CEIC Report

May 15, 2022

What follows is a summary of the Committee on Electronic Information and Communication's principal activities since the last IMU General Assembly, namely best practices documents and the Global Digital Mathematical Library. For minor topics (such as assisting the ICM Structure Committee), see the CEIC yearly reports.

1 Best practices documents

One task with which we started this cycle of CEIC was revising the existing CEIC recommendations. We prioritized this over writing new recommendations, since some of the existing documents were very much out of date. We moved some documents to a historical section, on the grounds that revisions are not needed (for example, *Call to All Mathematiciains to Make Publications Electronically Available* was an important message in 2001, but that battle has been won). The two documents we identified as being most in need of revisions were *Best Current Practices for Journals* (from 2010) and *Copyright Recommendations* (from 2001).

We revised *Best Current Practices for Journals* to bring it up to date and posted it to the CEIC website, while moving all outdated documents to a historical section.

After considerable discussion of alternatives, we decided not to revise *Copyright Recommendations*. The document itself was excellent for its time period, but by now it is more than twenty years out of date. During that time, the publishing landscape changed dramatically. The original document never mentioned the arXiv, Creative Commons licenses, or open access; instead, it focused on print publishing, while acknowledging electronic distribution or publishing as an emerging area. Much of the discussion, while not technically wrong, is of limited importance today.

Because of the evolving landscape, it was not feasible to update *Copyright Recommendations* through modest edits or additions. Instead, an up to date document would have to be written from scratch, and the key question was what would be most useful to convey. Basic information about copyright and its relevance to scholarly publishing is much more easily accessible today than it was twenty years ago. Instead, what's really needed is concrete recommendations for

what mathematicians publishing papers, journal editors, and publishers should do.

However, the committee found it difficult to reach consensus on these issues. There are many uncontroversial issues; for example, every mathematician publishing a paper should retain the right to post it to the arXiv. However, this battle seems to have been won already, and we are not aware of any mathematical publishers that object to use of the arXiv. What remains are much more subtle issues, such as:

- What are the minimal professional expectations of mathematicians publishing papers? For example, if someone chooses to publish in a closed-access journal, are they exercising their freedom to disseminate their work as they see fit, or are they harming the mathematical community? What if someone chooses not to make a preprint freely available, because they prefer for attention to go to the definitive, published version of record? Is there anything we can state every mathematician should do? If we don't, are we implicitly saying there is no professional obligation to do anything at all about access to the mathematical literature? If we do suggest a minimum, are we also indicating it's enough?

- What about editors? To what extent should they take responsibility for ensuring that the publisher's policies and business model serve the research community well? Or is this simply above their pay grade?

- Are certain business models unethical? For example, consider open access publishing based on article processing charges. Is this a perfectly reasonable approach, provided that editorial decisions are independent of the ability to pay and subsidies are available for those who cannot pay? Or is it an intrinsically flawed model? And, if so, is it better or worse than a model in which publishing is free but reading is expensive? What about the risk of predatory publishers that use low standards to collect as many fees as possible? Is this a disaster or merely an unfortunate side effect?

- Do we really understand which models are scalable and sustainable? For example, are diamond open access journals (no fees for anyone to publish or read) based on volunteer labor an altogether superior solution, or is skilled volunteer labor itself too limited a resource? Will article processing charges distort how publishing works? What about "subscribe to open" models? Will they save learned society journals, or eventually collapse?

It's valuable for CEIC to discuss and debate these issues, but for many of them it seems to be too early to reach a conclusion. Recommendations from CEIC are most useful when they reflect a strong consensus among a representative group of those with expertise in this area. That's the case with *Best Current Practices for Journals*, but there was considerable disagreement about what sorts of copyright recommendations are necessary or desirable. There are a number of overviews and proposals from other groups available on the web, and adding to

them did not seem likely to help settle these issues. Instead, they will require further discussion and experience over time.

1.1 Other potential topics for best practices documents

Another topic in a somewhat similar state is citation statistics. We can all agree that many applications of citation statistics to judging research productivity are bad, and some are truly terrible. The IMU has produced a report that highlights and explains some of these issues. One natural followup question is whether the IMU can give further guidance and help draw distinctions between better and worse uses of citation statistics, without seeming to endorse practices that, even in their best form, still cause serious concern. For example, can we communicate the message that unnormalized metrics are worse than normalized metrics, without seeming to endorse the use of normalized metrics? CEIC has not yet reached consensus on this topic. Important issues include openness of tools and data, legibility of algorithms, and the fact that behavior can adapt to evaluation methods and thereby ruin the quality of the evaluation.

