

On the Road to a Comprehensive Open Digital Mathematics Library

Darius Ehsani and Olaf Teschke (FIZ Karlsruhe, Germany)

In our previous column [BBHST], we discussed the feasibility of transforming scanned mathematics into formats which allow for automated digital processing. Obviously, the minimum requirement here is the existence of an openly available digital mathematical object, which is also what would be sufficient for most working mathematicians. Indeed, the comprehensive open digital availability of mathematics literature is the classical vision of the World Digital Mathematics Library (WDML), formulated more than 25 years ago. Since then, progress has been made through different approaches and policies. The aim of this column is to give a short overview on the current status.

Possible scope and extent

Given that mathematics is the language of exact science and is interconnected with so many diverse areas, it is almost impossible to precisely define the desirable extent of WDML. Throughout the decades, disciplines have undergone an evolution which is also reflected in publication patterns. Even supposedly uniform services like zbMATH have adapted their indexing policy several times – currently, it reads “published and peer-reviewed articles, books, conference proceedings as well as other publication formats pertaining to the scope defined by Mathematical Subject classification 2020 that present a genuinely new point of view”¹. If this cannot cover all fields in which mathematicians are involved through their research activities, it hopefully reflects most of the needs of the community and allows for interconnection with digital libraries in other disciplines.

Another aspect is the application of the scope of the definition – even services with a fairly similar approach show significant historical differences. E.g., MathSciNet and zbMATH have been found to have a historical overlap of just about 60% [IT]². In the following, we will work with zbMATH data, but it should be taken into account that this may leave a considerable amount of publications omitted.

How much mathematics is digitally available?

Digitisation efforts in mathematics already started in the first years of the internet. At that time, they were often only identified by their url, which quickly lead to the well-known problem of dead links. zbMATH still contains a significant number of hard urls mostly dating back to

these days, and experience shows that they are only occasionally useful (though they may sometimes help to trace back sources via the Internet Archive). For many years now, doi have been standard for accomplishing an (ideally) unique, sustainable referencing of digital publications. They are accompanied by stable IDs generated by large platforms and repositories such as arXiv, EuDML, Gallica, JSTOR, Math-Net.ru, or Project Euclid. This system of identifiers has proved its worth and is still incredibly reliable, although all kinds of issues occur on an individual basis³. Nevertheless, the number of working links remains relatively high; we checked the availability of digital objects given by these identifiers in the zbMATH database and obtained a success rate greater than 99.4%. This gives a much better picture overall in mathematics than the general analysis in [KB].

Thus, these IDs can be used for estimating the extent of digitally available mathematics. Below, we give the share of publications with sustainable digital identifiers in zbMATH.

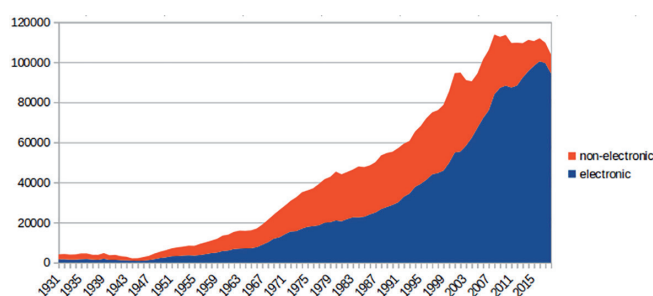


Fig. 1. Share of electronically available publications in zbMATH per publication year.

However, there is one important caveat: the number of publications does not tell the full story. Stable identifiers are much more prevalent for journals than for books. Historically, the latter have contributed up to 50% of the pages of published mathematical research (although this share has shrunk to about 14% recently), so the figures of digitally available pages look much less impressive. Based on [IT], we estimate that just above 60% of the

³ To name only a few that popped up during our availability checks at the time of writing this note: All doi for historical content of a classical maths journal did not redirect properly after a change of the publisher; all doi of a publisher did not resolve in some browsers due to cookie issues; doi assigned to new articles not registered more than one year after publication; same doi given to different articles; different doi resolving to identical digital resources, doi leading to an official dead landing page after change to a non-CrossRed member.

¹ Note that the formulation of the last addition is relatively new, triggered by the growth of (semi)trivial publications.

² Naturally, generic aggregators like Google Scholar differ by a much larger factor.

roughly 130 million pages of maths research since 1868 are digitally available.

What is the share of open accessible publications?

Open Access issues have been a recurring topic during the last decades, and have also been frequently discussed in this column (see, e.g., [T] for the various shades of OA in mathematics). Three approaches have contributed to the open availability of the literature: genuine OA publications, DML platforms like EuDML, Project Euclid, or Math-Net.ru, and green OA repositories like arXiv or HAL⁴. Overall, the growing numbers for all three solutions look promising, but they come with certain caveats. First of all, taking the previous remark into account, they do not so far apply to a large chunk of the mathematical publications. Books are not only less frequently digitised, but also much more rarely available open access; hence, the number of open available math pages looks much less impressive than the number of OA publications.

