New Developments in Electronic Publishing
AMS/SMM Special Session, Houston, May 2004
ECM4 Satellite Conference, Stockholm, June 2004
p. 161-169.

EMANI – Leader and Follower for the WDML

Bernd Wegner

Mathematisches Institut, Fakultät II, TU Berlin, Sekr. MA 8-1
Straße des 17. Juni 135, D – 10623 Berlin, Germany
e-mail: wegner@math.tu-berlin.de

Abstract. The rapidly growing activities in electronic publishing lead to the request to develop global repositories, which care about three fundamental activities: to collect and store the electronic material currently available, to pursue projects for solving the long-term archiving problem for this material with the ambition to preserve the content in readable form for future generations, and to capture the printed literature in digital versions providing good access and search facilities for the readers. Long-term availability of published research articles in mathematics and easy access to them is a strong need for researchers working with mathematics.

The article will describe some new developments for two main projects in this subject: the plan to develop a global World Digital Mathematical Library (WDML or DML) and the main project dealing with a coordinated archiving of digital documents in mathematics, the Electronic Mathematics Archiving Network Initiative (EMANI). For the core of the EMANI network, a co-operational system of reference libraries and content providers like publishers and editors has been be set up. Both systems share a lot of common work packages. Hence discussions in one project have impact on the other. WDML being more comprehensive integrates comments from a bigger group of mathematicians and librarians. EMANI mainly concentrates on obtaining first results for a preliminary core group. This leads to a smaller model, which is working already and may be taken by the groups promoting WDML as a start for a more comprehensive solution for installing the WDML. This report refers to the state of both projects at the end of 2004 and describes their mutual impact.

1. Electronic offers and their providers

As already described previously in [11] the impact of electronic devices on the daily life of researchers, teachers or other professionals results from a variety of tools and offers installed in

local machines or made accessible through the Internet. The part libraries are mostly involved in consists of electronic publications, or better electronic versions of printed publications. Some libraries already developed digital repositories containing retro-digitised publications, which had been obtained by scanning printed articles and books. But also offers, which could be published in electronic form only, become more and more important. In addition to this researchers and teachers increasingly take advantage of computer algebra systems and other computing software, and visualisation techniques using graphics software and image processing tools have become background for most of their presentations and publications. Finally, we should not forget that the Internet has been used to establish a communication infrastructure, which strongly facilitates their daily work and extends the possibilities for co-operation at distributed sites.

There is a wide range of providers of these offers, going from commercial publishers and learned societies to volunteers and single authors. Also the list of distributors and information brokers is a long one: libraries, databases and indexing services, internet-portals of different types, web browsers et al. Clearly, libraries try to transfer their system, they have developed for their printed holdings, to these new publications, and hence they still seem to be the most reliable information provider also with respect to electronic offers. But this role has to be acknowledged more widely and the offer has to be improved. Hence there are good reasons why libraries will be able to maintain their central role for distribution and storage of scientific information and succeed to extend this to the electronic media. They have developed precise and reliable access structures. Their service is free for their specific group of users, and this group is a large one in most cases. For external users they developed a good network of exchange facilities, which enables scientists to make their work really accessible for a wide community of users and to read the work of their colleagues without being confronted with bigger commercial barriers.

The starting point of the initiative to develop the WDML is the following (see also [6]). Mathematics is a science where the availability of electronic publications and retro-digitised documents leads to a considerable improvement of the conditions for research. Hence, though some of the subsequent arguments may apply to all sciences, they turn out to be of particular importance for mathematics: Mathematicians and professionals applying mathematics need quick, reliable and integrated access to mathematical publications. Long-term availability of publications is a particular need in mathematics. Digitising of print-only publications and the adjustment of these offers to the current facilities provided for electronic publications leads to a additional series of problems to be solved. Electronic publishing offers a variety of additional information in mathematics, which may be integrated into the access and display structures enhancing the traditional types of publications.

