Lecture Notes in Computer Science 2730
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Fengshan Bai   Bernd Wegner (Eds.)

Electronic Information and Communication in Mathematics

ICM 2002 International Satellite Conference
Beijing, China, August 29-31, 2002
Revised Papers

Springer

Volume Editors

Fengshan Bai
Tsinghua University, Faculty of Mathematical Sciences
1301 New Science Building, Beijing 100084, China
E-mail: fbai@math.tsinghua.edu.cn

Bernd Wegner
Technical University Berlin, Faculty II - Mathematics and Sciences
MA 8-1, Straße des 17. Juni 136, 10623 Berlin, Germany
E-mail: wegner@math.tu-berlin.de

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): H.3, I.2, H.2.8, H.4, C.2.4, H.5

ISSN 0302-9743
ISBN 3-540-40689-1 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10929070 06/3142 5 4 3 2 1 0

Opening Address or
Greetings from Karlsruhe

Dear Colleagues,

This conference, a satellite to ICM 2002 in Beijing, on Electronic Inform
Communication in Mathematics, is an important event in building and
bridges between our globally distributed cultures. The conference to
very clearly that we have to include a whole set of aspects to be
covered in a digital environment, and the responsible scientific comm
both organizers did a good job building up a comprehensive program.

As a member of one of the conference sponsors I want to thank espe
hosts from Tsinghua University for their excellent organization, which
ficient room for fruitful discussions (and meals) and for fun during
events.

It is certainly well received that Springer-Verlag offered to publish these
ings in the series Lecture Notes in Computer Science, and I personally
you will well remember our time in Beijing when rereading the prese
and you will be reminded of some of your fellow participants.

Now it is already time to think about the follow-up workshop to see wha
has been achieved or where we should concentrate forces to get commo
improved services and communication possibilities. The next ICM is i
us!

With my warmes
Prof. Dr.-Ing. Georg F. S
Genera
Fachinformationszentrum

At the end of August 2002 a workshop on Electronic Information and nication in Mathematics took place at the Tsinghua University in Bei title was “Find and Post Mathematics in the Web,” we were trying t contributions from the most important groups worldwide, that are wo this subject. Having date and location tightly connected with the ICM ternational Congress of Mathematicians), organized by the IMU (Inte Mathematical Union) in Beijing, the workshop benefited significantly f ing the ICM participants nearby. In particular it was possible to attra members of the CEIC (Committee for Electronic Information and Cor tion), which is an IMU committee in charge of the topic of the worksho the workshop succeeded in giving a comprehensive survey on the state in electronic information and communication in mathematics.

The detailed topics covered by the presentations were: models and s for information and meta-information representation; methods and tool search, discovery, retrieval and analysis; access to distributed and heter digital collections (interoperability, scalability, relevant information meta-information integration); intelligent user interfaces in electronic agent technologies, co-operative work on the data; advanced technolog ital collection generation in mathematics; authoring tools and prepa electronic publications in mathematics; business models for the genera distribution of digital collections; and data security and information p The contributions dealt with both sides, the current practical work fo duction and distribution of electronic information in mathematics and tural investigations for their management.

The papers of this volume cover all these aspects, though it was no to include articles from all participants. Two additional articles were from authors who could not participate in the workshop, but where th of this volume had the impression that the articles dealt with an i aspect of the programme. Taking into account the rapid developme area of electronic information and communication in mathematics, th decided to publish the proceedings after the workshop. This gave us th to include the most recent developments, some even partially obtaine after the workshop.

As organizers of the conference the editors want to thank all who co to its success. All participants appreciated the hospitality of Tsinghua U Special thanks should go to the sponsors who made the financing of possible. Finally we are very glad that Springer-Verlag agreed to pu proceedings in their series Lecture Notes in Computer Science.

April 2003 Fengshan Bai, Bern

The Development of E-mathematics Resources at Tsinghua University Library (THUL)

Guilin Liu, Lisheng Feng, and Airong Jiang and Xiaohui Zheng

Tsinghua University Library, Beijing, China
{liugl,fenglsh,jiangar,zhengxh}@lib.tsinghua.edu.cn

Abstract. This paper presents the development of mathematic electronic resources at Tsinghua University Library (THUL). For better service aimed to high-level education and research, THUL has made a strategic adjustment to enhance electronic academic resources including input quite many fine selected foreign databases and full-text E-Journals, and made some efforts to digitalize featured mathematic materials.

1 The Strategic Adjustment of Collection Development at THUL

Since 1995 a strategic adjustment of collection development has been going on at THUL. The adjustment was concerned with 2 aspects: One aspect, patron's requirement increases rapidly both in collection and services along with a new goal of the university development — this challenge made up necessary of an adjustment. Another aspect, two favorable factors: Networking environment and large available commercial electronic resources and increasing library budget — these opportunities made possibility of a strategic adjustment. Let's have some explanation.

1.1 About Increase of Patron's Real Needs

Any substantial development of an university library depends essentially on the development of the university itself: the library should meet real needs of the faculty and students. Tsinghua University (THU) is one of a few top universities within almost 1,000 Chinese universities and colleges, especially in science and technology fields. However there is a long way to go to catch up world level teaching and scientific research. In 1994 THU set up a new goal: To be ranked as one of world-class universities by the year 2011 (100th anniversary of the university then). This goal raises tremendous pressure to the library.

At present THU consists of 12 schools or colleges with 48 departments covering most fields and disciplines with 139 of Master's Degree and 107 of PhD's degree (see Table 1). Table 2 shows the statistics of the full-time students. THU has a tradition of recruiting outstanding students. About half of top students from the yearly National-wide College Entrance Examination are usually recruited in THU. Table 3 shows the undergraduate enrollment statistics from

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 1–13, 2003.

the past 6 years. They might be the best "raw material" while on the other hand they strongly need much more library resources for their self-study and personality development.

Table 1. 12 Schools in THU

Sciences	Social Sciences
Information Tech.	Law
Mechanical Eng.	Economics & Manag
Civil & Hydr Eng.	Public Management
Architecture	Arts & design
Software Eng.	
Medical	

Table 2. Enrolled Full-Time Students (2001)

Total student number	24,063
Undergraduate	13,861
Graduate	10,202
(3,200 are Ph.D. candidates)	

Table 3. Enrollment statistics from the past 6 years

Year	No.1-ranked students in the National College Entrance Examination in 30 provinces	Top ten students in the National College Entrance Examination in 30 provinces
1995	17	165
1996	27	182
1997	24	157
1998	41	190
1999	38	196
2000	39	226
2001	31	214

THU is a research-oriented university. Table 4 and 5 show some information of the scholars and staff, the research institutions or key laboratories. In terms of research funding THU ranked No.1 among Chinese universities in past many years. The graduate school has also been recognized nationally, ranking No.1 in the National Evaluation of Graduate Schools. However, to achieve more world level scientific research results, faculty and graduate students are anxious to obtain more comprehensive periodicals, proceedings, books (especially foreign publications), and much more rapid and convenient services from the libraries.

