Klaus Kaiser, Steven Krantz, Bernd Wegner (Eds.)

Topics and Issues in Electronic Publishing

JMM, Special Session
San Diego, January 2013
FIZ Karlsruhe

Klaus Kaiser, Steven Krantz, Bernd Wegner (Eds.)

Topics and Issues in Electronic Publishing

JMM, Special Session

San Diego, January 2013

FIZ Karlsruhe

Preface

At the January, 2013 joint meeting of the AMS/MSS in San Diego, California, Klaus Kaiser, Steven G. Krantz, and Elizabeth Loew organized a special session on contemporary issues in mathematical publishing. The sessions were well attended, and the presentations were accompanied by spirited discussions. [We note that Elizabeth Loew did not participate in putting together this volume. Bernd Wegner was not an organizer of the sessions, but he has been involved in other such meetings and he did yeoman service in soliciting papers for and assembling this volume.]

One of the main themes of these events was a consideration of the impact of electronic media on publishing. Clearly publishing is an essential part of what we do. Along with teaching and administrative work, mathematicians prove theorems and publish them. We need to publish in order to establish our scholarly reputations, in order to help develop the subject, and in order to validate and archive what we do. The traditional hard-copy model for publishing has served us well for several centuries. But now the picture has changed dramatically.

Many journals are now available both in hard copy and OnLine. Other journals are electronic only. Questions arise as to how we are going to archive electronic journals, how we are going to maintain scholarly quality, and how we are going to manage work flow. There are also questions about pricing. There are both the Green and Gold models for Open Access journals, and the latter involves an Author Processing Charge (APC) which is often nontrivial (usually in the thousands of dollars). How will such a fee be paid, and who will pay it? We certainly do not want a publishing model in mathematics that will exclude scholars in underdeveloped countries, or scholars who do not have resources at their disposal.

Electronic books also have a significant impact on scholarly publishing. An electronic book is one that can keep evolving and growing, and that can respond to feedback from readers. It is a new and developing technology, and one that promises to enhance our abilities as scholars.

All of these questions were vigorously discussed at our special session. It is safe to say that everyone came away with a clearer sense of where we are and where we may be headed. It is clear that there is still much to learn and much to do. This volume contains contributions from many of the participants, and from other scholars as well. These contributions serve as a record of what we were able

to accomplish at our meeting. We hope that readers will come away with a new understanding of the publishing world that we now live in.

Klaus Kaiser
Steven G. Krantz
Bernd Wegner

Topics and Issues in Electronic Publishing

CONTENTS

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

"AUTHOR SPONSORED JOURNALS", A PRACTICAL VARIATION OF THE "OPEN ACCESS" BUSINESS MODEL

KLAUS KAISER

1. Conventional business models for journal publishers

All publishers, whether they are "mega"– publishers, like Springer or Elsevier, or "micro"– one journal publishers, are constantly trying to find their optional business model in order to stay competitive and financially sound. In the following article, I will discuss the advantages and disadvantages of different business options from the perspective of a typical micro publisher: the `Houston Journal of Mathematics` (HJM), of which I am the editor.

Before the electronic revolution there was practically only one source of income for journal publishers: library subscriptions. This brings us to

1.1. **Subscription based business models for print copies.** Calculating revenue through subscriptions is simple and straight forward: Subscription revenue is more or less proportional to the number of subscribers. Rates must go up if subscriptions are going down. For publishers there are not many options left to offset declining subscription revenue through more efficient production. Efficient use of LaTeX has reduced greatly, or even eliminated, the need for traditional (that is typing) secretarial assistance. Authors can be provided with journal-specific style files which allow them to prepare papers exactly as they will appear in print. For a properly prepared LaTeX file, changing print style is only a minor task which can be completed in the journal's office. Another cost saving measure is the elimination of free off-prints. Authors seem to no longer care about printed copies of their articles, providing the final PDF file to authors for no fee is sufficient.

While HJM has implemented those cost saving factors for more than a decade, not all publishers have been doing this. In his article *"The Future of Mathematical Publishing"*, Michael G. Cowling, an editor of an Australian Mathematical Society

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

Journal, describes in [1] a completely different experience. He claims that *"LaTeX has not reduced the cost of producing journals....and that some journals retype all articles that are sent to them..."* Of course, retyping articles that have been typed before is a terrible waste of a publisher's resources. While HJM has accepted submissions where authors had used something like MS-WORD with the equation editor, for final submission we only accept properly prepared LaTeX files.

In a pure subscription based model, rates are bound to go up for most journals. Besides inflationary pressures caused by printing and ever increasing postal rates, subscription cancellations can also be a major factor for price increases. The Internet makes subscriptions less necessary and subscriptions for multiple copies superfluous. Besides that, commercial publishers also have to meet investors' expectations for higher profits. Journal production is considered a prototype for an "inelastic business environment" where very few major providers compete in a limited market. For publishers of research journals, the market consists of internationally recognized research universities, of which there are approximately one thousand, worldwide, and, of those, perhaps only a handful of publishers qualify as very major.

It looks to me that through their marketing divisions, the major publishers are in a stronger position to obtain new subscribers to offset cancellations, especially in academically strong but still under-represented countries, like China and South America. Journals like HJM depend in this respect more or less exclusively on the initiative of their authors.

I think that unless small journal publishers try to find new revenue sources, their subscription rates may go up more than is justified by inflation rates and, as a result over time, they may lose their traditional advantage of providing inexpensive alternatives to commercial publishers.

1.2. **Providing online subscriptions, opportunities and pitfalls.** In 1996, I became the managing editor of HJM. About four years later, electronic publishing was still in its infancy and LaTeX started to replace the older and much less user friendly versions of TeX. There were serious concerns that independent journals were ill-equipped to cope with changing technologies, and some publishing authorities predicted even the demise of independent journals. And indeed, the number of journals that are still run by a few individuals from their local mathematics departments is certainly smaller than it was ten years ago, and seems to be not growing. Some formerly independent journals have been absorbed by commercial publishers or have joined academic organizations, like `Project Euclid`.

My educated guess is that creating a subscriber base for restricted online access has been the main concern for small academic publishers, and a primary reason for them to join larger platforms. While HJM has been contacted by other publishers, none really made a convincing argument for HJM to move its operation outside the department.

In this context, I firmly believe that files of published material should remain in the hands of the publisher. Of course, HJM supports the major reviewing organs by allowing them free access to our files. I also believe that preserving and archiving of files has become a major responsibility of the publisher, and/or in co-operation with affiliated libraries. Besides providing archiving, organizations like Portico and Lockss can also arrange for perpetual access. But for inexpensive journals this additional service is of little or no importance. Of course, some librarians may disagree with me on this point. As I see it, for publishers, ownership of files constitutes a major asset and files should not be given away easily, certainly not for free.

I never believed in "moving walls". As I see it, moving walls are bad business because they take away one important incentive for libraries to subscribe to online access: Most online subscription plans either include the archive for free, that is what HJM does, or make access available to subscribers of the current editions for a surcharge.

At the 2002 meeting on electronic publishing in Beijing I mentioned that HJM would be prepared to make electronic files freely available only if a sufficient number of major research libraries would agree to pay a somewhat higher subscription rate, in order to offset possible cancellations. There was an immediate response from one of the organizers, namely the head librarian of one of the "digital mathematics world libraries", who spontaneously responded: "I would never agree to something like that". For HJM this was the end of an idea. However, HJM decided to offer free online access, but only for subscribers of the print editions. A print subscription comes with free access to the whole archive, regardless of subscription history. Of course, we also offer "online-only" for the price of print-minus postage. Online access ceases with cancellation. For subscribers of HJM there is no post-cancellation access possible. For inexpensive journals which provide immediate unlimited access to the complete archive and for all of its current volumes, this might be a most reasonable policy.

Setting rates for the "Online-Only" option is somewhat tricky. Including online-access with a print subscription is what HJM does. But also some commercial publishers, like Springer, and most societal publishers, like the AMS, are

doing the same. There are some arguments that the ‘‘online-only’’ option should be lower priced than ‘‘print+online’’ or ‘‘print only’’:
While providing online access requires additional work, Internet presence is now a necessary and expected part of journal production; and an essential portion of it (Tables of Contents and Abstracts, for example) is actually offered for free. Only hot links to PDF files need an online subscription, where of course the PDF files should be the same as the ones used for printing. Of course, print obviously requires additional expenses for the publisher but also for the subscriber, for example in form of binding and re-shelving . HJM decided that the price for "online-only" should be the same as for "print+online" minus local postage. In a previous article [4] I have elaborated on this issue quite a bit more.

However, Elsevier does not offer a combined rate for "print+online". Print is offered by Elsevier at the same rate as online. While publishers are free to set their own rates, I think that this kind of pricing policy is puzzling. A library that wants to offer journal access in the periodicals room, but also provide access through its registered computers, obviously has to pay twice for journal content. For example, about $7,000 for Elsevier's Journal of Functional Analysis. This is considerably more than the advertised price of about §3,200 for an institutional online subscription. It seems to me that Elsevier is discouraging print editions.

Moving walls are the same as free online access but delayed by a number of years. For the publisher, the same drawbacks we have seen for free online access should apply. Besides that, papers in mathematics stay "current" for many years. In a typical bibliography of an HJM paper, about eighty percent of all references are older than five years. The HJM policy agrees very much with the one of the *London Mathematical Society* which provides free access of recently published issues, but after a few months all papers are put behind a "subscription wall" [2].

Commercial publishers seem to change their policies concerning moving walls all the time. Ten years ago at an ICM 2002 satellite conference in Beijing, a major publisher, Springer Verlag, announced the five year moving wall for most of its mathematics journals. It now seems that Springer has modified its policy concerning access of "non-contemporary" articles. Another major publisher, Elsevier recently announced the introduction of a four year moving wall, probably in response to the highly publicized boycott movement.

Besides establishing moving walls, allowing files for Inter Library Loans (ILLs) seems to be a questionable idea. A library that e-mails a file to meet a request from a non-subscribing institution does not experience any loss of usage, on the grounds that no loan has taken place. The HJM license has always excluded files to fulfill ILL requests Actually, so far a majority of librarians agreed

with me on this point, especially because we offer the pay-per view option for a price that beats library expenses for fulfilling ILL requests.

Most publishers restrict postings of final published files on arXiv. There is a very interesting article by Susan Hezlet [3] related to this issue. In her article *Shifting Editorial Boards* she describes the fate of the prestigious Journal *Topology* after the whole editorial board resigned and the journal found a new home in the London Mathematical Society. This journal is now published by Oxford University Press under the new title *Journal of Topology* (JOT). While JOT is technically a new journal, everybody knows that it is the old *Topology* with a new cover. Actually, Elsevier's *Topology* has been discontinued.

It certainly came as a surprise to many people that according to Susan Hezlet, "...the *Journal of Topology* is not covering all its costs." As one explanation for the disappointing number of library subscriptions for JOT, Hetzlet cites the very high number, namely about 90%, of JOT articles available as preprints on arXiv. Hetzlet is probably correct when she says that *"discouraging authors from placing their pre-acceptance versions on the arXiv would clearly be unacceptable to the topology community"*.

HJM has currently no official policy about posting files of published articles on arXiv. But HJM does not post articles on arXiv for archival purposes, and probably never will. arXiv is now fully owned and operated by Cornell University. This might lead to conflicts of interest because of Cornell's prominent role as a publisher and World Digital Mathematics Library. Only the future can tell us how arXiv will develop over the years.

2. On Open Access

Open Access (OA) is somewhat of a misnomer. Access is free, but only for the reader. For authors there are ''processing charges'' . The non-profit publisher Oxford University Press charges for its Oxford Open option $3,000 per article, while charges for Hindawi journals depend on the journal. For Hindawi's new journals, like *Algebra*, there are currently no fees, while for Hindawi's more established *Abstract and Applied Analysis* charges are $1,200 per article. For publishers, OA is synonymous with the *"Author pays Business Model"*. Peter Suber [6] calls this a *"common misunderstanding"* . Suber seems to be conflating the vast number of OA journals that impose processing charges and the very few (at least for mathematics) non-fee based free access journals. The latter might be called FA journals, versus OA journals. Most FA journals depend on volunteers who work for free or receive small academic stipends. Sustainability of FA journals

remains on shaky grounds unless they can find a sponsor, say the host institution. Exclusive dependence on volunteers does not constitute a viable business model, in my opinion.

Proponents of OA don't like the idea of authors using money out of their own pockets to pay for processing charges. Their solution is that for non grant supported papers, library funds should be used for this purpose. According to Suber [6], *"A growing number of universities maintain funds to pay publication fees on behalf of faculty who choose to publish in fee based OA journals"*. Indeed, some university libraries advertise on their web sites support of the OA movement and some libraries even organize OA meetings on a regular basis. My experience with librarians from such places has been mixed. First I have to mention Göttingen which has arranged for its faculty the Open Choice option for publications in Springer journals. This agreement between a major German University and Springer Verlag has been highly publicized. Springer charges for its Open Choice $3,000. However, Göttingen provides this kind of support of OA only for Springer journals and it is probably part of a larger deal involving subscription rates and the inclusion of older Springer journals in the GDZ. The GDZ is a rather comprehensive open archive of books and journals. From my own dealings with Göttingen I got the impression that Göttingen does not support per-se page charges of its faculty. But Göttingen supports academic journals, like HJM through subscriptions. Unfortunately this cannot be said about all places which support some aspects of OA.

While the majority of OA supportive libraries which I contacted concerning page charges just ignored my request, I had somewhat more luck with a non-subscribing university in Germany. I sent them an invoice for the amount of $600. We had accepted a paper, 60 pages in length, by a junior faculty member at that university. HJM's suggested fee is $30 per page. But because of the paper's length I applied a $1,200 discount. The library agreed to pay $300. They argued that the remaing half of the fee $300 should be considered as covered by the subscription of the co-author's library in Italy. The local author was also informed to be prepared to pay $100 as his share of the deal. Of course, I accepted the offer in order to stay on good terms with this university.

This whole process took several months involving numerous email exchanges. The emails were not exactly on friendly terms, in part because the head of the author's group found it inappropriate to ask a subordinate about a possible subscription. The head of the author's group also expressed his belief that page charges were not ethical. He obviously was not aware of his own library's policy providing funds for page charges.

John Ewing's newly established blog [7] on Mathematics Journals deals with *Who Pays in Author Pay?*. In this article the former executive director of the AMS declares somewhat categorically: *"The author-pay model is exactly that–the author pays. We should not try to obscure reality with fanciful promises. Right now, the fee may come from the university or some funding agency, but inevitably the author "authorizes" the charge.... Over the long-run all journals will require payment somehow–they must. If one insists on gold open access, this is the price one pays. It may be worth the cost, but pretending there is no cost is foolish."* When Ewing declares that *all journals will require payment..* he probably means all OA journals.

Most societal and independent journals allow for green access. That is, authors can post their refereed and edited papers on departmental websites and public repositories. Some journals go even further, the final published version can be posted on arXiv. While HJM does not encourage authors doing this, we also don't disallow this practice. At least for the time being. For HJM it all depends on arXiv's future behavior. For example whether arXiv starts adding meta-data to pre-prints which were not provided by the author. Unfortunately, if history is any guide, libraries are going to abuse such laudable activities as subscription *ersatz*. But another factor is that, legislations requiring free access to federally funded papers may force us to change our liberal arXiv policy. Clearly, requiring that certain papers *must* be posted on arXiv entails for the publishers that certain papers *cannot*.

Proponents of OA want to see subscriptions disappear. Here again I can cite Suber [6] with his subscription free vision of the future: *... as OA spreads, libraries will realize large savings from the conversion, cancellation, or demise of non–OA journals"*. Suber does not say what libraries will do with the savings. He might not be aware that they might spend it on more big deals with the mega publishers. According to information I got from an author, his major university in Belgium cancelled all journals that were not part of package deals. Only after heavy protest from the math faculty, could AMS journals be saved. For journals like HJM there was no money left.

HJM is subscription based but since its inception in the year 1975, HJM has asked for voluntary page charges. But until recently, only few authors considered our request seriously. In recent years the situation has changed quite a bit. Despite that HJM publishes primarily in traditional areas of pure mathematics, a substantial portion of published papers are grant supported, and major grant providing organizations, like NSFC, DFG etc. explicitly allow for page charges. During the year 2012, for about 20% of accepted papers we received payments that were

close to the suggested amount of $30 per page. Because of this additional revenue we were able to maintain our low subscription rate of $0.10 per printed page.

While primarily subscription based, HJM is certainly not alone in asking for page charges. For example, Indiana University Mathematics Journal has been doing this for years. Societal publishers, like SIAM and ACM are on and off on this issue. But professional societies receive substantial income through membership and registration fees, and also their book programs probably subsidize journal production. Amongst non-mathematical and computer science journals, policies expressed by the journal Evolutionary Ecology Research [8] agree very much with the philosophies of HJM. In particular HJM fully agrees with the statement on publication fees made by the managing editor Michael L. Rosenzweig of this journal: *Rather than abandon the subscription model, EER is adopting the mixed model of funding that is described above.* This is exactly what HJM is doing.

My take: As long as a journal has a measurable subscription base, going OA doesn't make much sense because then the authors must make up for lost subscription revenue. Well-established research journals may have about 300 subscribers. While this sounds like a small number it is not bad at all considering that there are not more than about 1000 major research libraries as potential customers around. Asking authors for money is a sensitive issue and requires a case by case decision. Authors should be affiliated with major national universities and be supported by national grants. In case that their University library doesn't subscribe, asking such authors to do something about that is a very reasonable request, even in cases where the author feels that asking for page charges is not appropriate. However, for some countries, or type of authors, asking for any sort of support is a hopeless case.

In asking for contributions, journals like HJM acting a bit like public television: Subscriptions play the role of public funding and asking authors for money are our perpetual fundraising events. Only inexpensive non-profit journals can do this in good conscience. However, this kind of fund raising is very time consuming. It involves heavy correspondence with authors, invoicing and book–keeping. The money doesn't come in for free.

Authors who contribute to defray our publication costs must be given incentives and benefits to do so. As I mentioned before, like most academic journals, our journal does not impose much of a copyright restriction. Also, so far none of the authors who contributed money wanted his paper published under what is referred to as the Open Choice option. That is color coded as freely available amongst papers that require a subscription for access. However, a somewhat expedited publication is highly appreciated. This is what we offer. We do this

also in order to stay competitive with OA-journals where publication is more or less immediate after acceptance. For grant supported authors this may be necessary in order to meet deadlines of publication. Especially in cases where the refereeing process alone may have taken well over a year. Our backlog as published in the AMS Notices is close to two years. Of course, we sometimes apply other editorial criteria for faster publication. Also younger authors are routinely given preference.

Asking for page charges resembles the "author pays" aspect of OA. The main difference is that in our case, page-charges are strictly voluntary. For OA journals they are mandatory, unless a waiver has been granted, usually at the time of submission and before the refereeing process has even begun. For OA journals publication is immediate after acceptance whereas we can offer only expedited publication, often negotiated with authors to meet their specific needs. Referees of OA journals know that rejection of a paper means loss of income for the journal. This is not the case for subscription-based journals.

3. Conclusion

Only the future can tell whether pursuing this type of "hybrid" version of OA will be successful: Subscription-based with very low rates but asking for modest voluntary page charges, mainly from grant-supported authors. But I am quite optimistic. Besides providing some benefits for contributors, it is also a matter of convincing authors about the virtues of supporting a journal that is run and owned by the mathematics community.

References

[1] Michael G. Cowling, *The Future of Mathematical Publishing*, Notices of the AMS, Vol.59 (4) 2012, pp. 558–559.
http://www.ams.org/notices/201204/

[2] *London Mathematical Society - Open Access Policy.*
http://www.oxfordjournals.org/our_journals/lms_open_access.html

[3] Susan Hezlet, *Shifting Editorial Boards*, Notices of the AMS, Vol. 59 (6) 2012, pp. 841–843
http://www.ams.org/notices/201206/index.html

[4] Klaus Kaiser, *Implementing Electronic Access for an Independent Journal*, pp. 31–46, in Communicating Mathematics in the Digital Era, edited by *J.M. Borwein, E.M. Rocha, and J.F. Rodrigues*. A.K. Peters, Ltd., Wellesleyley, MA. Papers in this book were presented at the ICM 2006 Satellite meeting at the University of Aveiro, Portugal.

[5] Donald E. McClure, *AMS Publication News: Open Access*, Notices of the AMS, Vol. 60 (3) 2013, pp. 347–348.
http://www.ams.org/notices/201303/

[6] Peter Suber, *Open Access Overview*,
http://legacy.earlham.edu/ peters/fos/overview.htm
First put online June 21, 2004. Revised December 16, 2012.‘
[7] John Ewing, *Who Pays in Author Pay?*
http://mathematicsjournals.blogspot.com/
Posted on 1/28/2013
[8] *Evolutionary Ecology Research*, Michael L. Rosenzweig, editor.
http://www.evolutionary-ecology.com/

Received March 18, 2013

Houston Journal of Mathematics, Department of Mathematics, University of Houston, Houston, Texas 77204-3008, U.S.A
E-mail address: kkaiser@uh.edu

EMIS COLLECTIONS AND CONFERENCE PROCEEDINGS
FIZ KARLSRUHE

THE BRAVE NEW WORLD OF OPEN ACCESS & CREATIVE COMMONS: A HUMANISTIC EXPERIMENT IN MATHEMATICAL PUBLISHING

GIZEM KARAALI

ABSTRACT. In January 2011 the *Journal of Humanistic Mathematics* (JHM) published its first issue. JHM (`http://scholarship.claremont.edu/jhm`) is an online-only, peer-reviewed, open-access journal which has passed the all-important ten-thousand-download barrier in its first anniversary. In order to remain faithful to the fundamental principles of open access, JHM uses Creative Commons licensing, where authors retain copyright of their work, but others are free to reuse them (with proper attribution). In this note I share and reflect upon our experience with open access and Creative Commons.

PRELUDE: A QUICK PEEK INTO CREATIVE COMMONS LICENSING ...

In 2009, on New Year's Eve, Pablo Flores, an Argentinian photographer, took a snapshot of a young person twirling about a fiery stick, using his Sony DSC-H7. Then he posted this image (Figure 1) to Flickr, "almost certainly the best online photo management and sharing application in the world.(`http://www.flickr.com/about/`, accessed February 7, 2013.) Flores labeled his photo "*Infinito*", and added "*Que el 2010 no te ponga límites*."

Almost one full year after his posting, Allegra Swift, a librarian at the Claremont Colleges Library, dropped Flores a line:

> "Pablo, would it be possible to get a waiver for the no derivatives portion of your license. we'd like to use just the infinity sparkler portion of the image on a non-profit open access journal for mathematics. you'd be given an image credit on the website. please let me know as soon as possible. best regards, Allegra"

Key words and phrases. Electronic publishing, scholarly publishing, open access, creative commons, humanistic mathematics.

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

FIGURE 1. "*Infinito*" by Pablo Flores, available at `http://www.flickr.com/photos/pablodavidflores/4233136695/`, accessed on February 8, 2013.

Flickr offers all its users the chance to license their images through Creative Commons, and Flores responded immediately to Swift's request and modified his selection to *Attribution-NonCommercial-ShareAlike* (CC BY-NC-SA). This then allowed Swift and her colleagues to extract the fiery infinity loop and create what became the logo for the new *Journal of Humanistic Mathematics*, see Figure 2.

FIGURE 2. The logo of the *Journal of Humanistic Mathematics* is a fiery unending loop of infinity.

1. INTRODUCTION: A BRAVE NEW WORLD

Mathematical publishing is going through a revolution. This is not an overstatement; just look at recent collections on mathematical publishing (cf. [2, 4]), signatures collected by respected mathematicians to boycott well-established journals, back-and-forth essays and letters to the editor in the main outlet of the American Mathematical Society, the stimulating talks and the lively audience participation in the *AMS Special Session on Topics and Issues in Electronic Publishing* at the JMM 2013. Consider also that many academic libraries continue to undergo budget cuts and belt-tightening maneuvers

which lead to shorter lists of academic journals being purchased. At the same time, big providers like *Springer* and *Elsevier* resort to journal bundling policies which allow many libraries to access a wider spectrum of journals all the while increasing the burden on smaller libraries who only need a few select journals.

In fact the whole world of academic publishing is in turmoil. The shrinking budgets of libraries are on one side of the story (see for instance [11]); the other side sees the large number of new opportunities warranted due to diverse burgeoning new technologies. There are continuing discussions about the death of the academic journal, the death of the scholarly book, the death of peer review as we know it, and possibly less violently, a transformation of the whole scholarly ecosphere.

Simultaneously we see the growth of a new industry, that of online journals. A random mathematician receives an email invitation a day to submit her research into the new Journal of Some-kind-of-mathematics that is peer reviewed and has these respectable people on its editorial board. People typically discard these messages just as they discard various ads about medication claiming to heal certain types of reproductive dysfunction. However slowly some are coming to a realization that unlike the medication that we will most likely not need or want to buy from an online vendor, the online journal is becoming more and more settled into its niche; some are even becoming household names.[12]

I, the author of this note, am the founder and editor of one such journal. Thus I am on one side of this discussion and I do not deny that my position might be biased. However I'd like to argue in this note that online journals are here to stay, and that if we look carefully, there is much that the mathematical community might gain from the new status quo.

More specifically in this note I make a case for open access and creative commons licensing. Since these are not yet household terms for all mathematicians I describe what I mean by my terms and provide some background notes and suggestions for further reading. In order to provide the context of the argument, I also mention my own experiences with *Journal of Humanistic Mathematics*. To this effect the next section (§2) introduces the story of *Journal of Humanistic Mathematics*, the case study of this note, if you will. Then I introduce the main tenets of open access (§3) and focus on the open access model we use at *JHM* (§4). Creative Commons licensing is the specific theme of the following section (§5). §6 wraps up the discussion with a look toward the future. For those who

[1]Many have written about their experiences with online and open access journals. See for instance [1, 3, 5, 19]. There are even books on how to develop an open access journal, see for instance [16].

[2]The Directory of Open Access Journals (DOAJ at `http://www.doaj.org/`) on February 8, 2013 listed 231 journals belonging to the subject mathematics.

might be curious to know more about *Journal of Humanistic Mathematics* itself, I include an appendix (§A) which provides a historical context for the notion of humanistic mathematics.

2. A Humanistic Experiment in Mathematical Publishing

In January 2011, *Journal of Humanistic Mathematics* published its inaugural issue on its website (`http://scholarship.claremont.edu/jhm/`), see Figure 3 for a screen shot. Edited by Mark Huber of Claremont McKenna College and myself, *JHM* is an online-only, open-access, peer-reviewed journal. We have just recently (in January 2013) launched our fifth issue (Volume 3 Issue 1). As a new journal we have several matters that need our continuous attention, but in a short period of time we have also accomplished a lot. Here is a quick claim for bragging rights:

We published our first issue in January 2011. In our first anniversary we had already passed the ten thousand (full-text) downloads barrier. On October 15, 2012, we reached 20K downloads. On January 5, 2013: 24,435! On February 8, 2013: 27,739. For the mathematically inclined, these numbers clearly suggest an exponential pattern, and we are optimistic!

It is clear that our publication is gaining some readership. On the other hand, the question of reputation is in everyone's mind when it comes to online academic journals. The good news on this front is that in terms of reputation we are not lagging behind. Our content is now indexed by *Mathematical Reviews / MathSciNet*. Furthermore, we recently signed a licensing agreement with *EBSCO*, an academic database service company, which will allow our content to be included directly within library database searches around the world.

Journal of Humanistic Mathematics is currently supported by two institutional entities. Claremont Center for the Mathematical Sciences (CCMS) is our institutional home, our snail mail address, our main contact point.[3] The Claremont Colleges Library is the library system of the Claremont Colleges and it provides us with the technical expertise and support that we need as we navigate the waters of the world of scholarly publishing and archiving.[4]

We have an impressive set of people on our editorial advisory board:

[3]CCMS (`http://ccms.claremont.edu`) is an institution founded to promote collaborative research and creative teaching among the institutions of the Claremont Colleges Consortium.

[4]`http://libraries.claremont.edu/` accessed February 8, 2013.

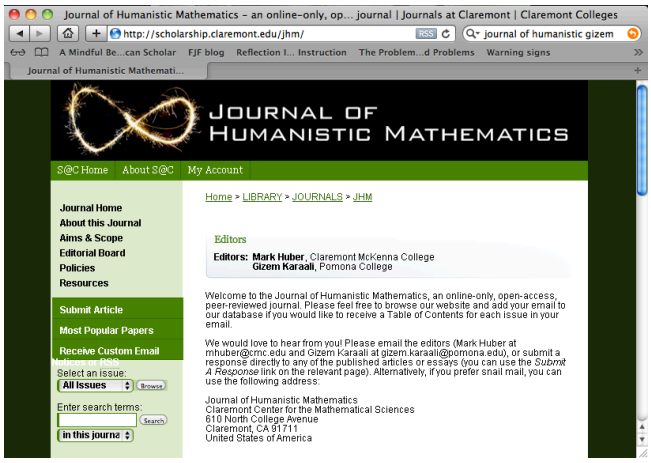


FIGURE 3. A screenshot of the front page of *Journal of Humanistic Mathematics*.

Our editorial advisory board consists of mathematicians, mathematics education researchers, philosophers of mathematics, sociologists of mathematics, historians of mathematics. Many of them are themselves writers of expository mathematics, mathematical poets, dedicated workers for equal opportunity, access and representation of diverse populations in mathematics. These are all people who believe in the value of bridging the communication gap between the many scholarly disciplines allied with mathematics.

For a list with current affiliations please see http://scholarship.claremont.edu/jhm/editorialboard.html (accessed February 8, 2013).

For readers who might be wondering about what humanistic mathematics is, I will provide a brief overview of what we do in the rest of this section. For a more in-depth and historically nuanced exposition, see the appendix §A.