In this context it may be worthwhile to repeat some more concrete figures on the age of citations in mathematical articles from [11]. The figures had been cited from an investigation by Joachim Heinze [8] pursued in the case of three journals in the year 2002. Most surprising (also surprising to mathematicians) are the numbers of citations before 1992. In the case of the most traditional mathematical journal from North-America, the Annals of Mathematics, 60 percent of the citations in the 35 articles published in that journal in 2001 had a publication date before 1992. Vice-versa, the number of citations from the volumes of 500 journals

published in 2001 to the Annals was about 4.500 and 82 percent of them were before 1992. Looking at one of the first journals, which published mathematics only (in contrast to journals which deal with several sciences), the Journal für die Reine und Angewandte Mathematik, founded as Crelles Journal in 1826, the first figure was 61 % and the second 65 %. Finally, these numbers still were high for a more "modern" journal which had been founded in the second half of the 20th century, the Inventiones Mathematicae: the first figure was 55 % and the second one 68 %. Such high numbers of older citations are not common for most of the other sciences.

2. The archiving problem

In the "paper world" the long-term preservation of publications was simple on the first view, though at a closer look it becomes obvious that a lot of problems had to be handled. They mainly came from the deterioration of the paper or the binding of a book or journal, and they appeared after a comparatively long period in which the physical situation of the document could be considered as stable. Also a wide distribution of documents to several locations worldwide was a factor of stability, protecting them against being all destroyed simultaneously by the impact of catastrophes, fires, wars etc.

For digital publications this period of stability turned out to be extremely small so far. What everybody experiences with his old releases of word-files, became true two years ago for the readers of PDF-files, for example. The release of the Acrobat-reader at that time could not handle files produced with the first release. The problem had been solved by a newer release, but nobody knows if it will come up again with other releases in the future. This is not a special problem with PDF, and it is only one problem. Another one is the stability of the physical carrier, where the data are stored. Furthermore with an electronic document a variety of facilities may be associated, which depend on additional software. Current releases of this software may have a short lifetime. What should we do with the document afterwards?

To solve this problem will be even more complicated when interactive documents in mathematics are considered, because they are most likely to have software depending enhancements. They will play an important role in the future. Projects like MoWGLI ([2]) will develop different types of structures enabling semantic mark- up of documents. Hence preservation will go far beyond caring about the displayed text only. Structures, links and other informational background provided with electronic articles will have to be taken care of, and all these tools are in permanent evolution.

3. The EMANI project

There is a period of more than 10 years during which electronic publications in mathematics developed from some offers in pioneering freely accessible journals to a first class publication facility with enhanced services in comparison to traditional printed publications. Hence the archiving of these publications became a increasingly urgent problem to be tackled. As a first step the Electronic Mathematics Archives Network Initiative (EMANI) had been founded in the first half of 2001 with the initial aim to develop models for the archiving of

electronic contents in mathematics. Having in mind that a distributed architecture would be more suitable and reduce the load on the partners for such a project, a network had been established, which also will be a more open approach for extending the project from an initially restricted group to a more comprehensive enterprise.

The core of the network consists of a co-operational system of reference libraries and content providers. In the ideal final version they are supposed to serve for a long list of purposes: The basic action is to store the digital content in mathematics from the content providers at the reference libraries. This already had been complemented by retro-digitising the printed publications in mathematics from the content providers at the reference libraries, covering a big part of publications in mathematics by electronic versions finally. On this basis first projects are pursued to care about the long-term preservation of this content in readable form. This refers to technical support for metadata production and access design and tools for the handling of TEX-files as is described by the article [7] in the current Proceedings.

The reference libraries for EMANI will serve as a backup system for other libraries, which want to store and provide part of the content on their own or refresh their existing offers by updated material. On the side of content providers further publications will be integrated and additional publishers will enter the system. These developments may provide a model how to organize the WDML as a whole and serve as a core for this.

