can be easily done because of the interactivity provided in the app.**

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 January 2022

https://doi.org/10.5256/f1000research.80816.r116162

?

Daniel W Hook
Digital Science, London, UK

We thank the authors for taking the time to improve their manuscript.

We make a number of optional suggestions for changes below that we believe would further improve and strengthen the paper:

1. The authors note that “Many tools, both commercial and free, exist to explore certain aspects of bibliographies”. While this is true it does not, perhaps, get at the core of what is intended. We suggest that the authors consider revision of this comment. We suggest that a text similar to the following may be appropriate: “Many tools, both commercial and free, exist to bring data bibliographic and bibliometric data together to meet a multitude of different use cases including: evaluation, compliance (e.g. OA), grant writing, literature review and keeping professional web profiles updated.”

2. The authors go on to state: “Some of the existing commercial tools including Elements (from the company Symplectic) and Dimensions, both of which are mainly intended for institutional use; but, especially Dimensions also offers functionalities for authors to explore their bibliographies”. We suggest that the authors consider revision along the lines of: “There is a distinct segment of research analysis and compliance tools that provides functionality to research institutions to allow research profiles to be curated to meet some, but not all, of the aforementioned use cases. Notably, these systems do not appear to have been designed (nor are individual subscriptions or versions available) to meet the needs of individuals who wish to curate their own individual profiles outside an institutional context.

A key use case that has emerged for contemporary research is to have a detailed understanding of their bibliometric rather than just their bibliographic profile. Given the growth of availability of bibliometric information and the growing need to use bibliometric information as a tool for academic decision making in recent years, the lack of software to bring this information together for the individual appears to be an omission in the tools landscape." It may also be helpful to note that tools such as EndNote, Papers and others also appear not to meet this need.

We believe that the authors should be more forthright in identifying the key use cases for their tool as this strengthens the rationale for their software.

3. The authors observe that "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." We suggest that it would be helpful to mention that most CRIS systems now include the ability to upload affiliation information to ORCID records from institutional systems; that Crossref/ORCID auto-update functionality (https://support.orcid.org/hc/en-us/articles/360006971293-Auto-updates-in-third-party-systems-Crossref) and that the ability to link ORCID to a Dimensions account (https://www.dimensions.ai/blog/digital-science-announces-dimensions-integration-of-orcid/) are all possible routes to updating data in ORCID. In addition, ORCID has released tools to ensure that full provenance is captured in these updates (https://info.orcid.org/documentation/api-tutorials/api-tutorial-add-and-update-data-on-an-orcid-record/#easy-faq-2692).

4. We believe that Vivo should be VIVO.

5. The authors may wish to consider mentioning that the new API from OpenAlex (https://openalex.org/) could be a useful addition to their data sources by providing improved article metadata and citation information within an appropriate data licensing context.

6. It would be healthy and helpful for the context of the article and of the software to include mention of DORA, the Leiden Manifesto and the responsible research evaluation movement in the motivation for the software. Bibliometrics can be interesting and informative but do not provide a comprehensive overview of an academic.

Is the rationale for developing the new software tool clearly explained?

No

Is the description of the software tool technically sound?

No

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

No

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 confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 11 Apr 2022

Reto Gerber, University of Zurich, Zurich, Switzerland

The authors note that "Many tools, both commercial and free, exist to explore certain aspects of bibliographies". While this is true it does not, perhaps, get at the core of what is intended. We suggest that the authors consider revision of this comment. We suggest that a text similar to the following may be appropriate: "Many tools, both commercial and free, exist to bring data bibliographic and bibliometric data together to meet a multitude of different use cases including: evaluation, compliance (e.g. OA), grant writing, literature review and keeping professional web profiles updated."

- **This part of the text has now been updated to be more specific.**

The authors go on to state: "Some of the existing commercial tools including Elements (from the company Symplectic) and Dimensions, both of which are mainly intended for institutional use; but, especially Dimensions also offers functionalities for authors to explore their bibliographies". We suggest that the authors consider revision along the lines of: "There is a distinct segment of research analysis and compliance tools that provides functionality to research institutions to allow research profiles to be curated to meet some, but not all, of the aforementioned use cases. Notably, these systems do not appear to have been designed (nor are individual subscriptions or versions available) to meet the needs of individuals who wish to curate their own individual profiles outside an institutional context. A key use case that has emerged for contemporary research is to have a detailed understanding of their bibliometric rather than just their bibliographic profile. Given the growth of availability of bibliometric information and the growing need to use bibliometric information as a tool for academic decision making in recent years, the lack of software to bring this information together for the individual appears to be an

omission in the tools landscape." It may also be helpful to note that tools such as EndNote, Papers and others also appear not to meet this need.

- **This part of the text has also been updated to be more specific.**

We believe that the authors should be more forthright in identifying the key use cases for their tool as this strengthens the rationale for their software.

- **One specific use case of pubassistant is to find publications in one public profile that are missing in another profile. This frequently happens when, for instance, the ORCID is missing (e.g., a collaborator forgets to add yours) and thus the public profile of ORCID will not show the respective publication although it might be recognized (automatically) by Google Scholar. A clearer description of the key use case has now been added to the manuscript.**

The authors observe that "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." We suggest that it would be helpful to mention that most CRIS systems now include the ability to upload affiliation information to ORCID records from institutional systems; that Crossref/ORCID auto-update functionality (https://support.orcid.org/hc/en-us/articles/360006971293-Auto-updates-in-third-party-systems-Crossref) and that the ability to link ORCID to a Dimensions account (https://www.dimensions.ai/blog/digital-science-announces-dimensions-integration-of-orcid/) are all possible routes to updating data in ORCID. In addition, ORCID has released tools to ensure that full provenance is captured in these updates (https://info.orcid.org/documentation/api-tutorials/api-tutorial-add-and-update-data-on-an-orcid-record/#easy-faq-2692).

- **The set of possible methods to automatically update data in ORCID has now been added to the manuscript.**

We believe that Vivo should be VIVO.

- **Vivo has been renamed to VIVO in the manuscript.**

The authors may wish to consider mentioning that the new API from OpenAlex (https://openalex.org/) could be a useful addition to their data sources by providing improved article metadata and citation information within an appropriate data licensing context.

- **Possible extension with OpenAlex has been added to the discussion.**

It would be healthy and helpful for the context of the article and of the software to include mention of DORA, the Leiden Manifesto and the responsible research evaluation movement in the motivation for the software. Bibliometrics can be interesting and informative but do not provide a comprehensive overview of an academic.

- **A note about DORA and the Leiden manifesto has now been added to the introduction.**

Competing Interests: No competing interests were disclosed.

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.