Table 4. THU Faculty (2001)

Faculty members	3,060
Full professors	932
Associate professors	1,800
Members of the Chinese Academy of Sciences	25
Members of the Chinese Academy of Engineering	24

Table 5. THU Academic Research Units

Research Institutions	39
Research Centers	32
National Key disciplines	49
post-doc programs	24
National key lab.s	13
Key lab.s of MST	1
Key lab.s of MOE	10

More over, due to historical reason THUL lost all disciplines of pure sciences, social sciences, Arts and humanities in year of 1952 when a national-wide reconstructing of higher education was initiated; then THU became a multi-disciplinary polytechnic university and lost large part of associated collections. Therefore when all such disciplines were recovered or rebuilt since ’80 of last century there was a big lack of library resources in most non-engineering/technology fields. To meet the new university goal such big lack should be filled in as soon as possible. So shortly, how to provide patrons with much more and new academic information, especially foreign publications, and also provide much better services becomes serious challenges to THUL.

1.2 About Networking Environment

CERNET, China Education and Research NETworking, was put in operation in 1995. It is the first national-wide academic networking in China. Its administrative center is located at THU. Till Spring of 2002 CERNET has linked to 800 universities covering more than 200 cities in all 31 provinces, i.e 80% of higher education institutions in China have linked to CERNET. The bandwidth of the backbone among 8 big area-center (see Fig. 1) is 2.5 G, while the international export is close to 300 Megabyte.

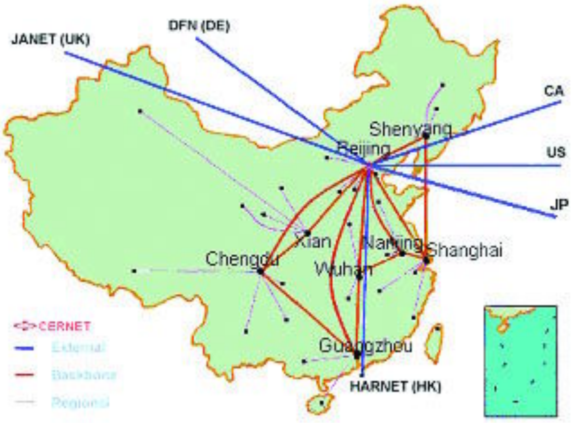


Fig. 1. The CERNET Topology Map

Tsinghua University Campus Networking (TH-CampusNet), starting in 1992, became more powerful after CERNET put in operation. Till Spring of 2002 it connects to more than 150 campus buildings with 1 Gbps speed including libraries, classrooms, laboratories, residential building and student dormitories (Fig. 2). Since more than 90% faculty and staff and more than 95% full time students live in the campus they may access to library E-resources from their own offices or apartments or living rooms via the TH-CampusNet.

1.3 Another Favorable Factor Is Budget Situation

Owing to “rely on Science and Technology” policy, Chinese government has been giving more financial support to higher education, especially to some key universities since 1995. As one of the results THUL’s budget has been obviously increased in past 6 years supporting to develop our library networking, to upgrade our library integration system into INNOPAC, to improve other related information infrastructure and also to purchase more collections. In period of 1999–2001 the yearly budget for collection development in THU was more than double comparing with the former 3 years.

1.4 Strategetic Adjustment of Collection Development at THUL

Considering the challenges and opportunities analyzed above THUL had two choices in working out its strategy for collection development: One was to in-

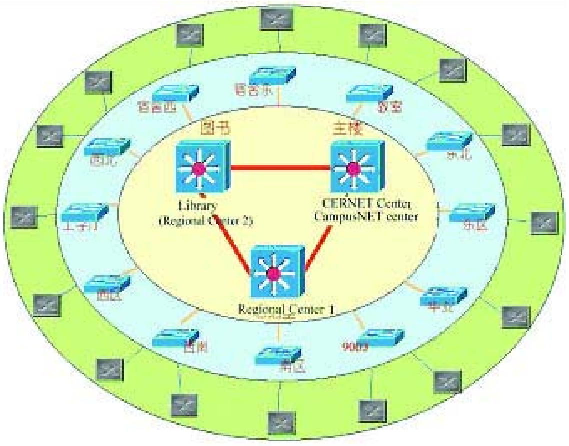


Fig. 2. TH-CampusNet Most faculty and students live in the campus

crease much more printed foreign Journals and proceedings which usually spent 2/3 of the total library budget. However even if double or 3 times of them up to 5–6 thousands titles that would be still a small number comparing with advanced universities abroad. Another choice is to pay more attention to, or give priority to electronic resources and Net-based services.

THUL has taken the later one as its strategy and has been successful till now.

1. In period of 1999–2001 the budget for E-resources has been up to 30% of the total budget of collection development.
2. For secondary literature, Abstract and Index databases, have already covered all research fields of THU, and most of them are in web-services via THCampusNet, such as OCLC FirstSearch, EI village, SCI, SSCI, A&H-Web of Knowledge, PQDD, etc. In math field, also include MathSciNet, Zentralblatt Math.
3. For first literature, THUL's strategy for providing full text consists of 3 points.
   (a) Keep certain the most important or frequent-use foreign printed Journals (Chinese printed Journals never be a problem) as a base;
   (b) Give more priority to increase foreign electronic resources;
   (c) Enhance ILL.

THUL conducted a large scale investigation in 1997 and 1998 to faculty members – asking them for selecting 10–20 titles of foreign printed Journals which are the most important or/and frequently-used they did think for respective discipline, and ensured that the library would keep those as much as possible. As a result of a summary of all feedback THUL got a solid baseline of the subscription of foreign printed Journals, about 2,000 titles. Taking these printed as a baseline THUL gave more priority to increase foreign electronic resources. At present the number of titles of full text foreign Journals is over 7,400, about 4 times of foreign printed Journals. They involve many famous commercial databases including Elsevier's SDOS, IEEE/IEE, Academic Press, EBSCO, JSTOR, Springer-Link etc. There are 5,600 titles of Chinese E-journals integrated in a commercial product CJN (see text below). The total full-text e-journals is over 13,000.

This strategic adjustment has not met any obvious obstacles from professors or from library staff at THU, which had ever happened in some western university libraries. Indeed, THU faculty, including old professors, warmly welcome large amount of E-resources because 1) the existed printed collection at THUL was far from their needs, a large scale increase of E-Journals may meet their requirements; 2) E-databases are much convenient in use; 3) THUL provides with ILL luckily the China National Library, the Library of Chinese Academy of Sciences, the National Library for Sciences and Technology, and many key university libraries including Peking University Library are just located nearby, within a range of 10 kilometers. These libraries host huge collections including about 20,000 titles of printed foreign academic periodicals.