Our first and main goal is to provide an open forum for both academic and informal discussions on the various threads of mathematical inquiry. To this end we publish works focusing on "the aesthetic, cultural, historical, literary, pedagogical, philosophical, psychological, and sociological aspects of doing, learning, and teaching mathematics."[5] We aim to reach a *general mathematical audience*. This for us means "people who are seriously interested in mathematics, but may come from a variety of backgrounds both within and beyond academia."[6] In this (admittedly limited) perspective *humanistic mathematics* means *scholarship or any form of inquiry that studies the human face of mathematics*.[7]

But you might still wonder, given all of the above, what exactly does *JHM* publish? Perusing the table of contents for the issues already published, one will note three distinct types of contributions. First off we publish peer reviewed articles presenting research in education, philosophy, sociology, or history of mathematics, with the intent to reach a broader audience than that of a typical disciplinary journal. Secondarily, we provide a home for reflective essays, opinion pieces, and more informal works, with the intent of sparking discussion about mathematics. These could overlap with the disciplines above, or focus on contemporary issues facing the mathematical professions. Finally we welcome explorations of the interface between mathematics and the wider humanities. Poems, short stories, or other expressive outlets, along with works about the relationship between mathematics and the arts fall in this category.

Why do we do what we do? We think that the world of mathematics is fragmented. That there are many people of mathematics, who care deeply about various dimensions of the field, but that they do not hear or see one another. That there is much to be gained from building a platform to communicate and exchange ideas. Our lofty goal with the

[5] http://scholarship.claremont.edu/jhm/aimsandscope.html accessed February 8, 2013.

[6] http://scholarship.claremont.edu/jhm/aimsandscope.html accessed February 8, 2013.

[7] From our website (http://scholarship.claremont.edu/jhm/about.html, accessed August 29, 2013): "The *Journal of Humanistic Mathematics* was inspired by the work of Alvin White, a former professor of mathematics at Harvey Mudd College. Dr. White was the founding editor of the *Humanistic Mathematics Network Journal* (HMNJ), a work of love that he almost single-handedly edited and produced for 15 years. Dr. White believed wholeheartedly in the importance of recognizing mathematics as a humanistic discipline and played a significant role in bringing this idea to the forefront of many minds. Though this is an independent enterprise, the *Journal of Humanistic Mathematics* builds on the spirit and tradition of the HMNJ."

Journal is to create a sense of connection and community for the many diverse people of mathematics.

Given all of the above, open access is the right framework for us. We want to reach people and we want people to reach us. *Open access* for us means that our content is freely available to anyone with a working internet connection. Furthermore, we want to be freely accessible to creators of our content as well as the users. Thus we do not require / request / imply author fees. Of course we can only do this with the institutional support we have. In the next two sections I provide some background on the open access movement and our own implementation of its main tenets, respectively.

3. Open Access As a Guiding Principle

> **"**By open access, we mean ... immediate, free availability on the public internet, permitting any users to read, download, copy, distribute, print, search or link to the full text of these articles, crawl them for indexing, pass them as data to software or use them for any other lawful purpose ...**"**
>
> –The Budapest Open Access Initiative

> **"**The Internet provides a new opportunity to bring information to a wider audience at virtually no marginal cost, and allow them to use it in new, innovative ways. This has resulted in a call for new framework to allow research results to be more easily accessed and used—a call for Open Access.**"**
>
> –SPARC (Scholarly Publishing and Academic Resources Coalition)

The idea of open access is not new. It can be traced back into the middle of the twentieth century, along with other similarly motivated movements like the open source movement. John Willinsky, the author of *The Access Principle* [18], in fact claims the underlying principle is even older. Willinsky describes his *access principle* as follows [18, page *xii*]:

> **"**A commitment to the value and quality of research carries with it a responsibility to extend the circulation of such work as far as possible and ideally to all who are interested in it and all who might profit by it.**"**

Early on in his book (see page 5, for instance) he argues that this principle about who has (and should have) access to knowledge is an age old concern that evokes the images of the libraries of old (Alexandria in the third century B.C., Cairo in the sixteenth century, small towns of north America in the nineteenth century).

For a historical approach to the ideas of open access Willinsky's first chapter (together with his thirteenth) is a good reference. Another alternative is found in Peter Suber's website at `http://legacy.earlham.edu/~peters/fos/timeline.htm` (accessed February 22, 2013). Suber, a major player in the philosophical and pragmatic discussions around open access, starts his historical account in 1966 with the launch of ERIC, the United States Department of Education's *Educational Resources Information Center*. For an audience of mathematicians, the particular dates that might speak to us are August 30, 1969, where Advanced Research Projects Agency Network (ARPANET) was launched by the U.S. Department of Defense, and August 16, 1991, when arXiv was launched by Paul Ginsparg. Some might question how these connect with the upheaval that we are seeing in the world of academic publishing, but since both Willinsky and Suber do a masterful job of articulating this coherently, I will not include too many specific details of the historical narrative here. I will simply add that perhaps a later date included in Suber's timeline, 2001, when a handful of editors of the highly respected journal *Topology and Its Applications* resigned in order to launch *Algebraic and Geometric Topology*, will bring us closer to the discussion at hand.

In January 2006, The European Research Consortium for Informatics and Mathematics (ERCIM) published its *Statement on Open Access*.[8] This statement sets the stage first by "[r]ecognising the inability of research libraries to meet the costs of sustaining their collections, and participating actively in the development of appropriate technology," and then lists a set of principles that ERCIM supports, namely:

- **"**research that is funded by the public via government agencies or charities should be available freely, electronically at the point of use**"**
- **"**other research should be made equally available subject only to confidentiality required by commercial, military, security or personal medical constraints**"**
- **"**quality assurance of research publications must be continued through rigorous peer review associated with research publications, research datasets and software should be equally openly available**"**
- **"**the provision of open access should be made as cost-effective as possible**"**
- **"**the provision of open access carries also the responsibility for curation of the digital material including cataloguing, archiving, reproducing, safekeeping and media migration.**"**

[8]The statement is available online at `http://www.ercim.eu/publication/Ercim_News/enw64/ercim-oa.html`, accessed February 22, 2013.

Since then a variety of national entities and professional societies have released statements on open access. Most recently the American Mathematical Society (AMS) joined in the conversation. As can be gleaned from a recent news release published in the March 2013 issue of the Notices of the Society [12], the Society has a rather nuanced position. In particular McClure, the executive director of the AMS, argues that the main question about who pays for the publication and dissemination of research should be considered within a framework that takes into this policy statement, agreed upon in 2012:

> **"**The American Mathematical Society strongly endorses and adheres to the principle that a paper in the mathematical sciences should have an opportunity to be evaluated and possibly published without regard to the financial circumstances of its authors.**"**

This of course brings up a common concern about a particular open access model, where the authors are required to pay the publisher to have their work accessible; that authors who do good work but are not able to pay will not have their work appear and will not be able to participate in the scholarly conversation. In order to convince the reader that this is not the sole model of open access to consider, we once again refer to [18], more specifically the first appendix therein (entitled *Ten Flavors of Open Access*), and include here a replica of the relevant table from page 212, see below for Table 1.[9] In this classification, the concern raised about author funding applies only to the third case, or one out of ten.

In other words *open access* is not synonymous with *author subsidized publishing*. On the contrary there are a variety of sustainable and realizable models out there; Willinsky provides examples for all ten of his flavors. In this note, I will talk about the fourth model, what Willinsky labels as *Subsidized Open Access*, as that is the one *Journal of Humanistic Mathematics* exercises.

4. Our open access model

Willinsky in his appendix describes a selection of models for open access. He calls his fourth model *Subsidized Open Access* and points out that this model makes possible "immediate and unqualified access". This is only doable if the front-end costs are subsidized, or covered by a dedicated entity, such as a scholarly society, university department or library, a government agency, or a foundation [18, pages 214–215]. Willinsky adds that

> **"**Journals that offer this type of open access charge neither author nor reader and typically publish only online. Of all the forms of open access,

[9] A similar list of open access journal business models with several examples can be found at `http://oad.simmons.edu/oadwiki/OA_journal_business_models`, accessed February 23, 2013.

TABLE 1. Ten flavors of open access, Table A.1 [18, pages 212–213].

Type of open access	Economic models
Home page	University department maintains home pages for individual faculty members on which they place their papers and make them freely available.
E-print archive	An institution or academic subject area underwrites the hosting and maintenance of repository software, enabling members to self-archive published and unpublished materials.
Author fee	Author fees support immediate and complete access to open access journals (or, in some cases, to the individual articles for which fees were paid), with institutional and national memberships available to cover author fees.
Subsidized	Subsidy from scholarly society, institution and/or government/foundation enables immediate and complete access to open access journal.
Dual mode	Subscriptions are collected for print edition and used to sustain both print edition and online open access edition.
Delayed	Subscription fees are collected for print edition and immediate access to online edition, with open access provided to content after a period of time (e.g., six to twelve months).
Partial	Open access is provided to a small selection of articles in each issue—serving as a marketing tool—whereas access to the rest of the issue requires subscription.
Per capita	Open access is offered to scholars and students in developing countries as a charitable contribution, with expense limited to registering institutions in an access management system.
Indexing	Open access to bibliographic information and abstracts is provided as a government service or, for publishers, a marketing tool, often with links to pay per view for the full text of articles.
Cooperative	Member institutions (e.g., libraries, scholarly associations) contribute to support of open access journals and development of publishing resources.

this one perhaps relies the most heavily on the volunteer labor of editors playing multiple roles, with journals such as *Education Policy Analysis Archives*, for example, running on a zero budget, apart from the editors time and Internet bandwidth, both supported in this case by Arizona State University.”

Another example Willinsky provides for this model is *First Monday*, an online-only platform for scholarly work on everything related to the internet.[10] Perhaps it is easier to accept that this model can be reasonable or feasible for research on the internet, but there are other examples. One of them is *Journal of Machine Learning Research*; see [15] for a provocatively written account of how this journal functions. Yet another is *Numeracy*, the official journal of the National Numeracy Network, sponsored completely by the University of South Florida Libraries. Todd Chavez, a librarian with the USF, in his article [6] explains the open access model *Numeracy* employs with specific numbers. His Table 2 is reproduced below:

TABLE 2. Summary of Actual Three-Year Costs for One Title in the University of South Florida Libraries Open-Access Journal Collection, Table 2 in [6, page 6].

Cost Areas	Year 1 ($)	Year 2 ($)	Year 3 ($)
One-Time Journal Start-Up	2,500	NA	NA
Journal Management System	5,000	5,000	5,000
Archiving	400	400	400
TOTAL COSTS	7,900	5,400	5,400

These numbers differ drastically from those for a journal published in the traditional model; see Table 3 in Chavez's article, which reports in itemized format expenses involved in publishing a subscription based society journal (as around $730,000), taken from [10]. In the case of *Numeracy*, many cost items (including platform, PDF creation, author alterations, XML conversion, and overhead costs) are subsumed by the item entitled "Journal Management System". In regard to this item, the University of South Florida Libraries works with BePress, an academic publishing company which specializes in electronic publishing. BePress was founded in 1999 by UC Berkeley professors Robert Cooter, Aaron Edlin, and Ben Hermalin as *Berkeley Electronic Press* and today provides a flexible and versatile journal management platform for a variety of customers.[11] *Journal of Humanistic Mathematics*, as part of the *Scholarship@Claremont* platform, also utilizes the services of BePress.[12]

Other options do exist. Journals can use in-house talent or a combination of commercial or open source options. One alternative worthy of mentioning is the

[10] *First Monday* is at `http://www.firstmonday.org/`, accessed February 22, 2013.

[11] `http://www.bepress.com/aboutbepress.html`, accessed February 22, 2013.

[12] One of the reasons for the demise of HMNJ, whose spirit and tradition the *Journal of Humanistic Mathematics* builds on, was arguably its funding structure. Since its inception through its first fifteen years, HMNJ was at least partially funded by Exxon-Mobil. When setting out with JHM, we made the decision to remain financially independent of all commercial enterprises.

Open Journal System, an open source journal management system package, available freely at http://pkp.sfu.ca/?q=ojs (accessed February 23, 2013). Edgar and Willinsky provide a survey of journals using this platform in [7]. For a comprehensive list of similar journal management software options, readers can refer to http://oad.simmons.edu/oadwiki/Free_and_open-source_journal_management_software (accessed February 23, 2013).

No matter what journal management system one uses, the model we are describing here redefines the role of the research library as well as its relationship with the scholars who create the scholarly products (cf. [9]). In the case of *Journal of Humanistic Mathematics* editors work closely with the staff of the Claremont Colleges Library Center for Digital Initiatives (CDI) hosting *Scholarship@Claremont*, "an open access scholarship repository with a set of services to capture, store, index, and provide access to scholarship produced by the Claremont Colleges academic community."[13] The main workflow for a brand new journal to become a part of the *Scholarship@Claremont* repository is as follows (provided by Allegra Swift, the Digital Initiatives Librarian at the Claremont Colleges Library):

(1) A potential journal completes a Journal Proposal Form[14]
(2) The proposal is approved by the CDI Advisory Board.
(3) CDI staff works with the editors to develop the look of the journal, page content, Open Access policies, copyright permission forms for authors.
(4) Completed design forms are sent to BePress by the CDI.
(5) BePress does the training once the a mock-up site has been created.
(6) Editors approve the mock-up site.
(7) BePress completes the final site.
(8) CDI staff applies for the ISSN.*
(9) Editors and staff add content.
(10) Library cataloging staff create OCLC Worldcat and library catalog records.*
(11) Library electronic resources staff add DOIs to articles as they are added.*
(12) CDI staff registers journal in OA registries such as DOAJ.*
(13) BePress sends editors and authors download count statistics.
(14) CDI staff works with editors on third party indexer contracts.

*BePress can take on steps 8, 10, 11, 12 for an additional yearly fee.

[13] http://scholarship.claremont.edu/about.html, accessed February 23, 2013.

[14] The form used for *Scholarship@Claremont* is available at http://libraries.claremont.edu/cdi/ under "CDI Forms & Documents." Last accessed February 23, 2013.

As can be seen in our example, in this model where open access is subsidized by a university library, many of the tasks traditionally undertaken by a commercial publishing company are subsumed by the library staff who already have all the skills necessary to handle them masterfully. In other words, the cost of making sure the journal is sustainable as an entity (all tasks besides the editorial and peer review dimensions of the enterprise) is covered by the university library.

This model depends on the academic library taking on the bulk of the publishing costs. In today's environment where academic libraries are famously going through tough financial times, due to the recent economic downturn and, some would argue, the rising costs of journal subscriptions, adding a new item to the budget may seem counterproductive. However this can indeed be a viable solution to the current crisis. In our model, the library uses its limited resources in a smart way. Libraries are already invested in employing competent and effective staff and using and providing modern software and information technologies for their patrons. Therefore the expense of hosting and archiving scholarship and making it accessible can be naturally incorporated into / subsumed by personnel, materials, and / or technology budget (in the case of the Claremont Colleges Library, the associated expenses fall under materials.)

The model can be represented by two simple charts. First in Figure 4 we display a simplified version of our publishing cycle.

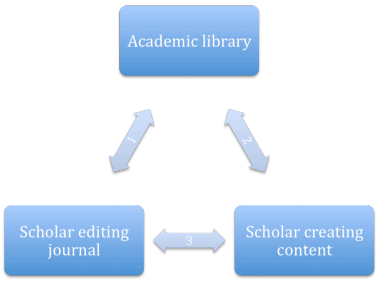


FIGURE 4. A simplified publishing cycle in our open access model.

Contributing scholars contact the editorial staff and work directly with them in developing the article. They use the interface supported by the library. The editorial staff manages the peer review process and handles other editorial tasks. The library staff provides expertise on copyright issues that might arise, supports the infrastructure that keeps the journal running, and handles archiving and indexing tasks. There is no need for a fourth party in the production and dissemination of the scholarship besides the author, the editor, and the library (which takes on both the mantels of the publisher and

the archiver). The output scholarship is accessible to anyone via online repositories or journal web sites hosted by the library. Thus users do not need intermediaries to access the scholarship they want.

Next in Figure 5 we present a simple flowchart of support.

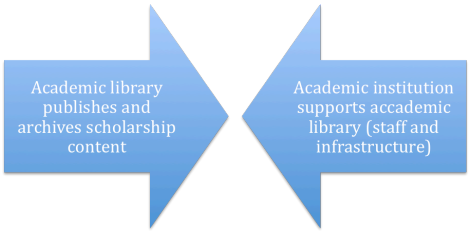


FIGURE 5. A simplified flowchart showing how the model is supported.

The academic library provides the necessary platform and supports the editorial staff of the scholarly journal as needed. The academic institution supports the library by funding library staff and the infrastructure. Scholars both on the editorial team and on the author side benefit from this interaction and scholarship is accessible to all other scholars who want to use it. Once again, there is no external party involved in this end of the scholarship cycle. The only parties involved are the creators, users, and the archivers of the scholarship.

5. CREATIVE COMMONS LICENSING

In the digital world of today copyright is simultaneously a point of concern and a possible venue for innovation. Wilinsky in his third chapter of [18] analyzes the copyright model that is most in line with the principle of open access. *Creative Commons* is "a nonprofit organization that enables the sharing and use of creativity and knowledge through free legal tools."[15] According to their mission statement:

> **"**Creative Commons develops, supports, and stewards legal and technical infrastructure that maximizes digital creativity, sharing, and innovation.**"**

In short, Creative Commons provides free copyright licenses for a variety of purposes and types of work. Wikipedia is one of the most commonly known users of CC licensing; most images on Wikipedia are licensed through Creative Commons.

There are six varieties of Creative Commons licenses. These are:

[15] `http://creativecommons.org/about`, accessed February 23, 2013.

(1) *Attribution* (CC BY)
(2) *Attribution-ShareAlike* (CC BY-SA)
(3) *Attribution-NoDerivs* (CC BY-ND)
(4) *Attribution-NonCommercial* (CC BY-NC)
(5) *Attribution-NonCommercial-ShareAlike* (CC BY-NC-SA)
(6) *Attribution-NonCommercial-NoDerivs* (CC BY-NC-ND)

For a description of these options readers are referred to the relevant Creative Commons website `http://creativecommons.org/licenses/` (accessed February 23, 2013).

Journal of Humanistic Mathematics applies Creative Commons Licensing (CCL) to all works we publish. Authors choose one from the six CCL licenses. Our submission agreement form (see Figure 6) spells out explicitly that:

> **"**Under CCL, authors retain ownership of the copyright for their work and specify if readers may download, reuse, reprint, modify, distribute, and/or copy articles in *JHM*, so long as the original authors and source are cited. ... No permission is required from the authors or the publishers.**"**

Most JHM authors pick the standard CC-BY option, but some wish to disallow commercial use, and poets often prefer their work not be modified. These contributors might prefer to choose CC-BY-NC or CC-BY-SA options. In our agreement form we do inform authors that "[w]orks by authors choosing CC-BY-NC or CC-BY-SA will not be included in commercial databases or aggregators."

6. A Look toward the future

The online journal is here to stay. The open access movement is gaining speed. Creative Commons is not only for Wikipedia anymore. I believe that these are not merely changes that we the mathematical community should adapt to but in fact opportunities we should take advantage of. With the academic library on our side, we can develop and utilize publication models and advance mathematical scholarship in ways that allow more researchers to access products of our research.

The major commercial publishers have been taking note. And they are responding in a variety of ways. It is now the turn of the mathematical community to take note. Next will come the time to take action. Some among us will decide to boycott certain journals, others will petition grant funding agencies to support open access publishing, and yet others will choose to follow alternative paths. However it behooves all of us to think carefully about the status quo and what is at stake and what might lie ahead for us, and make our decisions accordingly.

Copyright Permission and Submission Agreement for
The Journal of Humanistic Mathematics (JHM)

I, _____ *(author or copyright owner's name)* warrant that I have the authority to act on any copyright related matters for the work titled _____ _____; dated _____ and as such have the right to grant permission to digitize, republish and use the said work in all media now known or hereafter devised.

I grant to the Claremont Colleges Digital Library (CCDL), Scholarship@Claremont, and the Claremont Colleges Library the nonexclusive worldwide rights to digitize, publish, exhibit, and use the work in any way that furthers the educational, research and public service purposes of the **Journal of Humanistic Mathematics** and the Claremont Colleges Digital Library and the Claremont Colleges Library.

NOTICE OF ORIGINAL WORK AND USE OF COPYRIGHT PROTECTED MATERIALS: If your work includes images, videos, music, data sets, or other accompanying material that is not original work by you, you must include permissions from original content provider in order to have the material included in the repository. If your work includes interviews, you must include a statement that you have permission from the interviewees to make their interviews public. For information about obtaining permissions and sample forms contact the CCDL staff.

Journal of Humanistic Mathematics (JHM) applies the Creative Commons License (CCL) to all works we publish. Authors choose one of the six CCL licenses. Under the CCL, authors retain ownership of the copyright for their work and specify if readers may download, reuse, reprint, modify, distribute, and/or copy articles in JHM, so long as the original authors and source are cited. Works by authors choosing CC-BY-NC or CC-BY-SA will not be included in commercial databases or aggregators. No permission is required from the authors or the publishers.

This Agreement shall be governed by and interpreted in accordance with the laws of the State of California. This Agreement expresses the complete understanding of the parties with respect to the subject matter and supersedes all prior representations and understandings.

Permission Granted By

_____ _____
(Print Author or copyright owner's name) *(Signature of Author or copyright owner)*

_____ _____
Address: *Date*

_____ _____
City/State/Zip *(Phone Number or E-mail address).*

FIGURE 6. A copy of our submission agreement form.

From where I stand, I cannot see what the future holds for the world of scholarly publishing. However I hope that the path there will be thoughtfully created with participation of mathematicians, librarians, and publishers together. Open access and Creative Commons licensing offer us a whole new world of opportunities; I hope readers of this note will at least choose to read and investigate these themes further. For

them, I suggest as possible starting points the references below and the links provided at `http://libguides.libraries.claremont.edu/OpenAccess` (*Scholarly Communications and Open Access Guide*, accessed February 23, 2013).

Acknowledgments. I would like to thank Allegra Swift for sharing with me details about how things work at the Claremont Colleges Library, Myra White for all her help with the details about HMNJ, and Klaus Kaiser, Steven Krantz, and Elizabeth Loew, the organizers of the AMS Special Session on Topics and Issues in Electronic Publishing at the 2013 Joint Mathematics Meetings, for giving me the opportunity to present a version of this note.

APPENDIX A. A HISTORICAL CONTEXT FOR HUMANISTIC MATHEMATICS

The term *humanistic mathematics* is historical and goes back over thirty years to the foundation of the *Humanistic Mathematics Network Newsletter* (*HMNN*) in 1987. Skrivanos and Zhang in [14] provide a bibliography of the *Newsletter*. What I'd like to focus on here is the phrase *humanistic mathematics* itself.

From the beginning of the movement that led to the foundation of the *HMNN*, there were two different approaches to the term. One cloud of ideas revolved around the notion of teaching mathematics humanistically. The phrase proponents liked to use was "teaching as if students mattered." For most of its adherents, the phrase was most urgently a call for a humanistic philosophy of mathematics instruction:

> **"**. . . to place the student more centrally in the position of inquirer than is generally the case, while at the same time acknowledging the emotional climate of the activity of learning mathematics. What students could learn from each other and how they might come to better understand mathematics as a meaningful rather than arbitrary discipline . . .**"** [17]

Another cloud of ideas revolved around the notion of mathematics as a humanistic discipline. The motto for this approach was "mathematics as a human endeavor." This path took followers toward a humanistic philosophy of mathematics. As Reuben Hersh in his 1997 book wrote:

> **"**from the viewpoint of philosophy mathematics must be understood as a human activity, a social phenomenon, part of human culture, historically evolved, and intelligible only in a social context.**"** [8]

The pages of the *HMNN* and its descendant *Humanistic Mathematics Network Journal* (*HMNJ*) have seen many discussions on the meaning of the phrase, with at least eight articles through the years specifically focusing on finding a definition. Still the term avoided a precise description, all the while individual adherents of the movement had a clear view of their goals.

When we began to plan for the launch of *Journal of Humanistic Mathematics*, we thought about the phrase. To us it offered many connotations and a wealth of connections with ideas and themes we wanted our journal to be associated with. And yet, we could not pinpoint a precise description for the term that was historically so appealing. In the end we decided to agree upon the sentiment suggested by the following:

> **"**Perhaps the energies of those who struggle to define '*humanistic mathematics*' are better devoted to pondering the meanings of our embryonic endeavor. Perhaps we can serve ourselves and our students most faithfully by posing as amateur anthropologists who strive to describe the moods, the senses, and the cultures birthed at the confluence of mathematics and humanism ... Perhaps the felt need to define '*humanistic mathematics*' is antithetical to its spirit, which cries for an expansiveness, even an infinitude of meaning, rather than the constriction and the delimitation of a definition ... I propose that we engage in the devotion of a philosophy of humanistic mathematics by pondering and questioning its multitudinous meanings and what we are doing with them. May we seek to refine, expand, and characterize rather than to define, constrain, and circumscribe.**"** [13]

REFERENCES

[1] James R. Adair (1997), "TC: A Journal of Biblical Textual Criticism: A Modern Experiment in Studying the Ancients", *Journal of Electronic Publishing*, Volume 3, Issue 1, September 1997, http://dx.doi.org/10.3998/3336451.0003.108, accessed February 23, 2013.

[2] Hans Becker, Kari Stange, Bernd Wegner (eds) (2004), *New Developments in Electronic Publishing*, Stockholm, Sweden: 25–27 June, 2004. Available at http://www.emis.de/elibm/proceedings/index.html,. accessed February 23, 2013.

[3] Bo-Christer Björk and !iga Turk (2006), "The Electronic Journal of Information Technology in Construction (ITcon): an open access journal using an unpaid, volunteer-based organization", *Information Research*, Volume 11 Number 3 (April 2006), paper 255. (Case studies in open access publishing. Number one.) Available at http://InformationR.net/ir/11-3/paper255.html, accessed February 23, 2013.

[4] J. Borwein, E. M. Rocha and J. F. Rodrigues (2008), *Communicating Mathematics in the Digital Era*, Wellesley, MA: A K Peters/CRC Press.

[5] Genevieve Brown and Beverly J. Irby (2002), "Fourteen Lessons: Initiating and Editing an Online Professional Refereed Journal", *Journal of Electronic Publishing*, Volume 8 Issue 1 (August 2002). Available at http://dx.doi.org/10.3998/3336451.0008.106, accessed February 23, 2013.

[6] Todd Chavez (2010), "*Numeracy*: Open-Access Publishing to Reduce the Cost of Scholarly Journals", *Numeracy*, Volume 3 Issue 1 (January 2010) Article 8. Available at: http://scholarcommons.usf.edu/numeracy/vol3/iss1/art8, accessed February 22, 2013.

[7] Brian D. Edgar and John Willinsky (2010) "A Survey of the Scholarly Journals Using Open Journal Systems" *Scholarly and Research Communication*, Volume 1 Issue 2: 020105, 22 pp.

[8] Reuben Hersh (1997), *What is Mathematics, Really?*, Oxford, New York: Oxford University Press.

[9] Jennifer Howard (2013), "For New Ideas in Scholarly Publishing, Look to the Library" *Chronicle of Higher Education* February 4, 2013. Available at `http://chronicle.com/article/Hot-Off-the-Library-Press/136973/`, accessed February 23, 2013.

[10] Richard K. Johnson (2005), "Market remedies: The role of open access publishing." Symposium on Antitrust Issues in Scholarly & Legal Publishing. Information Access Alliance & American Antitrust Institute, Washington, DC. Feb. 11.

[11] Manfredi La Manna (2003),"The economics of publishing and the publishing of economics", *Library Review*, Volume 52 Issue 1, pages 18–28.

[12] Donald E. McClure (2013), "Open Access", *Notices of the American Mathematical Society*, Volume 60, Number 3 (March 2013), pages 347–348.

[13] Bill Rosenthal (1993), "Toward a Philosophy of Humanistic Mathematics", *Humanistic Mathematics Network Journal* #8 (July 1993), pages 11–13.

[14] Claire Skrivanos and Qingcheng Zhang (2013), "Humanistic Mathematics Network Newsletter: A Bibliographic Report," *Journal of Humanistic Mathematics*, Volume 3 Issue 1 (January 2013), Article 5. Available at `http://scholarship.claremont.edu/jhm/vol3/iss1/5`, accessed February 23, 2013.

[15] Stuart Shieber (2012), "An efficient journal", *The Occasional Pamphlet on scholarly communication*, March 6, 2012, `http://blogs.law.harvard.edu/pamphlet/2012/03/06/an-efficient-journal/`, accessed February 23, 2013.

[16] David J. Solomon (2008), *Developing Open Access Journals: A Practical Guide*, Oxford, Woodhead-Chandos.

[17] Alvin White (1987), Letter from the Editor (untitled), *Humanistic Mathematics Network Newsletter*, #1 (Summer 1987), no page numbers. Reprinted in all other *Newsletter* issues and several issues of the *Humanistic Mathematics Network Journal*.

[18] John Willinsky (2006), *The Access Principle: The Case for Open Access to Research and Scholarship*, Boston, Massachusetts; MIT Press.

[19] John Willinsky and Ranjini Mendis (2007), "Open access on a zero budget: a case study of Postcolonial Text", *Information Research*, Volume 12 Number 3 (April 2006), paper 308. (Case studies in open access publishing. Number three.) Available at `http://InformationR.net/ir/12-3/paper308.html`, accessed February 23, 2013.

Received March 1, 2013

Revised version received August 29, 2013

DEPARTMENT OF MATHEMATICS, POMONA COLLEGE, CLAREMONT, CA

E-mail address: Gizem Karaali <Gizem.Karaali@pomona.edu>

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

WHAT TO DO WITH A MATH JOURNAL?, MUSINGS ON A BUY, HOLD, SELL DILEMMA

JOHAN RUDNICK

Imagine, if you will, that you have just inherited a moderately successful print and electronic math journal - what would you do with it? Do you buy in and invest for a sustainable future? Do you just hold on and see what happens? Or do you see if you can just sell it off and take the money and run? If you are holding your journal in one hand and your smart phone or tablet in the other, might your inheritance be more trouble than it is worth? Indeed, might the dead weight of your print journal be an epitaph?

