The extension of zbMATH Open by arXiv preprints

Isabel Beckenbach, Klaus Hulek and Olaf Teschke

zbMATH Open has started a new feature – relevant preprints posted at arXiv will also be displayed in the database. In this article we introduce this new feature and the underlying editorial policy. We also describe some of the technical issues involved and discuss the challenges this presents for future developments.

1 Introduction

During the last three decades, arXiv has established itself as the main preprint repository for mathematics. Some years ago, in [6, 7], we analysed the share of publications in the zbMATH corpus available via arXiv versions. There we saw that, despite a very uneven distribution depending on the mathematical subjects, the share of arXiv coverage has been growing constantly. Today, the overall arXiv share exceeds a third of all recent mathematical publications, with the figures for several core areas of mathematics being significantly higher, often exceeding 50%. Moreover, the arXiv share is still growing, although some saturation effects can already be observed in certain areas (e.g., the percentage of about 80% of publications in algebraic geometry available on arXiv has not improved much recently). zbMATH Open has adapted its services to this situation, and has been providing links from published documents to available arXiv versions for more than ten years now. Obviously, this required a precise matching algorithm, taking into account the information provided by arXiv. This information, however, is somewhat unstable. It may vary with submission versions, and the citable source strings contain less information than the ones from published journals. Indeed, our success in matching zbMATH Open entries to arXiv versions has much improved since we started linking to it, and the figures (as evaluated on test sets definable by DOIs) now indicate a very mature state (see Section 4).

One obvious advantage is that this ensures the open availability of a significant share of mathematics research – indeed, despite the growing ecosystem of open and hybrid journals, platforms, and transitory deals, arXiv accounts by far for the largest share of mathematics publications available through open access [3]. But the advantages achieved through this matching go much beyond findability and access. Namely, this also opens up the path for numerous new investigations and thus new insights. Just to give an example: the links now serve as a proxy to estimate the time between submission and publication in various mathematical areas [1, 2, 5].

Especially the long period between submission and publication in many core mathematical areas, predominantly caused by a thorough peer review process, raises natural questions about the difference set: how large is the *unpublished* share of arXiv in mathematics? What are the chances of arXiv submissions to be finally published? And, perhaps most importantly, would it not be beneficial for a comprehensive information system in mathematics to integrate a well-defined subset of arXiv into zbMATH Open? In particular, it could serve to enhance the visibility of recent work, especially by young researchers whose career chances suffer from long publication delays. If so, how should one formulate a sound indexing policy that matches closely the zbMATH Open scope despite the lack of a formal peer review process at arXiv?

There can be few doubts about the relevance of unpublished arXiv preprints. Not just the share of the platform in recent work, but also the omnipresence of unpublished arXiv submissions in the references of already published documents underline its significance. So the principal decision to enlarge the information available in zbMATH Open by unpublished arXiv preprints was a very natural one. However, the exact realisation details are very involved. In this column we outline the editorial policies and the new features that define the arXiv subset now integrated into zbMATH Open.

2 The editorial policy

2.1 Aims and resulting editorial decisions

The scope of zbMATH Open, as defined by its editorial policy, has been to index "all available published and peer-reviewed articles, books, conference proceedings as well as other publication formats pertaining to the scope given above that present a genuinely new point of view".[1] This allows different interpretations

---

[1] https://zbmath.org/about/#id_1_1

and is subject to the dynamics of the field and its publication system. Nevertheless, the consistency of this approach has been ensured by the editorial board, the expert staff covering the diverse mathematical subjects, the ongoing maintenance and development of the applied classification system,[2] and the supervisory boards.

Unpublished arXiv preprints per se do not satisfy all conditions of the above definition. There is no refereeing process for arXiv, although a mature mediator and endorsement system is installed to ensure both the scientific relevance and formal standards of the accepted submissions. The first and foremost condition for integrating unpublished arXiv preprints into zbMATH Open must be that they are clearly distinguishable from published work. This now happens at two different levels: firstly as a new database type ('arXiv', in addition to Zbl, JFM, and ERAM); secondly as a new document type ('Preprints', in addition to books, and journal and collection articles). Filtering with both facets will currently provide identical results, although this may change with future developments (Figure 1).
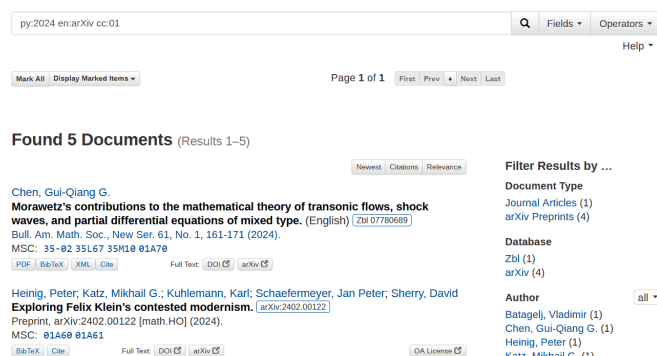

