UKSG

Measured in a context: making sense of open access book data

Open access (OA) book platforms, such as JSTOR, OAPEN Library or Google Books, have been available for over a decade. Each platform shows usage data, but this results in confusion about how well an individual book is performing overall. Even within one platform, there are considerable usage differences between subjects and languages. Some context is therefore necessary to make sense of OA books usage data. A possible solution is a new metric – the Transparent Open Access Normalized Index (TOANI) score. It is designed to provide a simple answer to the question of how well an individual open access book or chapter is performing. The transparency is based on clear rules, and by making all of the data used visible. The data is normalized, using a common scale for the complete collection of an open access book platform and, to keep the level of complexity as low as possible, the score is based on a simple metric. As a proof of the concept, the usage of over 18,000 open access books and chapters in the OAPEN Library has been analysed, to determine whether each individual title has performed as well as can be expected compared to similar titles.

Keywords

research impact; open access; academic book; impact measurement

RONALD SNIJDER

CTO/Head of Research

OAPEN Foundation

Introduction

For more than a decade, open access book platforms have been distributing titles in order to maximize their impact. Each platform offers some form of usage data collection, showcasing the success of their offering. However, the numeric representation alone is not sufficient to convey how well a book is actually performing. This is especially important when we take bibliodiversity[1] into account: books written in languages other than English might not always see the same usage numbers when compared with similar books in English.

Context is necessary to make sense of it all, and this article will propose a way to provide this, based on principles of transparency. Before the new metric is discussed, the next section will review the literature on citations and usage of open access books. It will mostly focus on the performance of whole collections, as not much is available on benchmarking for individual titles.

Literature review

The role of languages other than English as a main aspect of bibliodiversity is discussed by many. Especially in the humanities and social sciences – where books are an important publication format –national languages are commonly used. However, more and more publications in these disciplines are written in the English language.[2] Recently, Laasko[3] investigated the number of books that have been published in open access and whether preservation measures are in place. Based on this study's dataset,[4] 56% of the books published in 2022 were written in English. The infrastructure to find open access publications also tends to be optimized for English language publications and Berger[5] argues that this reinforces the imbalance between researchers in the Global North and the Global South.

> 'The infrastructure to find open access publications also tends to be optimized for English language'

While this article will mainly focus on open access for books, the journal impact factor (JIF)[6] should be mentioned. It has been used, and discussed, to assess academics for decades.[7] Currently, the JIF is part of the Web of Science platform, owned by Clarivate. A similar role is played by the Web of Science Book Citation Indices; the collection of these Book Indices is based on internal guidelines, and they tend to favour English language publications.[8]

Still, citation-based metrics such as the JIF do not work as well for books as they do for journal articles. There are fewer citations per title, and they take longer to accrue. That is one of the reasons why Linmans[9] proposes using library book holdings as an additional metric. Apart from typical book-related metrics, such as library book holdings, Torres-Salinas et al.[10] have looked at the use of altmetrics for books. These types of metrics can be categorized as mentions on online platforms such as Mendeley and Goodreads, on social media and within usage data from repositories or similar platforms. Some authors even go further and have devised a multi-level and multi-dimensional book impact evaluation system.[11] It is interesting to note the lack of literature on open access books in repositories, many of which are hosted and maintained by academic libraries. This is also illustrated by a recent book on open access policies in libraries, which barely mentions books.[12]

> 'It is interesting to note the lack of literature on open access books in repositories'

With the advent of (open access) book platforms such as JSTOR, Project Muse and the OAPEN Library, plus the online offerings of publishers, the interest in the usage of open access books has grown. An important aspect is the global uptake of freely available online books, which is significantly greater than the use of books behind a paywall.[13] The better availability of open access books also leads to a higher number of mentions on sites like Wikipedia and Mendeley.[14]

Open access books can be freely shared, and, as a consequence, these books will be made available in several places. While this helps readers to find books, it also leads to a multitude of sources of usage data, making it harder to get an overview of all the usage data attached to a specific book. In response, the HIRMEOS project[15] has developed a database collecting multiple sources of usage data.[16] Another illustration of multiple sources for usage data is the University of Michigan Press dashboard;[17] it lists OAPEN Library downloads, Google Books downloads, Google Books Views, JSTOR chapter downloads and Crossref Event Data. Each platform reports different usage numbers for the same period. See Figure 1.

