The Digital Mathematics Library as of 2014

Thierry Bouche

TEX and the Internet appeared in the 1980s. In 1991, Paul Ginsparg started a small HTTP server where people could upload their preprints written in TEX or LATEX rather than circulating them by postal mail. During the 1990s most scientific publishers started their electronic publishing platforms, and digitization programs such as Gallica, JSTOR, and GDZ were launched. Digital publishing at the end of the twentieth century could now be called the digital incunables as since then things have started to look pretty mature, evolving at a much slower pace. The basic layers such as TCP/IP, HTML, PDF, and XML were in place and have since then remained quite stable. At that time, the cliché of the World Wide Web as a universal library where all mankind's knowledge would rapidly be easily retrievable through hypertext links was well established. Ironically, Google itself was started in the context of a project aiming "to develop the enabling technologies for a single, integrated and 'universal' library" [7].

Thierry Bouche is director of Cellule Mathdoc (makers of NUMDAM, CEDRAM, mini-DML) and a member of the Institut Fourier at Université Grenoble Alpes, F-38000 Grenoble, France. He is the chair of the EUDML initiative, and a member of CEIC (IMU). His email address is thierry.bouche@ujf-grenoble.fr.

The author thanks Jim Pitman, Jiří Rákosník, and the column's editors for their help in the preparation of this text.

DOI: http://dx.doi.org/10.1090/noti1162

In 2001 every working mathematician was already supposed to have a personal homepage: The International Mathematical Union (IMU) emitted a "Call to all mathematicians to make their publications electronically available" [5]. This was the first effective, somewhat naive, attempt to create from scratch an open access universal library of the mathematical corpus. The idea was that a mathematician would maintain a webpage with a list of his/her published works, with links to full texts, either typeset (or rekeyed) with TEX or scanned. Supposedly, discovery and access would rely on generic Web crawlers and search engines.

At the same time, the project of digitizing twenty centuries of the published mathematical literature was discussed, first in Washington DC, but the discussion rapidly spread worldwide, filed under the name "The digital mathematical library" (DML) [4], [6]. An incarnation of this idea has been the DML list maintained by Ulf Rehmann [11].

Today, almost fifteen years later, this duality persists between the vision of the digital counterpart to the traditional library and the crowd-sourced constitution of a loose collection of digital content, based on availability rather than selection, with the hope that interlinking and advanced indexing, ranking or semantic computations will help users find the pearls.

Publishing and Library Businesses Differ

In today's glossary, a *digital library* is any collection of digital objects with a management plan. Thus the website of a commercial publisher constitutes a digital library as well as a linked webpage.

However, the functions and skills expected from publishers and librarians are not to be confused. Publishing is a business that aims at delivering new content, which is afterwards exploited over a limited period of time. It is risky because it requires a lot of investment prior to publication, while success is never certain. The library business has evolved a lot over the ages, but it is pretty well understood (as long as we stick to paper) that it involves content selection and acquisition, cataloguing, archiving, and preservation in order to maintain collections accessible to patrons. It is nowadays predominantly a nonprofit activity, typically a service offered by scientific institutions to their members.

It is amazing how many of the grand maneuvers in science, technology, and medicine (STM) publishing have been aimed at reducing the risks of publishing while removing the library functions from universities (see [9], [10]): journal Big Deals, subscriptions to book series, or, more generally, licensing access to bundled resources on the Web are effective methods to take away the selection process from libraries. The acquisition is limited to transferring an amount of money to license content hosted by the publisher. Gold open access publishing is in a sense the ultimate weapon in this direction as the publication costs are covered in advance and there is no longer any risk that an unsatisfied consumer will cancel the subscription.

Being the collection's host, the publisher should therefore take care of preservation and long-term access. Now even the discussions within the library buildings are being privatized through the development of so-called scientific social networks!

I think it is time to stop this trend and to balance the weight of libraries and publishers in the interest of science. Contrary to many colleagues, I don't think we should eliminate publishers. In fact the Internet has facilitated the creation of many publishing venues where publishing skills and professionalism are missing. Publishers should concentrate on producing new quality content and services while the archive should be selected and curated by nonprofit long-lasting institutions.

As a side note, I think it is also important to keep in mind the distinction between the library as a raw reservoir of original works and the encyclopedia as a gateway to highlighted and synthesized parts of that knowledge. An encyclopedia can be a very effective way to enter the library and to single out distinguished entry points to some subject. By design it cannot be neutral and exhaustive. In fact, the library is the infrastructure that stores original results, on top of which as many different gateways can be built as user communities require.

