



EMANI and related projects for the long-term preservation of electronic publications

Bernd Wegner

Scientific Co-ordinator, EMANI, Mathematics Institute, TU Berlin, Germany

Introduction

The internet has had a significant effect on the daily life of researchers, teachers and other academic information-using professionals. The most significant impact coming from the availability of ejournals and electronic versions of printed publications in digital libraries.

In addition to this, researchers and teachers are increasingly taking advantage of algebraic computer software, visualisation tools using graphics software and image processing tools.

Finally, the internet has opened up a communication infrastructure extending across distributed sites, allowing cooperation across and beyond the institution.

Electronic content is now readily available in many different formats (libraries, databases, indexing services, portals, web browsers) from a wide range of distributors (commercial publishers, learned societies, associations, single authors). In contrast to the "old world" of printed publications providers of econtent have different aims and it is not always clear from the user what to expect from new services. This is where libraries, perceived as the most reliable of 'information providers', must step in, managing electronic offerings to the same standard of service developed and maintained for print acquisitions.

There are many good reasons why scientific libraries particularly are especially keen to tackle the issues of storage, distribution and the long-term preservation of electronic publications:

- They understand the need for precise and reliable information access structures and systems.
- Free services remove any commercial constraints limiting the exchange of information across a wide community of users
- Scientific libraries may be subject specific but they must also be comprehensive in their holdings - resulting in a need for all acquisitions to be preserved and kept equally accessible despite the frequency of use.
- Scientific libraries have a three-fold obligation — to enable scientists to publish their work, to protect the information acquired, and to fulfil the needs of their users.

Mathematics and the need for long-term availability of resources

Mathematicians and professionals applying mathematics need quick, reliable and integrated access to mathematical publications, whether these are new ejournals or electronic versions of previously printed publications. Long-term availability is also particularly beneficial. Digitising print-only publications and assimilating these into an offering of ejournals raises problems to be solved not only with regard to access but as to how resources are displayed.

For non-mathematicians it is not clear why mathematics in particular requires the long-term preservation and availability of publications and resources — many in fact, do not understand what is involved in mathematical research and the unique way mathematical research is published.

When extensions and improvements to previous mathematical theorems are published they only include the additional work or achievement. The original theorem or previous thinking may be cited but proofs (the mathematical workings in detail) are not reprinted, even if the reading of these proofs are essential for understanding new results. Many proofs can be found in one place only. A new article therefore should be seen as the latest layer in a sequence of other articles which surround a central core of theorems, propositions, examples, models and proofs - all contributing to the current knowledge of a specific subject domain in mathematics.

[Monographs come closest to providing anything like a comprehensive review of the development of specific mathematical arguments. However, as can be seen by the variety of material found during research surveys in mathematics (the Itogi Nauki published by VINITI, for example), the more mathematical detail covered, the smaller the domain of knowledge which can be tackled.]

As a result mathematical research articles are commonly 'thin' and the references contained within essential to the complete understanding of the content, which is why past material must be preserved and made easily accessible.

In 2002 Joachim Heinze published the results of a review he had undertaken of three mathematical journals. One of the most surprising of his findings was the number of citations for articles published before 1992. In the case of the most traditional mathematical journal reviewed from North America, the *Annals of Mathematics*, 60 per cent of citations across 35 articles published in 2001 had a publication date before 1992. Of the 500 other journals referenced to the 2001 *Annals of Mathematics*, 82 per cent (4,500) of citations were for articles published before 1992. Reviews of other journals revealed similarly high incidents of pre-1992 citations: in the *Journal fuer die Reine und Angewandte Mathematik* 65 per cent of citations in 2001 volumes were prior to 1992, 61 per cent of citations in journals referenced to the publication; in the *Inventiones Mathematicae* 55 per cent of citations in 2001 volumes were prior to 1992 with 68 per cent of citations for articles before 1992 in journals referenced to the publication. Such high incidences of older citations are not common for most of the other sciences - it would be very interesting to have a more comprehensive comparison.