Figure 1. New filter functions allow to clearly distinguish additional unpublished arXiv entries.

Perhaps the most important distinction from classical zbMATH Open documents is that unpublished arXiv preprints will not be sent out for reviewing. This is due to the very nature of arXiv submissions – they are not stable versions, and may in fact vary significantly over time. To reflect this, they would require multiple reviews, which would both be beyond the resources of zbMATH Open and unsatisfactory for reviewers. Neither will there be multiple entries for (possibly quite different) arXiv versions, but only one, which is defined by the stable arXiv identifier – along with author and title information, the only metadata required for the entries. The preliminary nature is also reflected by the fact that, contrary to other new zbMATH Open entries, there will be no editorial classification – only MSCs already provided by authors will be

displayed. However, the entries *are* integrated into the zbMATH Open author disambiguation, to ensure that they are visible in the author profiles (more details are given below in Section 4.2).

Finally, once more emphasizing the preliminary nature of the arXiv entries, they will be merged with the stable version after publication. After that, the arXiv version will just appear as an additional link, precisely as it has been before the arXiv integration.

2.2 Defining the scope

While this approach aims at following both the perception of arXiv as a preprint repository and the user expectations, it also creates specific challenges to the scope definition. There are several possible errors which may not be completely ruled out, but whose effects need to be minimized. First, of course, matching errors may either create duplicates (in case of false negatives) or missing arXiv entries (for false positive matchings). As explained in Section 4.1, the zbMATH Open arXiv matcher is now sufficiently sophisticated to ensure that such effects are negligible. However, they cannot be completely ruled out – e.g., Saharon Shelah (the mathematician with most overall arXiv submissions) has submitted individual book chapters separately to arXiv, which – due to the lack of metadata – cannot be automatically matched to the review of the complete books in zbMATH Open.

Of course, there are also arXiv submissions which will, for a variety of reasons, never be published. In this case, the entries will remain forever in their incomplete form, and eventually form a group of permanent zbMATH Open entries which would not exist otherwise. However, as this reflects the reality of mathematical information, this should be considered an asset rather than a liability. It may also happen that arXiv submissions are published in a journal or conference outside the scope of zbMATH Open. In this case, the entry will, somewhat unsatisfactorily, likewise remain forever in its preliminary form – and, moreover, it may actually not fit very well into the zbMATH Open scope (as evidenced by its final publication venue). We aim to minimize both effects by choosing only a limited and carefully chosen arXiv subset for inclusion.

This may enlarge the risk to err in the other direction: excluding arXiv preprints, although their content would be relevant to zbMATH Open. Even when they become finally available though publication in a source indexed by zbMATH Open, the information is still missing for a significant length of time, and the authors concerned may feel unduly disadvantaged in comparison with other arXiv entries which were included. Obviously, there is no ideal solution to this dilemma. The only meaningful option is a thorough analysis of how much in the different arXiv categories and subcategories has actually been published within the zbMATH Open scope in the past, and then to define a practical threshold.

The arXiv math category has 32 subcategories currently in use (subcategories that are not in use anymore are irrelevant

to the question of which recent unpublished preprints should be included). Of these, 28 have an overlap of more than 60% with the zbMATH Open corpus, i.e., they have been published in sources indexed in zbMATH Open. Some subcategories even have a share of more than 70% of the zbMATH corpus, though the differences appear to be more correlated with different publication delays and publication behaviour than with actual scope differences.

The share for the categories math.GM (general mathematics), math.HO (history and overview) and math.IT is significantly lower, although for different reasons. math.GM and math.HO contain a relatively large amount of non-research mathematics (e.g., of educational nature), while math.IT has only been introduced in 2007 as an alias for the subject area information theory in computer science. For this reason it naturally contains numerous not primarily mathematical contributions.