The DML: Vision and Content

The initial DML vision was to digitize all published mathematics since Euclid up to the advent of electronic publishing in order to assemble a reference corpus, highly navigable and freely accessible. The motivation to do this is that verified mathematical results not only do not age but can find unexpected applications at any point in time, so a freely and easily accessible corpus would feed both current mathematical research and further innovation. Also, as mathematical writing is *cumulative* (new results are built upon old ones; they do not replace them), we feel uncomfortable as long as the reference corpus is incomplete. A pragmatic reason for urging in this direction at the end of twentieth century was a certain kind of academic Internet bubble: the belief that any move toward digital science would be heavily funded. But the bubble burst and the envisioned centralized project vanished, leaving us with a name and numerous local (many of them national) DML projects, each one with unique standards and policies.

Today, the portion of the mathematical corpus that exists in digital form is not far from exhaustive (of course the proportion of digital items is steadily approaching 100 percent as each new item is produced digitally, but there are still important references that have not been digitized). The landscape has been changed by the systematic digitization of backfiles by publishers, of entire shelf holdings from some large libraries, as well as a few dedicated projects. But most of this is unreachable to the working mathematician.

The IMU has advocated free access to the mathematical corpus after a reasonable time lag ("moving wall"), but this advice has not been followed beyond some academic circles. Well, it might have been heard in the form of understanding that the archive had value to mathematicians, hence could be sold, which is probably the dominant model currently. The DML content exists, but it is highly fragmented and locked behind a multitude of walls.

Typical barriers encountered by the working mathematician are:

- Digital content has been produced by a commercial provider and is behind a paywall. This also concerns very old content (including public domain).
- Digital content is available somewhere but practically invisible (badly indexed or unstable webpages, nonexistent or inaccurate metadata, no interoperability with those search systems the users are familiar with…).

These barriers move all the time: A Web crawler can suddenly index or forget a page; Elsevier freed access to four-year-old backfiles of fifty-three mathematical journals, which can be reverted whenever they see fit as there is no third party archive; JSTOR decided to provide open access to public domain texts after denying it for years; the *Annals of Mathematics* sought to withdraw the freely available content from its open access era when it reverted to the subscription model, etc.

Recent Developments

For almost ten years the only DML-related activity has been focused on creating and maintaining local DMLs as separate islands. Some consensus conferences have been held sporadically, typically in conjunction with a rumor of possible funding.

In 2008, the DML workshop was created by Petr Sojka as a companion to the Mathematical Knowledge Management (MKM) conference (it is now a track in the CICM series of conferences [3]). The main focus is on computer science research that could impact any aspect of a DML.

The Sloan Foundation supported a symposium on "The Future World Heritage Digital Mathematics Library: Plans and Prospects" held at the National Academy of Sciences, Washington, DC, in 2012, organized by the Committee on Electronic Information and Communication (CEIC) of the IMU. A committee on planning a global library of the mathematical sciences was formed at the National Research Council (NRC) right after and produced a report [8] which is summarized by Jim Pitman and Clifford Lynch in the June/July 2014 issue of the *Notices* ("Planning a 21st century global library for mathematics research," pp. 776–777).

Since 2003, the European Mathematical Society has tried to get some funding to advance the European chapter of the DML. The EuDML[1] project proposal was finally selected by the European Commission and the project started in February 2010 for a duration of three years. I have written a personal summary of the achievements of this project, to which I refer the reader for further details [1]. Earlier this year, the EuDML Initiative was established as the association of the partners taking care of the future of this effort [2].

EuDML was a breakthrough in a number of aspects.

- It was the first project to effectively build a transnational DML, built on local DMLs from eight countries.
- It set up the machinery to harvest and transform metadata from thirteen sources and exploit it on a single user-friendly website. It tried to make much more content visible by upgrading metadata at the article level.
- It defined and deployed a number of machine-oriented services that are probably not yet known to the extent they should be: it is possible to selectively harvest parts of the database content (using OAI-PMH protocol) but also to use it as a metadata hub (matching a citation against the database, then querying all identifiers associated to an EuDML item, etc.)
- Although most of the content is journal articles, all relevant content types are explicitly supported: edited books such as conference proceedings, monographs, dissertations, multivolume works.
- It didn't target simple textual metadata aggregation but experimented with many MKM techniques and results, making the mathematical nature of the content an asset rather than a liability. To name a few: conversion from TEX metadata to MathML, mathematical OCR, formula search.
- It served as a sandbox to try out more experimental features (such as linking content through mathematical content similarity, making mathematical content more accessible to dyslexic or visually impaired users, etc.)

Also, clear policies have been devised to build a reliable, sustainable system with nonvolatile content:

(1) EuDML content is scientifically validated and formally published in final form.
(2) The digital content is physically hosted at one of the partner institutions (local DML).
(3) It is freely accessible after a reasonable moving wall (which typically ranges from zero, or open access, to five years).

Starting from there, it should be relatively straightforward to enlarge the coverage to the point where EuDML becomes a valuable resource to the working mathematician as well as the educated citizen. The first front that will be engaged by the EuDML Initiative is to enlarge the content partnership beyond the project's partners (this already started with Italian and Serbian DMLs joining) and beyond European Union borders.