In aspect of input commercial foreign e-databases THUL usually unites other university libraries (sometimes the National library and the library of Chinese academy of sciences also be joined in) to form different consortia for better price and services from vendors. THUL also host several mirror sites, maintaining their servers and serving all members of respective consortium. Tab. 6 lists some examples to show enlarging member's number of respective consortium from time to time.

Table 6. THUL organizes consortia and number's increase

Setting Up Mirror Sites at THL	
Ei Village (1998):	11 —> 18—>28 —>31—>47
Elsevier Science (2000.3):	11—>28
CSA (1999.12):	9—>—>16—>
INSPEC (2002):	—>20
Springer Link (2002.10):	320
ACM (undergoing)	
Remote Access via Digital Island	
PQDD (2000.11.1):	—>34—>39
IEEE/IEE Electronic Library (2001.1):	—>12—>19
OCLC FirstSearch (1999.10.1):	—>62

Why set up mirror-sites? The reason is that according to present policy of CERNET although a campus user needn't pay for browsing and downloading from Internet within China and even sending message to abroad, but one should pay certain fee if receive information from abroad. In China's situation tuition fee usually is very low, may not cover such Internet fee. However to charge such kind of fee will greatly reduce student's usage of those foreign e-databases for which libraries have already paid much for subscription! Besides, how to charge each individual on site-on line is also a problem from technology view.

All practice of us till now has demonstrated that in aspect of input of foreign e-databases the scheme of "user consortia plus certain mirror sites" makes much lower cost and high usage; suits present china's situation.

A recent example is the Springer-Link China Consortium (SLCC) and related mirror site which locates at THUL. 320 of higher educational institutions and research Institutions over china (mainland) have already involved in this consortium. Students and faculties in each institution can access to the mirror server for unlimited use of Springer Link product 490 titles of Journals covering from 1998–now, including 82 titles of Math Journals-access controlled by respective IP address-section.

In aspect of Chinese electronic full text journals, CJN (Chinese Journals Networks) provides with E-version of 5,600 titles, most of current academic printed journals, covering all subjects, from 1994 to now (all back files will be followed later) as a web-based, daily upgrade (via internet) collection. A CJN mirror was set up at THUL for unlimited use within the campus.

2 Existing Electronic Math Resources at THUL

As described above, THUL subscribed 7,400 electronic full-text foreign journals and 5,600 Chinese journals. Among these foreign journals, 309 titles are relevant to mathematics as shown in Table 7; while among these Chinese journals, 58 titles are relevant to mathematics which has covered most, nearly all important Chinese math academic journals with 70,000 articles. Besides, many other math articles are dispersed in more than 100 university Journals. It should be mentioned that THUL also hosts a mirror site of Zentralblatt Math which contains articles from about 2,000 journals and serials, books, available proceedings of all conferences and other research papers appearing in any language. Also, users in THU can obtain useful information by remote access to MathSciNet database.

3 A Research Project of Chinese Math Collection Digital Library (CMDL)

Additionally to subscribed foreign and native electronic resources as stated above, THUL started to develop featured digital resources 3 years ago including THU dissertation (full-text) database, database for dissertation abstract of Chinese key universities, some books and periodicals associated with THU history, and a research study of Tsinghua Architecture Digital Library (THADL)[1]. THADL has digitized large part of the materials collected by SRCA – the Society for Research in Chinese Architecture which was established in 1930's by the famous professors Zhu Qiqian, Liang Ssu-ch'eng and Liu Dunzhen. They

Table 7. Foreign E-Mathematics Journals at THUL

Name of E-Database	Amount of Math E-Journals involved
Academic Press	22
Elsevier Science	53
Kluwer online	61
Springer link	68
JSTOR	15
EBSCO-Academic Search Premier	35
UMI-Academic Research Library	32
IEEE/IEE Electronic Library	9
American Physical Society	6
IOP Electronic Journals	4
Others	4
In Total	309

created the new way of Chinese architecture research which became a basis of the whole modern Chinese architecture science. SRCA took about 15 years in 1930's–40's to investigate 2,783 Chinese ancient architectures on site all over China, and achieved much rare materials including photographs, sketches drawing, etc. THADL is fulfilled by a cooperative group coming from 3 THU's units: library, dept. of computer science, school of Architecture.

Based on these experience THUL plans to have another research proposal: CMDL (Chinese Math Digital Library). In our proposal CMDL would be a collective study with THU's units (the library, dept of math. and dept. of computer science), math research institutions in China (Chinese Math Society, a few key universities) and EMANI (Electronic Mathematic Archiving Network Initiative)-as a partner of the international research project. The infrastructure of CMDL should be distributed and open. The proposed contents of CMDL may cover Chinese ancient math works, Chinese modern math publications and other featured collections.

3.1 Existing Chinese Ancient Mathematical Works B.C.2060–1911: Commentary and Amount

China, as one of the four ancient civilizations in the world, once made splendid achievements in science and technology including mathematics. China was the first country in the world to adopt the decimal place-value system. Since about 4th century B.C., the decimal place-value system and the fractions for the four arithmetical operations was used proficiently in ancient China. Some foreign scholars have given a high praise on ancient Chinese mathematics. D. E. Smith, an American historian of mathematics, remarked that China was among the pioneers in the establishing of the early science of mathematics[2]. Y. Mikami, a Japanese historian of mathematics, wrote that "It is for more than two thousand years that the development of mathematics had lasted in Chinese. The situation did not appear in other countries over the world. Mathematics developed in Greek merely for one thousand years from the 6th century B.C. to the 4th century A.D. In Arabic world it only flourished during the period from the 8th to the 13th century. Mathematics in today's European countries began in the 10th century. We can not say that the development of mathematics in China is not a rare example in the history of mathematics." [3].

The Earliest mathematical classic is the *Suan Shu shu* (*A Book on Arithmetic* , before 186 B.C.,). It is a book made of bamboo strips and was found in 1984 in a place near Jiangling in Hubei province[4], see Fig 3.