Current and future problems

The successful long-term preservation of printed publications may appear a simple task, but does in fact rely on maintaining two main functions: readability and access. In the printed world 'readability' can be affected by the deterioration of the paper on which text was written or the binding of a book or journal. Similarly, access can be affected by the distribution of texts over several locations to protect against the impact of fire or other catastrophes. In the digital world similar, yet more complicated considerations must be taken into account.

Digital versions of documents can become obsolete (as software programmes are upgraded or new electronic formats introduced) as can the systems in which electronic data is warehoused.

Current releases of software may only have a short life-time, what is to be done with the document after the means of access has expired?

Overcoming this problem is even more complicated when dealing with mathematical content as electronic resources are more likely to have software dependent enhancements. As digital preservation must go beyond protecting just the displayed text of the article, projects such as MoWGLI^[2] have begun work to develop different types of structures enabling the semantic mark-up of documents so that additional background information about the structure, author and location of resources can be preserved. These tools are in permanent evolution.

The best means of overcoming long-term preservation is to create an open and subject oriented approach and this is what the EMANI project has set out to achieve.

The EMANI project

It has taken ten years for electronic publications in mathematics to develop from pioneering freely accessible journals to a first class publication facility with enhanced services which can rival a traditional library of printed publications. Many digitisation projects (such as ERAM^[3] ^[8]) have increased the amount of previously print only content available

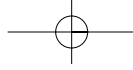
In the first half of 2001 the Electronic Mathematics Archives Network Initiative (EMANI) was founded as a special project to develop models for the archiving of electronic mathematical content. Understanding that a distributed architecture would be more suitable and reduce the load on the partners for such a project, a network was proposed, fostering a more open approach and allowing for potential scalability. By July 2002 the project was formalised during a workshop at Cornell University. The first set of project members are listed in the following section, the author of this article was made project coordinator.

At the core of the network a co operational system of reference libraries, content providers and editors has been set up. Their task is to store the digital content provided by participation publishers in a digital library repository. Alongside this content digitised versions of previously print only material are added. In the long run the repository aims to provide electronic versions of the majority of current and past mathematical publications.

During this endeavour, care has been taken to consider long-term preservation needs and the maintenance of content in a readable form. Simply storing content in a central repository would not solve the problems of long-term preservation. For example, at the most basic level, a print document scanned and saved in digital form in a library repository would have little usability. Plans were therefore made to introduce advanced linking and searching facilities, and to provide more convenient and affordable access to mathematicians and other professionals worldwide. Projects exploring the technological implications of such goals have begun.

It is hoped the network will serve as a reference system for other libraries which want to store or provide part of the content, or refresh their existing offers using updated material. When considering that publications provided through the network extend from the 19th century to current publications, it is clear a system of distribution agents will be needed. Extensions to the current group of providers may even bring older publications into the system. Overall a new business model for the distribution of content within a co operational enterprise between reference libraries and content providers is needed.

Online Information 2002 Proceedings



The starting point of EMANI: partners

It has been agreed that development must begin on a small scale first. Only after an architecture and action plan had been made sufficiently precise will an extension to the project scale be considered.

Initial libraries that are collaborating on the first steps to implement the initiative are:

- *Cornell University Library, Ithaca, New York*
Experienced in digitisation projects they also are involved archiving content for other sciences. In particular they are building a repository of mathematical eJournals for the Euclid project and serve as mirror site for EMIS [9].
- *State and University Library Goettingen*
Running some of the most important digitisation projects like ERAM (see [3] and [8]) and DIEPER. SUL Goettingen is also obliged to collect all publications in mathematics and as so have a high reputation as a centre for access to mathematical publications. They also serve as a mirror site for EMIS.
- *The Tsinghua University Library, Beijing*
This library has experience with the digitisation of Chinese publications. They are a Chinese centre of excellence for installing and offering electronic publications.
- *The Orsay Mathematical Library, Paris, in cooperation with the Cellule MathDoc in Grenoble*
The group in Orsay is coordinating a comprehensive consortium of French mathematical libraries. The strength of the partner in Grenoble consists in their excellent digitisation project NUMDAM [4].