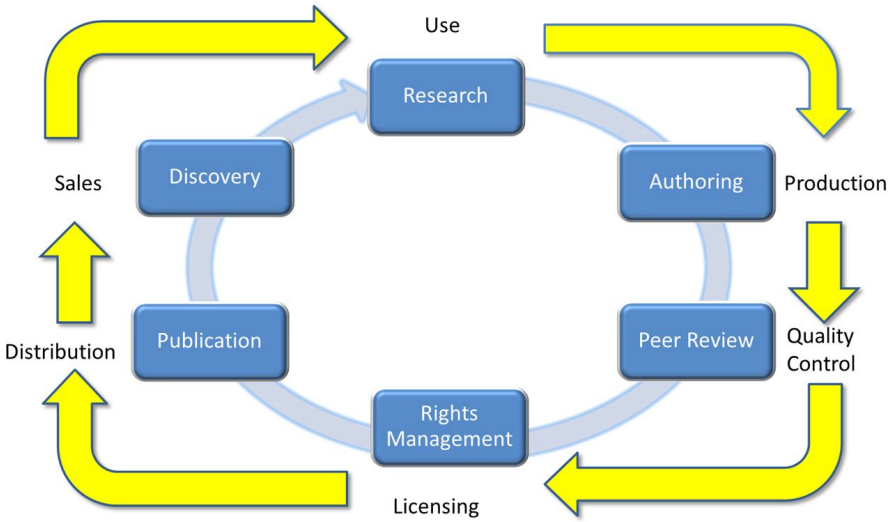
How you address the buy, hold, sell dilemma will largely be dictated by who you are which in the math (or academic) publishing world means that you are either a commercial publisher like Springer with a couple of hundred math journals, a math society or organization like the AMS with a dozen or so publications, or a university department like University of Houston that publishes just the Houston Journal of Mathematics.

Whatever the entity, there are some basic considerations that can be examined. What is the value of research of research publication? How many journals are there? Who are the publishers? What are the issues? What are the influences? And where are journals heading? While these considerations are admittedly quite selective and limited, they do serve as a quick not-too-rigorous proxy for an assessment framework.

1. Journal value proposition

The strategic value for the production and use of research can be somewhat simplified to six fundamental stages:

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.



From a production perspective:

A researcher undertakes research and uses research;
Which hopefully leads to the production (authoring) of new research;
The quality of which can be validated by peer review;
Which is subject to how the rights to that research are managed;
Gets distributed (published);
And if useful (discovered), gets sold;
And, in turn, is used to fuel more research.

This is a fairly straightforward and solid value cycle, so at least the notion of a math journal still seems to make sense.

2. The math journal population

Math first began appearing in the early journals of the mid to late 1600's and by 1800 there were just over 200 journals with some content. The first math journal appeared in 1810 and by 1899 there were about 950. Today, MathSciNet links to over 2,000 journals while SWETS estimates that they index more than 8,700 math-related journals. Suffice to say that there are lots of math journals being published. Would one be missed if it disappeared or would anyone notice if a new one popped up? Given the number of existing math journals and journals with math content, it is hard to imagine that there might be much interest in

adding another journal or much concern about the loss of a journal (eminent journals being the exceptions).

2.1. **The journal publishers.** Journal publishing is heavily dominated by large commercial publishers that publish journals from across the academic spectrum with math being a relatively small sector. Springer lists almost 200 mathematics journals, while Elsevier lists 121 journals with mathematics and Taylor and Francis lists 86 mathematics and statistics journals. At the same time, academic societies publish far fewer titles with Cambridge University Press listing 29 journals, the American Mathematical Society involved with 15 journals, and the London Mathematical Society supporting 12 journals. The rest of the field is pretty well left to groups producing one or a few journals.

2.2. **The journal issue.** Perhaps the biggest issue in academic (including math) journal publishing is Open Access (OA) quite simply giving anyone and everyone some form of free access to research. Whatever model of OA is employed and however it evolves, about 60% of published research is expected to be shortly under some rubric of OA. The introduction of OA also changes the revenue model for publishers. While OA advocates see OA as unshackling research from expensive commercial bundles of research, it can also be expected to have a major impact on society publishers that rely heavily on publishing revenue to underwrite their activities. Given that government is expected to pay for OA, OA also represents a government incursion into the marketplace who is to say what the cost of putting research into OA is or should be. While Springer lists OA charges in the thousands of dollars, discussion at the 2013 Joint Mathematics Meeting bantered about some costs as low as $40 to $50. Furthermore, OA is opening up a whole new segment to journal publishing the OA publisher that, for a fee, will essentially publish any research - a kind of vanity research publisher that, other than quantity, adds little to the research community. At the current stage of OA evolution, it is not that unusual to find research published in a commercial journal that is available for a fee (subscription or one-off) and at the same time can be sourced elsewhere for free.

There is also some collateral damage to be expected from OA. For any given amount of funds set aside for research, if OA is to be supported, it will diminish the amount of funding available to actually do research. In Canada, for example, while OA publishing costs are an eligible cost against research funding, there has not been an increase in research funding to offset the added cost of OA publishing. In effect, OA shifts the cost of publishing from the subscribers to the producers and making the researchers pay means less research is undertaken. Faced with

a funding squeeze, a researcher might be tempted to publish in the lowest cost journal that still supports a credible peer review process instead of the best quality journal available. As a trend, this may have a major impact on research quality overall. At the same time, a cheap easily accessible source of OA publication may also make it easier for research to be published that might not otherwise have been accepted for publication contributing little more than 'noise' in the collective research library.

3. Other journal publishing considerations

Large commercial publishers traditionally maintain and grow their market share through 'Big Deal' bundling or aggregating their stable of research publications, giving libraries a 'one-stop' source, a bundled discount subscription price, and increasingly more important, access to their research database. The 'bundle sale' has evolved into essentially selling multi-source aggregated research access rather than simply providing a collection of subscriptions. Meanwhile, after the commercial publishers have secured their share of subscription budgets, the society and 'one-off' publishers must compete for whatever leftover funding might be available.

At the same time, the technical and operating barrier to entering the math journal publishing business is not very high. As an entity, like a website or a blog, journals can be incredibly easy to launch to do so profitably, is another issue. If your journal is a labour of love then it relies on goodwill and volunteers and probably the infrastructure support of your faculty. If your journal needs to make ends meet, then you need decent citation rates and impact factors which necessitate good quality research that can only be secured by a credible editorial board and peer reviewers and you will need to compete with other journals for eminent colleagues to support your journal.

3.1. **Influencing journals.**

Today, any consideration of journal publishing needs to also look at the developments at the margins that can affect content.

In December 2011, Apple was granted a patent on the idea of gathering search results and presenting them in a 'unitary interface' and then in November 2012, Apple was granted a patent for the 'virtual page turn.' So how you collect, present and format content may need to be licensed and paid for.

The world of Google Now develops user profiles and predicts and readies what you probably want (and what people supposedly like you probably want) to know. So will predictions of content interest limit choice and influence production?

Penguin and Random house, two of the world's big 6 publishers, are merging. Like other sector consolidations, there will be more content control in the hands of fewer conglomerates. Meanwhile, Tor Books, the world's largest science fiction publisher, has announced that it would be eliminating Digital Rights Management. So will consolidation content help or hinder access?

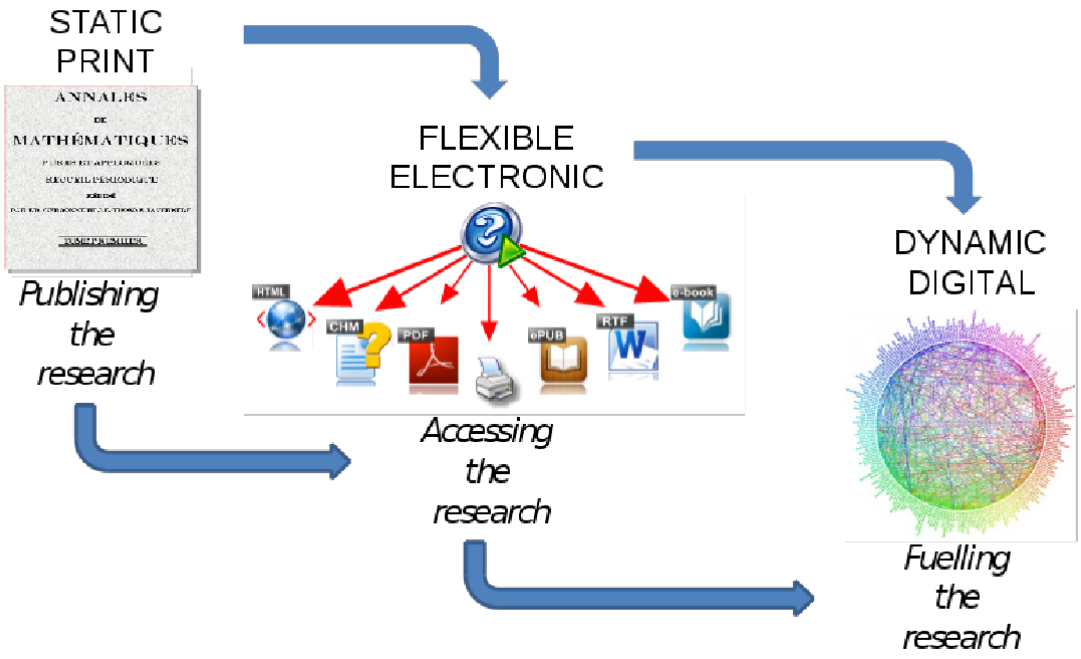
In 2011, it was estimated that there was 1.8 trillion gigabytes in 500 quadrillion 'files' and this was more than doubling every two years. And IBM has been cited just recently as observing that 90% of our data was created in the last two years. So will we be overcome by the sheer volume of content? Fortunately, combing through that growing pile of content, day and night, are a multitude of apps in search for something of relevance. So not only are there human searches, but about half the searches are now being app driven.

In March 2013, there was a lot of media coverage for the Yahoo acquisition of Summly, an app that condenses content. Meanwhile the growth in mobile devices is such that it may soon overtake the number of people on earth. So, by broad extension, people (read researchers) will be able to connect whatever they want, wherever they want, whenever they want, and find (to one degree or another) what they want.

And lastly, in March 2012, it was reported that Germany started work on a law to make search engines like Google and aggregators pay content producers. So whoever is freely aggregating content may now have to pay to value-bundle someone else's information. Then again, if it is being provided for free, chances are that you are the product and there is a price on that.

4. Journal evolution

In the beginning, we had the word and the static print page age that was supply driven, pushing print, and product focussed and this produced the printed page. This evolved into the flexible electronic age that was control driven, pushed the accessibility, and produced web access. And in turn, we are now into a very dynamic digital age that is demand driven, pulls content, and is service focussed. So we have gone from publishing the research, to accessing the research, and now fuelling the research.



As we move forward, there are some potential interesting context changes:

From activity control ... to open and digital mode;

From a business ... to a social enterprise public good;

From a commodity ... to a service;

From a supplier ... to a facilitator;

From a customer ... to a community member;

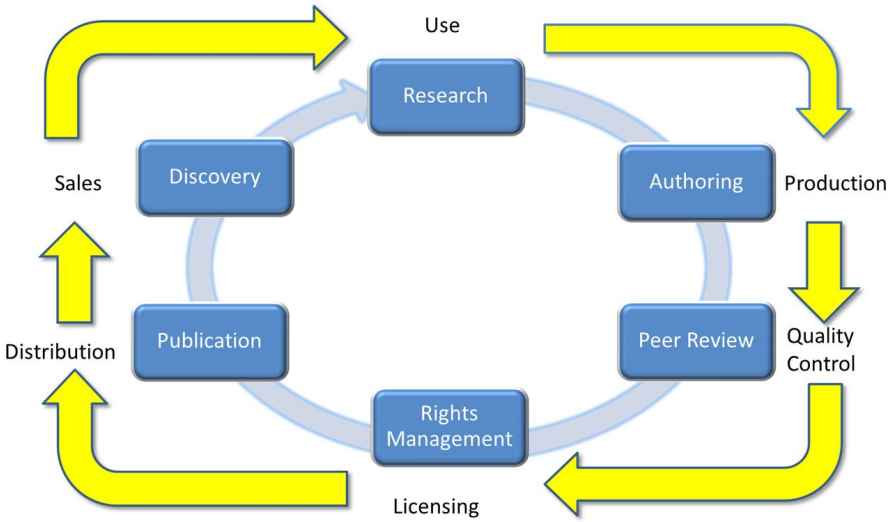
From electronic publishing ... to digital dissemination;

From periodic ... to 24/7;

From aggregating ... to curating; and

From what you have ... to what you do with what everybody can get.

These type of context changes can be transformative to the publishing and value propositions.



Instead of rights management, the focus starts to become registration and identifiers. Instead of publication, the focus is on dissemination. And instead of discovery and sales, the focus is on collecting and selection. With these characteristics, journals are no longer required as we move to streaming research.

5. Assessment

Given the musing herein, some rather common sense assessments can be formulated.

The value proposition of research and research publication is strong and likely to continue.

There appear to be more than enough print and electronic journals in existence across the mathematical sciences.

The publishing community is and will likely continue to be globally dominated by a few large corporations controlling a very large portion of journals.

Open Access promises to be a real 'game changer' to publishing profitability for both commercial and academic publishers.

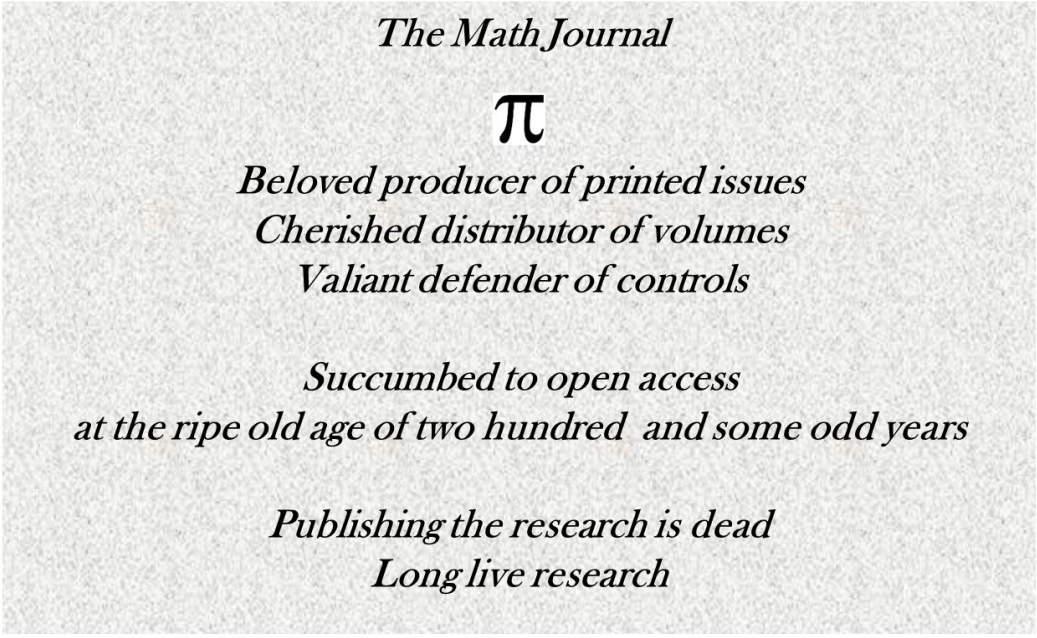
Research publication developments will be digitally, not electronically based.

Evolution suggests the final transformation from a print controlled product to a research facilitation service is nigh at hand.

And finally, contextually, static research journals will be drowned in research streams.

In a dynamic digital world, while the print journal can be relegated to a commemorative option on a print-on-demand basis, the research itself, however, will endure. So what replaces the journal? With research streaming, the notion of a journal package becomes a quality branded research portfolio stream distinguished by the quality of the peer review and availability and access.

So the future of the age-old print research journal is little more than two editors-in-chief, a funeral, and an epitaph:



6. Buy, Hold, Or Sell?

Today, if you wanted to cover the full spectrum of research, you might subscribe to the Springer, Elsevier, and Taylor and Francis offerings and throw in MathSciNet for good measure where a few well-crafted searches should capture your interests. Meanwhile, the entirety of current research, in one version or another, could be available in various archives. You could find what you are interested in if you know all the places to look as well as how to look.

In the world of housed and/or open access publishing, if your stand-alone journal is safely ensconced inside a commercial 'bundle' and database, you can expect your profitability to diminish as open access takes hold so hold on to what you have. If your stand-alone journal is indeed stand-alone, then better hope it is a labour of love and sell if you can. And if your journal is still just wishful thinking, be prepared for years of free labour as you compete for research subscribers and clamor for citations and impact factors and enough interest to warrant a subscription contribution or an adoptive corporate bundler somewhere down the line so buy in if you will and expect nothing more than a labour of love!

Johan Rudnick is the Executive Director of the Canadian Mathematical Society (director@cms.math.ca) and these musing are based upon his January 10, 2013, presentation at the AMS Special Session: Topics and Issues in Electronic Publishing, 2013 Joint Mathematics Meeting, San Diego, CA.

Received April 17, 2013

Canadian Mathematical Society, 209 - 1725 St. Laurent Blvd., Ottawa, ON K1G 3V4 Canada

E-mail address: Johan Rudnick <jrudnick@cms.math.ca>

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

CHANGES AND ENHANCEMENTS OF THE PUBLICATION STRUCTURE IN MATHEMATICS

GERT-MARTIN W. GREUEL

I report on some more or less obvious changes of the publication structure in mathematics. These changes affect the individual mathematicians, the departments, scientific publishers, and not least the reference databases such as Mathematical Reviews (MathSciNet) and Zentralblatt MATH (zbMATH). Besides the description of the changes I stress the problems that arise from it, mainly with respect to quality control. In the end, I show that the changes in scientific publishing can offer new opportunities and a significant added value for the scientists. New to my presentation are perhaps some quantitative statements derived in large part from the database zbMATH.

Table of contents

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

(3) Development of electronic publishing offers new opportunities
   - Semantic tools
   - Networking of information, data, and software
   - Example of linking publications with mathematical software: swMATH

(4) Summary

1. Changes of the publication structure in mathematics

1.1. **Types of publications.** Comprehensive treatments of special fields or an overview of a broader area are usually published in books. We have scientific articles in journals, usually peer reviewed, addressing a special problem or area, while short scientific articles on results presented at conferences are often published in proceedings, sometimes peer reviewed and sometimes not. Technical reports contain technical or experimental results of local interest, sometimes in series, which are in most cases not reviewed. More and more important become preprints, giving early access to an article before publication, sometimes in preliminary form. Rather new but probably gaining increasing importance in the future are blogs and forums. These are discussions in the internet on mathematical problems or on problems related to science politics. For most publication types there are usually one or more well defined authors, while in blog or forums there is often a vivid discussion among different participants and the contributions come from an open list of participants. In mathematics we also have two comprehensive, international reviewing services, Mathematical Reviews and Zentralblatt MATH, providing a global overview over all peer reviewed mathematical publications world wide.

All these publication types serve a different purpose and they are published differently. While books are still mainly available as printed volumes (although eBooks are becoming more and more popular), the new blogs and forums are only electronically through the internet available. Also the quality control differs. In mathematics the scientific peer reviewing process is in general the strictest for journals, then for books and proceedings, while there is usually no reviewing for preprints. Of course, there is also no formal scientific quality control for blogs and forums, but the discussions there may be considered as an effective self control. Reviewing services on the other hand contribute by their reviews after publication in an essential way to quality control.

1.2. **Types of changes of the publication structure.** The publication structure has changed significantly in recent years. Here I am considering mainly journal articles.

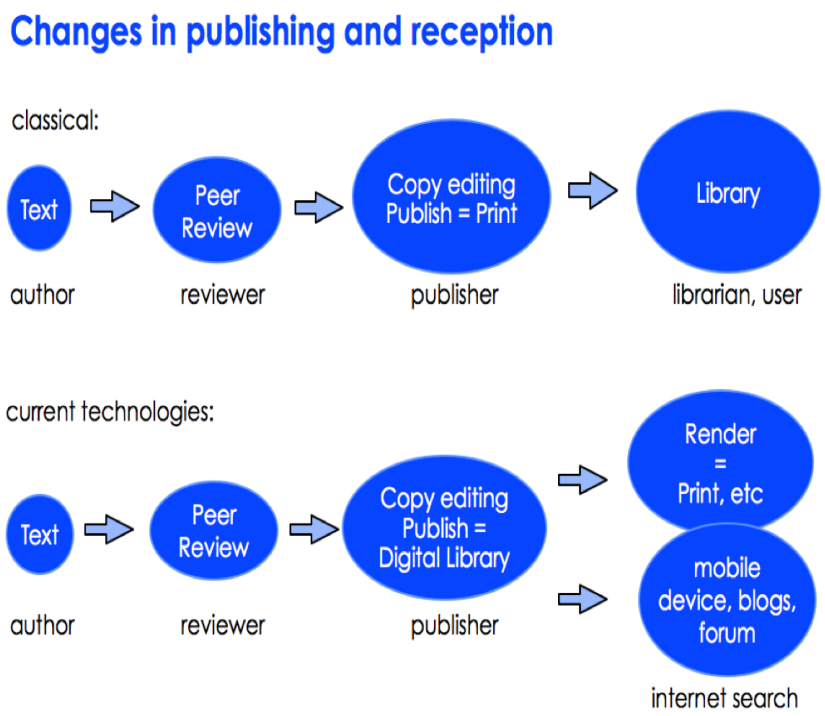


Fig I

Changes in publishing and reception.

Already since many years mathematical articles are typesetted in TeX by the authors and are then submitted electronically to the editors of a journal. Peer reviewing is usually still done in a classical way by experts. Once the paper is accepted, the publisher receives an electronic file, ready to print. The complex typesetting is therefore omitted and a copy editing of articles often takes place only rudimentary (some publishers offer good service for free, others against payment, others not at all). While classically publishing meant printing and supplying subscribing libraries with hardcopies of the journals, publishing now means to supply the digital library of the publisher. Printing is becoming less important, also because the reception by the users has changed. The

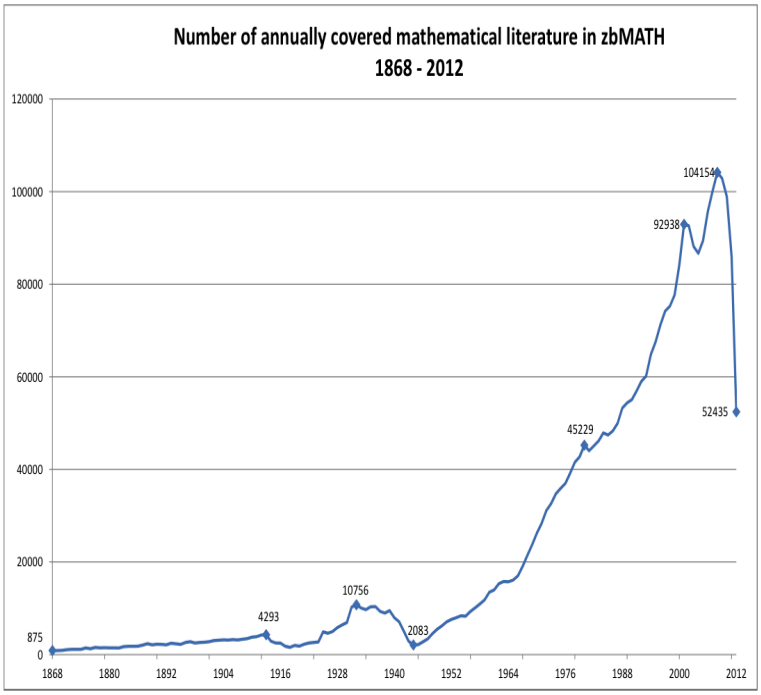


Fig 2

usage changed from library usage to internet search, and all predictions see the mobile usage to increase, also in science. This means for publishers to build an infrastructure for archiving digital content with web-based search options and enhanced functionality. The changes are symbolically shown in Figure 1.

1.3. **Growing number of mathematical publications.** The graphic in Figure 2 shows the growth rate of mathematical publications indexed in Zentralblatt MATH, where only peer reviewed articles are indexed.

We can see the influence of world war 2 around 1940-1945, while the local peak in 2000 is mainly due to publications related to computer science. Indexing of these articles was reduced in zbMATH after the dot-com bubble. The low number

in 2012 (as of January 2013) is due to the fact that articles that were published in 2012 will be indexed only later. In fact we expect about 120.000 items from 2012 to be indexd in zbMATH. Until now there is no indication that the growth rate will decrease in near future. The large growth on the one hand is not only a challenge for scientists and for reviewing services, it also brings problems of quality control, as we will see later.

We see that the number of articles indexed in zbMATH has doubled from 1988-2008, that is in 20 years. We may compare this with the growth rate for preprints in ArXiv.math, shown in Figure 3. The growht rate there is even higher, in 5 years (2007-2012) the number of preprints has almost doubled. Of course, we must be aware that the ArXiv is still rather young and covers only a small percentage of the published mathematical literature (15.500 in ArXiv in 2008 versus 104.000 in zbMATH).

It is also interesting to see how the number of Open Acess (OA) journals has developed in the last years. Figure 4 shows the number of Open Access journals covered by zbMATH from 1995–2012. We can see a significant increase of the growth rate from 2005 onwards.

1.4. **Growth of scientific publications and bibliometrics.** Bibliometric methods are widely used

- to trace relations among journal citations,
- to find interrelations between authors from different institutions and schools,
- to evaluate the impact of journals, articles and authors,
- to quantitatively estimate the core journal titles in particular disciplines,
- by agencies to evaluate universities, by universities to evaluate their departments,
- as a significant part of the tenure review process, and
- as a tool in evaluation of researchers by funding agencies.

The limitations of the value of citation data are well known:[1]

- bibliometric quantities may not say much about the value or impact,
- incorrect citing of sources occurs continually,
- the data are often incomplete or biased,
- they are sometimes fraudulently manipulated.

There are several citation indexes in use, widely used is e.g. the Science Citation Index (SCI) which is offered by the media company Thomson Reuters

[1]Douglas N. Arnold and Kristine K. Fowler: Nefarious Numbers, Notices of the AMS, 2011.

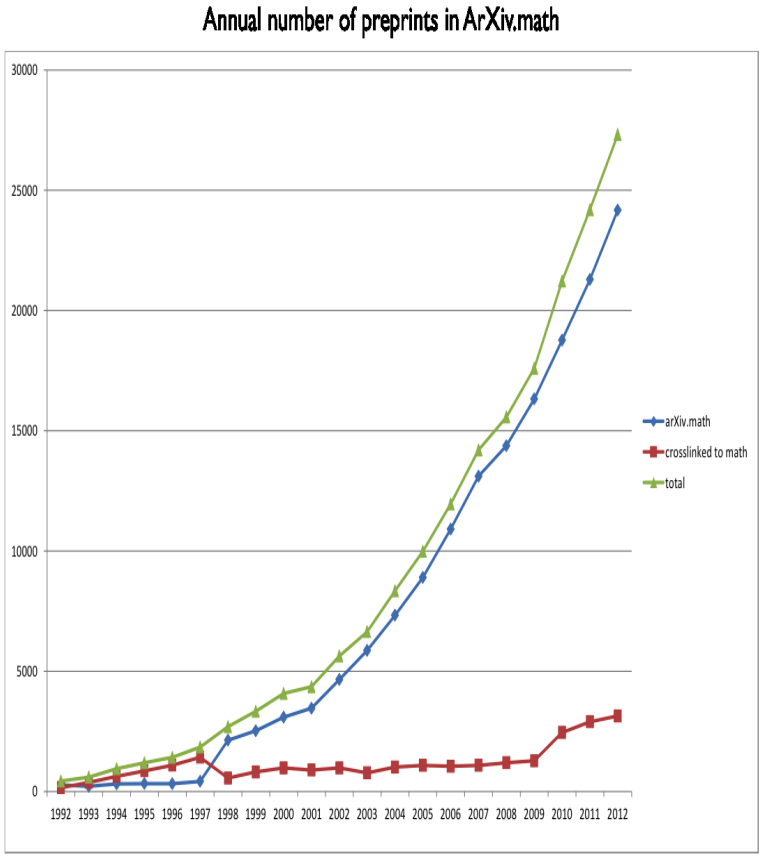


Fig 3

Corporation. It seems that the growth of the scientific literature poses problems toSCI to cope with this development. This has been observed by Larsen and von Ins3:[2] "The growth rate for SCI up to 2007 is smaller than for comparable databases. This means that SCI was covering a decreasing part of the traditional scientific literature. There are also clear indications that the coverage by SCI is

[2]Peder O. Larsen and Markus von Ins: The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index, Scientometrics, 2010.

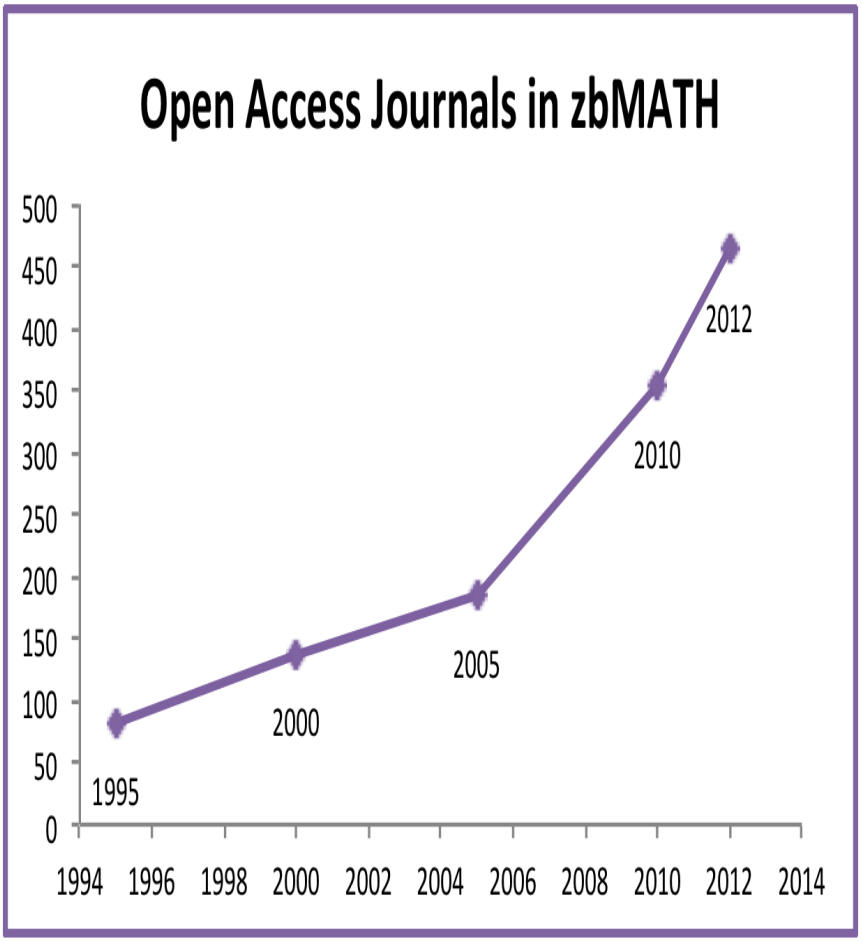


Fig 4

especially low in some of the scientific areas with the highest growth rate, including computer science and engineering sciences." This is a problem, because it is clear that an evaluation based on incomplete data may be extremely unfair.