The only borderline case turned out to be math.ST (which is an alias of statistics theory, stat.TH) with a share of about 50%, which mostly reflects the fact that it contains not just mathematical research in statistics but is also often cross-referenced from non-mathematical categories with descriptive statistical work. Although it is almost impossible to establish a perfect distinction here, further analysis showed that the overlap with the published mathematical research literature increases sufficiently if submissions cross-referenced from non-mathematical arXiv categories are excluded – so this is the criterion currently employed as the scope definition for math.ST. On the other hand, the overlap of mathematical physics (math-ph, or as an alias, math.MP) was sufficiently high to be included as a whole.

3 New features for the arXiv entries in zbMATH Open

3.1 Extent of the newly added information
Given these premises, slightly more than 200,000 unpublished arXiv preprints have now been added to the database – a figure that is likely to grow in the future due to the overall publication growth. But this does not represent a stable corpus in itself, since documents will be both added (through new preprints) and removed (due to their publication). About 83,000 documents are (preliminarily) assigned Mathematical Subject Classification numbers as provided by the authors. This figure could be expanded by a rough automated classification based on arXiv subcategories and semantic information, but no final policy decision has yet been taken in this direction. In any case, given the preliminary nature of the documents, there will be no further intellectual classification before publication.

With current figures, combinatorics (about 10,000 items) is the largest area among the newly added preprints, followed by partial differential equations (8,700), number theory (8,600), probability theory (8,300) and algebraic geometry (8,200).

It turned out that also the abstracts could be converted without too many problems, and are thus also displayed along with the entries. This came as a surprise as, in principle, arXiv submissions are free in their LaTeX definitions and do not necessarily fit into the zbMATH Open LaTeX framework [8]. The inclusion of abstracts allows for more efficient searches and a more informative display for these items. The entries are also integrated into the zbMATH Open author disambiguation (see Section 4.2). On the other hand, there is currently no reference extraction, since the citations formats are highly non-standardized within arXiv. Moreover, references may also vary significantly between different arXiv versions of the same submission. Therefore, these citations are currently not integrated into the zbMATH Open citation database (but see also the discussion in Section 5).

3.2 Features
As a result, the additional arXiv entries are now, with respect to most of their features, searchable and retrievable. The entries can be filtered by authors, submission year, or MSC subject (Figure 2).
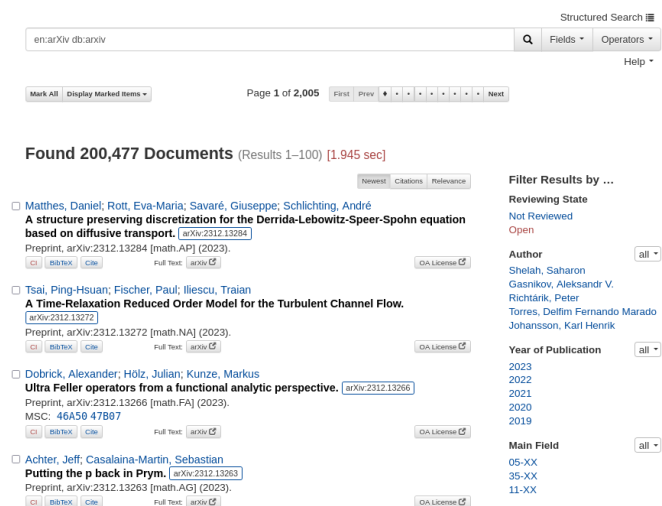

Figure 2. Filter functions for more than 200,000 newly added arXiv preprints.

Author signatures are linked to profile information, and the additional arXiv submissions are included into author profiles (Figure 3).

Another aspect is that, due to this addition, the share of open access documents available in zbMATH Open increased significantly, and approaches now about 1 million (i.e., >20%). Since these arXiv metadata are also included into the zbMATH Open APIs [4, 9], this creates additional opportunities for further research and infrastructure projects based on an open mathematics corpus.