4. The current state of EMANI

To set up the system in an efficient way EMANI started on a smaller well-controllable scale. The current partners on the side of the libraries are:

- The Cornell University Library, Ithaca, N.Y. (CUL): They have a good tradition in retrospective digitisation projects. In particular set up an excellent offer of a bundle of electronic journals in mathematics through project Euclid. They serve as a mirror site for EMIS (see [10]).
- The State and University Library Göttingen: Also there some important retrospective digitisation projects like ERAM (see [3] or [9]) and RusDML (see [12] or the article [5] in this volume). In addition to this the SUB Göttingen serves as a central reference library for all publications in mathematics. In this role they have a high reputation as a reliable provider of access to mathematical publications. Moreover they also serve as a mirror site for EMIS.
- The Tsinghua University Library, Beijing (THUL): This library has experience with the digitisation of Chinese publications. This refers to ancient mathematics in China and to recent mathematical publications. They are a Chinese centre of excellence for installing and offering electronic publications. For further details see the article [1] in this volume.
- The French partner mainly contributed through the group MathDoC (the Cellule MathDoc at UJF) in Grenoble. Their strength consists in their excellent retro-digitisation project NUMDAM ([4]).

The first group of content providers was given by Springer-Verlag and associated publishers like Birkhäuser on the commercial side and the electronic library ELibM offered through

EMIS, the European Mathematical Information Service (`http://www.emis.de`) on the open access side. The ElibM is a co-operation of several journals and editors on a voluntary basis bundling electronic offers in a worldwide mirror system of WWW-servers (see [10]). Through ERAM a co-operation with de Gruyter has started by digitising one of the best journals in mathematics, the Journal für die Reine und Angewandte Mathematik (Crelle's Journal). There are some smaller digitisation projects like the one of the Catalan Mathematical Society or the SwissDML, which asked EMANI to host and archive their content. On this level an extension of EMANI will happen during the next years.

For a reference of the structural results obtained by the EMANI working groups the reader may consult the article [11] or the EMANI website `http://www.emani.org`. These dealt with the following questions: import and presentation formats, workflow, metadata, access/navigation/design, copyright/licences, retro-digitisation, economic sustainability, outreach and dissemination of information, expansion into other disciplines, architecture of EMANI. To some extent they coincide with the corresponding recommendations developed for the WDML. Here later developments should be mentioned only where solutions still were missing when [11] was written. For example, one important step was the installation of the advisory board, helping to connect EMANI with the mathematical community and the mathematics librarians and to suggest further recommendations for EMANI to improve its scientific value.

One important development refers to the stepwise transfer of the available electronic content from the content providers to the reference libraries. The libraries should check if the files still can be used for the archiving, adjustments should be made in the case of files, which are unsuitable for this and recommendations will be developed how the content providers could care about a more convenient delivery in future cases. Taking care about journals will be easier than handling books. The tools to handle this evaluation procedure more or less automatically have been developed just recently and are described in the article [7]. The partners have agreed that the current formats to be preserved will be the TEX-files containing the TEX source code and the TEXmacros for every article. In addition to this PDF files will be preserved as a representation format.

The progress of EMANI, which is best visible, consist of the growing content of digital representations of mathematics, either produced electronically or retro-digitised from printed publications. In addition to the more than 60 journals offered by the ElibM in EMIS the four library partners make more than 100 journals accessible through EMANI: 27 from SUB Göttingen, 20 from CUL, 55 from THUL and 4 from MatDocC through NUMDAM. This includes journals like Mathematische Annalen, Mathematische Zeitschrift, Commentarii Mathematici Helvetici, Inventiones Mathematicae, Communications in Mathematical Physics, Journal für die Reine und Angewandte Mathematik in their period, when previously only a printed version was available.

5. The global digital mathematics library – WDML

WDML stands for the global initiative to have all mathematics digitally available and for the distributed repository itself. It has a lot of overlap with EMANI and both projects a tightly linked with each other. But in contrast to EMANI the global initiative at first will concentrate on retro-digitisation.

In 2001 John Ewing prepared his White Paper [6] in which a rough estimate has been made how much money would be needed to develop a global digital mathematics library containing all mathematics in digital form. This estimate was in the order of 100 million US Dollars. But that was not the main achievement of that paper. It contained a lot of structural considerations for such a library, and it also addressed the immense problems we will be confronted with when we really want to pursue such a project. As a caveat when reading this paper, one should be aware that it describes an ideal solution, and some parts like a central repository (by intention) do not reflect very well what has been developed already. For example, at present only a system of distributed repositories could be imagined, because proprieties and aspects of cultural heritage have to be respected. Furthermore, a distributed system can hook on existing providers like libraries, and this will be more efficient than the installation of an extra infrastructure to manage the WDLM, as far as the costs will be concerned.