> 'Open access books can be … available in several places … making it harder to get an overview of all the usage data'
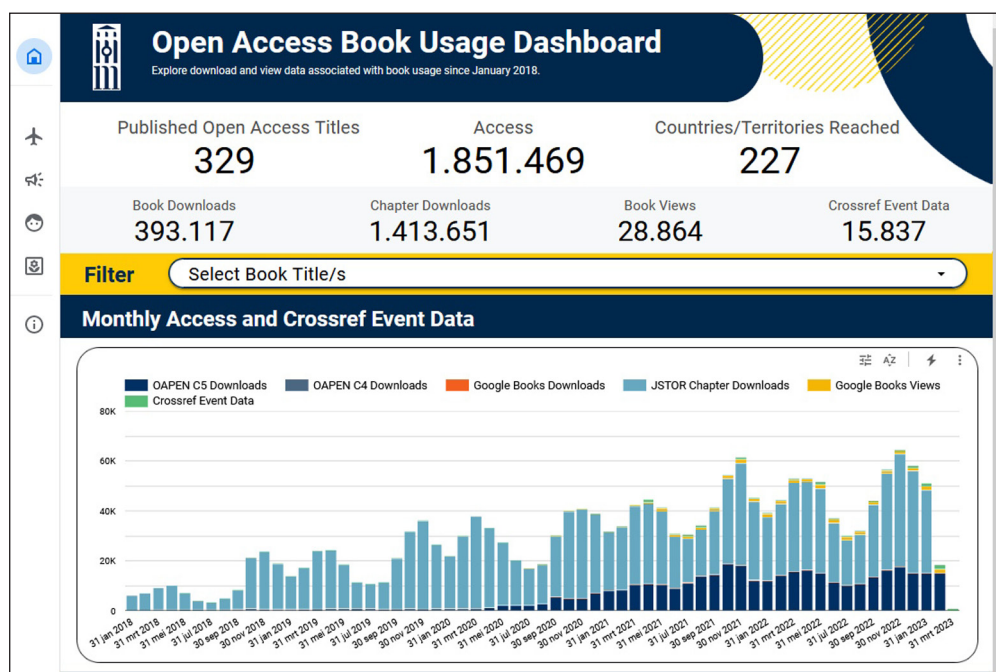

Figure 1. University of Michigan Press: Open Access Book Usage Dashboard

There are even large variations within the usage data of a single platform. For example, examining the titles hosted on the OAPEN Library platform shows that the numbers are directly affected by the subject matter and language of the publication.[18] Furthermore, Hellman[19] argues that analysis of the usage data should take into account that the data is distributed over a 'long tail', and that computing the arithmetic mean can be deceptive.

All things considered, it is not a total surprise that authors of open access books are confused about whether their books have made an impact. According to research by Wennström et al.[20] many authors do not know what data to use as a benchmark. Before the context-based metric is introduced in more detail, a citation from Fire and Guestrin,[21] 'First, these results support Goodhart's Law as it relates to academic publishing: the measures (e.g., number of articles, number of citations, h-index, and impact factor) have become targets, and now they are no longer good measures.'

> 'it is not a total surprise that authors of open access books are confused about whether their books have made an impact'

Metrics play an important role in academia and when they may affect many careers, there should be clear guidelines about their deployment. The Leiden Manifesto[22] discusses ways to do research evaluation in a responsible fashion. To achieve this, best practices have been codified in ten principles.[23]

Some of the guidelines discuss assessment in general and how academic institutions should practise them:

- quantitative evaluation should support qualitative, expert assessment
- measure performance against the research missions of the institution, group or researcher
- base assessment of individual researchers on a qualitative judgement of their portfolio
- recognize the systemic effects of assessment and indicators
- scrutinize indicators regularly and update them.