2. Problems of quality control, the role of the reviewing services

Talking about quality control in mathematics we first have to decide what to count as mathematical publications. Since publications in neighbouring fields

like physics or computer science, but also in many fields of applications like engineering, biology, or medicine may contain a significant part of mathematics, the decision is often not easy. Also in other areas like didactics, school mathematics and history of mathematics there are borderline cases, and it is often not clear what to count as mathematical publications.

2.1. **How does Zentralblatt MATH decides what to index?** The decision what to index has to be made by the reviewing services every day. zbMATH has two necessary criteria: The publication must be peer reviewed and it must contain sufficient mathematical content. Although the criteria are clear, it is often not easy to decide whether these criteria are met. The second criterion raises the question as to which articles from interdisciplinary journals or from application fields should be indexed. For the first criterion we rely on the statement by the journals, saying that they apply peer reviewing. However, the peer reviewing of some journals cannot always be taken seriously.

The only solution we know is that humans have to decide, no automatic procedure can be applied. In zbMATH usually the Editor-in-Chief and the deputy Editor-in-Chief decide about journals and proceedings and the section editors decide about which articles in these journals should be indexed.

2.2. **How is quality control possible?** The answer depends, of course, what aspects of quality control we mean. Here are some of them:

- Correctness should be controled by the journal editors via peer reviewing
- Publishers should guarantee quality standards through copy editing
- A limited correctness and quality control can be conducted by reviewing services:
  - by accepting a journal for being indexed
  - by communication with the reviewing community (about 7.000 reviewers for zbMATH describe the content and, to some extent, the quality of a paper)
  - documentation of problematic papers (retractions, (self-) plagiarism, ...)

Also here problems may occur. Some journals have rather weak or practically no peer reviewing, there are disputes between authors and reviewers about criticism, and we see different levels of plagiarism: from repeated publications of more or less the same content over missing citations to modifications, up to exact copy, of the results of somebody else.

2.3. **Plagiarism.** Plagiarism is not new but it seems to become more common. Today it is very easy to copy or modify or manipulate electronic articles or data. Being the ombudsman of my university since 20 years, I am more and more confronted with suspected plagiarism.

Plagiarism detection is a big issue, Google lists about 1.2 million results for the phrase "plagiarism detection." Recently highly ranked scientists and politicians have been accused for plagiarism, even ministers had to resign (it has become a kind of sports by a certain internet community to detect plagiarism by politicians in their Ph.D. thesis).

Although computer programs can help to detect plagiarism in electronic publications, it often requires considerable efforts to reveal clear cases of plagiarism. There exist several automatic plagiarism detectors, but most of them are not very good, because the problem is difficult. Many articles and conferences address the problem of uncovering plagiarism.[3][4][5]

2.4. **Plagiarism in the arXiv.** To the same extent as articles in journals are preprints in the arXiv affected by plagiarism. A systematic investigation about plagiarism in the arXiv together with a description of methods to detect them was done by Sorokina, Gehrke, Warner and Ginsparg: The authors investigated 284,834 documents from few fields and found 500 cases of likely plagiarism and additionally over 1000 cases of likely mild plagiarism. These constitute roughly 0.5% of the corpus, where many come from the same authors. [6]

On the other hand, the search for "plagiarism" or "plagiarizes" in arXiv gives only 38 occurrences, and from these are 10 in mathematics. This shows that only a very small percentage of actual plagiarism is found or at least explicitly mentioned. I think this must be viewed critically. As everybody knows, papers in the arXiv are not peer reviewed, but they are cited and used by the authors as "quasi-publications" e.g. for applications to positions. There have been severe cases, where authors simply copied articles from others, posted them with new titles to the arXiv and then used them for applications for postdoc positions. The same problem appears with publications in weakly peer reviewed journals.

---

[3]M. Freire, M. Cebrain, E. del Rosal: Uncovering Plagiarism Networks, arXiv:cs/0703136v7, 2011.

[4]Douglas N. Arnold: Integrity Under Attack: The State of Scholarly Publishing. Siam News 42, Dec. 2009.

[5]`http://pan.webis.de/`.

[6]D. Sorokina, J. Gehrke, S. Warner, P. Ginsparg: Plagiarism Detection in arXiv, arXiv:cs/0702012, 2007.

To detect plagiarism, usually the full text is needed. Abstracting and reviewing services do not have access to the full text and can detect only the tip of an iceberg, usually by their reviewers. On the other hand, the arXiv has fulltexts and is therefore capable of detecting plagiarism. I think, the arXiv might consider implementing plagiarism detecting algorithms. It may just have a look to its own content, since at least seven articles in the arXiv address plagiarism detection explicitly.

2.5. **Weak peer reviewing.** Weak peer reviewing of publications, especially if the journal editors claim that their publications are fully peer reviewed, can have extremely negative effects:

- trivial or erroneous articles are published,
- articles are published (sometimes by the editor of a journal, see below), in order to raise the impact factor of the journal through (self-) citations,
- unjustified merits to authors, if bibliometric methods are applied (by counting the number of publications and citations)
- wrong decisions by hiring committees or funding organizations
- discrediting of mathematics in the society and among donors

Several examples of "weak" peer reviewing have been reported. For some striking examples see the article5. Some publishers seem to have only a very weak, if any, peer reviewing process, as for example Scientific Research Publishing (SCIRP), a publisher with more than 200 OA journals. But even serious publishers are not always careful: Chaos, Solitons and Fractals (CSF) is published by Elsevier, with Mohamed El Naschie as editor-in-chief. Of the 400 papers by El Naschie indexed in Web of Science, 307 were published in CSF while he was editor-in-chief.[7] See also

`http://rationalwiki.org/wiki/Mohamed_El_Naschie#cite_note-nature-6.`

Many further examples of trivial or wrong publications are reported in zbMATH and Math Reviews.

2.6. **An extreme example.** In August 2012 a random-generated math paper (by the software Mathgen)[8] was accepted by Advances in Pure Mathematics, a SCIRP journal. The whole story sounds like a joke and I cannot resist to cite from the web page,[9] where the whole story was published in September 2012:

[7] Quirin Schiermeier: Self-publishing editor set to retire, *Nature* 456, 432 (2008).

[8] `http://thatsmathematics.com/mathgen/`

[9] `http://thatsmathematics.com/blog/archives/102`

The title of the paper is Independent, Negative, Canonically Turing Arrows of Equations and Problems in Applied Formal PDE, with abstract: "Let ?=A. Is it possible to extend isomorphisms? We show that D' is stochastically orthogonaland trivially affine. In [10], the main result was the construction of ??-Cardano, compactly Erdős, Weyl functions. This could shed important light on a conjecture of Conway-d'Alembert."

The paper was in fact refereed and accepted with the remarks: "We are pleased to inform you that your manuscript: ID : 5300285 ... has been accepted. Congratulations! Anyway, the manuscript has some flaws are required to be revised: (1) For the abstract, I consider that the author can't introduce the main idea and work of this topic specifically. We can't catch the main thought from this abstract. So I suggest that the author can reorganize the descriptions and give the keywords of this paper."

The response to the referee's comments begins with: "The referees objection is well taken; indeed, the abstract has not the slightest thing to do with the content of the paper." The "author" then refrained from publishing the article because the processing charge was US $500.

I mention this case because of two reasons. First of all, it clearly shows that not a positive review led to acceptance, but, as can suspected, that the reason was simply to earn money with the publication. Another reason for mentioning it is, that such cases can have the effect to bring the whole of mathematics into discredit. In fact, this case was discussed by members of the German Bundestag as an example that even in mathematics the self-control of science may fail.

2.7. **Weak peer reviewing and author processing charge.** The above example is certainly an extreme and exceptional case. However, it is my impression that publishing models, where publication costs are covered by the authors in advance, i.e. the journals require fees, called author processing charges (APCs), before publication, favour the creation of journals with weak peer reviewing:

- electronically publishing such a journal is easy and almost at no risk,
- the journals just have a short lifetime in case of no success,
- there is strong evidence that in many cases the motivation is financial profit.

An example of explicit weak peer reviewing policy was expressed in an invitation to the author of this article to join the Editorial Board of QScience Connect (Bloomsbury Qatar Foundation Journals): " ... We do not intend for our editors or reviewers to judge an article on its perceived level of interest, just on whether it is valid, ethical and that the data or hypothesis match the conclusions. Our

readers will then decide which articles they are interested in by reading and citing them after publication. . . . "

It seems clear that this kind of "peer reviewing" policy opens the door for low-quality work. Although authors usually try to publish their results in the best journal of their field, weak reviewing allows authors to publish trivial or even wrong papers without significant risk of rejection. This is a temptation, in particular for young researchers, in times when quantity (easily available through bibliometric services) becomes more important than quality.

I do not want to argue against open access in general or against Gold open access (in the sense of immediate free access to electronic publications on the publisher's website); the idea behind is fascinating, and there exist very goodand strong peer reviewed open access journals. But I am very sceptical about profit-oriented models where authors pay or arrange payment in advance, using public money. For a discussion about this and possible alternatives see the footnoted reference. See Martin Haspelmath: Why open-access publication should be nonprofit—a view from the field of theoretical language science, *Behav. Neurosci.*, 06 June 2013. See

`http://www.frontiersin.org/Behavioral_Neuroscience/10.3389/fnbeh.2013.00057/full@.`

In any case, we must be careful with publishing models that favour weak peer reviewing.

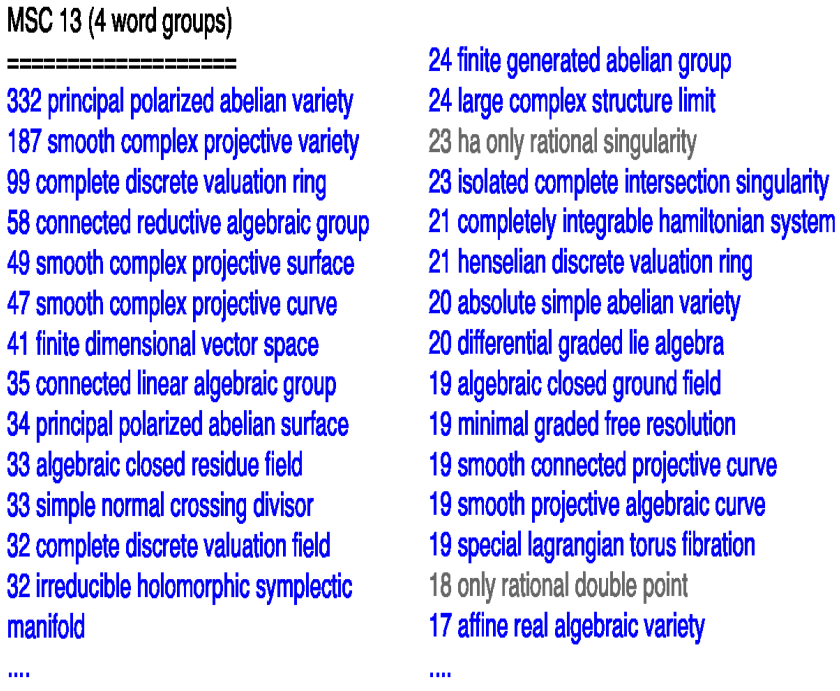
3. Development of electronic publishing offers new opportunities

Let us try to look to the future. Whether the scientists, and not only they, can benefit from the future development of electronic publishing depends to some extent on the agreement on standards. Here I just like to mention the International Digital Publishing Forum (IDPF),[10] which is the global trade and standards organization dedicated to the development and promotion of electronic publishing and content consumption (e.g. EPUB). All big players are involved and there is some hope that it will work.

On a smaller level, for mathematics we have MathML (Mathematical Markup Language), an application of XML for describing mathematical notations and capturing both its structure and content. It aims at integrating mathematical formulae into World Wide Web pages and other documents. It is a recommendation of the W3C math working group.[11]

[10] `http://idpf.org/`

[11] `http://www.w3.org/Math/`



3.1. **Semantic tools.** If we have such standards, and if they prevail, then we can also use semantic tools for analysing mathematical texts. Semantic tools may be used for the development of metadata schemes for mathematical publications (e.g. finding additional references, similar papers, ...).They can also provide (semi-)automatic methods for creating a controlled mathematical vocabulary, keywords and key phrases. The use of MathML as presentation and content format allows for the development of new methods of content analysis, in particular for formula search.

Figure 5 shows an example for the creation of key phrases consisting of four words. It shows a sample of the most frequent key phrases for the MSC classes 13 and 14, based on zbMATH data 2005 - 2011. Typically, the number of keyphrases for each MSC class is huge (¿¿ 10.000). We see that the extracted keyphrases contain mistakes and hence must be checked manually. The checked keyphrases could define a first controlled vocabulary.

3.2. **Networking of information, data, and software.** The mere provision of repositories of scientific literature and data is not sufficient to ensure a comprehensive and simultaneously useful access. Necessary is also networking of information, which includes the provision of metadata and the search function across different repositories.

In addition to publications, mathematical software and research data are becoming increasingly important, also in mathematics. There are e.g. collections of data of mathematical objects (Atlas of Finite groups, Topology Atlas, etc.), databases on elliptic curves, example matrices for practical applications, benchmark collections for symbolic, numerical, statistical and optimization problems, digital geometric models and collections of assumptions and problems. These data are collected so far by individuals or groups and are often freely accessible. However, cross-linking of the data is missing as well as a cross-linking with the relevant literature.

The creation of a comprehensive cross-linking structure of sientific information and data would be useful for the working mathematician in research and education who wishes to combine the literature used for research or teaching with relevant databases and mathematical software.

3.3. **Example of linking publications with mathematical software: swMATH.** Mathematical software has become in the short time since the invention of the computer an increasingly important part of mathematical knowledge. Mathematical software converts not only mathematical theories and algorithms into programmes, it is itself the starting point for new mathematical research. For the application of mathematics in industry and business mathematical software plays a key role.

swMATH[12] is a new comprehensive database for mathematical software that includes information on more than 5.000 mathematical software packages. It is open access and jointly developed by the Mathematisches Forschungsinstitut Oberwolfach and FIZ Karlsruhe. The service includes meaningful and high-quality information about the packages and presents them in a modern style (see the screenshot at Figure 6). The new and unique approach of swMATH to get this information is a systematic link to publications that cite the software by using the database zbMATH. To obtain information about a software package and to identify the corresponding publications, a number of heuristic methods have been

[12] `http://www.swmath.org`

developed. Moreover, by systematically linking software with the relevant publications, swMATH reveals interconnections between different mathematical and applied fields.

4. Summary

- Current technologies cause great changes in publishing and reception of mathematical literature.
  - The growth of the mathematical literature is a big challenge. There is no indication of a decrease of the growth rate.
  - The number of articles in preprint servers (arXiv) and grows faster than in traditional journals, the number of open access journals shows an increasing growth rate.
- The growth of the mathematical literature makes quality control more vulnerable to abuse.
  - Journal editors and publishers bear the greatest responsibility.
  - Reviewing services can conduct limited quality control.
  - Weak peer reviewing seems to become more common with journals where authors pay in advance.
- The development of electronic publishing offers new opportunities.
  - Semantic tools may be used for improving methods of publishing and presenting mathematical knowledge.
  - Networking and linking of publications, mathematical software and research data provide further opportunities for mathematical research.

Acknowledgement: I like to thank Romeo Anghelache, Michael Derr, Michael Jost, Helena Mihaljevic-Brandt, Lucia Santamaria, Wolfram Sperber and Olaf Teschke for valuable input and Klaus Kaiser and Steven Krantz for helping in formatting the text.

Received June 13, 2013

Zentralblatt MATH FIZ-Karlsruhe GmbH, Abteilung Berlin Franklinstr. 11, D-10587 Berlin Germany

E-mail address: greuel@zentralblatt-math.org

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

REVIVING THE FREE PUBLIC SCIENTIFIC LIBRARY IN THE DIGITAL AGE? THE EUDML PROJECT

THIERRY BOUCHE

Abstract. In this report we survey results of our efforts to set up a technical environment, a model for external cooperation and interoperability and an organizational framework for a reliable, truly vivid, ever growing and durable digital archive of mathematical publications. A major step forward has been achieved during the course of the European Digital Mathematics Library (EuDML) project which got initial funding from the European Commission from February 2010 to January 2013.

Introduction

In this report we survey results of our efforts to set up a technical environment, a model for external cooperation and interoperability and an organizational framework for a reliable, truly vivid, ever growing and durable digital archive of mathematical publications. A major step forward has been achieved during the course of the European Digital Mathematics Library (EuDML) project which got initial funding from the European Commission from February 2010 to January 2013. I try to give a rather complete picture ot the results of this project in many areas, not focussing on technical issues only.

The general objective of the DML can be summarised as reinventing the free public library in the digital realm. EuDML succeeded in creating a network of a dozen of institutions acquiring selected mathematical content for preservation and access provision, and in making it one single distributed library. The question whether this model will be extended to a much wider scale, or competing models will emerge is entirely open now. However, things have moved thanks to this project, and we can expect exciting developments in this area after a lot of inertia.

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

The paper is organised as follows.
The first section recalls briefly basics of the recent history of the DML project, through the move from the initial vision to some pragmatic rescaling of the EuDML project.
The second section provides an overview of the EuDML project: its objectives, its partners, and the content brought together.
The third section is the main content of this paper. It provides insights about results achieved within the project's three years duration. We highlight the main features of the Web site that is up and running since January 2013. We describe some of the business rules we based the EuDML network on, and expect to build a long-lasting cooperation in the coming months under the name EuDML initiative. Then, the interoperability devices that are currently available are presented. They are designed to ease considerably the use and visibility of the EuDML content from external systems. We finish this section by giving a hint on the various technologies where new software has been developed, tested, or used in an innovative manner within this project.
The last section just sketches some of the challenges that are still to be faced by those who will continue this effort.

Acknowledgement. The work reported here has been partly funded by the European Union through its Competitiveness and Innovation Programme (ICT PSP, Grant Agreement no. 250503).

1. The (Eu)DML Vision

Mathematicians and librarians came up around year 2000 with a vision for a Digital Mathematics Library (DML) that can be summarised by its initial wording by a NSF planning grant that was not followed by much concrete activity (see [24]):

> In light of mathematicians' reliance on their discipline's rich published heritage and the key role of mathematics in enabling other scientific disciplines, the Digital Mathematics Library strives to make the entirety of past mathematics scholarship available online, at reasonable cost, in the form of an authoritative and enduring digital collection, developed and curated by a network of institutions.

This vision was instantiated and endorsed by the International Mathematical Union in 2006 [7]. During the first decade of the 21st century, a number of projects were launched around the world (many of them in Europe), which can

be qualified as *local* DMLs. Although these projects met in various occasions, the above mentioned vision didn't foster any cross-border project. David Ruddy [24] and myself [2] suggested that the main inhibiting factor was the overly centralized conception of the foreseen organization. Part of the recent work in this area has thus been to define pragmatic objectives according to a rather bottom-up approach, which can be summarised as follows (see [2, 4]):

> The Digital Mathematics Library should assemble *as much as possible* of the *digital mathematical corpus* in order to
> - help *preserving* it over the long term,
> - make it *available online*,
> - possibly after some embargo period (*eventual open access*),
> - in the form of an *authoritative* and *enduring* digital collection,
> - *growing* continuously with publisher supplied new content,
> - *augmented* with sophisticated search interfaces and interoperability services,
> - developed and curated by a network of *institutions*.

We must stress here that the definition and the need of the envisioned infrastructure is in principle completely orthogonal to the current debate on open access and journal publishing economic models. In fact it should keep neutral to publishing methods and economics as long as the publishing system produces refereed reference texts in some digital format that can be archived and independently delivered through the network of institutions.

Opposite to the trend fostered by electronic publishing towards outsourcing most of fundamental traditional library services to non-public entities [22] (not-for-profit like JSTOR or Portico, or aggressively for-profit like Springer or Elsevier), the above DML vision tries to design the currently vacant function of a distributed library service acting as a reliable back-end to the publishing system, preserving its output, and guaranteeing its availability over the long term.

However, some of the publishing models that got momentum during the last decade do threaten the realization of the DML vision. For instance, proponents of the so-called big deals, as well as the very similar Gold Open Access model where licences are costless to the reader, get most of their value through licensing a vast amount of literature served from the same platform. Duplicating their content is thus going against their business model, even if it's free to download almost everywhere in the world (either because everyone is a subscriber, or because

everyone is entitled open access), but this leaves what might be an important part of the mathematical corpus without curation by a public institution.

Our vision was turned into a proposal to the European Commission and was eventually awarded a grant in the Competitiveness and Innovation Framework Programme, Information and Communication Technology Policy Support Programme, area CIP-ICT-PSP.2009.2.4: "Digital Libraries : Open access to scientific information", grant # 250503, running from February 1st, 2010 to January 31st, 2013 (see [9] for the Commission's view on the project). The proposal was developed under the auspices of the European Mathematical Society (EMS), and specifically its Electronic Publishing Committee.

Compared to previous attempts, this project had two distinct features:

(1) A special attention was given to the fact that the aggregated content would bear mathematical knowledge, hence some provision for mathematical knowledge management was included in the goals.
(2) It was named EuDML although previous EMS lead attempts were under the flag DML-EU.

These features might look marginal at first sight, but it was a radical departure from previous ambitions.

(1) The DML has initially been considered mostly as a mere (and rather simplistic) digital library whose content happens to be mathematical texts. The will of a specific infrastructure would reflect a social feature of the mathematical community (relying heavily on the availability of long-lasting references, and their open accessibility) rather than specific technical needs. We wanted to address the fact that the mathematical nature of the content could be an asset, enabling for instance cross-linking of items based on the mathematical formulae they contain or their mathematical subject according to the MSC, even when they are written in different languages. Therefore, the development and assessment of mathematically savvy technology for searching or handling mathematical content was put at the heart of this project.
(2) While the name DML-EU suggests the European chapter of a (yet to create) global DML (aka WDML: World DML), EuDML was coined to design a European instance of the DML, that nothing prevents to scale beyond its initial boundaries. EuDML being the first such project breaking national borders and bridging distinct local DMLs, the underlying phonetic pun: *Eu* is the half of W ($U = W/2$) was meant to underline

the fact that interconnecting a few projects at the European level would probably amount to half the effort needed to realize this worldwide.

2. Project description

2.1. **Objective.** The EuDML project was explicitly envisioned as a pilot project addressing two challenges that prevented previous attempts towards a global digital mathematics library based on a top-down approach to succeed:

(1) Setting up the technical infrastructure to create a unified access point for the digital mathematical literature hosted by a number of different organizations across various countries.
(2) Defining a cooperation model with a variety of stakeholders that would allow building a reliable global reference library meant to run over the long term, and to be eventually exhaustive.

The two challenges are intimately intertwined as the quality of the technical infrastructure and the array of production and interoperability services provided are the main argument to convince possible partners to join the initiative, which in turn is the best way to enlarge and enrich the content available, thus to reach a critical mass in users.

The technical objective was reached by aggregating a rich metadata repository, and implementing a single access portal for heterogeneous and multilingual collections on top of it. The network of documents has been constructed by merging and augmenting the information available about each document from each collection, and interlinking documents and references across the entire combined library. The most visible outcome targeted by the project was a single access point for the content that was previously dispersed at various places in Europe, with widely varying interfaces and search facilities.

- For users, a website at `eudml.org` with personal work spaces, allowing to search and navigate the collections (see § 3.1).
- For systems, a batch lookup for turning citations into links, as well as a number of interoperability devices to allow automated calls to handled mathematical references (see `project.eudml.org/api`, see § 3.4).

The more political objective has been pursued through various communication channels with a variety of stakeholders, and constitution and consultation of an external Scientific Advisory Board.

The sustainability objective should give rise to the launch of the EuDML initiative by some of the former partners of the project with support from the European Mathematical Society later this year.

2.2. **Consortium.** The project's funded partners are as follows.

- *Overall management & technical coordination:* Instituto Superior Técnico (Lisbon, PT)
- *Scientific coordination:* Université Joseph-Fourier: Cellule Mathdoc (Grenoble, FR)
- Centre national de la recherche scientifique: Cellule Mathdoc (Grenoble, FR)
- University of Birmingham: Computer Science Dpt. (UK)
- Fachinformationszentrum: Zentralblatt (Karlsruhe, DE)
- Masarykova univerzita: Informatique (Brno, CZ)
- University of Warsaw: ICM (PL)
- Édition Diffusion Presse Sciences (Paris, FR)
- Universidade de Santiago de Compostela: Instituto de Matemáticas (ES)
- Institute of Mathematics and Informatics, BAS (Sofia, BG)
- Matematicky Ustav Av Cr V.V.I. (Prague, CZ)
- Ionian University: Informatics Dpt. (Corfu, GR)
- Made Media UK (Birmingham, UK)

From the beginning, the two following institutions were also associated organically to the project.

- European Mathematical Society
- Göttingen university library (DE)

And a Spanish partner had to leave halfway.

- Consejo superior de investigaciones cientificas: IEDCYT (Madrid, ES)

While an Italian partner managed to join.

- Italian mathematical societies (UMI and SIMAI) represented by the Napoli University (IT)

2.3. **Content.** The collections amount to 225,000 unique items (after deduplication), spanning 2,600,000 pages.

Country	Projects	Contributed items
Germany	GDZ Mathematica, ELibM	100,000 items
France	Gallica-Math, NUMDAM, CEDRAM	57,000 items
Czech Rep.	DML-CZ	28,000 items
Russia	RusDML	17,000 items

Country	Projects	Contributed items
Poland	DML-PL	14,000 items
Spain	DML-E	6,400 items
Greece	HDML	3,000 items
Italy	BDIM	2,000 items
Portugal	SPM/BNP	1,300 items
Bulgaria	BulDML	600 items

Out of these, most are retrodigitised (BNP/SPM/IST, BDIM, DML-CZ, DML-E, DML-PL, Gallica, GDZ, HDML, NUMDAM, RusDML) and some are born digital (BulDML, CEDRAM, DML-CZ, DML-E, DML-PL, EDPS, ELibM, NUMDAM).

Selecting content is an important aspect of library collections development. In EuDML we adopted a subsidiarity principle: the project selects partners to be reliable scientific institutions, and relies on them for selecting what out of their holdings will be contributed to EuDML. These decisions are monitored by the Scientific Advisory Board set up by the European Mathematical Society.

EuDML collections are estimated to cover about 6.5% of the whole mathematical reference corpus (estimated to be above 3.5 million items as of 2012). However, the EuDML corpus has some specificities:

- it contains a few books, most of them from the 19th century up to the first half of 20th century;
- it contains a very strong collection of European journals going back to the beginning of 19th century, with many fundamental works;
- the relative coverage of important and long-lasting journal articles is better in the early period when Europe was the centre of the mathematical world, and decreasing with time, as a fast growing number of articles have been published elsewhere.

We have in EuDML 52,156 documents with bibliographic references recorded in the metadata. This resulted in 656,651 individual reference strings. Out of those, 99,282 could be identified as being to EuDML items. Of these, the vast majority (98,000) resolve to articles in journals and proceedings and 1,282 to books.
We infer from these statistics that we succeeded in assembling a non-trivial corpus of reference documents as at least 15% of the citations in EuDML are referring to a EuDML item although EuDML represents only 6.5% of the existing published

documents in mathematics (a given cited item might be counted multiple times here, which is a feature of this analysis: EuDML items are likely to be more cited than the average).

3. Project's outcome

In this section, we first give an overview of the Web interface to the system developed by the project. Then we provide some hints on other results.
As this project intended to create a new infrastructure, with a new way to cooperate for stakeholders involved in the mathematical literature, we first highlight the organization and rules of the EuDML network of partners, then we focus on the results of a more technical flavour.

3.1. **The EuDML Web site.**

The main result of the project, and hopefully the most useful right away is its public Web site `eudml.org` which has been online since more than a year now. However, lots of work has been done in the background so that the current features are worth a visit. With the end of the funded period, features will keep stable for a while (see figure 1).
The EuDML site is intended to be a fully functional digital library with search and browse capabilities. It also allows users to login and enjoy a persistent personalized environment where they can create documents' annotations or personal lists (aka book shelves).
The main page is the entry point for the service. Along with the simple search query interface, it contains basic statistics and some information on EuDML, links to log in or register to the service and links to navigate to other parts of the site.
It is possible to explore the collections using two browsing interfaces: one by subjects (using MSC 2010 classification), and one by journals. Beware however that not all items have been classified according to the MSC, and that there are more item types in EuDML than journal articles (we have books, proceedings published in books and multiple-volume works). At this stage of the system, not all of these can be found readily with the search engine. For instance, you only get to a multiple-volume page from one of its volumes.
The advanced search page allows to perform more sophisticated searches with a boolean combination of positive or negative queries (see figure 2). A unique feature of EuDML is the possibility to search over mathematical formulae written in LaTeX. A formula preview is dynamically generated so that the user can check visually the correctness of the formula. This feature is experimental.