Tao, Terence

Author ID:	tao.terence-c
Published as:	Tao, Terence; Tao, Terence C.; more...
Further Spellings:	Tao, Chi-Shen
Homepage:	https://www.math.ucla.edu/~tao/
External Links:	MacTutor · MGP · ORCID · Wikidata · Google Scholar · MathOverflow · Math-Net.Ru · dblp · GND · IdRef
Awards:	Clay Research Award (2003) · Fields Medal (2006) · Breakthrough Prize in Mathematics (2014)
Documents Indexed:	406 Publications since 1992, including 28 Books and 40 Additional arXiv Preprints 5 Contributions as Editor · 3 Further Contributions
Reviewing Activity:	111 Reviews
Biographic References:	10 Publications
Co-Authors:	156 Co-Authors with 258 Joint Publications 4,914 Co-Co-Authors

Figure 3. zbMATH author profiles with additional arXiv preprints.

4 Technical challenges and limitations

4.1 Matching

At a first glance it seems relatively easy to find a corresponding published version for an arXiv preprint. All one needs to do is to search for a document in zbMATH Open with the same title and author(s). Indeed, this approach leads to a lot of correct assignments. However, this will also miss a lot of matches as there is often a difference between the (latest) arXiv version and the published article indexed at zbMATH Open. The discrepancies include: different layout, numbering of theorems and sections, addition or deletion of content, or change of the title or abstract. Even the authors can differ in the published version. Thus, this naive matching approach would lead to a lot of duplicate entries after the arXiv integration. To avoid this problem, a precise matching procedure is needed. Our new approach is based on the DOI, title, authors and abstract of an arXiv article. It turns out that it has a higher accuracy as the naive comparison of title and authors.

Our matching algorithm has two steps. First, if an arXiv article has a DOI of a related published version given, then we search for an article with this DOI at zbMATH. If a unique article is found, we stop and return this article. Otherwise, a second matching step is executed, which is based on the title, author and abstract similarity. The idea behind this approach is that it might be that one value of those three types of metadata differs in the published version from the preprint, however, it is unlikely that all three differ simultaneously in a substantial manner.

In more detail, the second matching step works as follows. The title and authors of an arXiv preprint are used to search for a variable number of "candidate" zbMATH articles that might or might not match the input preprint. For each candidate, a three-dimensional vector is computed that contains the similarity of the title, authors and abstract of the candidate article to the input. Finally, a random forest classifier is used to decide whether a candidate matches or not. If more than one candidate matches, then the one with the lexicographic smallest similarity vector is returned

(the similarities are scaled to $[0, 1]$ and a smaller number means higher similarity).

A disadvantage of our new approach, as compared to the naive one, is its higher complexity and that training data are needed to train the random forest classifier. The training data consists of pairs of correct arXiv, zbMATH matches which were generated using DOI-matching. This leads to some wrong pairs in the training data as the DOI given by arXiv or by zbMATH might be incorrect. Overall, the DOI information is very precise, so the training data is good enough to generate a random forest with a good performance. On some test data, which was also generated using DOI-matching, the classifier matching gives a precision of 99.51% and a recall of 96.89%.

Does our matching approach also make a difference in practice? In total our new two-step matching algorithm found 250,425 matches (14th December 2023). 73,567 of those come from the DOI matching (first step) and 176,858 from the classifier matching (second step). From the 176,858 classifier matches, 144,825 have exactly the same title and authors (after some normalization steps). Which means that our approach gives us 32,033 new matches compared to a naive comparison of title and authors.

4.2 Author disambiguation

The release of the zbMATH Open interface with the additional arXiv entries is also internally coupled with a new version of the automated author disambiguation. Beside slightly improved matching results, its main advantage is mostly invisible to the outside – its much more modularized structure was a prerequisite to handle both the new amount and somewhat different structure of the newly added items. Naturally, the assignments of arXiv preprints are somewhat less reliable than for the published corpus, due to the higher occurrence of metadata errors (e.g., in separating author names). Also, the unstable nature of arXiv entries must be handled appropriately in the assignment tables. Moreover, there is also the option that arXiv papers may be withdrawn – currently, we follow the arXiv policy that they are nevertheless still available (Figure 4).

Figure 4. A withdrawn arXiv submission with a separator error.

Viergever, Anna M.

Author ID: viergever.anna-m ⧉
Further Spellings: Viergever, Anneloes

Documents Indexed: 2 Publications since 2023, including 2 arXiv Preprints
Reviewing Activity: 1 Review

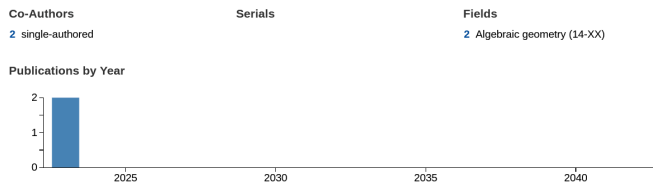| Co-Authors		Serials		Fields	
2	single-authored			2	Algebraic geometry (14-XX)

Publications by Year

Figure 5. A new author profile based on arXiv submissions.

