Designing Community Tracking Indicators for Open and Inclusive Scholarship

Altman, Micah Massachusetts Institute of Technology, USA | escience@mit.edu

ABSTRACT

There is evidence that scholarly processes have bias and create barriers to inclusion; more openness in scholarly communication is needed. Progress towards a better scholarly ecosystem requires comparable, reliable measures of the desired attributes of a better system. This paper describes an initiative in progress to produce standardized indicators that describe the volume and types of open science output systematically over time, using existing open data sources. We describe a replicable to clean, integrate, code, and analyze these sources to enable continuous publication of indicators. And we report on early results from this initiative, demonstrating how these indicators can go beyond 'overall impact' measures to advance the understanding of who is, and who is not, participating in open scholarship.

KEYWORDS

Open Access; Open Scholarship; Diversity, Equity, and Inclusion; Altmetrics; Measurement

INTRODUCTION

To make reliable progress toward a socially-desirable scholarly ecosystem the research community requires ongoing, systematic, and trusted measures of inclusivity, equity, durability, and sustainability. Environmental scans such as the Grand Challenges Summit (Altman *et* al. 2018), supported by the Mellon Foundation, and the ACRL report on Open and Equitable Scholarly Communications (Maron *et* al. 2019) have drawn attention to the need to measure and integrate equity and inclusion into the scholarly ecosystem. There is convincing evidence, based on point-in-time studies, that scholarly processes and outputs have substantial bias and/or create barriers to inclusion (Lee *et* al. 2013) and that more openness in science and scholarly communication is needed. Assessing progress towards a better scholarly ecosystem requires standard, reliable measures of the desired attributes of a better system.

While there is growing literature that uses bibliometric data to characterize inclusion in science, almost all of the work consists of one-shot analyses of specific dimensions of inclusion in a selected area of scholarship during a limited period. There are currently several projects that are produced measures for ongoing analysis of scholarly production (e.g. see Lepori, *et* al. 2009) and only two that measure scholarly inclusion. Both projects are prototypes, and target a narrow scope of scholarly content: BASE (Summann et al. 2020) aims to use OAI-PMH metadata harvesting to track statistics on the size of collections in institutional repositories worldwide, and reports activity by country. ORION (Stathoulopoulos et al. 2020) is a prototype for interactive visualization patterns of metadata describing publications in the life sciences in Microsoft Academic graph: it provides map visualization by gender and region.

While it is routine to use publisher-produced citation indicators for the 'impact' of scholarly communication, institutional decision making, and research policy, there is currently no comparable public data that summarizes diversity in who is citing, producing or accessing the same communications. Despite recent advances in making scholarly communication more openly available, few systematic measures are available to track, compare, or evaluate diversity and inclusion in open scholarship. As a consequence, both existing and proposed interventions to improve scholarly practices, norms of scholarly communities, and attitudes of scholars are often in dispute; and institutions lack benchmarks for their local communities and policies.

OPEN QUESTIONS ABOUT PARTICIPATION IN OPEN SCHOLARSHIP

Who is represented in open science and open scholarly communications? This question provides a necessary foundation for causal analysis and targets interventions in practice. This broad question can be divided into a three areas that are empirically measurable with the current state of available data:

- What is the prevalence of members of different scholars of different genders and nationalities in open-scholarship and open-science initiatives, and outputs?
- Where are open-scholarship and open-science outputs that are produced with and by group members used in the scholarly ecosystem?
- How does group prevalence in open-scholarship and science, and the use of open access products, vary within the scholarly ecosystem?

The first phase of the project aims to fully operationalize two measurable indicators for each question. The indicators be designed to inform researchers in analyzing trends in scholarship, and will inform leaders in developing institutional policy and strategies.

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

Quality Dimensions

The usefulness of potential answers to these questions depends on their temporal regularity, measured accuracy, comparability with other measures, and their reproducibility.

*Temporal regularity i*s required to detect trends (and seasonality) in the scholarly ecosystem and as a building block for measuring the effects of different interventions and events. The project has identified target a set of core data sources, described in the next section, hat are frequently or continuously updated and construct an automated retrieval and linking pipeline so indicators may be efficiently produced at regular intervals.