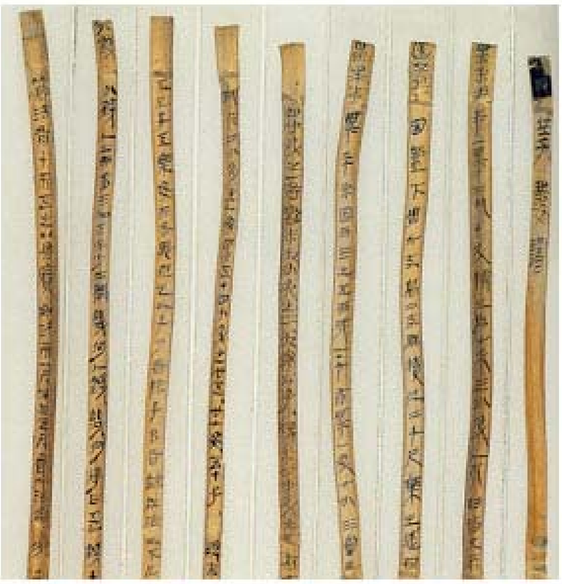
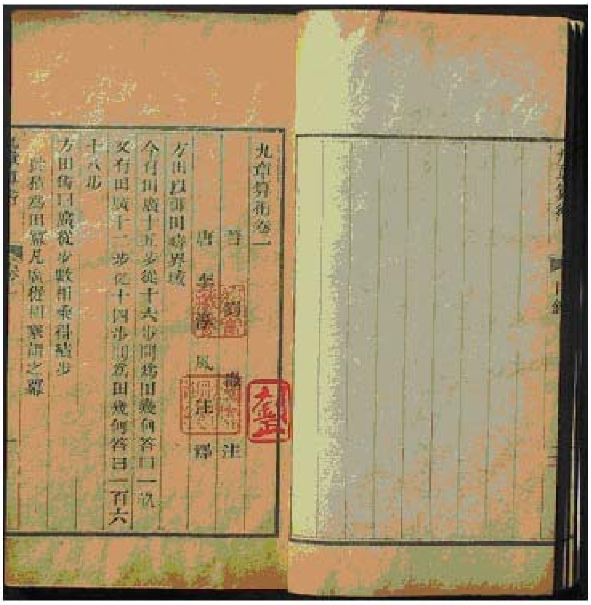


Fig. 3. Suan Shu shu (A Book on Arithmetic)

Jiu Zhang Suan Shu (Nine Chapters on the Mathematical Art) (Fig. 4), completed in the 1st century, is a well-known scientific work in the world. The Nine Chapters on the Mathematical Art played a central role in ancient Chinese and East Asia mathematics, somewhat similar to Euclid's Elements in the West. Negative numbers and their computation, plus solutions for simultaneous linear equations appeared in this book. It was one of the most advanced works on mathematics at that time. B. L.van der Waerden (in Mathematisches Institut der Universität Zürich), a famous modern mathematician, wrote in his book:

"It seems that the most faithful reflection of this oral tradition is found in the Chinese 'Nine Chapters'. Here we find a complete explanation of calculations in decimal system, including the simplification of fractions m/n by means of the Euclidean algorithm. This algorithm was known in China and Greece, but no



九章算術卷一

晉 劉徽 注

唐 李淳風 注釋

方田以御田疇界域

今有田廣十五步從十六步問爲田幾何答曰一畝

又有田廣十二步從十四步問爲田幾何答曰一百六十八步

方田術曰廣從步數相乘得積步

此積爲田冪凡廣從相乘謂之冪

Fig. 4. Jiu Zhang Suan Shu (An edition of the earlier time of Qing Dynasty, stored in Tsinghua University Library)

trace of it is found in any one of our numerous Babylonian and Egypt texts. In the "Nine Chapters" we also find a systematic method for extracting square roots and cube roots, and a completely clear exposition of the matrix method for solving sets of linear equations in an arbitrary number of unknowns. In these respects the Chinese text is unique." [5] From Qin Dynasty (B.C. 221–B.C.206) to the year 1911 many excellent mathematicians had written numerous treaties and contributed greatly to the development of mathematics both in China and in East Asian. Around 2,600–2,700 titles of traditional Chinese mathematical works were recorded. However, only 1,910 books (some other different editions of the same book are not counted) are existent. In year of 2000, Li Di, together with others, published a union catalogue, which covers both Chinese mathematical works stored in Chinese libraries and some European books translated into Chinese by either the 17th-century Jesuits or later foreign scholars. This catalog embodies the catalogs appeared from 186 B.C. to 1911[6]. Of total 1910 titles, Tsinghua University Library has owned 240 titles, 942 volumes (15 titles of rare book, 136 vol.s in total). Additionally, there exist around 1,000 mathematical works published in period of 1912-1949 excluding all kinds of textbooks for the primary and middle school.

3.2 CMDL's Scheme and Its Recent Progress

We have made a careful investigation on existing Chinese ancient math works in different historical periods, see Table 8.

CMDL will finally cover abundant ancient mathematical works and modern mathematical resources as much as possible. However as a research project we plan to deal with about 100 titles of typical works including 12 works of Han to Tang Dynasty and 20 works of Song & Yuan Dynasty as first step. 70 of them exist at THUL. Various problems should appear in process, such as CMDL metadata scheme, special OCR issue raising from Chinese ancient mathematical books (say, handwritten Chinese characters recognition), large Chinese character library, etc.

Table 8. Numbers of Existing Ancient Chinese Mathematical Works

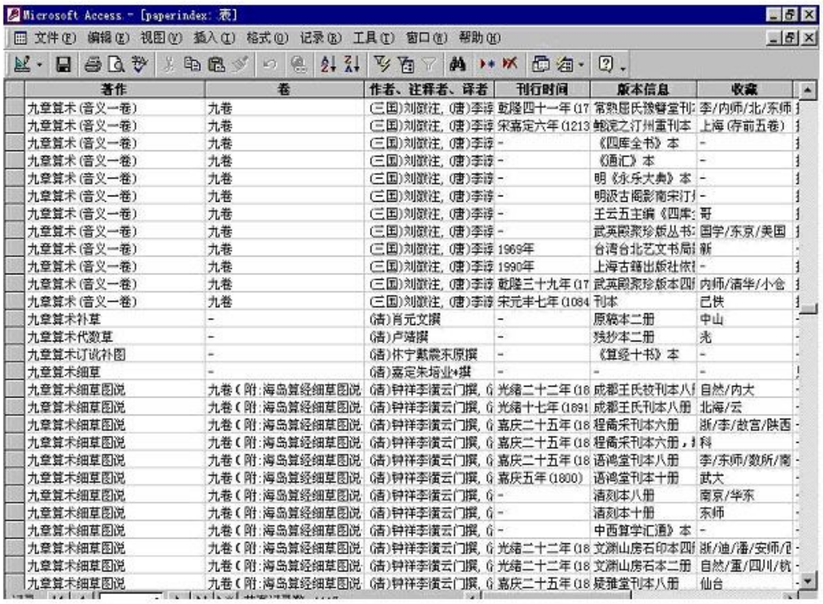
in Different Historical Periods (Amount to 1,910)

Han to Tang	Song to Yuan	Ming Dynasty	Qing Dynasty
(B.C.206–907A.D.)	960–1368	1368–1644	1644–1911
12 books	20 books	about 50 books	about 182

To facilitate use of investigated information two databases have been completed on a web platform. The first is a retrieval system for catalog which is based on Li Di's work[6]. This catalog database includes over 2000 titles with entries of volume number, author or commenter or translator, edition, publishing time and location (Fig. 5). Abstract will be added later. The second database is a reference database of research literature on Chinese mathematical works, including information of title, author, related periodical, publishing time of over 1400 research papers published in the period from 1906 to 1985 in Chinese. (Fig. 6).