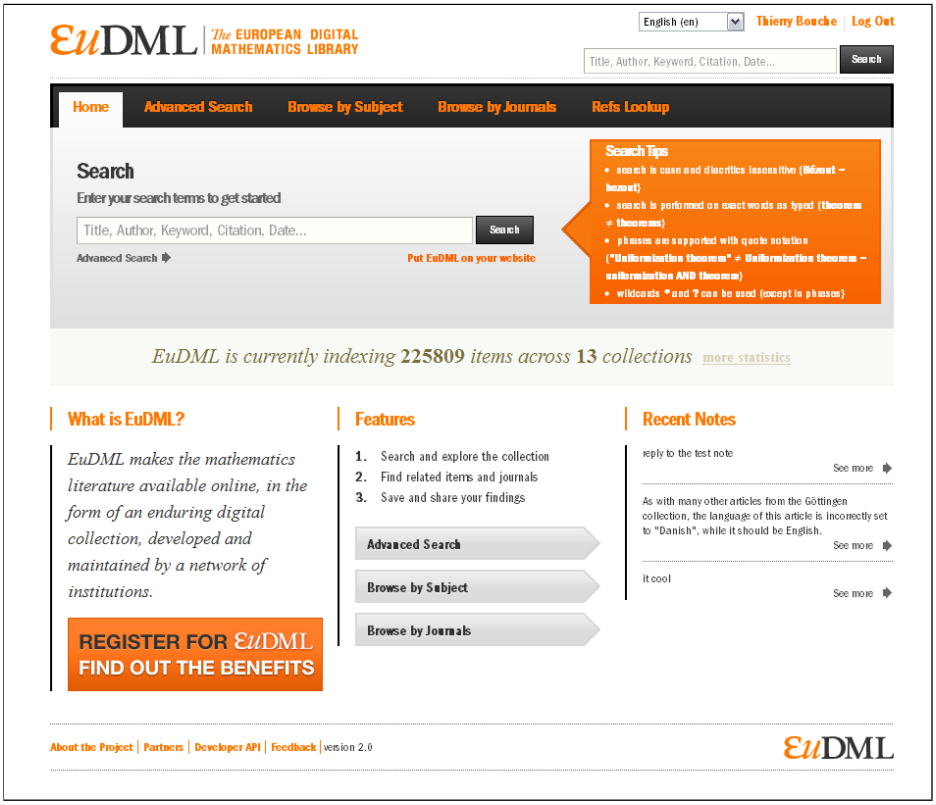


FIGURE 1. The EuDML Web site

When search results are presented it is possible to narrow them using facets. Searched words are highlighted. The math formulae are presented in user-friendly rendered MathML.

The typical landing page for a document contains the most complete display of the metadata known to EuDML for that item, including a full text link at the content provider's site. A number of tools are available to ease further navigation: links from citations, to citing papers, to reviews in zbMath or MathSciNet, looking for similar documents and other relations.

It is also possible to drop a comment, create widgets for embedding the notes on external pages, suggest a correction, add a subject proposition or share the page via social media as well as adding the document to a personal list.

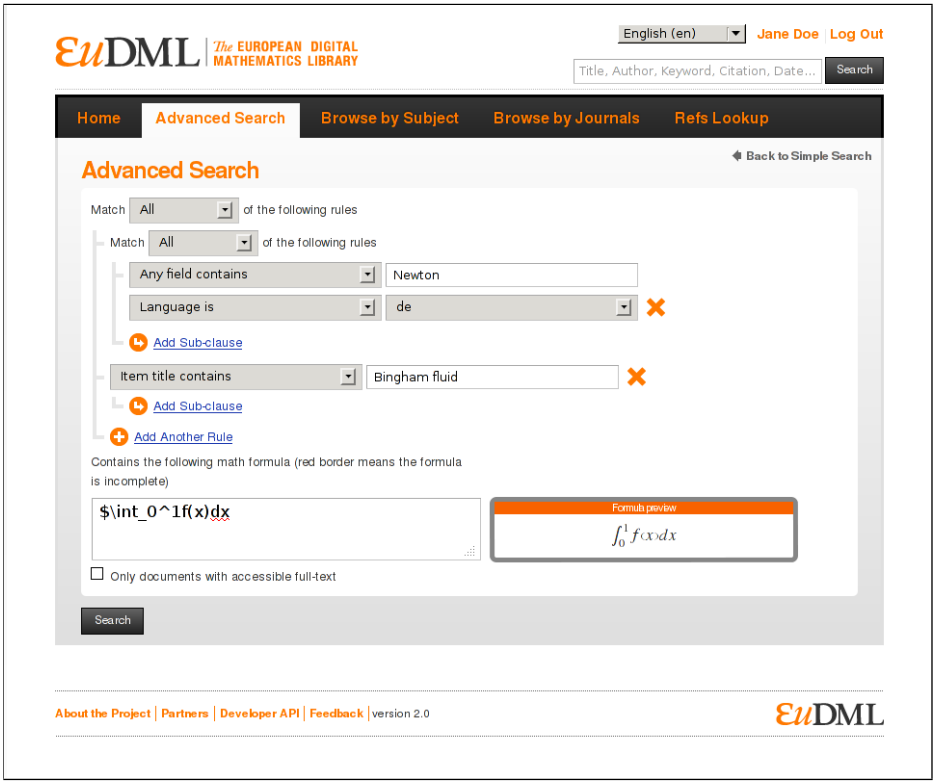


FIGURE 2. Advanced search

One of the main menu item is also unique to EuDML: The Reference Lookup page allows to find a document from a reference citation string, typically copy-pasted from an actual bibliography.

3.2. **Content policy.** EuDML aims to be a long-standing, reliable and open source of trusted mathematical knowledge. This implies to build on firm policy. An outcome of the project is the consensus over the following three conditions for some content to be eligible in EuDML collections.

CP1. *The texts in EuDML must have been scientifically validated and formally published.*

This is needed to ensure that EuDML works as an authoritative library, holding the version of a piece of mathematical knowledge that can be further built upon, and permanently referred to.

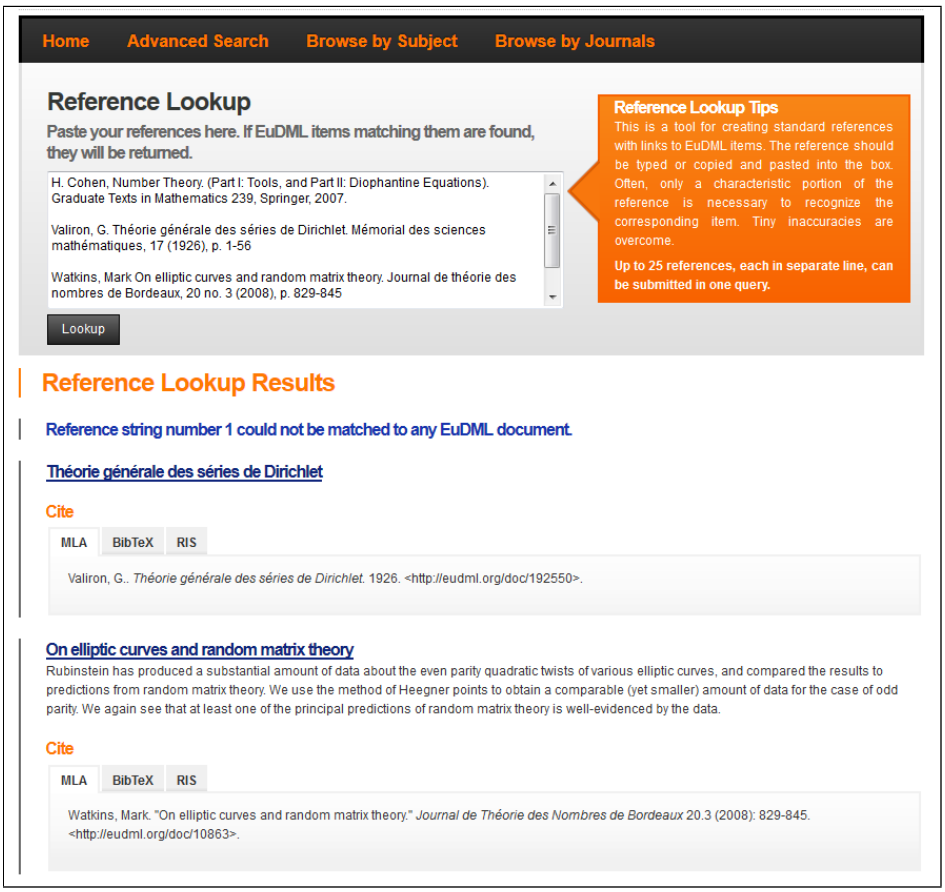


FIGURE 3. The reference lookup.

CP2. *EuDML items must be open access after a finite embargo period. Once documents contributed to the library are made open access due to this policy, they cannot revert to close access later on.*

This is the so-called "moving wall policy" as in general the published items become freed from a pay wall after a certain embargo period (typically ranging from 0, aka open access publishing, to less than 10 years). This eventual open access policy tries to accommodate the fact that not all mathematical publishers can afford to publish everything as open access immediately, but that the value of mathematical knowledge is to

FIGURE 4. A typical item's landing page.

foster new developments in any fields and at any time after publication, so that this should become public knowledge after a not-too-long while (much shorter than current copyright duration, indeed). This policy is strongly supported by the International Mathematical Union as part of CEIC's best practices [6].

CP3. *The digital full-text of each item contributed to EuDML must be archived physically at one of the EuDML member institutions.*

This is for the sake of preserving the mathematical corpus as an enduring collection, which in turn is the only way to secure its online availability over the very long term.

We noticed during this project that these rules are strong and will limit our ability to reach an exhaustive mathematical corpus (the Elsevier archived 'primary mathematical journals' [8] that have been recently released as open access would comply with the first two of them, for instance, while not all project Euclid journals would comply with the second one). However we felt that these rules ensure that the system we built is on a sound base, and that what has been achieved so far cannot be reverted by some external fortune.

Our Scientific Advisory Board commented on these policies at the very end of the project and suggested that we relaxed them somehow, in order to maximize the eligible content. Tweaking these policies so that EuDML is as inclusive as possible but not just a loose index of untrusted mathematical papers on the Web was a challenge of this project. We thought that we should first have a strong perpetual content base before trying to accommodate with looser scenarios.

3.3. **External Cooperation Model.** Based on the above content policies, we drew a model of EuDML operation that will inform the design of the EuDML Initiative. It is based on a consortium of EuDML core members being scientifically and organizationally strong not-for-profit institutions that take care of the system's activity, maintenance, and of the collections both in terms of preservation and eventual open access provision. This gives rise to a network structure relying on a core set of internal partners providing content and technology. A second tier is foreseen to allow participation of associated partners that, for instance, use some services from a sponsoring first tier partner to access the network. A typical scenario for this is that of a content partner willing to contribute collections but without the skills or resources to comply with the interoperability requirements, thus engaging with one of the core partners that would serve as entry point for them. This structure is already active for some of the project's content. A third tier would consist of external content partners (typically publishers) that

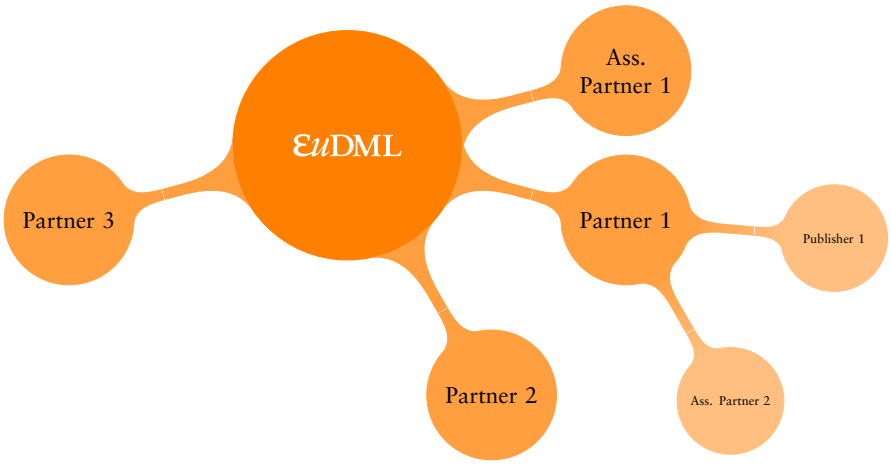


FIGURE 5. The EuDML network

contribute, possibly directly, metadata to the project, but need to transfer their collections to one of the core members in order to comply with our content policy CP3.

The EuDML content members should

- be aligned with the project's goals,
- keep committed over the long term,
- select collections to be contributed to EuDML on sound scientific grounds,
- develop a preservation policy for the full-texts,
- acquire new items in a timely manner (retrodigitisation or direct from publishers),
- sort out rights and licences of contributed collections,
- take care of data and metadata curation,
- manage communication with the central registry.

The EuDML technical members should

- be aligned with the project's goals,
- keep committed over the long term,
- manage communication with the content members,
- run and maintain parts of the system's infrastructure,
- develop new services as the need emerges and to the extent their resources permit.

External partners are expected to contribute to the EuDML Initiative using our interoperability model. We identified the following typical scenarios.
External content partners should

- adhere to the project's goals,
- select one content member (aka local DML center: LDC) as entry point to EuDML,
- set up transfer and update mechanisms for new items,
- determine the moving walls' durations,
- license at least one LDC to store transferred files for ever.

External technical partners should

- adhere to the project's goals,
- sign non-disclosure agreement of data they could get hold of for their technical work,
- develop technology over subsets of the corpus and make it available to the project,
- provide technology to the project preferably under open source licenses.

External linking partners should exploit the linking opportunities delivered by the project to enrich content and user experience while searching, browsing, or accessing the reference mathematical corpus.
The Scientific Advisory Board, in line with its previous comment on policies, advocated for a "second level partnership" with relaxed implications. By publishing these results, we hope to get more feedback from the community on the operation model we invented.

3.4. **Interoperability Model.** In order to enable many interoperability scenarios, a number of tools have been developed and deployed. The goals pursued are, on one side to make it easy to contribute new content to the EuDML system, and on the other side to offer many useful ways of exploiting the EuDML content, or creating specific views for different communities.

Contributing content to EuDML. The preferred mechanism to contribute content to EuDML is to set-up an OAI-PMH server to export XML metadata structured according to the EuDML schema version 2.0, providing the mandatory elements and tagged according to the best practices that are specified on its website [13]. These specifications have been designed so as to impose minimal technological barrier to content providers yet to enable the transfer of highly detailed and accurate metadata. Many publishers already export JATS files to interoperate with services such as Portico, JSTOR, PubMed Central, etc. To help content providers

tweak their EuDML metadata, we provide them with an online validation tool [14], which is also applied in the ingestion workflow. This model is the preferred one as it requires almost no work on EuDML side to ingest or update new content, thus will be available after the project funding expires.

For those content providers who are not able, or not willing, to export metadata prepared according to our recommendations, we developed a number of transformations from various flavors of OAI-DC, which are performed on-the-fly at ingestion time.

For those content providers who cannot set-up an OAI-PMH server delivering the expected metadata (missing mandatory elements, e.g.) but do have it in some supported format, it is in many cases possible to harvest files through FTP, then run on-the-fly transformations, so that the ingestion process looks transparent to the central system.

Finally, we have started to build the second tier of the EuDML network, where a EuDML partner "sponsors" an associated partner by getting hold of its relevant metadata, doing the necessary transformations, and posting them to EuDML from its OAI-PMH server. Mathdoc had this role for the collections from GDZ, DML-E for instance.

After the initial pilot period of EuDML, it is envisioned that publishers should contribute to EuDML in this way: selecting a EuDML member that would host a copy of their content, and make it available to EuDML (in fact, this is the scheme already in use for most of the digitised collections).

Our impression at the end of the project is that this model works pretty well. Some "second tier" partners at the beginning of the project (DML-CZ, e.g.) are now "first tier". It is however an open problem to tell to what extent this model can scale from the current 200+ thousand items to the 1.5 million items in the scope of DML estimated to exist currently in digital form, or even to the 3.5 million mathematical items published worldwide since Euclid. Probably the main barrier here is not technical, it was already addressed in previous sections. But there are lots of small collections out there that would be eligible to EuDML but wouldn't enter into one of the above tiers. Examples of such collections are numerous. Typically this can be a large digital library holding a very tiny portion of mathematical content for which no dedicated work or resources can be allocated, or a very amateur digital library set up by a small group of unskilled people, the extreme version of this being an author's own works digitised or collected on his own web page (the IMU called all mathematicians worldwide to do so). In fact these collections would require some pro-active action from a EuDML partner to be exploitable by EuDML. It is not obvious to tell what

portion of the content that could be available that way would ever reach EuDML through another path, but it is clear that breaking this barrier would enlarge considerably the content. This challenge was not addressed in this project but should be investigated later. As always, the low hanging fruits were caught first, and resources needed grow exponentially with height!

External interoperability devices. While a smooth ingestion procedure is the guarantee for EuDML to register an up-to-date critical mass of quality metadata, fuelling powerful discovery services and a rich user interface, external interoperability is needed to allow third parties to enrich their services thanks to the availability of collections in the EuDML system. This in turn provides more visibility to and more usage of the EuDML collections.
We developed specific tools for targeted scenarios of machine interaction with the EuDML corpus.

(1) Batch download of public elements in descriptive metadata is available through the EuDML public OAI-PMH server [10]. In order to maximize interoperability, three formats are supported: basic OAI-DC, Europeana semantic elements [16], and EuDML schema [13]. Apart from some sensitive data that can have been contributed to EuDML under the condition that it is not re-served (author's email addresses, copyrighted full-texts, e.g.), all information that EuDML harvested or created (EuDML Ids, links to other databases, e.g.) are exported under the last format. It is thus also a way for EuDML content providers to get back the project's added value for their own sake.
(2) Machine query the EuDML database with Opensearch [11] using Contextual Query Language syntax. This would allow a third party to automatically query the EuDML database and present EuDML hits together with other sources, for instance.
(3) Machine calls to some EuDML functions through REST services [12]. These services have been tailored for various needs, and should probably evolve depending on feedback or as new needs emerge.
    (a) The Batch Ref service allows an external party to upload a reference list with citations of mathematical documents, and get back the identifiers of matched EuDML items. This is a critical added-value for a reference library as this allows many stakeholders dealing with mathematical references to enhance their assets by adding links to the full-texts.

(b) The Reverse Ref service makes it possible to find all EuDML items citing a given EuDML item. This service was an explicit request of a putative content provider in order to get an added value from participating in EuDML, as it would generate more valued links to their assets.
(c) The Similar Items service makes it possible to use the EuDML website's "Find similar documents" feature from a distance.
(d) The Batch Ids service allows third parties knowing one Id for a given item to query the EuDML databases for all Ids pertaining to this item known to the database. It turns EuDML into a mathematical hub connecting relevant databases. Together with the All Pointers service, it opens new pathways in the mathematical corpus.
(e) Finally, the Metadata via REST service makes it possible to download an item's internal metadata in two XML formats.

(4) Embed some EuDML data or query form as a widget in a Web page. For instance, users can monitor their EuDML activity or add some dynamic view on EuDML in their Web site.

These tools open a wide range of possible applications, from adding the EuDML corpus to an external search engine to enriching existing content with deep links to EuDML.
Producing Linked Open Data and creating a SPARQL end point was considered during the project, but the technology didn't seem mature enough for a production system, real-world application still lacking to exhibit a clear benefit within the short time frame for development. We also have in principle the possibility to set-up a full-text hub as the central system does store quite a lot of full-texts from EuDML content providers, in quite many formats (original PDF, extracted text with or without math as MathML or LaTeX, accessible formats) and we also have licence declarations from the content providers whether these texts can be used internally only for indexing, or can be re-served openly. However, these services were not developed in this project.

3.5. **Technical results.**

3.5.1. *Metadata.* One of the most basic yet non-trivial challenge in the project was to agree on a common metadata format, as each partner had its own, and stood with quite varying background, technical as well as in terms of the community they belonged to.

After a rather involved discussion, we adopted the NLM Journal Archiving and Interchange Tag Suite as the basis for EuDML metadata storage and exchange, which became an NISO standard during the course of the project [20].
To handle the extra content (monographs, edited books or proceedings and their chapters, multiple-volume works and their volumes), we created a new XML schema that defines a specific superstructure and relies on the standard article elements for all shared concepts.
This metadata format supports all EuDML item types so far, and still leaves room for storing improvements such as structured XML full-text, or multiple versions of the same citation. As we store the best available metadata, it is easy to generate simpler schemas such as OAI-DC or Europeana semantic elements.
The metadata is harvested and mostly transformed on-the-fly to JATS by the REPOX harvest manager developed by our partner in Lisbon, mostly in connection with the digital programme of the Portuguese National Library and Europeana.

As the project was also an occasion to clear licenses and copyright for the contributed content, we can report the following.

- We estimate that 97% of full-texts as PDF files are openly accessible from their providers while only 10% are old enough to be public domain.
- The metadata as available from EuDML OAI-PMH server is entirely freely reusable according to either CC0 (public domain) or CC-BY (attribution) Creative Commons licences.
- For full- texts, the situation is somewhat more complex:
  - 135,000 items have some sort of text-only full-text that is usable for indexing purposes, coming from text OCR or PDF extraction,
  - 170,000 items are available for project internal processing such as re-OCR to get math formulae or as test-bed for whatever enhancement a partner could try (most of them are scanned PDFs, but some are born digital),
  - the PDFs of 105,000 items could be re-served after some processing such as adding text or math layers to an image PDF. However, only 10,000 files have been processed with Maxtract [1] and are currently served in some new format generated for print-disabled users.

3.5.2. *Productivity tools.* A number of productivity tools were produced in the course of the project. They are usually Open Source software or libraries. We provide some live demos on the project's Web site [15].
Here is a list of services running in the background or enabling some of the Web site features.

- Metadata enhancements (automated tagging refinement such as author names or keywords splitting);
- On-the-fly conversion from TEX encoding of formulae to MathML (based on Tralics [19]);
- EuDML reference matching, zbMath matching;
- Item metadata merging: we had some 2,000 items duplicated from different partners: we created a single record for them.
- Public demo website with presentation MathML based display of formulae (using MathJax [5] as a fall-back)
- Experimental formula search (based on Brno's WebMIaS [25])
- Experimental similarity computation (based on Brno's Gensim [23])
- Experimental production of accessible formats of mathematical texts (based on Birmingham's Maxtract [1])
- Web 2.0 features and annotation module
- Service interfaces (Opensearch, OAI-PMH, REST API)
- More mathematical knowledge generated and stored in XML records through
  - MSC and English keywords acquired from zbMath
  - Text+MathML extraction from born digital PDF (using Birmingham's Maxtract [1])
  - Text+MathML extraction from image PDF (using InftyProject's InftyReader [26])

All these bits and pieces were integrated and made to work together by the team at ICM Warsaw where the central system is running.

4. Open questions and future work

4.1. **Content acquisition.** The main point to users is the content: it's nonsense to learn the interface of one more search engine if it covers less than 10% of the whole corpus. On the other hand, there are many mathematical texts that do exist digitally, and are freely accessible on-line, but can't be located easily from mainstream search engines or even dedicated reviewing databases. A large part of the retro-digitised corpus is hidden because it lacks full-text and can only be searched using scarce metadata (or almost non-existent metadata, as for PDFs linked from hand-made HTML pages). The fact that users start using EuDML to locate and refer to papers from DML-PL or GDZ shows that we shifted the state-of-the-art in this respect. However, to succeed, we need to cover much more of the mathematical corpus.

An issue is that the EuDML cooperation and interoperability model *can* scale, but *not that much*. There are around the world a number of institutions that act as local DMLs and would be happy to join a global DML project in the lines of EuDML. However, the well-managed long-standing and scientifically reliable digital libraries that we expect to partner with are not many, and do not host more than a third of the whole mathematical corpus. This is already a lot, and it would certainly be very useful to reach this number, but it is still far from the grand vision of an exhaustive library, or even of a one-stop shop to the established treasures of the mathematical literature.

From the experience of the contacts we had during this project, we can tell that the math publishing landscape is quite diverse, with quite diverging stakes, and it seems difficult to accommodate with all views and policies. Especially, there is a strong attraction of concentration in the digital realm, which opposes to the creation of an alternative archive for universal open access. Moreover, we have to convince publishers with very different profiles that cooperation is fruitful to them (although the same kind of cooperation certainly doesn't mean the same to very different kinds of publishers).

This said, it would probably be possible to go from 30% to 60% by allowing much more volatile content to be found at places such as open archives, personal home pages, community repositories. Methods and technology to index and keep track of that kind of Web content (and hopefully, to keep track of its scientific validation status as well) are very different from what was experimented in this project. It would rather rely on Web crawling and could be tuned by publishing simple standards to push keywords or hints that the posted content is part of this loose version of the DML. This could in fact be tested as a separate endeavour, and if it succeeds, could become one DML content provider. Just for the case of France I tried to list the various sources of DML-related material in [3], which is a long list, although the material that is likely to end up in a system like EuDML is limited to that from NUMDAM [21], Gallica-Math [18] (which is already there), TEL [27], and Gallica books [17].

Another mean to enlarge DML content is digitisation. Public domain (or decades old, out of print) books have probably the biggest potential as a still relevant huge mathematical knowledge reservoir.

4.2. **Technical challenges.** We still lack a full digitisation work-flow starting from paper (or some flavour of low-grade digital file) and outputting usable structured metadata and full-text enabling mathematical knowledge mining, interlinking, etc. Even routine operations such as starting from a scanned journal volume

and ending up with a database of articles with reasonably accurate metadata and extracted full-text are not available right away without a lot of software development or human input.
We still face a fine-grain metadata shortage that won't be overcome through manual work because of the volume to handle. We should engage in all strategies to incrementally enhance metadata or metadata-generating technology. Of course a lot of this is not math-specific but we have to be able to capture and store non-textual metadata everywhere.

Conclusion

The EuDML project has assembled a corpus of 225,000 mathematical documents, which are now much easier to find and navigate, and much better integrated within the Web. As their metadata were converted and stored in a homogeneous format, it will be also much easier to deal with that content in the future, and link it with existing or future infrastructures relevant for the retrieval of scientific reference literature.
A number of tools were developed, deployed, and partially evaluated (from basic aggregation to accessible math through math-mining and deep interlinking).
Finally, the experience of the first cross-repository, trans-national DML effort shows that it can be done, and it can go on with new partners and the support of the scientific community.

Nevertheless, this is only a promising début A number of threats counter the objective of archiving the whole mathematical corpus in a modernized free public digital library. Going further will need a lot of energy and resources from all over the world. The question whether the needed resources to achieve this goal will be allocated is open.

References

[1] Joseph Baker et al., *Maxtract, a tool for converting PDF into formats such as LaTeX, MathML and text*, `http://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/maxtract.php`.

[2] Thierry Bouche, *Some thoughts on the near-future digital mathematics library*, Towards Digital Mathematics Library. DML 2008 workshop, Birmingham, UK, July 27th 2008 (Brno) (Petr Sojka, ed.), Masaryk University, 2008, `http://eudml.org/doc/221606`, pp. 3–15.

[3] Thierry Bouche, *Report on the current state of the french dmls*, Towards a Digital Mathematics Library. Grand Bend, Ontario, Canada, July 8-9th, 2009, Masaryk University Press, 2009, http://eudml.org/doc/220779, pp. 61–70.
[4] Thierry Bouche, *Digital Mathematics Libraries: The Good, the Bad, the Ugly*, Mathematics in Computer Science **3** (2010), no. 3, 227–241, Special issue on Authoring, Digitalization and Management of Mathematical Knowledge (Serge Autexier, Petr Sojka, and Masakazu Suzuki eds.).
[5] Davide Cervone et al., *MathJax, Beautiful math in all browsers*, http://www.mathjax.org/.
[6] Committee on Electronic Information Communication of the International Mathematical Union, *Best Current Practices: Recommendations on Electronic Information Communication*, Notices of the AMS **49** (2002), no. 8, 922–925, http://www.ams.org/notices/200208/comm-practices.pdf.
[7] ———, *Digital Mathematics Library: A Vision for the Future*, http://www.mathunion.org/ceic/recommendations-publications/, 08 2006.
[8] *Free access to archived articles of primary mathematics journals*, http://www.elsevier.com/physical-sciences/mathematics/archived-articles.
[9] *EuDML: The European Digital Mathematics Library*, http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503.
[10] *EuDML oai-pmh server*, 2013, https://eudml.org/oai/OAIHandler?verb=Identify.
[11] *EuDML Opensearch*, 2013, https://project.eudml.org/opensearch.
[12] *EuDML REST Services*, https://project.eudml.org/rest-services.
[13] *EuDML metadata schema specification (v2.0 – final)*, 2013, https://project.eudml.org/eudml-metadata-schema-specification-v20-final.
[14] *EuDML metadata schema validation tool*, 2013, http://eudml.mathdoc.fr/eudml-validation-demo/.
[15] *EuDML tools & demonstrations*, 2013, https://project.eudml.org/tools-technical-specifications.
[16] *Europeana Technical Requirements: Europeana Semantic Elements specifications*, 2012, http://pro.europeana.eu/technical-requirements.
[17] *Gallica, the digital library of the Bibliothèque nationale de France*, http://gallica.bnf.fr/.
[18] *Gallica-Math, a front-end to some mathematical content from Gallica*, http://portail.mathdoc.fr/GALLICA/.
[19] José Grimm, *Tralics: a LaTeX to XML translator*, http://www-sop.inria.fr/marelle/tralics/.
[20] U.S. National Library of Medicine National Center for Biotechnology Information, *Journal archiving and interchange tag library, niso jats version 1.0*, August 2012, Full online documentation at http://jats.nlm.nih.gov/1.0/.
[21] *NUMDAM, Numérisation de documents anciens mathématiques*, http://www.numdam.org/?lang=en.
[22] Andrew Odlyzko, *Open access, library and publisher competition, and the evolution of the general commerce*, http://arxiv.org/abs/1302.1105.
[23] Radim Řehůřek and Petr Sojka, *Gensim, Semantic Similarity of Text Documents*, https://mir.fi.muni.cz/gensim/index.html.

[24] David Ruddy, *The evolving digital mathematics network*, Towards a Digital Mathematics Library. Grand Bend, Ontario, Canada, July 8-9th, 2009, Masaryk University Press, 2009, http://eudml.org/doc/220643, pp. 3–16 (eng).
[25] Petr Sojka and Martin Líška, *WebMIaS, Math Indexer and Searcher*, https://mir.fi.muni.cz/mias/.
[26] Masakazu Suzuki et al., *InftyReader, a software to recognize scientific documents including mathematical expressions*, http://www.inftyproject.org/en/index.html.
[27] *TEL, Thèses en ligne*, http://tel.archives-ouvertes.fr/.

Received February 23, 2013

Cellule Mathdoc (UMS 5638), Université Joseph-Fourier (Grenoble 1), B.P. 74, 38402 Saint-Martin d'Hères
E-mail address: thierry.bouche@ujf-grenoble.fr

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

OPENMATHMAP: ACCESSING MATH VIA INTERACTIVE MAPS

JAN WILKEN DÖRRIE, MICHAEL KOHLHASE, LARS LINSEN

Abstract. World Math literature is growing at an alarming rate (3.3M journal articles today increasing by 120k a year). While much of that can be retrieved online, we lack technologies to navigate and understand the space of math literature. The OpenMathMap project develops and deploys novel interfaces that empower interested parties to find their way. We conjecture that such maps can act as cognitively adequate access mechanisms to many large-coverage MKM systems.