This article will not focus on these five principles. Its goal is not to create a tool for a complete assessment of a researcher's output. The aim is much simpler: an answer to the question of whether an open access book has performed well, in a clear context.

The Leiden Manifesto also contains principles that focus on the measurement itself:

- protect excellence in locally relevant research
- keep data collection and analytical processes open, transparent and simple
- allow those evaluated to verify data and analysis
- account for variation by field in publication and citation practices
- avoid misplaced concreteness and false precision.

The proposed metric – discussed in the next section – is based on these five principles. The bibliodiversity of scholarly output will be taken into account by analysing the usage data by subject and language. Furthermore, transparency and simplicity are key elements: the data used for the evaluation will be completely visible and accessible. The algorithm used is also extremely basic and can be easily checked. Lastly, there are only three possible outcomes: below average, average and above average. While these options are quite concrete, they are not measured in decimal places.

Introducing a context-based metric: TOANI score

As the review has shown, the literature focusses primarily on the 'open access citation advantage' for complete collections and additionally, for books, citations are not the primary impact measurement. Instead, other measurements – altmetrics, in which we include usage data from open access book platforms – are viable alternatives. Or at least, they are quicker to deliver results. The question is how to make sense of these numbers. Usage statistics

differ from platform to platform, and even the numbers within a single platform are hard to interpret. The usage depends on the subject and language, but also on a time period: not just from month to month, but also from year to year.

Here, the usage of over 18,000 titles will be analysed, with the goal of determining whether each individual title has performed as well as can be expected.

'Usage statistics differ from platform to platform, and even the numbers within a single platform are hard to interpret'

The dataset

The following table lists the usage data of a selection of titles hosted on the OAPEN Library.[24] Launched in 2010, the OAPEN Library hosts one of the largest collections of peer-reviewed open access books and chapters. In March 2023, the collection consisted of over 27,000 titles.

Our dataset is based on 18,014 books and chapters, of which 65% were written in English, 25% in German, while the remaining 10% consists of publications written in more than 30 other languages. The selected titles were added to the collection before 1 January 2022, and usage data for the 12 months from January to December 2022 has been captured. During that period, this collection of books and chapters was downloaded more than ten million times. Each title has been linked to one broad subject and the title's language has been coded as either English, German or Other language. Further details can be found in Table 1.

Subject	Language	Number of titles	Median downloads	Total downloads
A. The arts	English	814	326.5	520,359
	German	286	164.5	68,248
	Other	105	158	39,699
C. Language	English	636	259	353,019
	German	494	101.5	150,254
	Other	149	186	167,822
D. Literature & literary studies	English	821	266	488,074
	German	460	115.5	101,018
	Other	165	245	70,912
G. Reference, information & interdisciplinary subjects	English	408	431	388,402
	German	90	156	23,821
	Other	36	127.5	8,242
H. Humanities	English	2,350	325	1,465,656
	German	785	148	195,899
	Other	347	170	102,812
J. Society & social sciences	English	3,249	382	2,680,028
	German	1,392	209.5	492,086
	Other	720	94	194,182
K. Economics, finance, business & management	English	900	347	843,813
	German	467	115	120,521
	Other	110	33	12,726
L. Law	English	324	334.5	178,950
	German	240	87.5	42,894
	Other	38	135	7,425
M. Medicine	English	513	130	277,379
	German	75	238	69,809
	Other	14	138	5,906
P. Mathematics & science	English	560	202	380,040
	German	80	123	17,837
	Other	28	185.5	13,988

(Contd.)