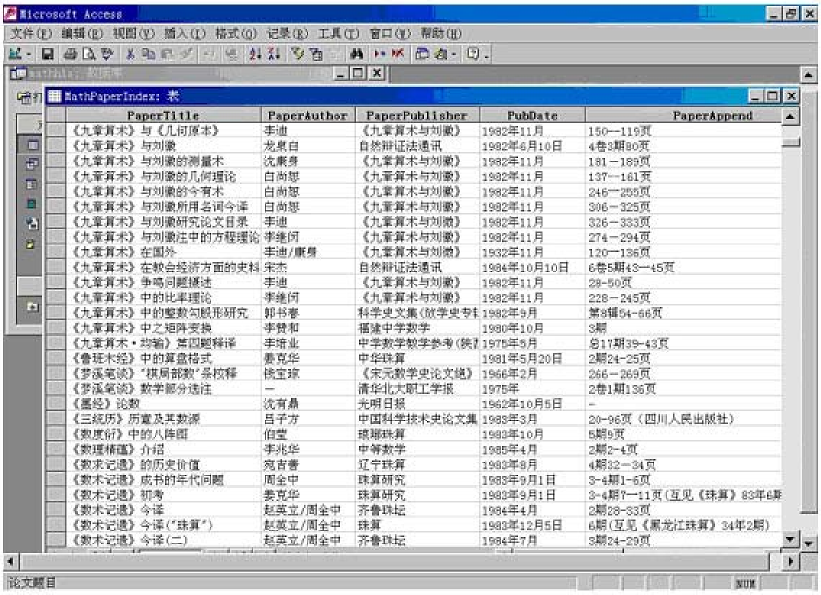
One edition of the *Nine Chapters on the Mathematical Art* has been scanned and made to be a full-text retrieval database.

To facilitate information exchange with other EMANI partners we should adopt national or international standards and intend to follow widely accepted methods. An extensive investigation has shown that several best practices are very useful for CMDL preservation system. For instance, OAIS reference model can be adopted as the system architecture and METS as the Metadata encoding standard. Table 9 is a preliminary metadata framework we developed for CMDL. Among the 5 modules, TechMD (technical metadata) is most elaborate, because the preservation function of the framework mainly lies on this module. It has


Microsoft Access - [paperindex: 表]
文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 记录(R) 工具(T) 窗口(W) 帮助(H)

著作	卷	作者、注释者、译者	刊行时间	版本信息	收藏
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	乾隆四十一年(17	常熟屈氏聚譬堂刊:	李/内师/北/东师
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	宋嘉定六年(1213	鲍浣之汀州重刊本	上海(存前五卷)
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	-	《四库全书》本	-
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	-	《通汇》本	-
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	-	明《永乐大典》本	-
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	-	明汲古阁影南宋汀	-
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	-	王云五主编《四库:	哥
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	-	武英殿聚珍版丛书:	国学/东京/美国
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	1969年	台湾台北艺文书局	新
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	1990年	上海古籍出版社依	-
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	乾隆三十九年(17	武英殿聚珍版本四	内师/清华/小仓
九章算术(音义一卷)	九卷	(三国)刘徽注, (唐)李淳	宋元丰七年(1084	刊本	已佚
九章算术补草	-	(清)肖元文撰	-	原稿本二册	中山
九章算术代数草	-	(清)卢靖撰	-	残抄本二册	光
九章算术订讹补图	-	(清)休宁戴震东原撰	-	《算经十书》本	-
九章算术细草	-	(清)嘉定朱培业+撰	-	-	-
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	光绪二十二年(18	成都王氏校刊本八	自然/内大
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	光绪十七年(1891	成都王氏刊本八册	北海/云
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	嘉庆二十五年(18	程矞采刊本六册	浙/李/故宫/陕西
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	嘉庆二十五年(18	程矞采刊本六册,	科
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	嘉庆二十五年(18	语鸿堂刊本八册	李/东师/数所/南
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	嘉庆五年(1800)	语鸿堂刊本十册	武大
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	-	清刻本八册	南京/华东
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	-	清刻本十册	东师
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	-	中西算学汇通》本	-
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	光绪二十二年(18	文渊山房石印本四	浙/迪/潘/安师/
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	光绪二十二年(18	文渊山房石本二册	自然/重/四川/杭
九章算术细草图说	九卷(附:海岛算经细草图说	(清)钟祥李潢云门撰, (	嘉庆二十五年(18	疑雅堂刊本八册	仙台

Fig. 5. Catalog Database of Traditional Mathematical Works (example)

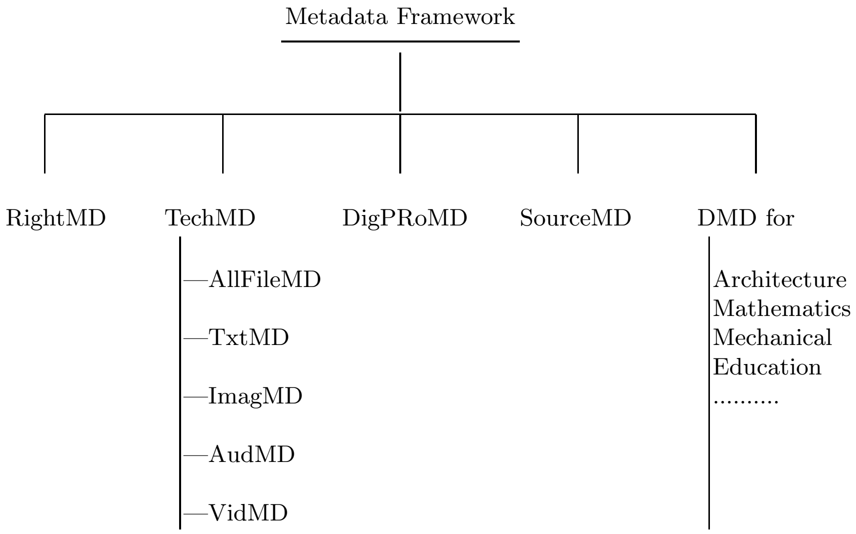

Microsoft Access
文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 记录(R) 工具(T) 窗口(W) 帮助(H)
MathPaperIndex: 表
论文题目
NUM