Measured accuracy is required to reliably distinguish systematic differences from statistically random variation. Accuracy will be managed using a total survey error approach (Groves, et. *al* 2010) that produces honest measures of uncertainty by managing and measuring error at stage of the estimation process, including related to measurement, linkage, coverage, and sampling.

Comparability is required to coherently combine the indicators with independent measures collected by other projects and surveys. While comparability is inherently contextual we aim align these indicators with other independent measurement frameworks such as those developed by the *Center for Open Science* and *National Center for Science and Engineering Statistics*.

Replicability is necessary both to ensure that the outputs are reliable, and to enable indicators and analyses to be updated efficiently over time. In the following sections we describe how an open source data pipeling that will enable core reports, indicators, and databases to be automatically, created, and updated.

UTILIZING OPEN DATA TO TRACK PARTICIPATION

The metadata describing open access and science is incomplete, scattered, and imperfect. (see for a review, Gregg et al. 2019) Notwithstanding, there is much that is openly available. Table 1 describes a core set of data sources that will be used to develop indices. We anticipate that other sources will be added in later project stages.

	ORCID	**DOAJ**	**DOAB**	**I40C**	**ROAR MAP**	**PLOS Articles**	**Open Editors**	**OSF.io preprints**
Overview	Largest registry of research identifiers	Largest database of open access journals	Largest database of open access books	Largest open network of citation information	The largest repository of institutional open access policies	Most detailed open article contributor information openly available	Tools for extracting journal editor information via web mining	Preprint database spans broadest range of fields
Directly Recorded Data	Researcher features: name, institution Researcher outputs	Journal features Journal article features	Open Monograph Features	Publication citation network Authorship of articles	Institutional open access policy	Researcher features: contributor roles; downloads.	Editor Names, Editor Affiliation, Related Journals, Editorial Role	Authorship of preprints
Mineable Features	Region, gender, race/ethnicity, career stage	Editorial board membership	Author gender, nationality	Researcher gender, nationality Bibliometrics	Policy related to equity and inclusion	Researcher gender, nationality	Gender, Nationality, Journal Characteristics (through linked data)	Researcher institutions engagement with preprints
How it links with other sources	Institution, journal, publication, researcher	Journal, publication, researcher	Author Name, work ISBN	Publication, researcher		Institution, journal, publication, researcher	Journal name, ISSN, Researcher Name	Researcher ORCID, Research Name

Table 1. Data Sources for First Phase

Each of these sources is well-established, regularly updated, provides documented APIs, and has committed to an open-license. (We have excluded sources such as Publons, and Scopus, that while informative, do not provide structure data available under open licenses.) While no single source is critical, in aggregate the databases capture a range of open outputs (reviewer activity, editorial activity, publications, software), forms of impact and recognition

(citations, grants, publication downloads), and contributor characteristics (contributor role, institution, region, gender, ethnicity, career stage).

Although a substantial amount of data on open access publications and activities is available without licensing fees, it is still complicated to interpret and evaluate without specialized skills because creating a reliable set of measures requires many steps: locating multiple data sources; interacting with different APIs and protocols; converting data across multiple formats; linking data with overlapping coverage, aggregated at different levels, and collected at different frequencies; and selecting and constructing comparable measures.

The proposed project addresses this complexity by encoding expert knowledge about each data source and measure in modular, open source, processing pipelines. We describe these in the next section.

AN OPEN SOURCE FRAMEWORK FOR RELIABLE DATA DASHBOARDS

The project is developing automated, repeatable data science pipeline to retrieve, clean, link, and normalize data from a set of open repositories of information. The data will be augmented through automated coding (e.g. application of gazetteer services to estimate region; and of name-matching to estimate contributor gender). Then the data will be run through a cross-sectional analysis to derive population-level statistics and estimate trends.