Significant progress is only seen when we restrict to digitally available journal articles. Below, we give a figure of the numbers for all three kinds of OA approaches by publication years derived from zbMATH (note that recent years are not yet fully covered):

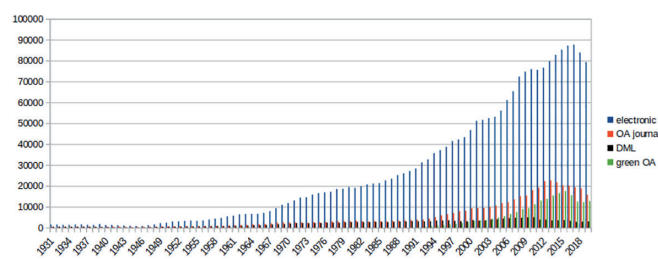


Fig. 2. Digitally available journal articles indexed in zbMATH, and those OA available directly via OA journals, green OA, and DML platforms.

This looks like a relatively successful picture, especially since one is tempted to count the three OA approaches cumulatively. However, the numbers do not tell us much about the overlap. E.g, arXiv overlay journals will automatically appear both as OA journal publications and green OA, and all other combinations are likely to happen as well. A detailed breakdown of the respective shares reveals a more granular picture (see Fig. 3).

A perhaps surprising takeaway is that historical publications in fact have a larger OA share than recent ones, mainly thanks to both the open DML platforms and open society journals with a rich tradition. There is a stable overlap of still existing OA journals available through the platforms, though, also driven by the existence of moving wall OA journals on the platforms (which explains the diminishing role of DML for recent publications). The shrinking share of OA for publication

⁴ An analysis of recent zbMATH publications with unpaywall [https://unpaywall.org] confirms that the arXiv is still by far the largest sole source for green OA in mathematics, although other repositories gained shares during the past years.

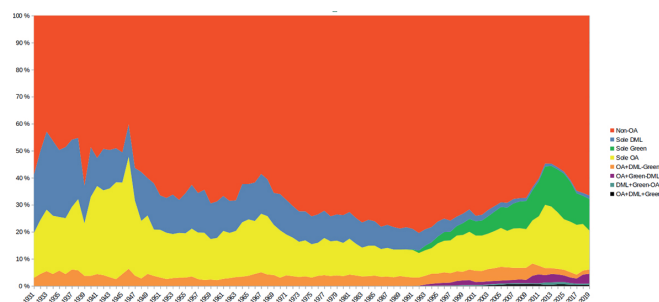


Fig. 3. Share of different OA resources for digitally available journal articles indexed in zbMATH.

years up to 1992 illustrates the concentration dynamics involving large commercial publishers. It is clearly visible that the creation of arXiv was a game changer in the 90s, reversing the trend. The relevant and stable share of papers both openly available at arXiv and at OA journals is not just due to overlay journals, but reflects a general OA-friendly community in several areas.

A more ambiguous trend is the sudden spike of papers in sole OA journals from about 2008 (and their decline after 2012). This reflects the boom of both APC and nickel OA journals which started around this time (see [T] for a more detailed discussion). Numerous examples in the following years indicate that their formal peer review process might not have always have been sufficient to live up to the classical zbMATH standards. This resulted in the tightened indexing policy mentioned above, coming into effect in 2017.

Such effects can be omitted when we restrict our analysis to core mathematics journals, which we define as journals indexed in zbMATH as Cover-to-Cover and belong to the top two internal categories [T]. They make up about 40% of zbMATH indexed electronic journal articles, with a growth of about 30% during the last decade (compared to about 50% overall). As discussed in [T], APC journals, which are responsible for most of the growth of sole OA journal publications, play almost no role in core mathematics journals. Hence, the figure for the relative share of the OA solutions looks a bit different when restricted to core math electronic journals:

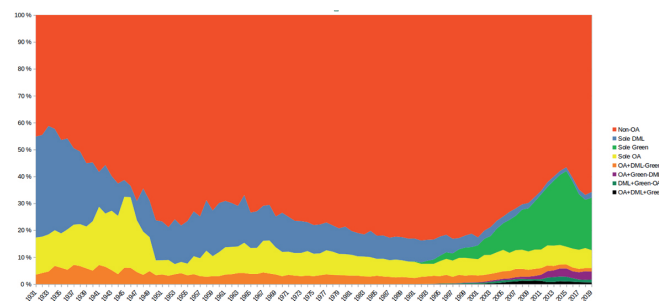


Fig. 4. Share of different OA resources for digitally available core math journal articles.