PaperTitle	PaperAuthor	PaperPublisher	PubDate	PaperAppend
《九章算术》与《几何原本》	李迪	《九章算术与刘徽》	1982年11月	150—119页
《九章算术》与刘徽	龙杲白	自然辩证法通讯	1982年6月10日	4卷3期80页
《九章算术》与刘徽的测量术	沈康身	《九章算术与刘徽》	1982年11月	181—189页
《九章算术》与刘徽的几何理论	白尚恕	《九章算术与刘徽》	1982年11月	137—161页
《九章算术》与刘徽的今有术	白尚恕	《九章算术与刘徽》	1982年11月	246—255页
《九章算术》与刘徽所用名词今译	白尚恕	《九章算术与刘徽》	1982年11月	306—325页
《九章算术》与刘徽研究论文目录	李迪	《九章算术与刘徽》	1982年11月	326—333页
《九章算术》与刘徽注中的方程理论	李继闵	《九章算术与刘徽》	1982年11月	274—294页
《九章算术》在国外	李迪/康身	《九章算术与刘徽》	1932年11月	120—136页
《九章算术》在教会经济方面的史料	宋杰	自然辩证法通讯	1984年10月10日	6卷5期43—45页
《九章算术》争鸣问题概述	李迪	《九章算术与刘徽》	1982年11月	28-50页
《九章算术》中的比率理论	李继闵	《九章算术与刘徽》	1982年11月	228—245页
《九章算术》中的整数勾股形研究	郭书春	科学史文集(数学史专	1982年9月	第8辑54-66页
《九章算术》中之矩阵变换	李赞和	福建中学数学	1980年10月	3期
《九章算术·均输》第四题释译	李培业	中学数学教学参考(陕	1975年5月	总17期39-43页
《鲁班木经》中的算盘格式	娄克华	中华珠算	1981年5月20日	2期24-25页
《梦溪笔谈》"棋局都数"条校释	钱宝琮	《宋元数学史论文集》	1966年2月	266—269页
《梦溪笔谈》数学部分选注	—	清华北大职工学报	1975年	2卷1期136页
《墨经》论数	沈有鼎	光明日报	1962年10月5日	-
《三统历》历意及其数源	吕子方	中国科学技术史论文集	1983年3月	20-96页(四川人民出版社)
《数度衍》中的八阵图	伯堂	琼期珠算	1983年10月	5期9页
《数理精蕴》介绍	李兆华	中等数学	1985年4月	2期2-4页
《数术记遗》的历史价值	宛吉春	辽宁珠算	1983年8月	4期32—34页
《数术记遗》成书的年代问题	周全中	珠算研究	1983年9月1日	3-4期1-6页
《数术记遗》初考	娄克华	珠算研究	1983年9月1日	3-4期7—11页(互见《珠算》83年6期
《数术记遗》今译	赵英立/周全中	齐鲁珠坛	1984年4月	2期28-33页
《数术记遗》今译("珠算")	赵英立/周全中	珠算	1983年12月5日	6期(互见《黑龙江珠算》34年2期)
《数术记遗》今译(二)	赵英立/周全中	齐鲁珠坛	1984年7月	3期24-29页

Fig. 6. Reference Database of Research Literature on Chinese Mathematical Works

5 submodules: AllFileMD, TxtMD, ImagMD, AudMD, and VidMD. AllFileMD includes technical metadata common to all kinds of digital files. Each of the other 4 sub modules of TechMD relates to technical features specific to one of the following types of file: text files, image files, audio files or video files. For DMD (descriptive metadata) and RightMD that are heavily influenced by legal and cultural environments, we'd better make an agreement on a minimal list of core elements for EMANI project.

Table 9. CMDL Metadata Framework



References

[1] Xing Chunxiao, etc: THADL: A Digital Library for Chinese Ancient Architecture Study, Global Digital Library Development in the New Millennium – Fertile Ground for Distributed Cross-Disciplinary Collaboration. Tsinghua University Press. (2001)

[2] Smith, D.E.: History of Mathematics, Vl.1, Ginn, New York, p. 33

[3] Mikami, Y.: Special Characteristics of Chinese Mathematics. Toyo Gakuho (Reports of the Oriental Society of Tokyo), **15**(no.4) (1926) period for nonautonomous Hamiltonian systems.

[4] Peng Hao: Commentary of Suan Shu Shu in Han Dynasty's bamboo strips excavated from Zhang-jiashang, Science Press, 2001. Color Photo. 2

[5] Van der Waerden, B. L.: Geometry and Algebra in Ancient Civilizations. Springer. (1983)

[6] Li Di (ed.): Union Catalogue of Chinese Mathematical Books. App. Vol. 2. Beijing Normal University Press.(2000)

MOWGLI – An Approach to Machine-Understandable Representation of the Mathematical Information in Digital Documents

Andrea Asperti[1] and Bernd Wegner[2]

[1] Dipartimento di Scienze dell Informazione, Universita degli Studii di Bologna
Via di mura Anteo Zamboni VII, I - 40127 Bologna, Italy
[2] Fakultät II, Institut für Mathematik, TU Berlin
Sekr. MA 8-1, Straße des 17. Juni 135, D - 10623 Berlin, Germany

Abstract. The acronym MOWGLI stands for "Mathematics On the Web: Get it by Logic and Interfaces". MOWGLI is a European Project founded by the European Community in the "Information Society Technologies" (IST) Programme. The partners are the University of Bologna, INRIA (Rocquencourt), the German Research Centre for Artificial Intelligence (DFKI, Saarbrücken), the Katholieke Universiteit Nijmegen, the Max Planck Institute for Gravitational Physics (Albert Einstein Institute, Golm), Trusted Logic (Paris) and TU Berlin.

The aim of the project is the study and the development of a technological infrastructure for the creation and maintenance of a virtual, distributed, hypertextual library of mathematical knowledge based on a content description of the information. Currently, almost all mathematical documents available on the Web are marked up only for presentation, severely crippling the potentialities for automation, interoperability, sophisticated searching mechanisms, intelligent applications, transformation and processing. The goal of MOWGLI is to overcome these limitations, passing from a machine-readable to a machine-understandable representation of the information, and developing the technological infrastructure for its exploitation.

The project deals with problems traditionally belonging to different scientific communities: digital libraries, Web publishing, automation of mathematics and computer aided reasoning. Any serious solution to the complex problem of mathematical knowledge management needs a coordinated effort of all these groups and a synergy of their different expertise. MOWGLI attempts to build a solid co-operation environment between these communities. The current paper will describe the objectives and first achievements of MOWGLI in some detail.

1 Aims and Mission of MOWGLI

After a ten years period of electronic publishing in mathematics we are still confronted with slightly enhanced electronic versions of printed publications. Almost all mathematical documents available on the Web are marked up only for presentation, if such an enhancement is available at all. Only a minority of documents

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 14–23, 2003.

try to care about some of the potentialities for automation, interoperability, sophisticated searching mechanisms, intelligent applications, transformation and processing. But these approaches could be considered as first preliminary steps towards an electronic document providing all these facilities. Hence, the goal of MOWGLI is to overcome these limitations, passing from a machine-readable to a machine-understandable representation of the information, and developing the technological infrastructure for its exploitation.