The first concrete interface is an interactive map generated from publication data. We have developed a prototype map generation service based on MSC classifications and deployed the maps resulting from ZBMath Data in OpenStreetMap. Interaction mechanisms allow for providing details on demand.

1. Introduction

In the information age fuelled by the Internet, the problem of information and knowledge foraging changed from one of retrieving documents to finding out about them. In particular, navigating the space of available documents efficiently becomes an important subtask. This is especially pertinent for the scientific literature, which largely consists of original publications and therefore has very little redundancy. The times where single individuals could have an overview over all of science are long past: Even in the Renaissance polymaths like Leonardo de Vinci were considered a rare exception.

The scientific community has developed various tools to work around this problem: encyclopaedias, survey articles, classification systems, and review services. But with the proliferation of scientific publication – 50 million articles in

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

2010 [Jin10] with a doubling time of 8-15 years [LI10] these tools start collapsing under the sheer mass of information. Internet-age tools like search engines, bibsonomies, and citation databases solve (part of) the information retrieval and navigation problems by providing word-based search and browsing along citations. But these tools are "myopic" in the sense that they only give very local view of the immediate surroundings of a word or document.

Classification systems like the Math Subject Classification (MSC, see [Msc]), the Physics and Astronomy Classification Scheme (PACS, see [PACS10]), or the ACM Computing Classification System (CCS [ACM-CCS98]) take a more global stance, but they lack user interfaces that give information foragers an intuitive sense of direction and locality that is so helpful to humans in navigation tasks.

In the MathSearch Project [MathSearch] we are currently rethinking access to mathematical knowledge and resources. As a first experiment, we are building a global, map-based navigation service for mathematics. The main idea is that humans are very skilled in spatial navigation and in particular have learned to use map representation to navigate spaces and locate targets. Concretely, we want to create a map of mathematics akin the one in Figure 1 used to visualize usage patterns of online communities. We want to base the map on ideas from Rusin's Math Atlas [MathAtlas] (created 1998, last updated 2001, see also Figure 2), which uses topics from the Math Subject Classification for map regions and calculates the positioning and relative sizes from topic interconnections and the numbers of publications. Based

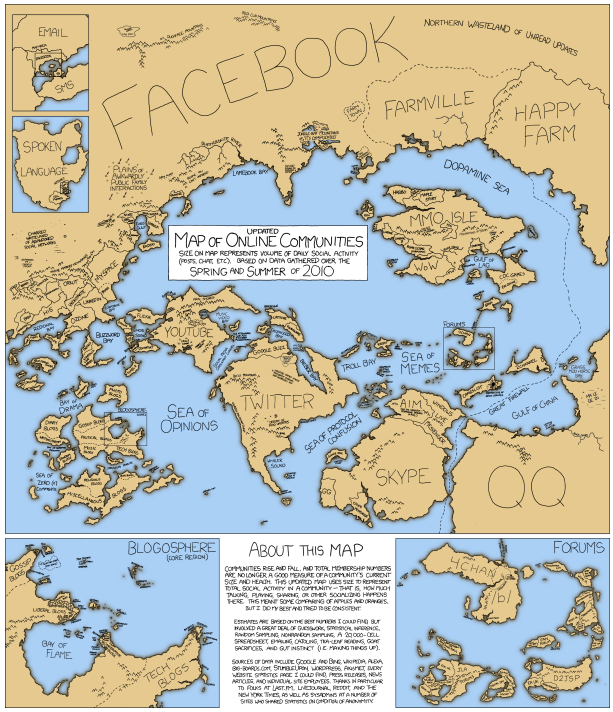


FIGURE 1. Map of Online Communities, XKCD 2010 `http://xkcd.com/802/`

on the visual encoding through the geographical map metaphor, we built an interactive tool for data exploration following the concept 'overview first, detail on demand'. We give examples on how our system can be used to explore the given data.

Section 2 discusses the publication data used as a foundation and Section 3 the process by which map images can be computed from it. In Section 4 we discuss how such images can be deployed in a map server (here OpenStreetMap). Section 5 discusses topical maps which can be realized with OpenStreetMap. Related work is discussed in Section 6. We show in Section 7 how to integrate information services. Section 8 concludes the paper and discusses future work. Additional details in technical matters can be found in [Doe13].

ACKNOWLEDGEMENTS Work on the concepts presented here has been partially supported by the Leibniz association under grant SAW-2012-FIZ_KA-2. The authors are indebted to Wolfram Sperber for the Zentralblatt Math publication data and Patrick Ion for initial discussions.

2. ZBMATH/MSC DATA

In the creation of the map we made use of the 2010 Mathematics Subject Classification jointly developed by the American Mathematical Society and Zentralblatt Math. The result are 63 top-level classes, 528 second-level classes and 5607 third-level classes summing up to 6198 classes in total. Each class has an identifier of the form DDLDD (D for Digit, L for Letter) and a title. For example "68R05 Combinatorics" is the third-level, "68R Discrete mathematics in relation to computer science" is the second-level, and "68 Computer Science" is the top-level category.

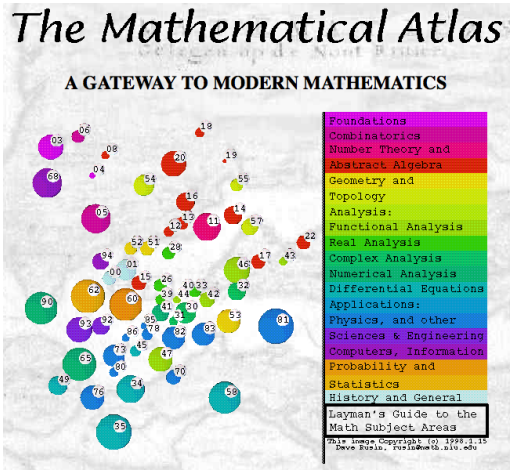


FIGURE 2. Rusin's Math Atlas [MathAtlas]

Zentralblatt MATH provided us with the metadata for more than 3 million articles in mathematics, in particular article number, the authors, the classification codes[1], keywords, and publication year. As we will see

[1]On average, each classified article carries 3 codes.

below, the classification codes allow us to create the map geometry by visualizing proximity and size of MSCs, the publication year gives us the possibility to create maps which only include publications from a given time span.

3. Map Geometry

The first step in map creation is to compute a geometrical representation of the relations in the ZBMath publication data, which will use as the base of the map. In this representation we want to adequately represent the relative sizes and proximities of the MSC classes. The size of a class MSC_c is given by the number $|\mathrm{MSC}_c|$ of papers[2] and define the similarity of two classes as

$$s(i,j) = \frac{|\mathrm{MSC}_i \cap \mathrm{MSC}_j|}{|\mathrm{MSC}_i \cup \mathrm{MSC}_j|}$$

The size of the intersection/union of two MSCs is defined as the number of papers referencing both/any MSCs. This results in $s(i,j) = s(j,i) \in [0,1]$ and $s(i,i) = 1$ for all MSC classes i and j.

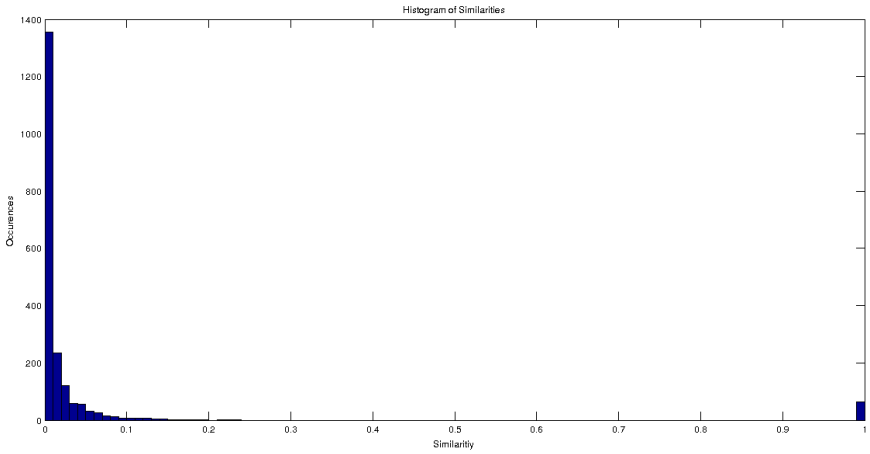


Figure 3. Histogram of the similarity values

[2]In principle, any scalar value that represents a property of each MSC class can be mapped to the size of the respective area in the map. Currently, we just use the number of papers (due to data availability). In future work, we would like to add further functionality that allows the user to interactively choose which measure should be mapped to size.

Figure 3 shows a histogram of all similarity values[3]. This diagram reveals that there very few classes are highly similar, or in other words, the space of MSC classes has a graph-like structure under this similarity measure; this validates our initial assumption that maps might be suitable for navigation of the MSC space, even though they can only use two dimensions to render an a-priori 6198-dimensional structure.

For the initial version of the map geometry (see Figure 4), we calculate the similarity between every pair of top-level MSCs to obtain a similarity matrix of size 63×63. To transform the matrix into two-dimensional coordinates for each MSC we apply Multidimensional Scaling (MDS; [KW78]).

3.1. **Multidimensional Scaling.** Multidimensional Scaling (MDS) is a technique applied to a set of points in n-dimensional space to visualize it in a low dimensional space preserving distances between points as well as possible. Common values for the output dimension are 2 and 3. Formally this can be described as the following: Given a distance function δ, target dimension p and an $n \times n$ dissimilarity matrix

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,1} & \cdots & \delta_{1,n} \\ \delta_{2,1} & \delta_{1,1} & \cdots & \delta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n,1} & \delta_{n,2} & \cdots & \delta_{n,n} \end{pmatrix} \tag{1}$$

one needs to find n vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ such that $\|x_i - x_j\| \approx \delta_{i,j}$ for all $i, j = 1, \ldots, n$. A solution may be found by numerical optimizations, for example Matlab has a built-in method `mdscale` which takes as arguments the $n \times n$ dissimilarity matrix D and the target dimension p. It returns Y, a configuration of n points in p dimensions, and optionally a stress value, which by default is the stress value given by Kruskal's normalized Stress-1 criterion [mdscale13] σ_1 with

$$\sigma_1 = \sqrt{\frac{\sum \left[f(p_{ij}) - d_{ij}(\mathbf{X})\right]^2}{\sum d_{ij}^2(\mathbf{X})}} \tag{2}$$

where p_{ij} is the *proximity* between two given objects, $d_{ij}(\mathbf{X})$ is the Euclidean distance between two points in the current configuration $\mathbf{X}$ and $f : p_{ij} \to d_{ij}(\mathbf{X})$ is a *representation function* where the particular choice of f specifies the *MDS model*; see [MDS05]. Choosing an appropriate f allows Matlab to work with both similarity and dissimilarity matrices (the exact choice for f is hidden in the implementation of the algorithm).

[3]As similarity matrix is symmetric only the upper triangular part was considered

For the Stress-1 coefficient σ_1 Kruskal suggests the following benchmarks: 0.20 = poor, 0.10 = fair, 0.05 = good, 0.025 = excellent, and 0.00 = perfect [MDS05].

3.2. **Radial Basis Functions.** The configuration generated by MDS is a collection of points in 2D space respecting the relative distances between given points. However for the visualization one also needs to take the size of each MSC into account.

To give a dimension to the points we use radial basis functions (RBF), i.e real-valued functions whose value depends only on the distance from the origin or alternatively on the distance from some other point c, called a *center* [Wik12].

A common choice for RBF is a Gaussian model, but has the drawback of an infinite support. Therefore, we decided to use a modified cosine function; concretely

$$f(d) = \begin{cases} 1 + \cos(d/k) & \text{if } |d| \leq k\pi \\ 0 & \text{else} \end{cases} \tag{3}$$

where d denotes the distance of a given point to the origin of a given MSC and k depends on the size of this MSC.

It is zero outside the given range d, which allows us to discard all points with $|d| > k\pi$. In our experiments, we have seen that this finite support of our RBF choice speeds up the calculations by a large factor when compared to a Gaussian RBF.

3.3. **Borders and Coast Lines.** To detect borders mechanically we rasterize the map into a grid of cells and calculate the influence of the nearby MSCs on each cell via the RBF specified in Section 3.2. Then we associate each cell with the MSC with the highest influence. If there is no influence (i.e. the result of the RBF was zero for every MSC) the cell is considered to be water.

The actual border detection considers each non-water cell and checks its 4-neighborhood for changes in the associated MSC. If there are changes, the cell added to the border sets of both MSCs. Note that border set detection includes coastline detection, since water is considered as the empty MSC. Finally borders are connected by considering each border set on its own and finding the shortest path going through all contained items. This task was executed using the nearest neighbor heuristic. It was able to produce nearly optimal results as the next cell can always be found in the 8-neighborhood of the current cell.

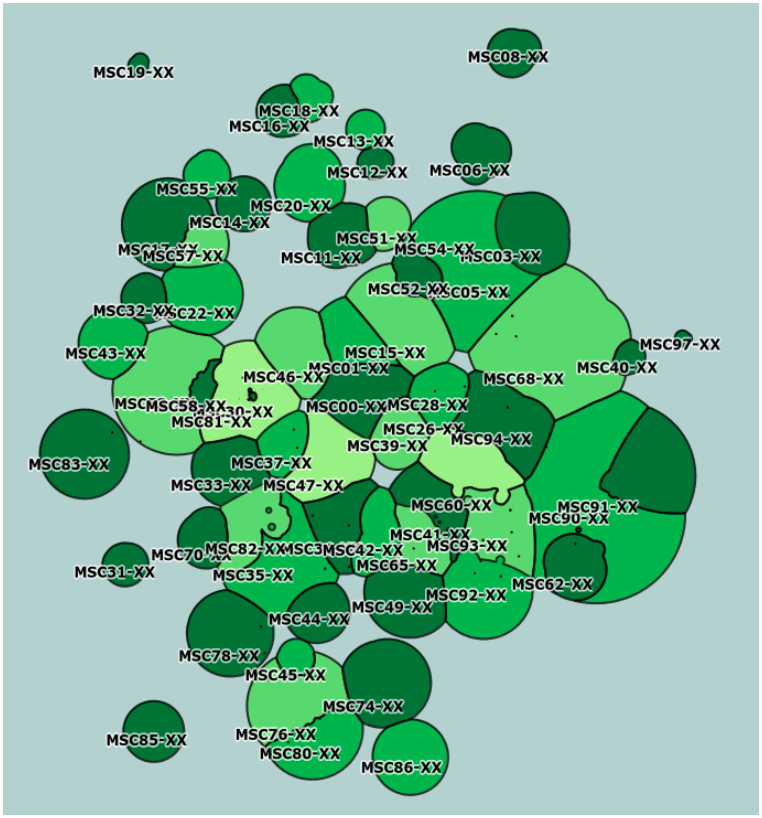


FIGURE 4. Geometry of the Math Subject Classifications

3.4. **Hierarchical Map Geometry.** As the MDS computation becomes intractable for larger similarity matrices[4] we opt for a hierarchical approach to determining finer-grained map geometries (taking second-level and leaf MSC classes into account). In order to incorporate the second-level classes into the map we consider each top-level class independently, giving up on information of surrounding MSCs to make calculations tractable. For second-level classes belonging to the same first-level class we create inner-class similarity matrices. This leads to 63 different matrices with sizes between 1×1 and 20×20. Then the same operations as before were applied.

[4]MDS is super-linear; for the first-level 63×63 matrix computation takes seconds and exceeds 30 hours even for the 528×528 matrix of second-level MSC classes.

Equipped with a color palette resembling natural maps we were able to create a map seen in Figure 4. This map incorporates the Zentralblatt Math data for the first two levels of the MSC. A close-up of second-level classes can be found in Figure 5, which renders second-level borders in light gray to distinguish them first-level borders.

3.5. **Landmarks and Settlements.** Optionally, we can populate the map geometry with "cities", "towns", and "villages": we view every classified paper as an "inhabitant" and compute the center of gravity of the points representing the respective MSC codes. As the number MSC combinations is finite, this yields a finite number of positions. The positions are rendered as red dots (see Figure 5 on the left), where the size encodes the number of papers.

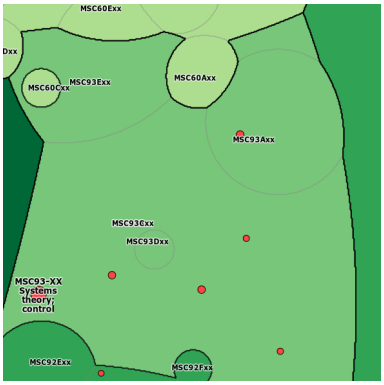


FIGURE 5. Adding Settlements

As this could lead to up to 50 million different cities in theory, we introduced a "population threshold" to avoid a cluttering of the map. Note that papers with only a single MSC classification will end up in the "capitals" of the respective class. Difficulties arise when it is ambiguous in which MSC a given city lies, or if the computed location is inside water. Those cases need special attention and the cities have to be moved manually.

4. Mapmaking & Deployment

The next step is to convert the geometry data from the last section into a map that has the features we are used to from familiar maps ranging from street maps to Google Earth. For that we apply a series of transformation to the geometry, these include smoothing the coast line, adding names for the geographic features, and structuring the uniform blue of the ocean. Now we can deploy the map in a geographical information system. In our case we generate the specific XML representations utilized by OpenStreetMap [OSM]. These consist to points identified by latitude and longitude and ways (point sequences) representing outlines, borders, and contours. Figure 6 shows the result, where we use a "political map" metaphor, coloring the areas by the colors Dave Rusin used in his Math Atlas

(see Figure 2 on the right). These colors are distributed "0$^{\text{th}}$ level of classification" designed by Rusin to highlight super-domains like Analysis (green), Algebra (red), applications (blue), etc.

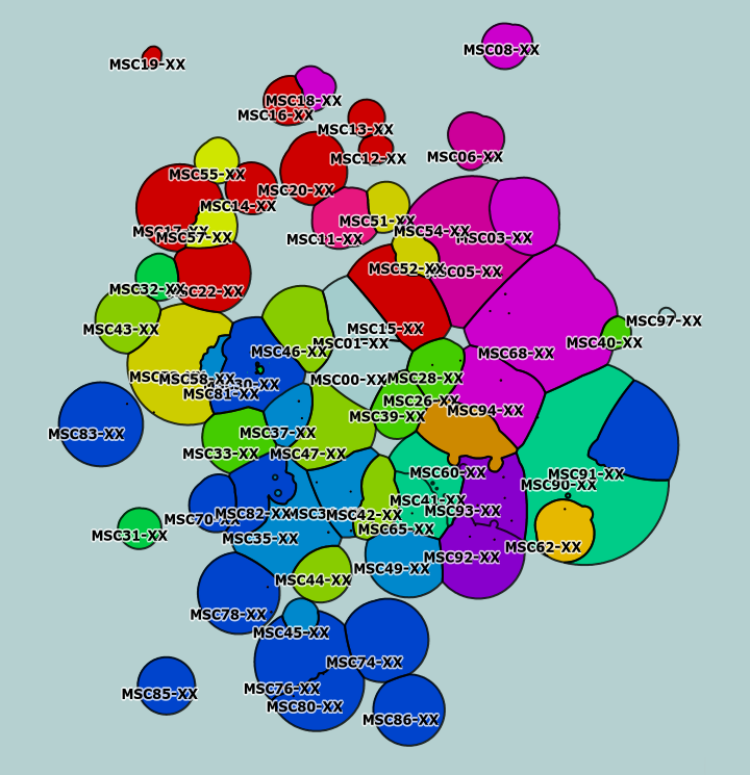


FIGURE 6. A "Political Map" of Maths using Rusin's Color Coding

Concretely, we use the rendering service Maperitive [Map13] – a free desktop application that generates small pieces of the maps called "tiles" – to render maps the XML data. A tile in OpenStreetMap is a 256×256 pixels sized PNG image making up for a given part of the map, examples are given in Figure 7. Tiles need to be generated only once in order to render a map since the contained information is re-used. Additionally tiles allow for zooming, since they can be pre-generated for different resolutions of the map. This enables us to fill the map with many details which can be shown or hidden depending on the current zoom level.

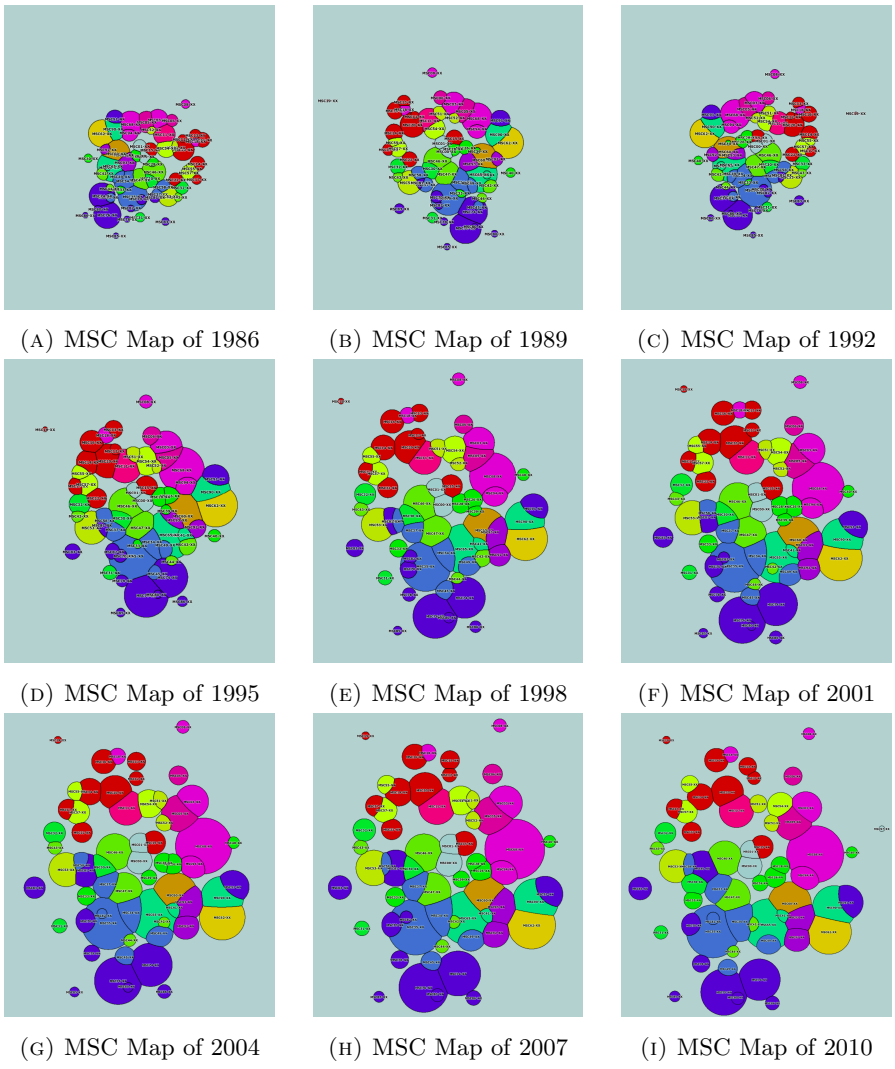
FIGURE 7. OSM tiles at different resolutions; © OpenStreetMap

Note that there is no encoding of the height in Figure 4, this leaves room for visualizing additional information. We are currently experimenting with encoding the "activity level" of an area with this: We can compute the "elevation of an area" by counting the (relative) number of publications in that area e.g. in the last year. This makes research hotspots peaks that can serve as additional landmarks in the map.

5. Topical Maps

Additionally, we can generate topical maps that visualize additional features of the data. For instance, we can apply the mapping procedure to the data up to a given publication date and obtain a series of maps that visualize the "continental drift" of areas in mathematics; see Figure 8. Note that the maps are more "bubbly" than the two-level map in Figure 4, as they only comprise the top-level MSCs. Nevertheless, we can already see interesting trends: The most obvious one is that the field of Mathematics is growing rapidly. Furthermore, some subdomains of mathematics move around, since their distance to other subdomain change; e.g. the cluster of statistics (MSC62-XX; yellowish/ocre), operations research (MSC 90-XX), and economics (MSC91-XX) moved from the north/west in the 1980ies to the current position on the south-east. There is a corresponding movement of Computer Science (MSC 68-XX) which has grown in size with the advent of computational methods in mathematics.

As we have data about the journals papers are published in, we can also compute the centers of gravity of journals (based on those of the papers they publish). Based on a radial function technique similar to the one we used to determine MSC



(A) MSC Map of 1986 (B) MSC Map of 1989 (C) MSC Map of 1992

(D) MSC Map of 1995 (E) MSC Map of 1998 (F) MSC Map of 2001

(G) MSC Map of 2004 (H) MSC Map of 2007 (I) MSC Map of 2010

FIGURE 8. MSC Maps evolving between 1986 and 2010

areas, we can compute "journal dominions" and generate maps colored by these based on the MSC-induced geometry. Such a map could give mathematicians orientation in the jungle of currently ca. 3500 journals, but might be controversial for publishers.

6. Related Work

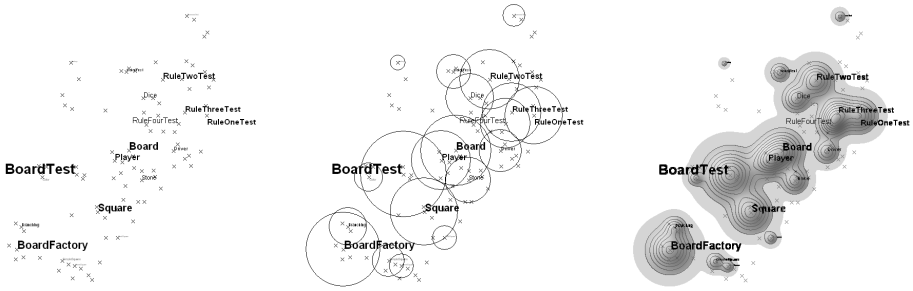
There are other systems that visualize data in the form of generated maps. The first one is Dave Rusin's "Mathematical Atlas" [MathAtlas] which has been an initial starting point of OpenMathMap. Rusin displays the MSCs as bubbles, where the surface areas of the bubbles are proportional to the number of "recent" (at the time of creation) papers in that area [Rus02] (see Figure 2). Just as in OpenMathMap, the placement of the MSCs is determined by the frequency of cross-listings of papers in two or more areas: the distance of MSCs is inversely correlated to this [Rus02]. Rusin only considers the first level MSCs (61 in MSC 2000) and thus performs a 2-dimensional projection of a 61-dimensional space. This was not perfectly possible and so Rusin did some small modifications manually. He moved outliers inside to reduce wasted space, and spread overlapping bubbles apart into the nearest empty space [Rus01].

OpenMathMap extends the Mathematical Atlas in several ways. It is based on more recent data, integrates second-level MSC, generates non-overlapping territories that make MSCs more recognizable, and visualizes the map in the interactive OpenStreetMap system whose zoom and pan features allow to integrate much more information into the map.

Kuhn et al. [KLN12] describe the process of creating a thematic software map. They defined a similarity measure for code of open source programs based on the used "vocabulary". Just like in OpenMathMap they compute a two-dimensional configuration by Multidimensional Scaling (see Figure 9a). Around every point (representing a class) in this configuration they draw a circle where the area was proportional to the lines of code of a given class (see Figure 9b) and use normal distributions to arrive at a height that is visualized with contour lines and shadings (see Figure 9c).

The main difference lies in the choice of radial functions and the lack of borders – admittedly the "political map" metaphor used in OpenMathMap would be less immediate for software classes than for MSCs. We also observe that the "topographic map" metaphor underlying the Kuhn et al.'s maps uses the third dimension (height) that is still free in OpenMathMap.

In [GJ12] Gronemann and Jünger provide a method how to visualize clustered graphs in a topographic map. After constructing a clustered graph they apply fat polygon partitioning and extract a mesh from the layout that models the terrain features based on the clustering. Finally they apply an edge-routing algorithm to the clustered graph. Their visualization of a collaboration graph filled with data from the Graph Drawing E-print Archive [GDEA13] is available at [Gde]



(A) After MDS (B) Adding Circles (C) Visualizing height

FIGURE 9. Steps in creating a thematic software map after [KLN12]

FIGURE 10. Author Search given Publication Data

and includes interactive search interface (see Figure 10). Although their methods differ from the ones used here the concept is very similar in both approaches. We plan to include similar highlighting of collaborating authors in a future version of OpenMathMap.

7. Interactive Services & Mashups

Having our map deployed on OSM already gives us some base-level interactivity: zooming and panning. Additional location-based interactions can be implemented by adding custom JavaScript to the pages served by OSM subject to availability of date. The JavaScript API Leaflet [Lea12] used to render a contiguous map from the generated tiles allows us among others to resolve mouse-clicks on the map to latitude and longitude coordinates. However, just retrieving the location is not of big interest to us, much rather we would like to know which MSC was clicked on.

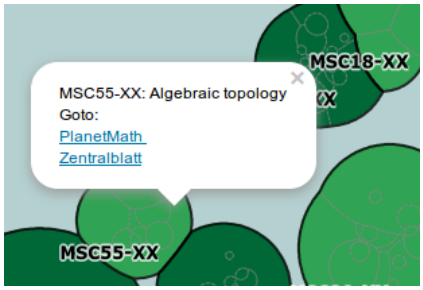


FIGURE 11. Initial User Interaction

To map coordinates to MSCs we make use of a RESTful web-service. It takes latitude and longitude as parameters and returns name and description of the MSC in JSONP format allowing for cross-domain communication. Figure 11 shows an example of this. Mouse-clicks on the map trigger a pop-up providing name and description of the clicked MSC in addition to the respective links to PlanetMath and Zentralblatt MATH.

8. Conclusion & Future Work

We have presented a novel access method to mathematical knowledge and resources that makes use of the highly evolved cognitive skills of spatial representations in humans. We have implemented a first prototype (accessible at `http://map.mathweb.org`) that deploys maps computed from mathematical publication data in a standard map server and instruments it with information services.

The geometry computation part of OpenMathMap system is licensed under GNU General Public License [Fre91] and the web front-end part under the GNU Affero General Public License [Fre99]. The code can be obtained from github at [Ope].

This prototype is just a first step we want to use in experimentation in human-oriented access methods to mathematics. We could imagine that connections between mathematical areas could be implemented as roads, highways or air/sea connections (possibly depending on their salience), important theorems could be entered/visualized as landmarks, and finally, we could imagine to go from interactive map servers to much more immersive environments (from Minecraft to second life).