Subject	Language	Number of titles	Median downloads	Total downloads
R. Earth sciences, geography, environment, planning	English	343	316	242,296
	German	62	133.5	14,833
	Other	58	172.5	18,854
T. Technology, engineering, agriculture	English	462	173	287,523
	German	74	155	36,070
	Other	55	238	37,305
U. Computing & information technology	English	285	308	273,597
	German	16	260.5	6,355
	Other	3	95	289
Grand Total		**18,014**		**10,398,943**

Table 1. Dataset: OAPEN Library titles by language and subject

Median downloads vary considerably between subjects. The same holds true for languages. The number of median downloads of English-language titles is, in most instances, much higher than those of titles in other languages. Additionally, there is a large variation between German and the Other language category. In order to make more sense of these numbers, it would be helpful to have a guideline that takes into account this diversity linked to the different subjects and languages.
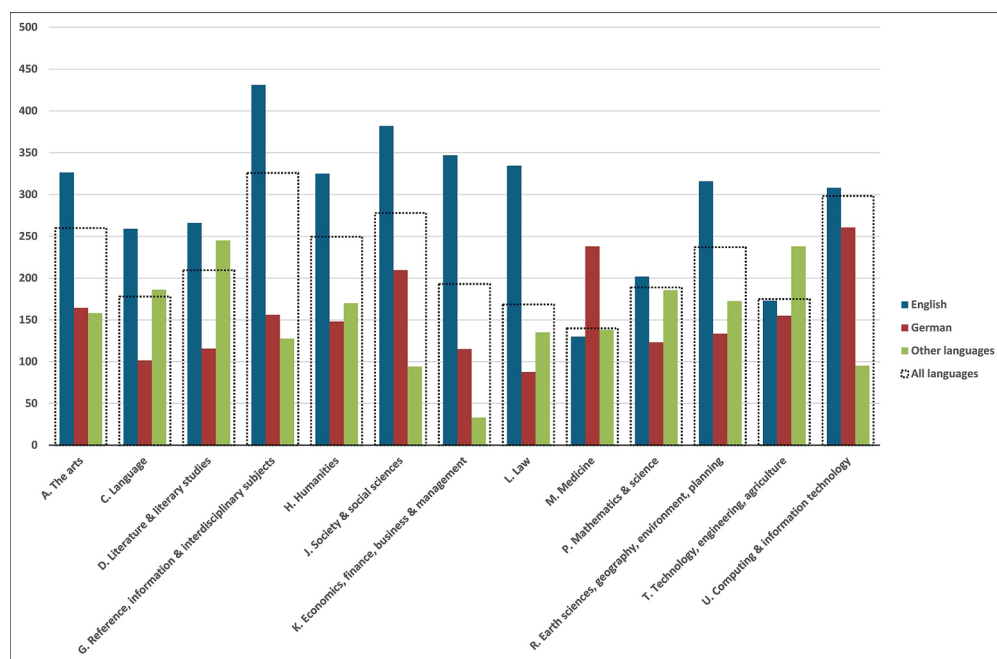

Figure 2. Median downloads of titles in the OAPEN library, 2022

When the median downloads per subject are represented in Figure 2 – especially when the median of the different languages is compared to the median of all languages – the differences are striking. In most cases, books and chapters in English are downloaded more, but the divergence of titles in German and other languages is quite large. For instance, the median number of downloads of titles on Literature & literary studies in German is roughly half of those in English or Other languages. In the case of Medicine, this is quite the opposite.

Additionally, the median downloads per subject themselves also differ to a large degree. Titles discussing Reference, information & interdisciplinary subjects have a median number of downloads of 326, while Medicine has 140. All in all, even when the usage data is simplified to sets based on broad categories, it is impossible to give a simple answer to the question of whether a certain number of downloads is a 'good result' or not.

'it is impossible to give a simple answer to the question of whether a certain number of downloads is a 'good result' or not'

The TOANI score

As a possible solution, we introduce the TOANI score. The acronym stands for Transparent Open Access Normalized Index. The transparency is based on the application of clear rules and by making all the compiled data visible. The data is normalized using a common scale for the complete collection of an open access book platform. Additionally, there are only three possible values to score the titles: below average, average and above average. This index is set up to provide a clear and simple answer to the question of what impact an open access book has made. It is not meant to give a sense of false accuracy; the complexities surrounding this issue cannot be measured to several decimal places.