In order to reach this goal MOWGLI has to deal with problems traditionally belonging to different scientific communities: digital libraries, Web publishing, automation of mathematics and computer aided reasoning. To our knowledge, MOWGLI is the first attempt to build a solid co-operation environment between these communities. In principle, any serious approach for providing good tools for mathematical knowledge management needs a co-ordinated effort of several partners from the above-mentioned communities and a synergy of their different expertise. The choice of partners for MOWGLI took this condition into account, as can be seen in Section 4.

The goals of MOWGLI largely overlap with the aims of the so-called “Semantic Web” [14]. Associating meaning with content or establishing a layer of machine-understandable data will allow automated agents, sophisticated search engines and interoperable services and will enable higher degree of automation and more intelligent applications. The ultimate goal of the Semantic Web is to allow machines to share and exploit knowledge in the Web way, i.e. without central authority, with few basic rules, in a scalable, adaptable, extensible manner. However, the actual development of the Semantic Web and its technologies has been hindered so far by the lack of large-scale distributed repositories of structured content oriented information. The case of mathematical knowledge, the most rigorous and condensed form of knowledge, is paradigmatic. The World Wide Web is already now the largest single resource of mathematical knowledge, and its importance hopefully will be increased by the emerging display technologies like MathML.

Machine-understandable information will make possible to offer added-value services like:

- Preservation of the real informative content in a highly structured and machine-understandable format, suitable for transformation, automatic elaboration and processing.
- Cut and paste on the level of computation (take the output from a Web search engine and paste it into a computer algebra system).
- Automatic proof checking of published proofs.
- Semantic search for mathematical concepts (rather than keywords).
- Indexing and Classification.

Due to its rich notational, logical and semantic structure, mathematical knowledge is a main case study for the development of the new generation of semantic Web systems. The aim of the MOWGLI project is both to help in this process, as well as pave the way towards a really useful virtual, distributed, hypertextual resource for the working mathematician, scientist or engineer.

2 Standards and Tools

Current standards for electronic publishing in mathematics are mainly presentation oriented. New tools for the management and publishing of mathematical documents are in development like MathML [3], OpenMath, OMDoc ([17],[18]) and integrated with different XML technology [7] (XSLT [8], RDF [4], [5], SOAP [6]). All these languages cover different and orthogonal aspects of the information and its management; the aim of MOWGLI is not to propose a new standard, but to study and to develop the technological infrastructure required for taking advantage of the potentialities of all of current standards and those, which are likely to be established in the near future.

MOWGLI makes an essential use of standard XML technology, aspires to become an example of "best practice" in its use, and a pioneering leading project in the new area of the Semantic Web [12]. In particular, the potentialities of XML will be deeply explored in the following directions:

- Publishing. XML offers sophisticated publishing technologies (Stylesheets, MathML, SVG, etc.), which can be profitably used to solve, in a standard way, the annoying notational problems that traditionally afflict content based and machine-understandable encoding of the information.
- Searching and Retrieving. Metadata will play a major role in MOWGLI. New W3C languages such as RDF (Resource Description Framework) or XML Query are likely to produce major innovative solutions in this field.
- Interoperability. Disposing of a common, machine-understandable layer is a major and essential step in this direction.
- Distribution. All XML technology is finally aimed to the access of the Web as a single, distributed resource, with no central authority and few, simple rules.

MathML [3], introducing for the first time a content mark-up layer in parallel with a presentational one, has indubitably been a pioneering project towards the mining of the mathematical treasure available on the Web. Still, its limitations are evident as well:

- MathML is merely focused on mathematical expressions. However, in order to bring the idea of a Semantic Web of Mathematics to its full potentialities, other layers of mathematical information must be considered as well. In particular, we need a clean, microscopic description of proofs, a mark-up for mathematical "objects" (theorems, lemmas, corollaries, examples, etc.), a mark-up for "structured collections" of these objects (documents, theories, etc.), possibly "functors" between these collections, and finally a good "metadata" layer.
- MathML is just a (important) piece in a much wider technological puzzle. Passing from content to a good presentational format requires sophisticated operations; on the other side, these transformations are themselves a basic component of the whole mathematical knowledge (like mathematical fonts). XSLT [8] provides here the right technology, opening the way to the creation

of well maintained and documented libraries of mathematical style-sheets [11].

Similarly, the creation and maintenance of the library as a distributed repository, and the crucial aspect of managing the information in the "Web way" requires a light but powerful communication protocol, overcoming some of the limitations of HTTP (SOAP [6] looks as a promising solution).

Metadata will eventually require a fairly sophisticated model, much beyond what is currently offered by typical metadata models as the Dublin-Core system [1]. Here, RDF (Resource Description Framework, [4], [5]) looks as the right framework for developing the model, providing a general architectural model for expressing metadata and a precise syntax for the encoding and interchange of these metadata over the Web.

The fact of encoding also the microscopic, logical level of mathematics opens the possibility to have completely formalised subsystems of the library ([9],[10]), which could be checked automatically by standard tools for the automation of formal reasoning and the mechanisation of mathematics (proof assistants and logical frameworks ([15],[16]). At the same time, any of these tools could be used as an authoring system for documents of the library, by simply exporting their internal libraries into XML, and using style-sheets to transform the output into a standard, machine-understandable representation, such as MathML content mark-up or OpenMath. In MOWGLI we shall use the Coq Proof Assistant of INRIA [13] as a paradigmatic example of these applications.

An alternative route for the creation of content-based mathematical information from standard digital repositories by means of a suitable LaTeX-based authoring system will be explored by the Albert Einstein Institute. They publish the "Living Reviews in Relativity" [2], a solely electronic journal on the Web, which provides refereed, regularly updated review articles on all areas of gravitational physics. AEI will develop a LaTeX-based authoring tool interfacing with MOWGLI, and serve as a showcase to demonstrate how content-mark-up in mathematics improves the usability and information depth of electronic science journals.

3 A Minimal Technological Infrastructure

It is clear that the creation and maintenance of large repositories of content-based mathematical knowledge can only be conceived as a cooperative and distributed process, comprising not only the creation of documents, but also libraries of notational rules, metadata and management tools. The crucial point is to build a minimal infrastructure to start up this process, so that more and more tools can be added by interested parties. All these considerations lead to two basic requirements for the developments in MOWGLI:

– Information must be accessible with few basic rules an no central authority (the Web way).

– Make extensive use of standard XML technology and tools, even when it would be easier or more efficient just to develop an ad-hoc solution.

In this way, we put no barrier to third party development and, every time a standard technology or tool is improved, we can simply benefit of the new implementation with minimal effort.