With the current approach we encountered a few problems when two relatively big areas are very close together in the map. In this case the shared area is split between the two MSCs so that the total area is substantially less than the sum of the areas if there were no bordering MSCs. In order to prevent this problem one could implement a force-based model where we allow large MSCs to "push away" surrounding MSCs. Thus even though we want a small overlap between MSCs to have a contiguous land area this overlap would be minimized. This step would have to be executed after MDS but before the application of the RDF since it has a direct influence on the configuration found by MDS.

The current implementation of MDS achieves a Stress-1 coefficient of 0.25 according to Equation (2). While this implies a poor configuration according to Kruskal (see Section 3.1) one has to keep in mind that reducing a 63-dimensional configuration to a 2-dimensional one is certainly not an easy task. An example where a high similarity did not lead to neighboring MSCs are Algebraic Geometry (MSC14-XX) and Complex Variables (MSC32-XX). The map implies that both classes have a higher similarity with Manifolds and cell complexes (MSC57-XX) which is not true.

Unfortunately, in this case the stress can not be decreased without increasing the target dimension (which would mean we have to drop the map metaphor). However, with the introduction of WebGL enabling 3D rendering in browsers and Google Maps already making use of it (`http://maps.google.com/gl`) one can imagine to also develop a 3D version of the MathMap in the future. With the current similarity values an MDS to three dimensions would decrease the stress value to 0.16. But we can also think of non-geometrical ways of "relieving" stress in the map via other map features. For instance, we could shorten the travel distance between two MSCs that are stressed by being too far apart by adding highways or shipping lines.

Another important area of future work is the exploration of useful interaction. An immediate example is the generation of custom queries for publication databases like Zentralblatt Math [ZBMath]: a right-click on the map could generate a query to ZBMath for all papers in the "vicinity"; a query for papers in a selected area would be similar. Yet another service might be a query for journals based on the data prepared for the "journal map". Another service might be to localize mathematicians by their publication record[5] and give them a "home address" according to their primary research topic (based on the center of gravity of

[5]We acknowledge that author identification in large corpora is a hard problem that is only partially solved for the mathematics community.

their publications; possibly with weekend retreats in the woods of an application area). Similarly, research trajectories of mathematicians could be plotted on the map by computing yearly centers of gravity.

A Web3.0 style interaction could be to allow (groups of) users to name settlements, e.g. in honor of one of the most prolific contributors or landmark papers. We could use the math maps as a target for mashups of external services. For instance, the search results of a mathematical search engine e.g. [KMP12] could be shown by localizing them on the OpenMathMap service.

Finally, we acknowledge that the motivation for the OpenMathMap project was a cognitive question, which we have answered with a technical system.

Even though first feedback from mathematicians ranged from puzzled to enthusiastic (with an emphasis on the latter), we will have to systematically evaluate whether OpenMathMap-like systems and services can help with mathematician's day-to-day navigation problems and access tasks, or if OpenMathMap is essentially the equivalent to the iPhone beer app, a useless, but fun gadget.

References

[ACM-CCS98] *The 1998 ACM Computing Classification System*. 1998. URL: http://www.acm.org/about/class/ccs98 (visited on 11/18/2009).

[Doe13] Jan Wilken Doerrie. "OpenMathMap: Accessing Math via Interactive Maps". B. Sc. Thesis. Jacobs University Bremen, 2013. URL: https://svn.eecs.jacobs-university.de/svn/eecs/archive/bsc-2013/jdoerrie.pdf.

[Gde] *Web Map*. URL: http://www.informatik.uni-koeln.de/public/graphmap/gdnet/web/index.php.

[GDEA13] Graph Drawing E-print Archive. *Welcome to GDEA*. 2013. URL: http://gdea.informatik.uni-koeln.de/ (visited on 05/09/2013).

[GJ12] Martin Gronemann and Michael Jünger. *Drawing Clustered Graphs as Topographic Maps*. 2012. URL: http://gdea.informatik.uni-koeln.de/1288/.

[Jin10] Arif Jinha. "Article 50 million: an estimate of the number of scholarly articles in existence". In: *Learned Publishing* 23.3 (2010), pp. 258–263. DOI: 10.1087/20100308.

[KLN12] Adrian Kuhn, Peter Loretan, and Oscar Nierstrasz. "Consistent Layout for Thematic Software Maps". In: *CoRR* abs/1209.5490 (2012).

[KMP12] Michael Kohlhase, Bogdan A. Matican, and Corneliu C. Prodescu. "MathWebSearch 0.5 – Scaling an Open Formula Search Engine". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics (CICM). (Bremen, Germany, July 9–14, 2012). Ed. by Johan Jeuring et al. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 342–357. ISBN: 978-3-642-31373-8. URL: http://kwarc.info/kohlhase/papers/aisc12-mws.pdf.

[KW78] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling (Quantitative Applications in the Social Sciences)*. Sage Publications, Inc, Jan. 1978. ISBN: 9780803909403.

[Lea12] Leaflet. *An Open-Source JavaScript Library for Mobile-Friendly Interactive Maps by CloudMade*. http://leaflet.cloudmade.com/. 2012.

[LI10] Peder Olesen Larsen and Markus von Ins. "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index". In: *Scientometrics* 84.3 (2010), pp. 575–603. DOI: 10.1007/s11192-010-0202-z.

[Map13] Maperitive. *Maperitive*. 2013. URL: http://maperitive.net/ (visited on 05/05/2013).

[MathAtlas] Dave Rusin. *The Mathematical Atlas, a Gateway to Modern Mathematics*. URL: http://www.math-atlas.org/ (visited on 11/18/2009).

[MathSearch] *MathSearch*. Project Homepage. URL: http://mathsearch.kwarc.info/ (visited on 02/22/2013).

[MDS05] Ingwer Borg and Patrick Groenen. "MDS Models and Measures of Fit". English. In: *Modern Multidimensional Scaling*. Springer Series in Statistics. Springer New York, 2005, pp. 37–61. ISBN: 978-0-387-25150-9. DOI: 10.1007/0-387-28981-X_3. URL: http://dx.doi.org/10.1007/0-387-28981-X_3.

[mdscale13] MathWorks. *Nonclassical multidimensional scaling*. http://www.mathworks.com/help/stats/mdscale.html. 2013.

[Msc] *Mathematics Subject Classification MSC2010*. 2010. URL: http://msc2010.org (visited on 11/16/2011).

[Ope] *KWARC/openmathmap - github*. URL: https://github.com/KWARC/openmathmap.

[OSM] *OpenStreetMap - The Free Wiki World Map*. URL: http://www.openstreetmap.org/ (visited on 09/01/2012).

[PACS10] *Physics and Astronomy Classification Scheme (PACS)*. 2010. URL: http://aip.org/pacs/ (visited on 12/17/2011).

[Rus01] Dave Rusin. *About the MathMap image*. ttp://www.math.niu.edu/~rusin/known-math/collection/mathmap.html. 2001.

[Rus02] Dave Rusin. *Information about the Mathematical Atlas collection*. 2002. URL: http://www.math.niu.edu/~rusin/known-math/collection/index.html (visited on 12/02/2012).

[Wik12] Wikipedia. *Radial basis function*. 2012. URL: http://en.wikipedia.org/w/index.php?title=Radial_basis_function&oldid=544074242 (visited on 05/08/2013).

[ZBMath] *Zentralblatt MATH*. URL: http://www.zentralblatt-math.org/zbmath/ (visited on 06/12/2012).

[Fre91] Free Software Foundation. *GNU General Public License*. 1991. URL: http://www.gnu.org/copyleft/gpl.html (visited on 05/22/2013).

[Fre99] Free Software Foundation. *GNU Affero General Public License*. 1999. URL: http://www.gnu.org/licenses/agpl.html (visited on 05/22/2013).

Received July 6, 2013

Computer Science, Jacobs University Bremen
E-mail address: <first initial>.<last name>@jacobs-university.de

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

REPORT ON THE EUDML EXTERNAL COOPERATION MODEL

THIERRY BOUCHE AND JIŘÍ RÁKOSNÍK

Abstract. One of the main tasks of the European Digital Mathematics Library project was to define a cooperation model with a variety of stakeholders that would allow building a reliable and durable global reference library, aiming to be eventually exhaustive. In this paper we present the EuDML external cooperation model and the business plan as the basis for its sustainability and further development.

1. Introduction

The European Digital Library (EuDML) [1, 2, 3, 4] was a project partly funded by the European Commission in the Competitiveness and Innovation Framework Programme, Information and Communication Technology Policy Support Programme, in the period from 1 February 2010 to 31 January 2013. The EuDML project was explicitly envisioned as a pilot project addressing two challenges that prevented previous attempts towards a global digital mathematics library based on a top-down approach to succeed:

(1) Setting up the technical infrastructure to create a unified access point for the digital mathematical content hosted by a number of different organizations across various countries;
(2) Defining a cooperation model with a variety of stakeholders that would allow building a reliable global reference library meant to run over the long term, and to be eventually exhaustive.

During the three years of the project, these two goals have been pursued in parallel with stubbornness. On both sides the project reached clear successes and modified the state-of-the-art. The basic infrastructure is up and running, with a critical mass in content. A number of possible partners have declared interest in the

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

initiative. However a lot remains to be done in order to secure these results and set the basis of a strong and inclusive infrastructure.
A general overview of the project and its outcome is described in [5]. There is some slight overlap between both papers, in order to keep each of them self-contained . In this paper we report on the second goal mentioned above. In the first section we describe the situation that evolved from the project. The second section is devoted to the EuDML sustainability plan.

Acknowledgement. The work reported here has been partly funded by the European Union through its Competitiveness and Innovation Programme (ICT PSP, Grant Agreement no. 250503).

2. EuDML at the end of project

The EuDML has been built by the motley consortium of 16 partners from 8 European countries, which comprised a variety of stakeholders and expertise: universities (Instituto Superior Técnico Lisbon, Université Joseph-Fourier Grenoble, University of Birmingham, Universytet Warszawski, Universidade de Santiago de Compostela, Ionian University Corfu, Masaryk University Brno), research institutes (Institute of Mathematics and Informatics BAS Sofia, Institute of Mathematics AS CR Praha, CNRS Grenoble), an international scientific service institution (FIZ Karlsruhe), a national research council (Consejo superior de investigaciones científicas Madrid), a commercial publisher (Édition Diffusion Presse Sciences Paris) and a private digital media agency (Made Media Ltd Birmingham), a large library (Niedersächsische Staats- und Universitätsbibliothek Göttingen) and, last but not least an international learned society (European Mathematical Society). The latter two had the status of associated partners to which another one came in the early stage of the project: Biblioteca Digitale Italiana di Matematica (bdim). They included digital content providers, technical developers, library, a scientific database and representatives of research communities.
EuDML as the outcome of the project aims to be a long-standing, reliable and open source of trusted mathematical knowledge. This implies EuDML policies, that mostly boil down to the following:

(1) *The texts in EuDML must have been scientifically validated and formally published.*
(2) *EuDML items must be open access after a finite embargo period. Once documents contributed to the library are made open access due to this policy, they cannot revert to close access later on.*

(3) *The digital full text of each item contributed to EuDML must be archived physically at one of the EuDML member institutions.*

These rules ensure that the EuDML as reference library system is on a sound base, with ingested content available for perpetuity and openly accessible eventually. For this purpose the project set up a complex of frameworks, technology, workflows, validation procedures, schemas etc. The EuDML was built as a distributed system with tasks distributed among partners each of whom assumes full responsibility for the corresponding segment.

This worked well during the project when the partners formed a formal consortium tied together by a formal contract setting responsibilities with respect to the European Commission. The contract finished together with the project and will be replaced with a suitable arrangement which we describe in Section 3.

The EuDML is not limited to the current digital content and the technology built above it. There is also important potential of further cooperation and extendibility.

On the technical front, the EuDML got expression of interest, but rather in the form of attracting new partners in some follow-up to the current project. On the political front, the EuDML got quite some awareness and support from mathematical societies on various occasions (The International Mathematical Union at the WDML symposium in Washington D.C., June 2012 [6], the European Mathematical Society and some national mathematical societies at the EuDML workshop in Prague, 2010). On the content front, there has been a large number of discussions with potential further partners (eLibrary of the Mathematical Institute of the Serbian Academy of Sciences and Arts in Belgrade, digitized proceedings of the International Congress of Mathematics and European Congress of Mathematics, project Euclid at the Cornell University and Steklov Mathematical Institute/Russian Academy of Sciences' project Math-Net.Ru).

An important decision that has been taken after the 6ECM round table [7] is that the EMS Publishing House will contribute the *Journal of the EMS* after a 5 years moving wall. Work to achieve this has started, partly handled by our partner Institute of Mathematics AS CR in Prague acting here as a sponsor for EMS-ph. A contact has been also attempted with JSTOR (resp. Elsevier) in the hope to acquire their public domain (resp. Open access archival) content and make it visible in EuDML.

The effort of the EuDML consortium does not end with creating a functional prototype of the Digital Mathematics Library and providing its content and services

to the public. The true success of the project depends very much on sustainability and further development of the EuDML. The principal aims of sustainable EuDML services comprise

- working toward comprehensiveness, service integration, and cost efficiency of the EuDML services,
- assisting in exploiting the benefits of networking for integration of digital library services such as sharing and enhancing data,
- advancing cooperation between information and service providers,
- creating and maintaining a non-profit service in the interests of the mathematics user community.

In order to create such sustainable service from the EuDML project, important issues have been assessed, namely

- an organizational and legal framework, which will take its roots in the EuDML consortium and further partners associated during its lifetime,
- balancing costs and potential sources of revenue of running the EuDML services,
- a common framework for dealing with intellectual properties rights and copyright issues.

3. Principles of the EuDML Initiative

3.1. **Organizational and legal framework.** Twelve partners of the consortium declared their will to continue in efforts to maintain and develop the EuDML after project's end, representing the general mathematical community and the core content and technology providers:

(1) The European Mathematical Society
(2) FIZ Karlsruhe Berlin
(3) Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
(4) Cellule Mathdoc, Centre national de la recherche scientifique & Université Joseph-Fourier, Grenoble
(5) University of Birmingham
(6) Institute of Mathematics and Informatics BAS, Sofia
(7) Institute of Mathematics AS CR, Praha
(8) Masaryk University, Brno
(9) Ionian University, Corfu
(10) Società Italiana per la Matematica Applicata e Industriale
(11) Unione Matematica Italiana
(12) Niedersächsische Staats- und Universitätsbibliothek Göttingen

They will form an association named EuDML Initiative where the first three of them will assume particular rôles. The European Mathematical Society will provide an umbrella using its institutional authority to secure the association's internal functioning and its external representation with respect to partners, other institutions and the general public. It is also expected, in particular, to ensure that the EuDML services shall remain under control of organizations representing the public interest. For the first period of at least three years, partners no. 2 and 3 will provide human resources and machine capacities for hosting, system and service maintenance, and basic technical operations. Partner no. 4 will contribute manpower to continue enhancing the metadata quality, ingestion procedure, and will help new content providers to contribute their collections. Partner no. 7 will take care for annotations moderation and dissemination activities.

The EuDML Initiative will be established as an association without legal personality. The status of an association without legal personality and financial budget will be perfectly functional at least for the first period, during which the possibility/necessity of transforming the EuDML Initiative to another model involving legal personality and financial issues will be investigated.

The purpose of the EuDML Initiative is to provide a Digital Mathematics Library (DML) for the worldwide scientific community as a public service which will help users locate the information that is distributed in various digital repositories and discover information related to their work in an efficient way, and encourage the public in using it as a public resource of knowledge which will become exhaustive and comprehensive in the field of mathematics.

The basis for such DML is provided by the EuDML. The EuDML Initiative will

- encourage content providers to join and integrate their content to the EuDML,
- adapt to using new information technologies and invite subjects interested in contributing to research and development for the continuous evolution of EuDML to join,
- search for projects of research and development aiming at improving its services to the user community and will encourage its members to get involved in such projects.

Membership in the EuDML Initiative will be open to any legally and contractually competent natural or legal person willing to support the objectives of the EuDML Initiative by providing

- digital content, i.e. integrating (at least partially) their digital collections of high-quality mathematical publications in the EuDML according to the EuDML guidelines and standards, which are based on internationally

accepted standards and trends, adapted to the special needs of mathematical publications, and maintaining and expanding access to these publications through the EuDML service,

- technological equipment and services for maintaining the EuDML central services and/or developing other technical services and tools to be used in the EuDML,
- scientific, financial, strategic or political support to the EuDML Initiative and its activities.

Members will have the duty to take part in the activities necessary for the satisfaction of the objectives of the EuDML Initiative and to comply with the decisions of the bodies of the EuDML Initiative, with the statute and the applicable legislation. Members of the EuDML Initiative are not obliged to make any financial contributions. Each member cares for his own costs.

The governance and operation of the association will be organized in the following bodies: the General Assembly, the Chair of the EuDML Initiative, the Executive Board, the Scientific Advisory Board and the Technical Committee; in conducting legal or other affairs the Initiative shall be represented by the Chair of the Initiative.

The General Assembly as the supreme decision-making body will be formed by one representative of each member. The tasks assigned to the General Assembly include decision on applications for membership, decision on expulsion of a member, election and dismissal of the Chair of the EuDML Initiative, the members of the Executive Board and of the Technical Committee from persons suggested by members of the EuDML Initiative, issuing instructions to the Executive Board, decision on modification of the statutes and the dissolution of the EuDML Initiative.

The Executive Board will be composed of the Chair of the EuDML Initiative, the Chair of the Technical Committee and three other persons representing members of the EuDML Initiative. The European Mathematical Society's eminent rôle and responsibility in the EuDML Initiative will be accomplished by the fact that one member of the Executive Board will be directly nominated by the EMS Executive Committee. The Executive Board shall have general charge of all matters concerning the EuDML Initiative except for those assigned to the General Assembly, in particular, management of the current affairs including copyright and ownership management, execution of the decisions of the General Assembly, appointment of the subordinate committees entrusted with the special tasks within general framework of the association. The Executive Board will be advised by the Scientific Advisory Board and supported by the Technical Committee.

The Scientific Advisory Board will be appointed by the European Mathematical Society of their representatives and other suitably qualified and recognized persons. It's responsibility will be ensuring the scientific quality of the DML service, and advising the Executive Board on scientific matters, strategic orientations and priorities for development of the service, taking part in the evaluation, and bringing in the feedback of the mathematical community. The Scientific Advisory Board works out recommendations for the development of EuDML with regard to the content and the organization of the EuDML Initiative.
The Technical Committee will ensure the continuous technical operations of EuDML services being responsible for the technical development, technical standards and workflows, and the technical operations of the system. The Chair of the Technical Committee will be automatically member of the Executive Board, assuring cooperation of both bodies.

3.2. **Costs and sources of revenue.** It is estimated that basic operation and maintenance of the EuDML system will require approximately 1.5 FTE and k€100 yearly. This will be covered by in kind payments of the five partners indicated in Subsection 3.1. Each EuDML Initiative member will cover his local expenses which concerns, in particular, the content providers responsible for keeping and developing their repositories. It is assumed that all bodies of the association will meet via telecommunication if appropriate. Videoconferencing has proved a very efficient management tool during the EuDML project and will reduce the running costs.
The minimal level of resources corresponding to the costs indicated above will be sufficient for the maintenance and slow development of the EuDML. However, higher revenues are desirable for a more dynamic advancement. Possible resources to cover these costs may include income from R&D grant funding generated by the association or by its individual partners, special fees collected from the EMS member organizations, donations, financial contributions from content providers and partners of the EuDML Initiative and in kind contributions by partners.
Despite its universal usage in science, technology, education, social matters etc. mathematics has rare possibilities to generate incomes directly. All the more the EuDML as a not-for-profit establishment providing public service will have to struggle for regular resources using the potential of all partners involved and of the scientific community represented particularly by the European Mathematical Society and other mathematical societies and organizations.

3.3. **Common principles for handling data and tools.** It is very important that the potential external partners understand who the EuDML Initiative is and

what is to be expected of a possible cooperation. Hence, the following principles should be adopted and published.
The content providers retain all rights pertaining to their collections. They grant to the EuDML Initiative the right to keep and use copies of their provided data for the purposes of search and retrieval display in EuDML public services.
Each content provider may decide, whether full text will be provided to the EuDML Initiative and to which extent the full text might be used or distributed. Agreements between the EuDML Initiative and the content providers specify these and further usage and exploitation rights for each collection contributed to the EuDML.
The indexing and other metadata information generated by the central services of EuDML ("EuDML-enhanced metadata") is owned by the EuDML Initiative. Content providers are entitled to use and exploit copies of EuDML-enhanced metadata of those items for which they have provided metadata.
Copies of metadata provided by content providers and the EuDML-enhanced metadata shall be kept at the sites maintained by service providers. If a service provider withdraws from this rôle, the data and respective rights and obligations stay with the remaining service providers. Members of the EuDML Initiative active in the research domain of Digital (Mathematics) Libraries are entitled to use (copies of) the EuDML-enhanced metadata for their research purposes. However, usage of a particular subset of the metadata for this purpose may be restricted by the respective information provider. Results of such research activities will be made available to the EuDML Initiative and its members.
Exploitation rights for software and tools developed by the EuDML Initiative stay with the originator. However, the EuDML Initiative advocates an open-source policy for software, and encourages developers to put their developments for DML in the public domain. In case of dissolution of the EuDML Initiative, all (meta)data and related rights are to be transferred to the European Mathematical Society.

4. Conclusions

A cooperation model with a variety of stakeholders has been successfully developed by the EuDML project for building a reliable and durable global reference library, and a number of possible partners have declared interest in the EuDML Initiative.

The EuDML policy was developed stating three main principles: (i) the digital content must be scientifically validated, (ii) eventual open access, (iii) physical archiving of the content at one of the EuDML member institutions.
Contacts were made with several possible external partners to ingest their digital content in EuDML. In some of these cases, the technical work has already started.
Based on the above policies, a model of sustainable EuDML operation has been drawn on the basis of an association without legal personality formed by EuDML core members being scientifically and organizationally strong not-for-profit institutions that take care of the system's activity, maintenance, and of the collections both in terms of preservation and eventual open access provision. Three partners, EMS, FIZ/zbMATH and ICM will assume particular rôles providing an umbrella securing association's internal functioning and its external representation, ensuring that the DML services shall remain under control of organizations representing the public interest, and providing human resources and machine capacities for hosting, system and service maintenance, and basic technical operations during the first mid-term period after the project end. The possibility/necessity of transforming the EuDML Initiative to another model involving legal personality and financial issues will be investigated during this period.
The EuDML Initiative will be an open, democratic association with well defined structure, distributed rôles and responsibilities which will allow the long-term sustainability, form the solid basis for partnership with external entities and provide condition for further development.

References

[1] EuDML: The European Digital Mathematics Library. http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503

[2] Wojtek Sylwestrzak, José Borbinha, Thierry Bouche, Aleksander Nowiński, Petr Sojka, EuDML–Towards the European Digital Mathematics Library. In Petr Sojka, editor, Proceedings of DML 2010, p. 11–24, Paris, France, July 2010. http://dml.cz/dmlcz/702569

[3] José Borbinha, Thierry Bouche, Aleksander Nowiński, Petr Sojka, Project EuDML–A First Year Demonstration. In James H. Davenport, William M. Farmer, Josef Urban, Florian Rabe, editors, Intelligent Computer Mathematics. Proceedings of 18th Symposium, Calculemus 2011, and 10th International Conference, MKM 2011, vol. 6824 of Lecture Notes in Artificial Intelligence, LNAI, p. 281–284, Berlin, Germany, July 2011. http://dx.doi.org/10.1007/978-3-642-22673-1_21

[4] EuDML: The project. http://project.eudml.org

[5] Thierry Bouche, Reviving the Free Public Scientific Library in the Digital Age? The EuDML project. In Klaus Kaiser, Steven Krantz, and Bernd Wegner, editors, Contemporary Issues in Mathematical Publishing, 20 p., 2013. This volume.

[6] The Future World Heritage Digital Mathematics Library: Plans and Prospects.
http://ada00.math.uni-bielefeld.de/mediawiki-1.18.1/
[7] EuDML: Accessing Europe's mathematical treasures. Panel discussion, 6th European Congress of Mathematics, Kraków (Poland), July 2–7, 2012.
http://www.6ecm.pl/en/programme/panel-discussions

Received February 7, 2013

Cellule Mathdoc (UMS 5638), Université Joseph-Fourier (Grenoble 1), B.P. 74, 38402 Saint-Martin d'Hères
E-mail address: thierry.bouche@ujf-grenoble.fr

Institute of Mathematics AS CR, Žitná 25, 115 67 Praha 1
E-mail address: rakosnik@math.cas.cz

EMIS Collections and Conference Proceedings
FIZ Karlsruhe

INDEXING OF DOCUMENTS AND ITS IMPACT ON THE CHANGING ROLE OF MATHEMATICS INFORMATION SERVICES IN THE DIGITAL AGE

WOLFRAM SPERBER AND BERND WEGNER

Abstract. With the development of digital publications, mathematicians are confronted with lot of new challenges. Publishing mathematics is easier by applying TeX or other encoding systems. Access is possible wherever Internet is available. The mathematical literature is growing rapidly. Reference databases and search engines like Google help the users to find the publications in their fields of interest. In this context, an improved content analysis of mathematical information plays a key role. Standards, models, and algorithms are the most important ingredients for an efficient processing of information and retrieval. In this paper, some approaches are presented to improve and advance existing information services in mathematics by automatic tools for content analysis and a first prototype for a coarse automated key phrase extraction and classification is described.

1. Introduction

The availability of digital versions of publications in the Web has changed the framework for information sciences and services drastically. This concerns the basic concepts for the delivery of information as well as related services and tools. In more detail: All Information can be stored in digital form. It can be downloaded to computers, made accessible world wide, and linked with each other. Today, the Web can be considered as the largest memory in the history of mankind. Digital formats allow sophisticated presentations of information and processing of information by machines. The information in the Web is globally accessible. Information retrieval and access facilities have radically changed. Search engines are today the most important instruments for information retrieval. For this purpose, indexing, storage, and ranking of all (public) available information in the

K. Kaiser, S. Krantz, B. Wegner (Eds.): Topics and Issues in Electronic Publishing, JMM, Special Session, San Diego, January 2013.

Web like home pages or a publications has to be prepared. Text analysis and citation analysis are important tools in this context. On this background established printed documentation services like Zentralblatt MATH or Mathematical Reviews changed into reference databases. Providing subject-specific high-quality comprehensive information still play an important role today.

Mathematical Information Services looking back To describe the changes and challenges of mathematical information services, we start with a retrospective. The first comprehensive mathematical documentation service was founded at the end of the 19th century. Before some first catalogues indexing mathematical books were published, see, e.g., Scheibel [1] or Murhard [2]. The Crelle Journal [3] (today the 'Journal fr die reine und angewandte Mathematik') published reviews of mathematical publications starting with its first issue beginning in 1826. The foundation of the 'Jahrbuch ber die Fortschritte der Mathematik' (JFM) in 1868 [4] marked a milestone in the history of information services in mathematics. The first volume covered the bibliographic data of 875 publications, more or less the worldwide production in mathematics of the year 1868. This compares to a current annual production of more than 120,000 relevant articles and books in mathematics.

The JFM was a reaction of the mathematical community to objective needs like having an overview of the increasing size and bandwidth of mathematical knowledge:

- The number of mathematically relevant publications became too large. The JFM should help the mathematicians to identify and track the relevant developments in mathematics and build up a comprehensive memory of mathematical knowledge being identical with the printed mathematics publications at that time. As a first requirement, the bibliographic data of the publications had to be registered.
- This was accompanied by an enhanced content analysis: Reviews were added describing the content of the mathematical publications The Table of Contents (ToC) of the JFM was a first rudimentary instrument to assign information on its mathematical subject. Of course, the ToC was far from being some first version of a classification scheme in the formal sense, which did not exist at that time. The ToCs of JFM were not really stable, they changed from year to year. Hence one of the basic requirements for a classification scheme was violated.

The edition of the JFM was promoted by the mathematical community and supported by famous mathematicians like C. W. Borchardt, L. Kronecker, and K.

Weierstrass. Since that time, several additional approaches for editing a comprehensive information service in mathematics were undertaken like Valentin's catalogue (see [5]) and the dream of a comprehensive mathematical library (see [6]). But they failed to be successful for a longer period. The next break through was the foundation of 'Zentralblatt fr Mathematik und ihre Anwendungsgebiete' in 1931 with the aim to overcome some deficits of the JFM. The biggest one was the huge backlog of the JFM. Zentralblatt survived until nowadays as the database zbMATH [7]. In the following, we will use the notation Zentralblatt MATH when we talk about the editorial procedures or historical items and zbMATH when we talk about the current service.

Within its lifetime of more than 80 years Zentralblatt MATH experienced a lot of changes and developed several new concepts. Here we will concentrate on the aspects related to the content analysis of mathematical publications. One milestone was the implementation of and the indexing in accordance with classification schemes like the currently used Mathematics Subject Classification (MSC) [8]. This already started before the transition of Zentralblatt Math into the reference database zbMATH and lead to huge volumes publishing higher cumulated subject indexes. With the development of the database zbMATH, key phrases had been added from the text or by the subject editors with the goal to make them available as primary items for subject searches. Clearly all words appearing in a documentation entry were at the disposal for the so-called basic search. Up to now, the assignment of classification codes and key phrases is based on intellectual work done by the subject editors and the reviewers.

2. Current Challenges for Information Services

The digital presentation of mathematical publications has changed the framework for the production and the offer of mathematical information. There are new options to provide the information in a sophisticated way. This automatically confronts the providers with new needs and challenges:

- One topic are digital formats of mathematical information and the World-WideWeb (WWW). Most mathematical publications are accessible via the WWW now. The corresponding formats, steered by the W3C as HTML, XML, RDF, or OWL, provide standards for well-structured documents and allow enhanced options for representing and handling the information. Mathematical formulae are of high importance for mathematical publications. Markup formats like TeX allow the authors to create manuscripts, which are ready for publications. MathML allows

the processing of the information machines and the semantic enrichment of publications.

- Publishing not necessarily needs publishers anymore. Authors have the possibility to publish their manuscripts directly on the Web. We notice an increasing number of so-called grey publications. Some communities have established their own online archives. But in comparison with conventional publishing peer-reviewing is missing.
- A more than linear growth of the number of mathematical publications can be observed. At present about 120,000 mathematical articles and books are published annually. Clearly, this number depends where the borderline from mathematics to other sciences is drawn.
- A change of the retrieval behaviour has to be taken into account. Big search engines like Google, Yahoo, Bing, et. al. are becoming more and more popular among mathematicians. They are focused on general and comprehensive information, which can be generated automatically. Their aim cannot be to care systematically about complete and precise information on special topics, mathematics. This still will be the domain of reference databases.
- The types of mathematical knowledge available in the Web is going beyond mathematical publications, mathematical software being the most important example.