The TOANI score is based on the following principles:

- select only titles that have been available for at least 12 months
- use the usage data of the same 12-month period for the whole collection
- assign each title one – high-level – subject
- assign each title one language
- group all titles based on subject and language
- the groups should consist of at least 100 titles
- make the following data available for each title:
  - platform
  - total number of titles in the group
  - subject
  - language
  - time period used for the measurement
  - minimum value, maximum value, median, first and third quartile of the platform's usage data
- based on these principles, classify the titles as:
  - 'below average' – first quartile, 25 % of the titles
  - 'average' – second and third quartile, 50% of the titles
  - 'above average' – fourth quartile, 25 % of the titles.

There are several reasons behind these principles. The TOANI score is based on the usage data of a particular platform. Other platforms might be measuring different things, and this could lead to different figures. For example, the titles of Michigan University Press are made available on Google Books, reporting book views, while JSTOR reports chapter downloads and the OAPEN Library reports COUNTER-conformant downloads of the complete books. As a result, the numbers from these platforms are hard to compare. There are also seasonal differences, with less usage in the months of June, July and August. Another time-related issue is that usage might differ across several years. Hence, the selection of the twelve-month period.

The influence of subject and language are profound, which is reflected here. However, it is also very important to keep things simple. In line with the Leiden Manifesto principles, we have aimed to account for the variation in usage data that is tied to diversity in subject and language. On the other hand, it is also important to enable the verification of the TOANI score. This is achieved in several ways. Firstly, by consistently simplifying – books can only be part of one subject and one language group, the groups themselves are large, leading to fewer classifications, and the TOANI score is based on quartiles, instead of an opaque formula. Secondly, all data must be made visible to enable scrutiny.

'The influence of subject and language are profound'

Another principle of the Leiden Manifesto is the avoidance of misplaced concreteness and false precision. By only allowing for the three options 'below average', 'average' and 'above average', the TOANI score adheres strongly to this. It also makes clear that these scores are

based on a specific platform. Different platforms might not only lead to differently measured figures, but they might also vary in regional reach. For example, a Portuguese-language book discussing a local Brazilian subject will most likely find more readers on the Brazil-based SciELO Books platform[25] compared to the Zendy platform[26] which focuses on the Middle East and North Africa.

Applying the TOANI score: OAPEN Library usage

When the TOANI score is applied to the books and chapters in the dataset, we see that 4,520 titles have usage data that is below average, 8,992 titles have average usage data and 4,502 titles performed above average. In other words, the 25%, 50%, 25% division of the previous subsections. However, visualizing the usage data show shows large differences between subjects and languages. Books and chapters in English mostly see the highest usage, but the range of usage leading to an average score differs widely per subject.

As an illustration, as shown in Figure 3, a German language book on Humanities with 300 downloads is doing better than average, while an English language book on Humanities would need to have reached at least 652 downloads to reach the same level. Another example is the difference between titles on Language in German versus Other languages. Here, German-language books downloaded more than 250 times are scoring better than average. For books in Other languages the bar is much higher at 385.
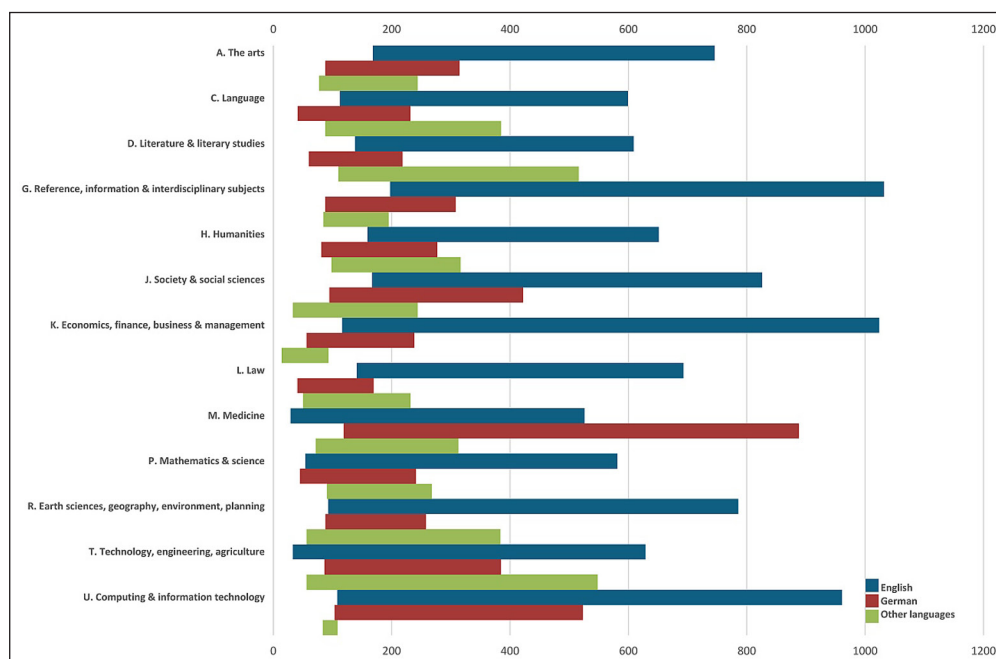