As a consequence a planning grant had been given to Cornell University by the National Science Foundation of the U.S.A. to make a feasibility study for the WDML. This had been discussed during two workshops in Washington D.C. at the end of July 2002 and at the SUB Göttingen in May 2003. The results have been written down in a report, which can be found under the address `http://www.library.cornell.edu/dmlib/`. The technical details of this report have been accepted by the CEIC of the IMU as a first standard for the digitisation of papers in mathematics though further development will be needed to get a more comprehensive and up-to-date recommendation. Unfortunately, the strict guidance of the initiative available during the NSF-project is missing at present, which will slow down the further structural developments. In contrast to this the digital content for mathematics is rapidly growing at present.

In contrast to EMANI the following question occurs on the global level of the WDML. How can we determine what has to be considered as the content of mathematics? The answer is not urgent, because at the beginning the patches for the WDML are developed according to the funding available by the group, which succeeded to get the funding. Several parameters have to be taken into account for determining the ideal content: time, quality, location, integration of applications etc. But do we really have the chance and the interest to cover all mathematical publications world-wide by the DML? If not so, the selection criteria have to be discussed. Not everything could be done immediately. In principle, a time schedule for building up the WDML step by step has to be set up and this requires an order and hence a selection of the publications. But these are theoretical considerations, which only could be implemented when global funds for mass digitisation will be available. Registries for digitised material will play an important role, and for this at least two solutions are in progress.

There is also a cultural dimension. Though the global approach is a challenging idea, the

development of the repositories should take national interests and funding possibilities into account. Hence distributing the content to single projects has to respect what had been covered already and what should remain under the guidance of a special mathematical community. Only the remaining content may be open for adoption for retro-digitisation. To distinguish this will be one of the main tasks during the content determination and it will be a delicate task, because very quickly there may be the impression that one party wants to buy out the mathematical heritage from another one.

6. Digitisation projects

In principle, the current content of the WDML consists of the electronically produced publications and the repositories of digitised mathematical articles and books. It will be too long to address all these offers here. Only a survey on the projects dealing with retro-digitisation should be given here.

On the free level I want to highlight ERAM. The acronym ERAM stands for "Electronic Research Archive for Mathematics". The aim of the project is the installation of a (digital) archive of articles relevant for mathematical research, full searchability and access through a database, captured from the "Jahrbuch über die Fortschritte der Mathematik" (1868–1943). The Jahrbuch database has been finalised and provides access to a digital archive of almost 1.2 million pages built up within the project. The articles are offered in a good presentation format. Facilities for text searches are in development. For more details on the background of this project see the references [3] and [9]. The current status can be checked on the ERAM-homepage under `http://www.emis.de/projects/` clicking on the box for the Jahrbuch.

To have the digitised content of a journal stored somewhere will not be sufficient. On the basic level retrospective digitisation leads to accessible scanned images only, which are enhanced by metadata for the access and a browsing structure for reading the paper like in the printed original. As an important enhancement it is necessary to improve the usability of the retro-digitised publications by introducing advanced linking and searching facilities. Making the text searchable is a first step. Providing reference links the next one. At present most of the current digital archives in mathematics and digitisation projects are providing such facilities or are on the way to install them. For example, on the open access level NUMDAM [4] has already installed these offers. They are partially available for the content digitised at CUL. The offer at THUL operates with the images only. Since most of the articles in their journals are in Chinese, the text search will be difficult to manage in a system where English is the lingua franca.

Since the report [11] has been written a lot of further digitisation activities had been initiated. Several mathematical journals in the USA added the digitised back volumes to their electronic offers. Under the supervision of KISTI 16 mathematical journals from Korea have been digitised. The Catalan Mathematical Society provides their three journals completely in digital form. Supported by DFG and RFBR (Russian Foundation of Basic Research) the project (see [12], [5]) has started in June 2004 with goal to develop a core for a virtual library of Russian mathematics. The major Polish journals in mathematics are offered by a digital

repository kept at the ICM in Warsaw. National digitisation projects are on their way or in discussion for Portugal, Spain, Italy, Switzerland, Czech Republic, Serbia and Bulgaria.