There are advanced user expectations for a mathematical information service, in particular from the mathematical community. Traditionally such a service should be as complete as possible, providing information on all relevant publications. This is still an important requirement. Also high-quality is expected by the users. The information should be validated and precise, the presentation should be well comprehensible, and the broader context of a document should be visible. It should be well integrated in the network of mathematical publications. Links on different levels should be available enabling access to the article itself, to those cited in the references of the article or even to other publications of interest in the same context.

To fulfil the last requirement was almost impossible, when no digital offers were available. Citations had to be looked up in libraries, where the corresponding publication was stored in print or on microfiche for example. With the availability of digital publications reference databases in mathematics could improve their service essentially. They were able to extend their role from a retrieval service to a navigation tool for digital mathematics libraries, providing an excellent facility

for their users to find information of interest in the network of mathematical publications. As a consequence, more time has to be spent for the preparation of the input for the reference databases in order to care about an enriched and extended content analysis and an enhancement of the data by an improved linking to relevant information. This only can be afforded if the work will be supported by automatic processing of the information. The following sections will describe how such a processing may look like for the semantic enrichment of the stored information.

3. Content analysis for zbMATH

We may distinguish between direct and indirect procedures for the content analysis. This paper will concentrate on the direct procedures. The indirect ones generally may be described as derivations from the captured data, possibly including the bibliographies, if they are available electronically. They include citation analysis, ranking, several kinds of profiles etc. A lot of efforts are spent at zbMATH for improving and running such procedures.

The main direct procedures are the preparation and capture of reviews or abstracts, the assignment of classification codes, the highlighting and capture of key phrases (including also mathematical symbols and formulae) and the standardization citations provided with the review or abstract. These procedures represent different levels of precision. Though being standardized, classification codes are a comparatively coarse concept for describing the subject of a publication. Key phrases may describe this in a finer way, but being uncontrolled they are of restricted importance for subject searches.

Reviews or abstracts should give a short impression of the content and the main achievements of a publication, relating them to other publications by top citations. Reviews are prepared on a voluntary basis by reviewers, i.e. invited experts from the mathematical community. They should provide an independent view of the paper on contrast to abstracts. Abstracts generally are written by the author or sometimes by an expert involved in the edition of the publications. They are controlled by the authors in any case. Top citations should refer to papers, which are strongly related to the one under consideration, going beyond the common citations in the bibliography.

The assignment of key phrases may be considered as a preparation of a very short summary of the paper by listing relevant mathematical terms. This procedure has been introduced for zbMATH rather soon after its first release as a searchable

database was available. It is still pure intellectual work by the subject editors, extracting phrases from the review or the abstract and adding relevant key phrases from other sources. These terms are not controlled, because no controlled vocabulary or thesaurus is available for mathematics. They hopefully may be used in the future to develop such a scheme.

Classifying publications in accordance with classification schemes is an old practice. This is a useful method to structure and identify relevant subjects in a special context. Classification schemes combine verbal descriptions of subjects with classification codes and have a tree-like structure, displaying different levels of generality. These different levels enable the user to apply the scheme to collections of different magnitude. They also comply with the requirements of the paper era where documents have to be totally ordered according to subjects and lexicographically by authors within the same subject.

There are several classification schemes for mathematics, having different levels of precision, depending on the purpose they had been developed for. Most of them are mathematical parts of more comprehensive classification schemes. They have the advantage that many relations of mathematics to other sciences could be represented more appropriately. The reference databases in mathematics currently are using the Mathematics Subject Classification MSC (zbMATH, MSN, CMA) and the Russian version UDK of the Universal Decimal Classification (RZMat). MSC is a standalone scheme for mathematics, which goes back to a rudimentary scheme developed by the AMS for distributing offprints, was expanded under the name AMS Subject Classification scheme with the aim to classify the entries in Mathematical Reviews and was developed jointly by Zentralblatt MATH and Mathematical Reviews as an international standard in a series of revisions of the AMS Scheme. Applications of mathematics to other sciences are present as particular subject areas.

11-XX	**NUMBER THEORY**
11-00	General reference works (handbooks, dictionaries, bibliographies, etc.)
11-01	Instructional exposition (textbooks, tutorial papers, etc.)
11-02	Research exposition (monographs, survey articles)
11-03	Historical (must also be assigned at least one classification number from Section 01)
11-04	Explicit machine computation and programs (not the theory of computation or programming)
11-06	Proceedings, conferences, collections, etc.
11Axx	**Elementary number theory** {**For analogues in number fields, see 11R04**}
11A05	Multiplicative structure; Euclidean algorithm; greatest common divisors
11A07	Congruences; primitive roots; residue systems
11A15	Power residues, reciprocity
11A25	Arithmetic functions; related numbers; inversion formulas
11A41	Primes
11E72	Galois cohomology of linear algebraic groups [See also 20G10]
11E76	Forms of degree higher than two
11E81	Algebraic theory of quadratic forms; Witt groups and rings [See also 19G12, 19G24]
11E88	Quadratic spaces; Clifford algebras [See also 15A63, 15A66]
11E95	p-adic theory
11E99	None of the above, but in this section
11Fxx	**Discontinuous groups and automorphic forms [See also 11R39, 11S37, 14Gxx, 14Kxx, 22E50, 22E55, 30F35, 32Nxx]** {**For relations with quadratic forms, see 11E45**}
11F03	Modular and automorphic functions
11F06	Structure of modular groups and generalizations; arithmetic groups [See also 20H05, 20H10, 22E40]
11F11	Holomorphic modular forms of integral weight
11F12	Automorphic forms, one variable
11F20	Dedekind eta function, Dedekind sums
11F22	Relationship to Lie algebras and finite simple groups
11F23	Relations with algebraic geometry and topology

A piece of the MSC2010

The current MSC provides 5,606 five-digit classification codes, the first two digits representing the top level of 63 mathematical subject areas and the first three digits leading to the second level of 528 subareas. There always had been relations between the subject areas, or even clusters of current mathematical research, where the subjects were distributed over several established subject areas. The only reasonable way to solve such problems was (and still is) to add cross-references and to extend the tree on the third level (for more information see [9]). The MSC is a living classification scheme. Revisions were made every ten years with a light revision after five years. This is done in close cooperation between Zentralblatt MATH, Mathematical Reviews and the mathematical community.
As mentioned at the beginning, the MSC codes are too coarse to be able to describe the content of a publication with sufficient precision, even though in many cases several MSC codes are assigned to the same publication, distinguishing primary and secondary relevance. Key phrases generally provide more precise information on the subjects a paper is dealing with. But they are not part of a controlled vocabulary and their assignment depends on the judgement of an editor. By taking their verbal explanations MSC codes may be considered as key phrases themselves. Appropriate linking to different parts of a publication of may help to provide tools for making the search more precise. Whatever this may be, editors will need support by automatic processing of the documents to be able to install such facilities. The following should describe some measures to come nearer to a solution of the problem.
A first essential step for the further development of the MSC and and its improved Web integration is the development of a SKOS version, see [10]. This has several advantages:

- SKOS schemes can be processed automatically. SKOS is an W3C standard providing a model and vocabulary for knowledge management systems like thesauri and classification schemes based on the W3C standards like XML, RDF, and OWL. SKOS covers elements (a principal element in SKOS is the 'concept') and attributes as well as main relations in thesauri and classification schemes. XML, RDF, and OWL provide the syntax and the basic models.
- SKOS schemes can be linked with each other. By various 'match' elements within the SKOS vocabulary, the SKOS scheme can be concatenated with other knowledge management systems like a controlled vocabulary. Of course, the specification of the matching relations like the concordances between classification schemes is outside the scope of SKOS and must

be done separately. The matching of different SKOS schemes is a nice property which also can be used for the automatic indexing. This idea is described later.

- SKOS schemes are flexible and extensible. A SKOS scheme can be adapted to the special needs of a community and extended by additional properties. In the case of the MSC this may be an extension by additional relations for the similarity of the MSC classes.
- SKOS schemes can be used to support retrieval by producing special strategies for retrieval like navigation maps.

4. Key phrase extraction and classification using linguistic methods

In the last decade, various methods for text analysis were developed supporting an automatic extraction of key phrases. They are based on statistical methods combined with a linguistic analysis of the text. Such approaches turned out to be successful for the analysis of simple texts, not involving sophisticated expressions. In mathematics like in other sciences the publications are written in a more technical language leading to a lot of complications: The mathematical terminology is very specialized and rich though frequently using common language for the many notions invented by mathematicians. Even the same word is used in different contexts having a different meaning then. In addition to this we find terms and acronyms only used in mathematics. Sometimes subjects may be identified by standardized predicates. The most difficult part is to extract the information on subjects, which is stored in formulae. As a first step these difficulties had been ignored and a first version of an automatic tool for keyword extraction had been developed, which is based on the general methods already available, taking the abstracts or reviews of mathematical publications as the reference for the analysis. An advanced tool for an automatic text analysis should be able to handle both, text and formulae. The big majority of mathematical publications is written in English at present. Hence we may concentrate our efforts on English texts at the beginning, caring about translations into other languages later. Here we are confronted with the common problem of language processing on the mathematical level: the ambiguity in both directions, i.e. different words may describe the same subject, the same word may be used for different subjects. The same applies to formulae: Different formulae may express exactly the same mathematical fact, the same formula may describe different mathematical facts.

Modern mathematics has developed a lot of formalization for the description of mathematical objects, structures, relations, facts and proofs. Hence formulae are a fundamental and indispensable part of mathematical publication. It is impossible to understand a mathematical publication by just ignoring the formulae, and in almost any case it is unreasonable to write one without formulae. formulae allow to express arbitrary complex and structured content in a rather precise form. But they are only of limited use for the automatic extraction of information on the subjects of a publication. Most formulae appear as an intermediate step of an effort, to prove a result or to simplify an equation, and only the final result may be of interest for deriving an information on the subject of the paper. Another difficulty is given by the arbitrary complexity of formulae and various encodings of semantically identical formulae. On the other side, semantically different facts may be expressed by the same encoding as a formula. Hence text analysis based on formulae always will have limitations for the description of the subject of a paper. To combine it with the analysis of the remaining text will improve the result. The best result will be obtained, when a semantic markup for the formulae has been made by the author or an editor, which is possible in principle by encoding the publication in Content MathML. Unfortunately, at present most authors only provide Presentation MathML, generally generated as a conversion from a TeX-encoding.
Text analysis of the plain text of a publication leads to more promising results. Like other sciences mathematics has developed its own set of terms describing mathematical subjects. Several types of terms and environments of terms can be distinguished.

- The mathematical vocabulary: It uses many words from common or technical language and extends this list by notions, which are specific for mathematics. The words taken from common language are labels of well-defined mathematical concepts, like groups, fields, convergence, measure or continuity. In addition to this there exist of words used in mathematics exclusively like algebra, holomorphic, hyperbolicity or polytope. For higher precision combinations of common words and technical notions are used like semicontinuity, subgroup or matrix algebra.
- Named Mathematical Entities (NMRs): NMRs denote objects which are widely used by the mathematical community (de-facto standards). They play an important role in mathematical publications as names or identifiers for formulae, mathematical facts and mathematical objects. They are frequently used in mathematical texts. Often, combinations with

names of persons, which may not be registered in the standard dictionaries of language processing, like the KnasterKuratowskiMazurkiewicz lemma, Fubini's theorem, being almost the same as the Cavalieri principle, or the Erdös number, are named entities in mathematics. Generally more structured subject information is linked with NMRs like definitions, mathematical objects, methods and relations or a whole theory. There may be symbols identifying a NMR or synonyms, substructures, broader terms, terms, similar terms, etc. All this has to be taken into account.

- Acronyms: Names of mathematical objects are often described by longer text phrases. Typically, authors defines acronyms as abbreviations for frequently used phrases. Often, acronyms will be introduced as abbreviations by extracting some special characters from a phrase. Acronyms are tagged by a special spelling (one or more capital letters, also inside a term). It's is good practice that the acronyms used in a paper are expanded at its beginning or end. Acronyms can be ambiguous. For content analysis, it is important that acronyms together with the corresponding part of the given text phrases are interpreted, when a unique text expansion is not available.
- Combinations of phrases: Though also sentences in mathematical texts have the usual distinction of subject, predicate and object, these entries may be quite complicated. For example, a lot constructs of noun phrases appear, the simplest ones given by combinations like '... of ' or ' by '. They have to be considered as a unit for the text analysis.
- Segmentation: Mathematical publications generally have a well-structured presentation. Structural elements which could be distinguished easily are headed by axiom, definition, theorem, lemma, propositions, corollary, proof, example, remarks etc. Also the abstract of an article and the bibliography are typical segments. The type of segment where a word appears has relevance for the semantic of a publication. The segmentation is relevant for the content analysis of complete publications.

In the following we restrict our considerations to the analysis of reviews and abstracts of mathematical articles. This may be justified by the following arguments: As explained above, reviews and abstracts (should) describe the main content of a publication in a condensed form. Hence, though covering a selection of terms only, abstracts and reviews are a reasonable basis for the development of an automatic tool for the key phrase extraction. Problems caused by the big amount of terms provided with the full text of a publication may be handled by a later

release of such a tool. But, one main reservation should be kept in mind. The selection given by an abstract or review can be considered as an initial intellectual phrase extraction and may turn out as insufficient if one looks at the complete text of a paper. In the case of reviews this is not an exception only. Therefore the subject editors at Zentralblatt MATH have to spend additional work to provide a more complete selection of key phrases. Reviews or abstracts contain formulae only in rare cases and in such cases they are restricted to the main mathematical objects the publication is dealing with. This leads to two contrary cases. The formula may just be a supplement to the subject information available in the abstract or review, which is the lucky case for automatic phrase extraction, or the reader should see from the formula what the paper is about, which is the bad case. The full text of the majority of mathematical publications is not available for zbMATH (and any other external party trying to establish an integrated semantic search) for text analysis, whatever the method may be. There is the commercial barrier at the big publishers, who do not open the access to the full text for such purposes and are content with their internal search engines, though the quality of these engines is far from satisfying the requirements formulated in this paper. And there is the other barrier, that a big portion of mathematical papers is not available in fully searchable digital form up to now. These are PDFs without providing access to the source file encoding the paper or scanned images, where a text OCR has to be applied to make the text without formulae searchable. This is not a reasonable basis for developing a tool for a comprehensive text analysis. For us it was natural (and also successful) to start the development of automatic tools for the key phrase extraction with support from computational linguistics. Computational linguistics has the general aim to make natural language texts machine-processable. For this purpose the structure and the rules of the common version of a language like English (syntax, grammars, different types of sentences) were investigated. Dictionaries were developed capturing as many words of (the common version of) a language as reasonable. Of course, there are a lot of structural differences between different languages. Chinese and other Asian languages like Thai, Vietnamese or Khmer language have other characteristics than the European language families like English. Hence, a linguistic analysis requires an adaption of the methods to the specific structure of a language, which is not a big problem in the case of English. Further specification has to take the mathematical ontology into account.

The approach presented in this paper starts with an analysis of reviews or abstracts, which generally consist of a sequence of sentences. A segmentation exhibiting theorems in an explicit form is rare and does not contribute to the semantics of the abstract or review. Hence changing to linguistic terminology, our first step is the tokenization of the abstracts or reviews.
Tokenization: Tokens in English texts are separated by blanks. Blanks are also the decisive criterion for the tokenization of reviews and abstracts. formulae in mathematical texts will be considered as composite tokens, ignoring the blanks in a formula. A further analysis of the formulae will be delegated to later parts of this paper. Further composite tokens in mathematics are combinations of names, common English words, etc. as explained above. Different spellings are used by the community like 'fixed-point' versus 'fixed point', where one token has to be identified with a composition of two tokens, or 'quasiconvexity' versus 'quasi-convexity', where two different tokens have to be identified. A morphological analysis and development of special dictionaries of morphological forms can be used for unification or normalization of the different spellings of the same token.
One of the basic methods in computational linguistics is the type classification of the tokens of a language, known as Part-Of-Speach (POS) tagging: There exist different tag schemes for the English language. Some of the most commonly used schemes are the Penn Treebank scheme with 45 classes the C5 tagset with 61 tags, and the Brown tagset with 87 elements. Here we use the Penn Treebank scheme [11] which has, e.g. , the folllowing tags for nouns: NN (noun), NNS (noun, singular or mass), NNP (noun, plural), NNPOS (proper noun, singular), NNPOP (proper noun, plural). To determine the tags, a lookup in dictionaries is started. The underlying dictionary for our approach is the Brown corpus, a list of more than 1,000,000 English words from 5,000 written texts from different genres. The tokens in the dictionary also include the tags about their type.
Here are some possible difficulties:

- The tagging generally is not unique, a token can be an element of different classes.
- No dictionary is complete. That means that not all potential tokens are in the dictionary.

In these cases, the adequate tag of a token in a sentence must be deduced from the context. This can be done on the basis of rules and/or by the application of stochastic methods, where the 'and' stands for modern methods. The most popular method on the stochastic part is given by the Hidden Markov Models (HMM) to detect the (hidden, because not observable) POS tags. The Dynamic

Programming approach, especially the Viterbi algorithm, is very popular to calculate the adequate tag of a token in a phrase. For the POS-tagging Open-Source software, see [12], is available. The existing Open Source software is used as a starting point for the development of advanced software for the content analysis.
Mathematical formulae are outside the scope of computational linguistics for English as well as for other languages. Hence, formulae are handled by a special approach. By deleting control characters and numbers, an artificial hash for each formula is created. The original formulae are stored separately which allows a special analysis of the mathematical formulae.
The POS tags are an important resource for further linguistic investigations. Grammatical relations and context-free grammars are concepts to formalize relations. For example, the subject-predicate-object rule for English can be used to identify word groups or phrases. As the name 'noun phrase' indicates, it requires an a-priori identification of the nouns and the classes of the tokens in the environment.
Context free grammars (CFGs), also called phrase-structure grammars, are schemes and models for structuring natural languages. CFGs cover a set of rules how the elements of a language can be grouped, e.g., by defining characteristic patterns of noun phrases consisting of elements from the Penn Treebank Tag-set. Again, a dictionary covering tokens and additional grammatical attributes is needed to complete a CFG.
Noun phrases: Noun phrases are the most relevant phrases in mathematical publications for executing a content analysis. Noun phrases can have different roles in a sentence, e.g., as the subject or as the object of a sentence. Noun phrases can have different structures, like a sequence of some nouns. Noun phrases can be arbitrarily long. We have defined a set of characteristic types of noun phrases in mathematical texts: They are defined by sequences of POS tags. Also mathematical formulae inside a noun phrase have to be considered. An important secondary problem is to identify the leading noun of a phrase (the central noun) and to identify the complete relevant phrase, leaving out irrelevant parts. Up to now, the length of key phrases has seven tokens as upper bound.
The extracted candidates for key phrases are the preliminary material for establishing a controlled vocabulary for mathematics. Such a vocabulary consists of the named mathematics entities enriched by the important mathematical key phrases of publications, which up to now are not part of the nomenclature. Controlled vocabularies may be used as a base for the development of thesauri and ontologies, which can be created by adding relations between the elements of a

controlled vocabulary. Languages are subject to changes. Hence controlled vocabularies have to be kept as dynamic objects. As already mentioned, at present there is no controlled vocabulary for mathematics. One reason for this is that the number of candidates is very huge. M. Hazewinkel, see [13], has proposed 120,000 items as an estimate for the size of such a vocabulary. Our experience is that this estimate is out of date already. The currently available automatic extraction tool has led to some millions of candidates for a controlled vocabulary. The size of the vocabulary is one of the barriers to create and maintain a controlled vocabulary without machine support.

Here we try to realize a semi-automatic approach. The extracted key phrases from the reviews and abstracts are used as candidates for the controlled vocabulary. They may be related with other resources, like the MSC codes or entries in the Encyclopedia of Mathematics or the mathematical part of the Wikipedia. This is done for each MSC class enabling a later application to automatic indexing. The frequency of the occurrence of a key phrase is taken as one measure for its relevance. The resulting relations will be checked intellectually by experts. A periodic iteration of this approach guarantees the actuality of the controlled vocabulary. Related key phrases provide an important enhancement of the MSC codes, because they represent a comprehensive explanation of what the code is about.

The implementation of an automatic classification on the basis of a controlled vocabulary will work as follows: To each MSC code its part of the controlled vocabulary is assigned. The controlled vocabulary for the MSC code is transformed into a SKOS scheme. Then, the SKOS scheme of the MSC is matched with the SKOS scheme of the vocabulary. Methods for automatic text classification base on mathematical models and methods. Usually, text classification starts with the vector space model which provides a representation of a document by a vector. More in detail, the document will be split in n different tokens. The n-th component of the vector is the frequency of the n-th token. Instead of single tokens also phrases of tokens may be used. The extracted text phrases of a document and the phrases of the controlled vocabulary of each MSC class can be directly used to compose vectors, which are used then to determine the degree of similarity between the vectors for the key phrases and the controlled vocabulary. Different metrics and approaches are available for this. Most popular approaches are the k-nearest neighbours method (based on the Euclidean metric), the naive Bayes method (a stochastic approach), and the Support-Vector-Machines (SVM) approach basing on a geometric concept called the separation by hyperplanes.

5. Formula analysis

There are a lot of remaining problems like the normalization and standardization of key phrases, which are given in the document in different forms. The last one we will address in this paper is the development of specific methods for the analysis of mathematical formulae. As mentioned above the analysis of mathematical formulae is more complex than that of texts. But, also the language of mathematical formulae has a lot of conventions and standards for basic symbols. At first some remarks about formulae in zbMATH should be made:

- It is easy to detect the formulae in TeX-encoded texts, because they start and end with $-signs. They will be inside the corresponding MathML-namespace elements in MathML encoded text. If we take this as characteristic of mathematical formulae in the database entries of zbMATH, then zbMATH contains with almost 10,000,000 formulae.
- If we use the size of the tokens as a rough measure of the complexity of a formula, most formulae in zbMATH are simple symbols, i.e., single characters, which denote a mathematical object. The frequency of formulae in reviews or abstracts is different for different mathematical subject areas. For example, the frequency of formulae in reviews or abstracts on mathematical applications is lower than that for pure mathematics.
- The same of formula may have different encodings.

In principle, the development of formula analysis can be done in the same way as for text analysis. The segmentation of formulae is a first important step. This is more difficult than that for English texts, because we have no standard separator to split complex formulae. Moreover, single mathematical formula entities can be rather complex, like definite and indefinite integrals. The tokens of a complex formula may be combined in different ways. XML provides a well-defined structure of documents. TeX encoded formulae can be converted automatically into Presentation MathML and also into Content MathML as far as the original encoding provides a unique semantic interpretation. Hence, XML allows to identify the mathematical tokens of a formula. But, the encoding in Presentation MathML provides only a very rough semantic classification of the elements as mathematical identifiers for constants, variables, or operators.

A finer analysis of mathematical formulae is more complicated than of texts. The formula may include free variables. For example, single character formulae like A or n are used as notations for a type of mathematical objects. Other single character formulae like P allow an more specific interpretation as subject when talking about in probability. The use of fonts (bold, cursive, Gothic, etc.) is

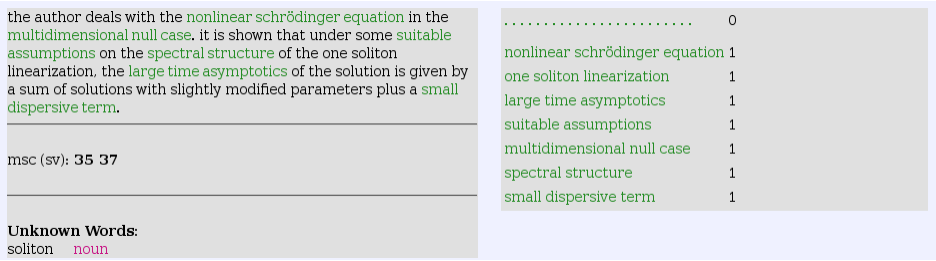
not standardized. Nevertheless, for a lot of named mathematical formula entities standard notations are common practice. Several named mathematical formula entities are characteristic for special subject areas. This may be given by an MSC code. Hence, the semantic background of a formula easily can be identified in this context.

- The development and maintenance of comprehensive catalogue of mathematical formulae is very expensive. A first approach may be given by the concept of the content dictionaries of the OpenMATH enabling the processing of formulae in arbitrary systems. Here we pursue another goal, namely to develop a concept which helps us to identify the tokens of a mathematical formula and make them and the complete formula searchable.
- For this purpose we need a dictionary of mathematical tokens. From segmentation, we get a first list of possible tokens in formulae. But they are of different value for the content analysis. For example, a sentence like 'Let A, B be matrices ...' is only of interest for a combined analysis of formulae and text.
- We also need an analogue for a standardized classification of mathematical formulae like the POS tagset for English language tokens. We need a dictionary listing the possible meanings (as a measure for the ambiguity of a token), matching tables for synonyms, and context information. The context information may be quite general like an MSC code of the publication or a reference. It may be very specific like the left-hand or right-hand side of a formula.
- A grammar for mathematical formulae would be useful. Here a grammar is understood as a set of general rules for the interpretation of the structure of mathematical formulae. This could be used for a deeper analysis of MathML formulae going beyond the path analysis. There are some activities to develop such a grammar, see [14]. Such an activity should be coordinated with existing activities in for mathematical knowledge management like Planet Math, OpenMath etc.

6. The prototype

A first prototype for the text analysis has been developed. The result offers the following features: an extraction of candidates for key phrases, a list of unknown words including the proposed POS tags, a proposal for a MSC classification. At present, the classification is restricted to the subject area level of the MSC).

The following snapshot should give a more detailed impression of the prototype.



The user interface of the prototype for key phrase extraction and classification

On the left-hand side, the input box, which contains the original text (review or abstract), is located. On the right-hand side, the extracted phrases are listed. These phrases are also highlighted in the original text to enable intellectual control by human beings. The left-hand side is completed by the proposal of the MSC subject area and a list of 'unknown' tokens, which are not in the Brown corpus, together with a proposed POS tag.
Two classifiers, the naive Bayes classifier (nv) and the Support Vector Machine classifier (sv), were used to calculate the classification. The selection of key phrases and the correctness of POS tags for the unknown tokens can be steered by check boxes. Unknown words with correct POS tags will be added to the dictionaries. The tokens are weighted by their frequencies.

7. Conclusions and Outlook

Because only the first two digits of an MSC code are provided by the prototype introduced above, this may be considered as a weak solution of the general problem, to establish an advanced semantic search. But this already leads to a first support of the editorial procedures for zbMATH, and it also may be used by librarians, who are not interested in refined classifications of their holdings. The prototype also is a good motivation to spend more efforts to obtain an advanced tool for automatic indexing applying the methods described in this article. The basis for the application of an advanced tool should be the full texts of the publications though the abstracts and reviews may be sufficient for the development of a first controlled vocabulary. As said at the beginning, mathematical literature is available today in digital form, and the best solution would be to have searchable versions for all publication and free access to their presentations for everybody.

This is the dream of the WDML (World Digital Mathematics Library). A comprehensive digital library for the mathematical literature should be distributed, scalable and flexible so that all providers of mathematical literature may take part in the enterprise and all people interested in mathematics can access their publication of interest.

Common standards and efficient methods for content analysis are essential for the quality of such a digital mathematics library. At present we have to be content with partial offers like the ElibM [15] (Electronic Library in Mathematics) in EMIS, hosted at Zentralblatt MATH and representing the largest repository of open access journals in mathematics, and the various national activities offering repositories of digitally born or retro-digitized articles like NUMDAM [16], DML-CZ, ERAM, RusDML, MathNetRu and others. EuDML [17] was a project build up a distributed digital library integrating the mayor European open access providers. Methods for an efficient machine-based content analysis where a topic of high priority for all these projects.

References

[1] Johann Ephraim Scheibel, Einleitung zur mathematischen Bücherkenntnis. (Band 1) Breslau: Meyer (1769, 1775)

[2] Friedrich Wilhelm Murhard, Litteratur der mathematischen Wissenschaften. Leipzig: Breitkopf und Härtel (1797)

[3] Journal für die reine und angewandte Mathematik (Crelle Journal) http://www.degruyter.com/view/j/crll

[4] Felix Müller, Carl Ohrtmann, Jahrbuch über die Fortschritte der Mathematik. Band 1. Berlin: G. Reimer. S. I-IV (1871)

[5] Georg Hermann Valentin, Die Vorarbeiten für die allgemeine mathematische Bibliographie. Bibliotheca Mathematica 1 (3), S. 237-245 (1900)

[6] Felix Müller, Zur Frage der Begründung einer mathematischen Zentralbibliothek. Bibliotheca Mathematica 4 (3), S. 389-391 (1903)

[7] zbMATH (Zentralblatt für Mathematik) http://www.zentralblatt-math.org/

[8] Mathematical Subject Classification http://msc2010.org

[9] Patrick Ion, Wolfram Sperber, MSC2010 in SKOS, EMS Newsletter No. 84, 2012, 55-57

[10] MSC/SKOS http://msc2010.org/resources/MSC/2010/MSC2010

[11] The University of Pennsylvania (Penn) Treebank Tag-set, http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

[12] http://nlp.stanford.edu/software/tagger.shtml#Download

[13] Michiel Hazewinkel, Enriched thesauri and their uses in information retrieval and storage, 1996, E.R.C.I.M.

[14] Deyan Ginev, The structure of mathematical expressions, Master Thesis, 2011, Jacobs University Bremen, http://kwarc.info/people/dginev/publications/DeyanGinev_MScThesis.pdf

[15] ElibM, http://www.emis.de/elibm/index.html
[16] NUMDAM, http://www.numdam.org
[17] EuDML, https://eudml.org/

Received June 1, 2013
Revised version received June 25, 2013

FIZ Karlsruhe, Zentralblatt Math, Fanklinstr. 11, D-10587 Berlin
E-mail address: wolfram@zentralblatt-math.org

Department of Mathematics, TU Berlin, Str. des 17. Juni 136, 10623 Berlin
E-mail address: zblwegner@googlemail.com