Open access funding models is also a contentious issue that has not yet reached the point where a specific IMU recommendation is desirable. Funding models include article processing charges and read/publish agreements, as well as more traditional approaches such as subscriptions (or subsidization in other ways). It would not make sense for the CEIC to make specific recommendations for funding models before the community gets more experience with how these approaches play out in practice, since their financial viability and scholarly side effects are not yet fully clear.

2 The Global Digital Mathematical Library

2.1 Introduction and background

In 2006, the International Mathematical Union (IMU) General Assembly endorsed the idea of a Global Digital Mathematics Library (GDML), which would provide seamless access to the mathematical literature and make as much of it as possible freely available. This project is clearly important for mathematics and for society more broadly, but it is an ambitious undertaking, and there has been little progress in creating such a library.

The biggest step so far has been the creation of the International Mathematical Knowledge Trust (IMKT) in 2015, which resulted from an initiative of the IMU's GDML working group. Currently, the IMKT has important projects in several areas, such as special functions, formalization of mathematics, and document analysis. However, most of these projects are focused on adjacent issues, rather than short-term progress towards a GDML exposing the subject's literature to wider view. This is because it was possible to encourage community interest in these issues.

In 2014, a committee of the Board on Mathematical Sciences and their

Applications issued a report on the GDML, called Developing a 21st Century Global Library for Mathematics Research. This 143-page book does an excellent job of laying out the justification for a GDML and its potential benefits, as well as the issues and challenges that must be confronted, and it concludes with a strategic plan. Unfortunately, nobody has been found to take a leadership role in funding and implementing the plan, and there seems to be little momentum in that direction.

From our perspective, part of the reason why this project has stalled could be seen as prioritizing ambitious goals over taking seemingly obvious initial steps. One explanation is that funding agencies continue to be more likely to fund smaller research projects with aspirations over efforts that can be seen as development and deployment of known methods, particularly if the resulting public good is going to have to be maintained over the long run. Many of the people in this area are excited by questions about how to store and present mathematics online, how to integrate it with mathematical software and formal proof systems, how to search for mathematics and return useful results despite differences in terminology or notation, how to assist users in navigating the literature, etc. These are deep and fundamental questions, which will occupy mathematicians, librarians, and data scientists for generations to come. If we try to solve them definitively before creating a GDML, then we shall not see one in our lifetimes.

Instead, we're in favor of a pragmatic approach, starting with whatever would create the greatest short-term benefits while building momentum towards a full GDML. We should avoid controversy, difficulties, and waiting for further research, while focusing on low-hanging fruit.

2.2 GDML_0

As explained in more detail in the document *zbMATH as* $GDML_0$ (by Patrick D. F. Ion and Mila Runnwerth), CEIC proposes looking at the GDML as a combination of several contributions:

- We need an initial interface and index, which one might call GDML_0.
- We need as many papers as possible to be freely available.
- We need to build tools and interfaces on top of GDML_0. For example, how should search work? Can it handle formulas and changes of notation or terminology? These sorts of questions are far more subtle and complex, but solutions can evolve over time once the foundation is in place.

2.2.1 zbMATH Open

The best currently existing foundation for GDML_0 is zbMATH Open, which is the online version of Zentralblatt and is now freely available worldwide. Its transition to an open access database was explicitly motivated by the IMU's GDML vision, as acknowledged on their website:

> "After a concerted effort by zbMATH stakeholders, the Joint Science Conference (Gemeinsame Wissenschaftskonferenz) agreed in the first week of December 2019 that the Federal and State Governments of Germany would support FIZ Karlsruhe to transform zbMATH into an open platform. As a result zbMATH Open has become a freely accessible research tool for the mathematical community worldwide since January 2021.
>
> This step was inspired by the International Mathematical Union's 2014 vision of a Global Digital Mathematics Library: 'to provide a coherent and sustainable open platform in which all mathematics-relevant information and data can be brought together, comprehensively accessed and used free of charge under a uniform interface'."