Finally big portions of mathematics are covered by JSTOR or packages of back files developed by commercial and academic publishers. Those from Elsevier appeared first, and there was a long discussion about the quality of the resolution of their initial scans. Just recently SIAM offered the back files of their journals in a package against charge including those again, which are already in JSTOR. Springer-SBM as an enterprise combining the previous Springer-Verlag and Kluwer Academic Publishers has set up a project to digitise all their journals. The package probably will be ready in the middle of 2005.

Taking all these developments together, not depending if they provide a charged service or open access, a big part of mathematics published in journals is available in digital form already. Hence they form a part of the WDML. It is an open question, if the WDML ever will be able to represent more than a collection of collections. It should be more than this and provide an integrated access. On the other side the reference databases provide an integrated access already. The voluntary cooperation in EMANI shows that an integration is possible. In that sense EMANI has a pioneering role for the WDML. But it should not be forgotten that at present a professional steering of the further development of the WDML does not exist, though the members of the CEIC do their best to move things forward.

References

[1] Liu, Guilin; Lisheng, Feng; Jiang, Airong; Zheng, Xiaohui: *The development of E-mathematics resources at Tsinghua University Library (THUL)*. Bai, Fengshan (ed.) et al., Electronic information and communication in mathematics. ICM 2002 international satellite conference, Beijing, China, August 29–31, 2002. Revised papers. Berlin: Springer. Lect. Notes Comput. Sci. **2730** (2003), 1–13.

[2] Asperti, Andrea; Wegner, Bernd: MOWGLI – *A new approach for the content description in digital documents*. Ninth International Conference "Crimea 2002" Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 8–16, 2002, Volume 1, 215–219.

[3] Becker, Hans; Wegner, Bernd: ERAM – *Digitisation of Classical Mathematical Publications*. Proc. ECDL 2000, Lecture Notes in Computer Science **1923** (2000), 424–427.

[4] Berard, Pierre: Presentation at the San Diego DML-meeting, Joint Mathematics Meeting, January 2002 (see also `http://www-mathdoc.ujf-grenoble.fr/NUMDAM/`).

[5] Evstigneeva, Galina A.; Wegner, Bernd: *Recent progress with RusDML*. See these Proceedings.

[6] Ewing, John: *Twenty Centuries of Mathematics: Digitizing and disseminating the past mathematical literature*. (See `http://www.ams.org/ewing/Twenty_centuries.pdf`).

[7] Fischer, Thomas; Pokutta, Sebastian; Törner, Günter: *TEXDocC: A Service Center for the Use of TEX Documents in Academia and Libraries*. See these Proceedings.

[8] Heinze, Joachim: Presentation at the EIC-Satellite Conference to the ICM 2002, Tsinghua University, Beijing.

[9] Wegner, Bernd: ERAM – *Digitalisation of Classical Mathematical Publications*. Seventh International Conference "Crimea 2000". Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 3–11, 2000, Volume 1, 268–272.

[10] Wegner, Bernd: ELibM *in* EMIS – *A Model for Distributed Low-Cost Electronic Publishing*. Eighth International Conference "Crimea 2001". Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 9–17, 2001, Volume 1, 317–320.

[11] Wegner, Bernd: EMANI – *a project for the long-term preservation of electronic publications in mathematics*. Bai, Fengshan (ed.) et al., Electronic information and communication in mathematics. ICM 2002 international satellite conference, Beijing, China, August 29–31, 2002. Revised papers. Berlin: Springer. Lect. Notes Comput. Sci. **2730** (2003), 178–188.

[12] Wegner, Bernd: DML *and* RusDML – *virtual library initiatives for covering all mathematics electronically*. Proceedings of the Digitisation Workshop at Bansko (Bulgaria), August/September, 2004.

Received November 25, 2004