The framework incorporates dozens of open-source components—too numerous to include here. Figure 1 provides an overview of the categories of software component:

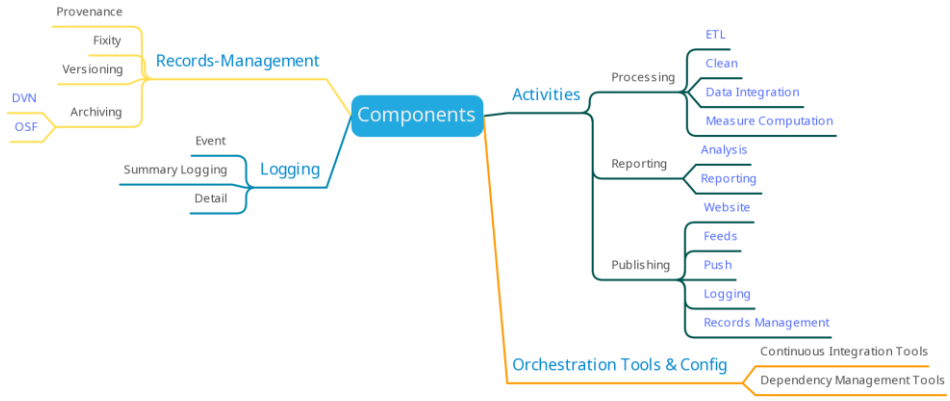

Figure 1. Categories of Open Source Software Components for Processing and Publishing Framework

Our approach to technical implementation is based on high-level orchestration tools for continuous integration (see Krafcyzk *et* al. 2019 for a review), the 'tidy' data science framework (Grolemund and Wickham 2017) for data processing, and interactive publication using executable scientific notebooks (see Konkol *et* al. 2020), and rendering through open data visualization libraries such as plot.ly (see Sievert 2020).

EARLY RESULTS—GOING BEYOND CITATION TRADITIONAL CITATION METRICS

The following examples illustrate the type of activities that can be characterized using the data sources described above, and some associational patterns emerging from descriptive analysis. Specific methodology is described in detail in the referenced articles. Preliminary work analyzing authorship in open journals and open-monographs (Altman 2021; Altman & Cohen 2021) were produced using a subset of the open-source tools and open data targeted in the project. The resulting self-contained, reproducible, open-source publications demonstrate the creation of pilot indicators, as well as the capability to support interactive data tables and data visualization. These examples show the potential for a continuously updated set of integrated data to examine patterns suggested by point-in time analyses such as (Lee *et* al. 2013).

The exploratory analysis described in Altman (2021) suggests that that women have been consistently underrepresented as authors of open monographs since 2011 (Table 2) and that author-paid book-publishing fees have declined substantially in the last decade.

Any Female Authors	Opened Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Totals
false		60.4%	57.5%	58.1%	64.8%	62.4%	63.0%	69.8%	60.0%	68.7%	56.9%	64.4%
true		39.6%	42.5%	41.9%	35.2%	37.6%	37.0%	30.2%	40.0%	31.3%	43.1%	**35.6%**
Totals		**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 2. Monograph Authorship Trends by Gender

The summative analysis in Altman & Cohen 2021, integrates multiple data sources to impute the gender of hundreds of thousands of journal editors, and creates indicators of the diversity of editorial boards in over fourteen thousand journals. This analysis, suggests that open-access journals are associated with lower gender diversity and more international diversity than their closed-access counterparts; editorial boards on average are disproportionately male and US/UK-centric; and diversity of editorial boards varies substantially by discipline (Figure 2).