Currently, zbMATH Open contains approximately 4.2 million database entries for mathematical papers and books. All of the data compiled or created by zbMATH itself is available under a Creative Commons BY-SA license (an open access license allowing reuse with attribution, provided that the results are also made available under the same terms), while some bibliographic data supplied by publishers is more restrictively licensed.

Among freely available databases,[1] zbMATH Open has by far the best and most accurate coverage of mathematics, and it would be an ideal portal to the GDML. Many mathematicians access the literature through tools such as Google Scholar. These tools are useful, but they suffer from several disadvantages: they produce automated results of mixed quality, and they are under the control of for-profit companies (whose interests may not always align with the mathematical community's interests in the future). Now that zbMATH Open is freely available, we expect that it will become an increasingly important tool for mathematicians, as well as the foundation for the GDML.

2.2.2 OpenAlex

OpenAlex (named after the Library of Alexandria) is an intriguing database that was founded in 2021 to replace Microsoft Academic Graph (Microsoft's competitor to Google Scholar, which was discontinued at the end of 2021). It remains to be seen how important OpenAlex will be for mathematics, but it is well funded and ambitious, and it has the potential to be of considerable use, particularly for the interface between mathematics and its applications. This is an area that both zbMATH Open and MathSciNet struggle with, since identifying and indexing mathematics across a huge range of journals in adjacent fields is even more challenging than dealing with mathematics journals per se. In the long run, the GDML may end up combining a core database such as zbMATH Open with supplementary data and references from broader indexes such as OpenAlex.

[1] Its primary competitor, MathSciNet, is not freely available.

2.2.3 The Sophize Project

The Sophize Foundation, founded by Abhishek Chugh, is another noteworthy attempt to address this issue. Sophize is a software development effort aimed at creating something like a GDML. Specifically, Chugh proposes the following as the goals of the Sophize Project:

> "Develop an open state-of-the-art academic platform with the following capability to:
>
> - Aggregate knowledge from Mathematics resources including encyclopedias, research papers, mathematical reference books, mathematical databases, and formal systems to make it easier to search and explore Mathematical content.
> - Formalize the largely unexplored network of information that connects mathematical objects to create rich research experiences. These deep connections will also be used to connect other significant efforts such as FABstracts, DLMF, Lean, MMT, Flexiformal Mathematics etc.
> - Allow researchers to easily run Mathematics-related computations online in languages like Python (Sage), Java, R, Mathematica.
> - Facilitate knowledge-based group discussions and academic collaborations."

This is an ambitious and well-informed software development effort, which we hope will contribute to a future GDML.

The IMU had $3,689.45 left over from a Sloan Foundation grant aimed at GDML-related activities. We gave this money to the Sophize Foundation to fund the creation of a prototype Online Structured Content Extraction Tool, which is now available as the Structured Data Extractor. Both the code and a project report are available online. While this tool is only a small part of a full GDML, it's a valuable step forwards.

2.2.4 Other digital libraries

There are also a number of other digital libraries with nontrivial coverage of mathematics, including:

1. EuDML, https://eudml.org/
2. Réseau National des Bibliothèques de Mathématiques, https://www.rnbm.org/
3. HathiTrust, https://www.hathitrust.org/
4. Göttinger Digitalisierungszentrum, https://gdz.sub.uni-goettingen.de/

5. ISTEX, https://www.istex.fr/
6. SciELO, http://www.scielo.br/
7. Digital Mathematics Registry, https://mathscinet.ams.org/dmr/
8. Dissemin, https://dissem.in/
9. Sci-Hub (which of course has legal issues with its approach)

While most of these are not intended to be as broad in scope as the GDML, they represent important experience dealing with the same issues.

2.3 Access to back issues

The most challenging aspect of the GDML is ensuring access to mathematics papers. One key step would be access to back issues via a moving wall, in which all papers published in mathematics journals are eventually made freely and publicly available under a suitable license. This license should enable not just access from the publisher's website, but ideally also third-party hosting and distribution, for example as part of the GDML.

The American Mathematical Society (AMS) and Elsevier both make papers from their mathematics journals available for free after at most five years,[2] although neither one currently uses a license as permissive as we'd like. The fact that free access to back issues is compatible with the business models of both the AMS and Elsevier is a very encouraging sign: anything these two very different publishers can agree on should be broadly acceptable within the mathematics publishing industry. A five-year window seems like an uncontroversial choice, which would not overly constrain publishing or funding models.