The benefits of the EuDML are already obvious, even to users unaware of its existence, through the better visibility of many sources in Web search engines. However, it won't be the one-stop shop for all mathematics literature as long as the underlying corpus is restricted to a limited number of benevolent content providers with sufficient resources to make their systems compliant. On

[1] *EuDML: The European Digital Mathematics Library; see* `http://eudml.org`.

this front, my guess is that we can cross a new boundary if every scientific editor of every live journal or book series requires from its publishers cooperation with a local DML to archive its content and arrange a continuous flow for new material. Beware that all *three* conditions in EuDML policy above should be checked or else this could be reversed afterwards! If the publisher refuses, move to a more sympathetic publisher. There are many of them, and they will typically be nearer to the community, thus serving your publication better.

What Next?

One new challenge is the number of non- (or hardly) archivable mathematical contents like Polymath (relying on a combination of a blog and a wiki), personal collected works on professional homepages after leave or retirement of the author, collective books that are indeed a version control repository (like the Univalent Foundations book from Princeton), PlanetMath, MathOverflow, etc. While these are properly indexable and crawlable by Web search engines, it is not at all obvious how they will be preserved and referred to in decades to come, if ever this happens.

Nevertheless, I hope that in the near future we will see clear progress in the following areas:

- The EuDML Initiative will establish a strong organization (Europe-based, but with worldwide scope) which will convince more stakeholders to join and adhere to its goals and methods. The number of items available through the EuDML should reach a million, the point of no return.
- The NRC committee's suggested creation of a linked open mathematical concept network will be implemented by a dedicated team, using freely available digital resources as a testbed.

Then, glueing the bits and pieces together, we can expect to advance the DML vision to the point where doing mathematical research and interacting with the mathematical corpus might be completely different and much more powerful in ten years than what we experience today.

References

[1] THIERRY BOUCHE, Reviving the free public scientific library in the digital age? The EuDML project, in *Topics and Issues in Electronic Publishing*, proceedings of the AMS Special Session on Topics and Issues in Electronic Publishing at the 2013 Joint Mathematics Meetings, January 9–10, 2013, San Diego, (Klaus Kaiser, Steven Krantz, and Bernd Wegner, eds.), FIZ Karlsruhe, 2013, pp. 57–80. See http://www.emis.de/proceedings/TIEP2013/05bouche.pdf.

[2] THIERRY BOUCHE and JIŘÍ RÁKOSNÍK, Report on the EuDML external cooperation model, in *Topics and Issues in Electronic Publishing*, proceedings of the AMS Special Session on Topics and Issues in Electronic Publishing at the 2013 JMM (see Ref. [1]), pp. 99–108. See http://www.emis.de/proceedings/TIEP2013/07bouche_rakosnik.pdf.

[3] DML workshop and CICM DML track. Information: http://www.fi.muni.cz/~sojka/dml-2011.html, http://www.cicm-conference.org/. Proceedings: http://dml.cz/handle/10338.dmlcz/702563.

[4] John Ewing, Twenty centuries of mathematics: Digitizing and disseminating the past mathematical literature, *Notices of the AMS* **49**(7), 2002, pp. 771–777. See http://www.ams.org/notices/200207/fea-ewing.pdf.

[5] IMU, Call to all mathematicians to make publications electronically available, endorsed by the IMU Executive Committee on May 15, 2001, in its 68th session in Princeton, NJ. See http://www.mathunion.org/fileadmin/CEIC/Publications/Call_to_All_Mathematicians_to_Make_Publications_Electronically_Available.pdf.

[6] Allyn Jackson, The digital mathematics library, *Notices of the AMS* **50**(8), 2003, pp. 918–923. See http://www.ams.org/notices/200308/comm-jackson.pdf.

[7] NSF, The Stanford Integrated Digital Library Project. See http://www.nsf.gov/awardsearch/showAward?AWD_ID=9411306.

[8] National Research Council, Committee on Planning a Global Library of the Mathematical Sciences, *Developing a 21st Century Global Library for Mathematics Research*, The National Academies Press, Washington, DC, 2014. See http://www.nap.edu/catalog.php?record_id=18619 or http://arxiv.org/abs/1404.1905.

[9] Andrew Odlyzko, Open access, library and publisher competition, and the evolution of the general commerce, preprint, 2013. See http://arxiv.org/abs/1302.1105.

[10] ———, Are libraries and open access becoming irrelevant? *Notices of the AMS* **61**(4), 2014, pp. 390–392. DOI 10.1090/noti1093.

[11] Ulf Rehmann, DML: Digital mathematics library, retrodigitized mathematics journals and monographs. See http://www.mathematik.uni-bielefeld.de/~rehmann/DML/dml_links.html.