Figure 3. Dataset TOANI score: average usage between the first and third quartile

All the data describing the TOANI score for each title in the dataset plus all other relevant data are available at the link found in the data accessibility statement at the end of the article.

Discussion

The TOANI score is designed to provide a simple answer to the question of how well an individual open access book or chapter is performing. We have also seen that language and subject greatly affect the usage, and thus the answer must allow for this context. To keep the level of complexity as low as possible, the score is based on a simple metric: the quartiles of the usage per group of similar titles.

'The TOANI score is designed to provide a simple answer to the question of how well an individual open access book or chapter is performing'

As a proof of concept, a collection of books and chapters in the OAPEN Library has been assessed and the considerable differences between subjects and languages have been shown. However, it is not proven if this diversity is also visible in other platforms. To ensure a comparable scoring, the groups of titles should be based on the same language and subject selections. This requires a categorization that is not dependent on one particular platform. A possible option is to use the OpenAlex knowledge graph.[27] This large and openly available resource contains a list of 19 high-level concepts, which could be applied to all publications on the different platforms.[28]

Apart from the question of handling multiple platforms, another aspect to consider is the possibility of using the TOANI score in an inappropriate manner, for instance by misinterpretation. It is important to be clear about what is measured, and what is not measured.

Additionally, usage is not the same as quality. All books and chapters in the OAPEN Library are subject to peer review, and therefore all publications conform to academic standards. As we have seen, usage depends on factors that are not inherent in the quality of a title, but which have a strong correlation with the platform's reach.

> 'usage depends on factors that are not inherent in the quality of a title, but which have a strong correlation with the platform's reach'

The usage of books also depends on which topics are currently being debated and so the usage patterns may change over time. Some patterns will significantly change over several years, while we have also seen that in the months of June, July and August fewer books were downloaded from the OAPEN Library.

Would it be possible to 'game' the TOANI data? The results are based on download patterns which are the responsibility of the platforms. In the case of the OAPEN Library, all reported usage adheres to COUNTER Release 5 rules. Crucial to COUNTER reporting is removing any usage data that is deemed to be unintended by a human user.[29] In theory, the outcome could be affected by changing the groups the score is based on. Books on a niche subject would probably attract less usage than books discussing a popular subject. If the less popular titles were separated into a separate group, this could possibly 'improve' the scores for those titles. This can be mitigated by complete transparency.

> 'the TOANI score aims to give a simple answer to a complicated question … how well is my open access book doing compared with others in the same language and subject?'

In conclusion, the TOANI score is not a tool to support full qualitative or quantitative measurement of research performance. By providing context, the TOANI score aims to give a simple answer to a complicated question, i.e. looking at a certain platform, how well is my open access book doing compared with others in the same language and subject?