Content providers who are collaborating from the start include Springer-Verlag; Birkhaeuser Verlag; Teubner Verlag; Vieweg Verlag and the electronic library ELibM offered through EMIS, the European Mathematical Information Service (<http://www.emis.de>). The four publishers involved have a long tradition in publishing mathematics and produce several of the best journals in mathematics. In contrast to this, the ELibM is a cooperation of journals and editors working on a voluntary basis, bundling electronic offers in a worldwide system of WWW-servers (see [9]).

The starting point of EMANI: first steps

Agreements on the architecture of the system had to be made. It was agreed the content stored would be copied and deposited in all the reference libraries forming the network as a matter of safety. Refreshed versions would then be exchanged accordingly. It was also made clear that all partners should work toward the improvement of the system, making known any developments which would further the aims of the whole project.

Next clarification was needed as to how available electronic content would be transferred from content provider to reference library. Libraries agreed to check if files received could be used for archiving and make adjustments where necessary. They also agreed to make recommendations to content providers where necessary, as to how they could develop their offerings to provide more convenient delivery.

Archiving related metadata needs also had to be defined, so that an integrated access structure could be built which would satisfy the needs of all kinds of expert users. Though links from the reference databases will satisfy some of the needs of users, the professional management of the archives will require more than just metadata-driven services.

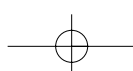
For the further progress of the EMANI project, working groups have been established to address specific tasks, such as, input formats; investigating MathML as a mark-up language. Progress made in these areas will then be used to build a more concrete version as a prototype for the long-term preservation of digital publications.

Some other archiving projects

To keep this paper brief only a rough survey of other archiving projects is given.

Clearly, there are several approaches in dealing with the digital archiving problem. Whatever the approach they must involve all the key stakeholders — publishers, libraries and authors. Thankfully only a few people still believe that simply copying and storing articles where they are posted is a promising solution and in 2001 several cooperative projects were launched looking to find more advanced solutions using different models. Most projects have been driven by libraries. Here are some examples:

- most prominent is the cooperative project between Elsevier and the libraries at Yale University which aims to digitally preserve all the publications of the publisher
- Harvard University is working on a similar project with Wiley, Blackwell Science and Chicago University Press



- LOCKSS is a system of archiving sites coordinated by Stanford University
- project Harvest, undertaken by Cornell University aims to archive agricultural publications
- MIT is dedicated its efforts towards the preservation of dynamic documents and publications including multimedia
- the New York Public Library is working on the digital preservation of arts journals
- and the American Institute of Physics and the American Physical Society have established an archiving system for their publications which involves the automatic conversion of files when a new release of the reader is distributed.

What all of these projects have in common is that they represent a first approach only - nobody has yet developed a comprehensive solution.

Digitisation projects in mathematics

When it comes to the digitisation and usage of older documents, searchability becomes the most important factor, otherwise how will researchers find their way around the huge knowledge base of mathematical achievements?

No current search engine is able to locate mathematical results. Abstracts and classification codes for specific subject areas can help considerably restrict the number of documents returned as relevant, but still it will be difficult for users to find specific resources relevant to his or her own research. Therefore literature databases (abstracts) for the classical period of mathematics are required. To include this functionality to modern mathematics databases is the starting point for the ERAM project; also known as the Jahrbuch project.

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 with full searchability and access through a database, captured from the "Jahrbuch ueber die Fortschritte der Mathematik" (1868-1943). The most comprehensive current literature database in mathematics, Zentralblatt MATH, was founded about ten years before the end of the Jahrbuch period. All data from the Jahrbuch will have been re-keyed by the end of 2002.