Perhaps most importantly, many new profiles for mostly young researchers have been created who have currently only arXiv submissions, but not yet published articles. Their enhanced visibility will be one of the most immediate effects of the integration (Figure 5).

5   Future challenges

As discussed in Section 3, the additional arXiv entries are partially integrated into the whole zbMATH Open framework, but not all features are currently available. The addition of an automated preliminary classification for the items that have not yet an MSC assigned by the authors is technically feasible and would not be very difficult to incorporate. Also, for some arXiv submissions we already collect software information, although it is not yet displayed. Once fully available, this would further allow for the integration of these entries into the mathematical software information service swMATH. Likewise, the interlinking with the Online Encyclopedia of Integer Sequences (OEIS), the Digital Library of Mathematical Functions (DMFL), and the question-and-answer platform Math-Overflow – as provided for stable zbMATH Open documents – is not yet implemented, and requires further development work. Probably the most challenging task will be the integration into the zbMATH citation database, due to the various formats and unstable version entries of arXiv references.

On the positive side, the availability of open LATEX sources for arXiv documents opens up a considerable future potential. Depending on the licence, a multitude of full-text features could be developed and implemented, ranging from full-text formula search to features involving large language models based on a reliable mathematical corpus. With the information available on the zbMATH Open APIs, we invite the community to develop interesting features which include the newly added information!

References

[1] A. Bannister, K. Hulek and O. Teschke, Das Zitationsverhalten in mathematischen Arbeiten. Einige Anmerkungen. *Mitt. Dtsch. Math.-Ver.* **25**, 208–214 (2017)

[2] A. Bannister and O. Teschke, An update on time lag in mathematical references, preprint relevance, and subject specifics. *Eur. Math. Soc. Newsl.* **106**, 41–43 (2017)

[3] D. Ehsani and O. Teschke, On the road to a comprehensive open digital mathematics library. *Eur. Math. Soc. Newsl.* **118**, 76–78 (2019)

[4] M. Fuhrmann and F. Müller, A REST API for zbMATH Open access. *Eur. Math. Soc. Mag.* **130**, 63–65 (2023)

[5] K. Hulek and O. Teschke, How do mathematicians publish? – Some trends. *Eur. Math. Soc. Mag.* **129**, 36–41 (2023)

[6] F. Müller and O. Teschke, Will all mathematics be on the arXiv (soon)? *Eur. Math. Soc. Newsl.* **99**, 55–57 (2016)

[7] F. Müller and O. Teschke, Progress of self-archiving within the DML corpus, with a view toward community dynamics. In *Intelligent Computer Mathematics (CICM 2016)*, Lecture Notes in Comput. Sci. 9791, pp. 63–74, Springer, Cham (2016)

[8] M. Schubotz and O. Teschke, Four decades of TEX at zbMATH. *Eur. Math. Soc. Newsl.* **112**, 50–52 (2019)

[9] M. Schubotz and O. Teschke, zbMATH Open: Towards standardized machine interfaces to expose bibliographic metadata. *Eur. Math. Soc. Mag.* **119**, 50–53 (2021)

———

Isabel Beckenbach studied mathematics at FernUniversität in Hagen, Technical University of Berlin, and Free University of Berlin. From 2013 to 2019 she worked at Zuse Institute Berlin in the Network Optimization department where she completed her PhD in combinatorial optimization and graph theory. In 2020, she moved to FIZ Karlsruhe, and now she works on the transition of zbMATH towards an open information platform for mathematics.

isabel.beckenbach@fiz-karlsruhe.de

Klaus Hulek is professor emeritus at Leibniz University Hannover, working in algebraic geometry. He was editor-in-chief of zbMATH Open from 2016 to 2023.

hulek@math.uni-hannover.de

Olaf Teschke is the Head of the Mathematics Department of FIZ Karlsruhe, he is also the managing editor of zbMATH Open and chair of the EMS Committee on Publications and Electronic Dissemination.

olaf.teschke@fiz-karlsruhe.de