References

1. Elea Giménez Toledo et al., "Bibliodiversity-What It Is and Why It Is Essential to Creating Situated Knowledge," *Impact of Social Sciences Blog*, 2019, http://eprints.lse.ac.uk/103440/1/impactofsocialsciences_2019_12_05_bibliodiversity_what_it_is_and_why.pdf (accessed 13 September 2023).

2. Ana Balula and Delfim Leão, "Multilingualism within Scholarly Communication in SSH. A Literature Review," *JLIS.It* 12, no. 2 (15 May 2021): 88–98, DOI: https://doi.org/10.4403/jlis.it-12672 (accessed 13 September 2023); Emanuel Kulczycki et al., "Publication Patterns in the Social Sciences and Humanities: Evidence from Eight European Countries," *Scientometrics* 116, no. 1 (1 July 2018): 463–86, DOI: https://doi.org/10.1007/s11192-018-2711-0 (accessed 13 September 2023).

3. Mikael Laakso, "Open Access Books through Open Data Sources: Assessing Prevalence, Providers, and Preservation," *Journal of Documentation* 79, no. 7 (1 January 2023): 157–77, DOI:
    https://doi.org/10.1108/JD-02-2023-0016 (accessed 30 September 2023).

4. Mikael Laakso, "Dataset for 'Open Access Books through Open Data Sources: Assessing Prevalence, Providers, and Preservation,'" *Zenodo*, (8 November 2022), DOI:
    https://doi.org/10.5281/zenodo.7305477 (accessed 13 September 2023).

5. Monica Berger, "Bibliodiversity at the Centre: Decolonizing Open Access," *Development and Change* 52, no. 2 (2021): 383–404, DOI:
    https://doi.org/10.1111/dech.12634 (accessed 13 September 2023); Laakso, "Dataset for 'Open Access Books."

6. Eugene Garfield, "The History and Meaning of the Journal Impact Factor," *JAMA* 295, no. 1 (4 January 2006): 90–93, DOI:
    https://doi.org/10.1001/jama.295.1.90 (accessed 13 September 2023).

7. Erin C McKiernan et al., "Use of the Journal Impact Factor in Academic Review, Promotion, and Tenure Evaluations," ed. Emma Pewsey, Peter Rodgers, and Björn Brembs, *ELife* 8 (31 July 2019): e47338, DOI:
    https://doi.org/10.7554/eLife.47338 (accessed 13 September 2023).

8. Kim Quaile Hill and Patricia A. Hurley, "Web of Science Book Citation Indices and the Representation of Book and Journal Article Citation in Disciplines with Notable Book Scholarship," *The Journal of Electronic Publishing* 25, no. 2 (10 December 2022), DOI:
    https://doi.org/10.3998/jep.3334 (accessed 13 September 2023).

9. A. J. M. Linmans, "Why with Bibliometrics the Humanities Does Not Need to Be the Weakest Link," *Scientometrics* 83, no. 2 (13 August 2009): 337–54, DOI:
    https://doi.org/10.1007/s11192-009-0088-9 (accessed 13 September 2023).

10. Daniel Torres-Salinas, Nicolas Robinson-Garcia, and Juan Gorraiz, "Filling the Citation Gap: Measuring the Multidimensional Impact of the Academic Book at Institutional Level with PlumX," *ArXiv:1710.00368*, 1 October 2017,
    http://arxiv.org/abs/1710.00368 (accessed 13 September 2023).

11. Qingqing Zhou and Chengzhi Zhang, "Impacts towards a Comprehensive Assessment of the Book Impact by Integrating Multiple Evaluation Sources," *Journal of Informetrics* 15, no. 3 (1 August 2021): 101195, DOI:
    https://doi.org/10.1016/j.joi.2021.101195 (accessed 13 September 2023).

12. Karen Brunsting, Caitlin Harrington, and Rachel Scott, *Open Access Literature in Libraries: Principles and Practices* (Chicago: ALA Editions, 2023),
    https://ir.library.illinoisstate.edu/fpml/163 (accessed 13 September 2023).