In the first instance new data has been made accessible on the web and though for many items enhancements such as English keywords are still missing, the database has found a lot of grateful users. The JFM database will provide access to a digital archive to be built up within the project. Currently there are no conversions of scanned documents into text files, which limits text searches across all data. However this is planned for a later phase of the project.

A first step in this direction has been made by a project based on the cooperation of experts from Japan, Germany and the United States [7].

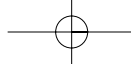
In cooperation with EMANI, ERAM has digitised all the back volumes of *Mathematischen Annalen* and *Mathematische Zeitschrift* and these journals have now been posted electronically. Most of the journals published electronically via the EMIS (European Mathematical Information Service) also agreed that to have their print-only back volumes digitised within the ERAM project, and now this also has been achieved. In ERAM, about 800,000 pages have been scanned so far and the project capacity is such to reach 1.2 million pages. For more details see the references [3] and [8] or the ERAM homepage <http://www.emis.de/projects/> clicking on the box for the Jahrbuch.

ERAM is part of a global initiative to make all mathematical information digitally available. This initiative is called the Digital Mathematical Library (DML). The only current activity has been much discussion and the formation of a planning group to work out how such a project could be arranged. DML has a lot of overlap with EMANI, but in contrast to EMANI, the global initiative will concentrate primarily on the preparation of digital versions of texts which are not yet digitally available. Long-term preservation is a secondary aspect of the DML at present. Clearly, in addition to ERAM there are several other digitisation projects on the way, general projects like JSTOR, DIEPER, and the Elsevier backfiles system, and projects in mathematics like NUMDAM [4] or the national heritage activity in Colombia by Victor Albis [1].

In 2001 John Ewing, Executive Director of the American Mathematical Society, prepared his White Paper [5] which serves as a basis for discussions on the DML. The article contained a lot of structural considerations for the DML and also addresses the immense implementation problems for 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 can be imagined and implemented because proprietaries 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 DML, as far as the costs will be concerned.

Online Information 2002 Proceedings



References

- 1 Victor Albis, Conservacion del patrimonio matematico colombiano.
<http://www.accefyn.org.co/historia-matematica/histmatcol.htm>;
<http://www.accefyn.org.co/proyecto/conservacion.htm>;
<http://168.176.37.80/matepro.html>
- 2 Andrea Asperti; Bernd Wegner: 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 Hans Becker, Bernd Wegner: ERAM - Digitisation of Classical Mathematical Publications, Proc. ECDL 2000, Lecture Notes in Computer Science 1923, 424-427 (2000).
- 4 Pierre Berard: Presentation at the San Diego DML-meeting, Joint Mathematics Meeting, January 2002 (see also <http://www-mathdoc.ujf-grenoble.fr/NUMDAM/>).
- 5 John Ewing: Twenty Centuries of Mathematics: Digitizing and disseminating the past mathematical literature. http://www.ams.org/ewing/Twenty_centuries.pdf
- 6 Joachim Heinze, presentation at the first EMANI workshop in Heidelberg, February 2002 (article to appear in the Proceedings of the EIC-Satellite Conference to the ICM 2002, Tsinghua University, Beijing)
- 7 Gerhard Michler: How to build a prototype for a distributed digital mathematics archive library. Proceedings MKM 2001, Linz, <http://www.emis.de/proceedings/MKM2001/>.
- 8 Bernd Wegner: ERAM - Digitalisation of Classical Mathematical Publications. Seventh International Conference Crimea 2000O 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.
- 9 Bernd Wegner: ELiBM in EMIS - A Model for Distributed Low-Cost Electronic Publishing. Eight International Conference Crimea 2001O 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

Contact

Professor/Dr Bernd Wegner
Fakultät II
Institut für Mathematik
TU Berlin
Skr. MA 8-1
Strasse des 17. Juni 135
D - 10623 Berlin
Germany

wegner@math.tu-berlin.de