Here, we do not see an APC spike; on the contrary, the share of OA journal articles has been remarkably stable throughout the last decade. This does not imply stagnation – indeed, as we are all aware, there are numerous ini-

tiatives – it just says that they did not outpace the general growth. In comparison, the impact of green OA is even more significant here. It basically accounts for all progress made in the OA share during the past two decades.

The diagram may allow some conclusions about feasible approaches to further expand the share of OA publications. First of all, the strength of DML platforms, which already provide us with a large share of the literature until the 1960s, should also be used to facilitate the integration of more recent publications (note the dip from the mid-1960s until 2000 in the diagram!). This could be achieved by implementing broad moving wall policies, accompanied by both forcing suitable open licenses for this content and allocating resources for the platforms, which enables them to preserve it sustainably. A similar approach to enabling the integration of publications from OA journals would also be to enable DML functions for the recent literature, where the share is still relatively small.

In particular, this would address both the problem of possibly limited sustainability of sole OA journals⁵ [LMJ] as well as limited machine readability (see, e.g., [KBS]). Hence, while there is obviously a need to expand DML services further, the platforms have not been much in the focus of recent OA initiatives, and the resources made available for them do not seem to quite match these tasks.

On the other hand, the progress of green OA seems almost undamped, and is the single most important driving factor eating into the share of non-OA publications. Since there are still no large indications of saturation, encouraging green OA via feasible platforms still seems to be the most effective measure to achieve broader OA in mathematics (note that this seems to be quite different from many other subjects). Perhaps the only visible tendency in green OA is a recently growing share of OA journal articles available as green OA (and a corresponding smaller share of sole green OA articles). While this is positive in general, since it provides more alternatives in a sustainable way, it may also indicate that the foundation of new OA journals in core mathematics during the last few years has mainly been addressing a community which is already quite OA-minded, hence achieving less with respect to reducing the overall non-OA share.

The conclusion of [T] that APC OA journals are no feasible way to propagate OA in core mathematics has only been reinforced once more by the diagram. Such enterprises have done nothing to significantly enlarge the OA share during the past years, although quite considerable funds have been made available in several countries in the past years by transforming subscription to APC resources. For core mathematics, implementation of policies supporting OA throughout APCs appear to be a misallocation of resources. It remains to be seen whether the implementation of transformative agree-

ments like the “Project Deal” agreements in Germany have a broader impact in the future.

This leaves the question about how to open up the significant share of recent publications which will not be available by green OA in the foreseeable future. The subscribe-to-open model as recently backed by, e.g., the EMS Press <https://ems.press/subscribe-to-open>, appears to be a new and attractive model to address this issue. It will be interesting to see how its implementation will affect the OA share in the future.

References

- [BBHST] Marco Beck, Isabel Beckenbach, Thomas Hartmann, Moritz Schubotz and Olaf Teschke, Transforming Scanned zbMATH Volumes to LaTeX: Planning the Next Level Digitisation. *Eur. Math. Soc. Newsl.* 117, 49–52 (2020)
- [KB] Martin Klein, Lyudmila Balakireva, On the Persistence of Persistent Identifiers of the Scholarly Web. e-print, arXiv:2004.03011
- [KBS] “Who is asking? Humans and Machines experience a different Scholarly Web” <https://osf.io/pgxc3/>
- [IT] Patrick Ion and Olaf Teschke, Continuing toward a Global Digital Mathematics Library. AMS Special Session on Mathematical Information in the Digital Age of Science at the Joint Mathematics Meetings (San Diego, CA) (2018)
- [LMJ] Mikael Laakso, Lisa Matthias and Najko Jahn, Open is not forever: a study of vanished open access journals, e-print arXiv: 2008.11933
- [T] Olaf Teschke, Green, gold, platinum, nickel: on the status of open access in mathematics. *Eur. Math. Soc. Newsl.* 110, 60–63 (2018)



Dariush Ehsani has worked as a mathematician at several universities within the United States and Germany, specializing in the field of Several Complex Variables. In 2020, he joined FIZ Karlsruhe to facilitate the transition of zbMATH to an open access platform, with a special focus on the integration of digital mathematics libraries.



Olaf Teschke studied mathematics at Humboldt University Berlin, and completed his PhD in algebraic geometry there. He moved to FIZ Karlsruhe in 2008, and has been working there since 2009 as the head of the Mathematics Department and Managing Editor of zbMATH (including a short intermediate term as Editor-in-Chief). Since 2017, he has been serving as the Vice Chair of the EMS Committee on Publications and Electronic Dissemination. His main occupation is information infrastructure for mathematics.

⁵ This issue has also been discussed in a recent IMU Newsletter, but it should be noted that the two vanished OA journals listed in [LMJ] as mathematics journals did not fulfill zbMATH indexing requirements.