13. Cameron Neylon et al., "More Readers in More Places: The Benefits of Open Access for Scholarly Books," *Insights* 34, no. 1 (1 December 2021): 27, DOI:
    https://doi.org/10.1629/uksg.558 (accessed 13 September 2023).

14. Michael Taylor, "Open Access Books in the Humanities and Social Sciences: An Open Access Altmetric Advantage," *ArXiv:2009.10442 [Cs]*, 22 September 2020,
    http://arxiv.org/abs/2009.10442 (accessed 13 September 2023).

15. "Hirmeos," OPERAS,
    https://operas-eu.org/hirmeos/ (accessed 13 September 2023).

16. Javier Arias, "Collecting Inclusive Usage Metrics for Open Access Publications: The HIRMEOS Project," in *ELPUB 2018* (Toronto, Canada, 2018),
    https://hal.archives-ouvertes.fr/hal-01816811 (accessed 13 September 2023). DOI:
    https://doi.org/10.4000/proceedings.elpub.2018.11

17. "Advance Social Impact," University of Michigan Press Ebook Collection, 2023, University of Michigan Press,
    https://ebc.press.umich.edu/impact/#oa-book-usage (accessed 13 September 2023).

18. Ronald Snijder, "Books in a Bubble.: Assessing the OAPEN Library Collection," *JLIS.It* 14, no. 2 (15 May 2023): 75–92, DOI:
    https://doi.org/10.36253/jlis.it-498 (accessed 13 September 2023).

19. Eric Hellman, "The Open-Factor: Toward Impact-Aligned Measures of Open-Access Ebook Usage," preprint (LIS Scholarship Archive, 18 October 2019), DOI:
    https://doi.org/10.31229/osf.io/7npdf (accessed 13 September 2023).

20. Sofie Wennström et al., 'The Significant Difference in Impact: An Exploratory Study about the Meaning and Value of Metrics for Open Access Monographs' (ELPUB 2019 23rd edition of the International Conference on Electronic Publishing, Marseille, France, 2019),
    https://hal.archives-ouvertes.fr/hal-02141879 (accessed 13 September 2023). DOI:
    https://doi.org/10.4000/proceedings.elpub.2019.9

21. Michael Fire and Carlos Guestrin, "Over-Optimization of Academic Publishing Metrics: Observing Goodhart's Law in Action," *GigaScience* 8, no. 6 (1 June 2019): giz053, DOI:
    https://doi.org/10.1093/gigascience/giz053 (accessed 13 September 2023).

22. Leiden Manifesto for Research Metrics,"
    http://www.leidenmanifesto.org/ (accessed 13 September 2023).

23. Diana Hicks et al., "Bibliometrics: The Leiden Manifesto for Research Metrics," *Nature* 520, no. 7548 (April 2015): 429–31, DOI:
    https://doi.org/10.1038/520429a (accessed 13 September 2023).

24. 'OAPEN Library – Online Library and Publication Platform', OAPEN Foundation, 2021,
    https://www.oapen.org/ (accessed 13 September 2023).

25. "Books," SciELO,
    https://books.scielo.org/ (accessed 13 September 2023).

26. "Your Online Library," Zendy,
    https://zendy.io/ (accessed 13 September 2023)

27. Jason Priem, Heather Piwowar, and Richard Orr, "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," *ArXiv:2205.01833,* 16 June 2022, DOI:
    https://doi.org/10.48550/arXiv.2205.01833 (accessed 13 September 2023).

28. "Concept Object," Concepts – OpenAlex API documentation, February 2023,
    https://docs.openalex.org/api-entities/concepts (accessed 13 September 2023).

29. "7.0 Processing Rules for Underlying COUNTER Reporting Data," COUNTER, Project Counter,
    https://www.projectcounter.org/code-of-practice-five-sections/7-processing-rules-underlying-counter-reporting-data/ (accessed 13 September 2023).

Corresponding author:
Ronald Snijder
CTO/Head of Research
OAPEN Foundation, NL
E-mail: r.snijder@oapen.org
ORCID ID: https://orcid.org/0000-0001-9260-4941