We hope this idea achieves critical mass and spreads universally among mathematics publishers. At CEIC's recommendation, the IMU Executive Committee has established a Committee on Permissions to study this idea, discuss it with publishers, and offer concrete recommendations. We hope this will be an important step forward.

Of course we don't see this proposal as an ideal long-term solution. We hope the research community eventually ends up with immediate open access to all research papers, and we believe the world is generally headed in that direction. However, it's a difficult issue. For example, immediate open access is not compatible with a subscription model for journal funding, and it's not clear how to replace subscriptions with a funding model that would be well suited for mathematics. It isn't clear what consequences different funding models will have for scientific publishing. Instead of trying to resolve this issue, we should focus at once on the much less controversial topic of back issues.

Another tricky topic is books, for two reasons: the way books are sold is even less compatible with open access than journals (publishers are more reliant on a long tail for sales of popular books), and authors generally receive royalties. This

[2] Four years in the case of Elsevier.

does not amount to much money for the vast majority of authors of mathematics books, but it does for a few, and even those not making much in royalties might be legitimately unhappy to lose this income. In the long run, the GDML will have to deal somehow with the issue of books. However, that is best postponed until easier and more important issues have been settled, and it can be approached separately; there are possibilities for digital forms of older books to be included in the GDML_0 presently envisaged.

Another issue is where to draw the lines between mathematics journals and those in computer science, physics, statistics, data science, or other fields. There is no principled answer to this question, and the best we can do right now is to work with publishers to draw the lines somewhere. Ultimately, all of academia needs to address this problem, but mathematicians are in an excellent position to take the first steps.

2.4 Benefits of focusing on mathematics

Mathematics is an ideal test case for opening up back issues, for two reasons. One is that mathematics papers remain relevant to current research for an exceptionally long time, far longer than in most other disciplines, so mathematicians care more about back issues than just about anyone else. The other is that mathematicians are particularly knowledgeable and enthusiastic about open access. We're in the same league as physicists, and far ahead of most of the sciences and humanities. This makes it quite a bit easier to convince the community to work for the sort of changes we're discussing.

The flip side of this is that publishers won't be giving up anything of great value in their overall scheme of things. We believe that's why Elsevier has been willing to open its archives: mathematicians consider this access valuable, but it's not worth a lot of money in other fields, so there's no risk of financial disaster if the idea spreads beyond mathematics.

3 What remains

Thanks to initiatives such as Plan S, many mathematicians have been thinking about open access and the future of publishing. A window-based compromise that could enable the GDML without causing unnecessary disruption. From this perspective, there are three aspects that will engage the Permissions Committee, CEIC, and the broader community:

1. Reaching consensus that mathematics papers should not be locked behind paywalls for the entire term of copyright (typically the life of the author plus 50–70 years), but should instead be available to all readers within some reasonable span of time. This is an ethical issue on which we anticipate strong support from mathematicians and little opposition from publishers.

2. Agreeing on what a reasonable span of time might be. Individual opinions might range from zero (for open access advocates) to decades, but based

on the examples of the AMS and Elsevier, we hope four to five years could be a widely accepted compromise. Of course it would only establish a baseline, and some papers could be available much earlier.

3. Discussing legal technicalities involving licenses specifying what users can or cannot do with the papers. Standard options include Creative Commons licenses, with a proliferation of nonstandard licenses as well (e.g., Elsevier's). Users often don't think about licenses or terms of service (such as zbMATH's), but services such as the GDML must comply with them. One advantage of standardization and predictability is that they make it easier and less risky for users to neglect these issues. (Nobody becomes a mathematician because they want to think about copyright licenses, and we aim for a world in which mathematicians do not have to think about this issue individually if they don't want to.)

One might ask whether this issue matters, given that many papers are freely available on the arXiv and many more are illegally available via Sci-Hub. One issue is that the Global Digital Mathematical Library cannot primarily rely on sites that are illegal and move periodically to avoid disruption, or sites that are incomplete and contains copies that may be out of date. More generally, mathematics papers are not fungible: access to 70–80% of the papers you need, or even 95–99%, does you little good if you can't get your hands on a key reference. From that perspective, every increase in accessibility is a victory, even if many papers are already available. In addition, it's a matter of equity: all mathematicians should be treated as first class citizens in the mathematical community, rather than relegated to partial or illegal access. A compromise based on time since publication would not fully solve the equity issue, but it would be a significant step forward.