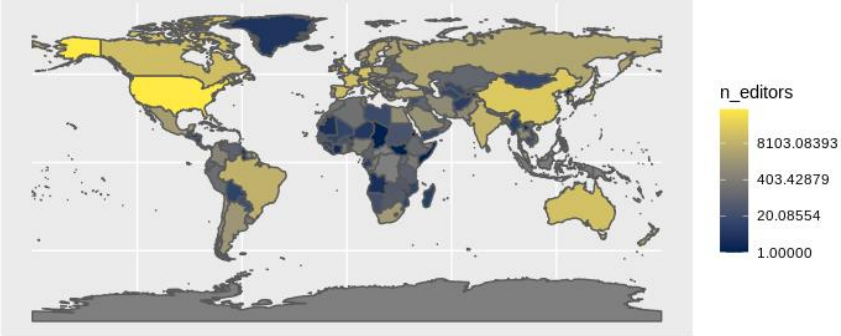

Figure 2. Concentration of journal editors geographically

Methodology for imputing gender, although used widely in bibliometric studies, is evolving and requires careful evaluation and validation. This method is intended for aggregate analysis and not for individual-level analysis—e.g. the assignment of a pronoun to an author. Although Building a robust, comprehensive, comparable, reliable set of indicators will require more development including: developing measures that are standardized across data sources, computing and monitoring a standardized set of data quality indicators, cross-validating the results and generating reliable measures of statistical uncertainty, automatically monitoring data sources for changes, packaging reusable code as public libraries and disseminating them through open archives , developing documentation, engaging in training and outreach, and tracking usage of the data, reports, and tools for evaluation. What this exploratory analysis demonstrates is that open data sources can be used to analyze inclusion in an open and reproducible way—and that such analysess can yield new and important insights.

CONCLUSION

Diversity, equity, and inclusion are core values of librarianship and the field of information science. For over a decade practitioners and scholars in this field have been leaders in the advance of the open access movement. Increasing adoption of open access and open science has highlighted the need to understand how 'open' these practices are to participation from a diverse community of scholars. In this paper, we summarize work-in-progress toward the creation of open ecosystem-wide measures of diversity and inclusion in open scholarship. This work demonstrates the potential for open data and open software to go produce systematic indicators that go beyond measures of overall production and impact to show how participation in scholarship varies over time and across discipline.

ACKNOWLEDGMENTS

Authors gratefully acknowledge support from IMLS (#LG-250130-OLS-21).

REFERENCES

Altman, Micah. Exploring the Public Evidence on Open Access Monographs. (2021) CREOS White Paper. < https://hdl.handle.net/1721.1/129690 >

Altman, Micah, et al. "A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science." (2018). Pubpub. < https://doi.org/10.21428/62b3421f >

Altman, M., & Cohen, P. N. (2021). Openness and diversity in journal editorial boards.

Gregg, Will, Christopher Erdmann, Laura Paglione, Juliane Schneider, and Clare Dean. "A literature review of scholarly communications metadata." Research Ideas and Outcomes 5 (2019): e38698.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. Public opinion quarterly, 74(5), 849-879.

Grolemund, G. and Wickham, H., R for data science. O'Reilly & Sons. 2017.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. "Bias in peer review." Journal of the American Society for Information Science and Technology 64, no. 1 (2013): 2-17.

Lepori, Benedetto, Philippe Larédo, Thomas Scherngell, Diana Maynard, and Massimiliano Guerini. "Exploring knowledge production in Europe. The KNOWMAK tool." In 17th International Conference on Scientometrics and Informetrics, ISSI 2019, vol. 2, pp. 2561-2562. International Society for Scientometrics and Informetrics, 2019.

Krafczyk, Matthew, August Shi, Adhithya Bhaskar, Darko Marinov, and Victoria Stodden. "Scientific tests and continuous integration strategies to enhance reproducibility in the scientific software context." In Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems, pp. 23-28. 2019.

Konkol, Markus, Daniel Nüst, and Laura Goulier. "Publishing computational research-a review of infrastructures for reproducible and transparent scholarly communication." Research integrity and peer review 5, no. 1 (2020): 1-8.

Nancy Maron, Rebecca Kennison, Nathan Hall, Yasmeen Shorish, Kara Malenfant. Creating a More Inclusive Future for Scholarly Communications: ACRL's New Research Agenda for Scholarly Communications and the Research Environment. ELPUB 2019, Jun 2019, Marseille, France..

Sievert, Carson. Interactive web-based data visualization with R, plotly, and shiny. CRC Press, 2020.

Stathoulopoulos, Kostas, Zac Ioannidis, and Lilia Villafuerte. "Orion: An interactive information retrieval system for scientific knowledge discovery." Talk presented at AKBC 2020

Summann F, Czerniak A, Schirrwagen J, Pieper D. Data Science Tools for Monitoring the Global Repository Eco-System and its Lines of Evolution. Publications. 2020 Jun;8(2):35.