The MOWGLI architecture is essentially based on three components, which are distribution sites, standard browsers and plug-outs, and active components, such as XSLT processors, to elaborate the information. Distribution sites are simply HTTP and FTP servers, widespread throughout the world; user browsers are HTTP clients and run on the user host. We do not require any other components to run on a specific host. Active components must provide answers to browsers, requiring an HTTP server interface; they must also ask data to distribution sites, acting as HTTP clients. Hence, MOWGLI is essentially conceived as an HTTP pipeline.

The module client of the distribution sites is the "getter", which maps URLs to URLs and hence documents, offering functionalities similar to the APT packet management system (http://www.debian.org).

The main active component is the XSLT style-sheet manager, whose typical functionality is the application of a list of style-sheets (each one with the respective list of parameters) to a document. However, other components may be added in a completely modular way. This is exactly the content-based architectural design of future Web system enabled by XML technology.

4 The Contributions from the Participants

The practical background for the work in MOWGLI is represented by the activities at the participating institutions. Though details could easily be obtained from the MOWGLI Web-page (http://www.cs.unibo.it/mowgli) some short remarks on this background should be made here.

The Department of Computer Science at the University of Bologna is the only educational institution in Italy to be affiliated to W3C. They care about the coordination of MOWGLI. The HELM project (Hypertextual Electronic Library of Mathematics, http://www.cs.unibo.it/helm, see also [12]) is active in Bologna since 1999. It is one of the systems of reference mentioned in the previous section. In the initial phase the team in Bologna cared about different conversion schemes.

INRIA (Institut National de Recherche en Informatique et Automatique) is a French institution located in Rocquencourt and Nice. They pursue two projects of importance for MOWGLI: the Lemme project, introducing and developing formal methods for use in writing scientific computing software, and the LogiCal project, which developed the Coq proof assistant (see [13]).

The German Research Center for Artificial Intelligence (DFKI) is based in Kaiserslautern and Saarbrücken. Its main mission is technology transfer, i.e. to move innovations in Artificial Intelligence from the lab to the market place. One main MOWGLI-related prototypical product of DFKI is the Web-based learning environment ActiveMath that integrates several external services. They also are in charge of OMDoc ([17], [18]), which will serve as an intermediate format suitable for HTML/MathML rendering

The Subfaculteit Informatica of Katholieke Universiteit Nijmegen hosts a broad experience in logic, formal methods and theorem proving. They are involved in several research activities in this domain as the EC sponsored Network "TYPES", the FTA project (Fundamental Theorem of Algebra), the EC Working group Calculemus which also deals with OpenMath et al.

The role of the Albert Einstein Institute (MPG, Golm) near Potsdam has been described above already. They provide a test bed with the Living Reviews, which will represent the important link to the domain of mathematical publishing. This also is the main concern of the partner TU Berlin, which is formally associated to AEI caring about the exploitation and information dissemination for MOWGLI.

Trusted Logic makes the group complete. This is a French start-up company, which offers a wide range of efficient and secure solutions of smart cards and terminals in a wide range of areas. Their development methodology includes a permanent concern of quality and security aspects for software.

5 Current Practices and First Results

MOWGLI started formally in March 2002. According to the action plan some deliverables have been made available after one year, when this report has been written. One important initial step is provided by the requirement analysis, which is based on the following ideas.

Providing mechanisms and a system for presenting and 'doing' mathematics on the Web MOWGLI has to deal with mathematical papers (including mathematical physics and other applied areas), formalised mathematics (in the Coq system), and computer science correctness (also in the Coq system), like Trusted Logic does. In the applied areas the expressions often are semantic-less (things like path-integrals may have a semantics, but the authors in physics may not want to be bothered by it). To accommodate this, the system should be able to deal with un-interpreted symbolic expressions. One wants to interact with a computer algebra system, probably not with proof assistants.

Formalised mathematics and computer science correctness are related, but in computer science proofs are very big, very trivial, and not to be looked at by humans. One wants support for 'managing' correctness proofs on a high level. Here we also are almost outside mathematics. In formalised mathematics, the user wants to be able to see or have access to all details of the development. Apart from the special view concerning the requirements developed by the project partners, MOWGLI has to take into account the requirements the mathematical community in general has in this relation.

The team in Bologna has a lot of experience in translating formal Coq developments to MathML based Web documents. These documents are not ‘textual’, but more ‘graphical’, closely representing the structure of the formal development. They have a lot of knowledge of this technology (XML, stylesheets, etc.). The DFKI and the team in Eindhoven as a part of the Nijmegen group use OpenMath and OMDoc to encode mathematical objects and mathematical documents. Eindhoven uses this in its interactive course notes, which are textual mathematical documents extended with the possibility to interact with a computer algebra system through OpenMath objects. DFKI has created a mathematical database, where the mathematics is encoded via OMDoc. The resulting initial requirement to make a choice, whether to use MathML or OpenMath for MOWGLI, has become obsolete meanwhile. MOWGLI will be based on the following translation procedures: XML to OMDoc and then via presentation XSLT to HTML/MathML. The conversion of Coq into OMDoc is available already.

Mathematical expressions with structure but no semantics are useful for presentation, exporting equations to a computer algebra system or manipulation and editing of these expressions. With semantics they may serve for interaction with a proof assistant or faithful (meaning preserving) exchange of mathematical expressions between different systems. Having the next generation of enhanced Web-documents in mind, physics and mathematical editors should be interested in expressions with semantics.

Another principle of distinction is that between static and dynamic mathematical documents. In static documents, the mathematics cannot and should not be changed by the reader. These documents should satisfy the permanence principle for mathematical publications. Hence, they also should not be altered by the author. The Web version of a printed or printable document is a typical example. The advantages of this as a reliable and citable document are obvious. Dynamic documents in mathematics deal with any kind of representation that can in some way be interacted with. Typical for this are documents with Java applets to illustrate or animate something. Such documents have a different and new impact on research and education in mathematics than the conventional static documents.

Authoring mathematical documents for the Web can be done in three ways: a) Writing a document in the usual way, with a text processor. There should be some additional features to make the mathematics dynamic. This will serve for structured mathematical expressions only. b) Writing a document interactively with a specially designed program, like Mathematica. This will result in a structured mathematical document, with structured expressions that have some semantics. c) Generating a document from a completely formal development, in Coq for example. This provides for a structured mathematical document where all the math expressions have a completely determined semantics.

Authoring tools depend on the content. Here we distinguish between research articles, classroom documents and expository mathematics. Research articles are printable documents and hence they are static. Accessibility is the most important issue. For authors it is important to be able to easily generate a Web version of their research paper that is well accessible, readable and printable. Classroom documents are characterized by being used both on the Web and in printed form. In this case easy installation facilities are important on both levels, in the Web and in printable form. Accessibility is of secondary importance, is sometimes deliberately restricted, but as an exception one usually does not want to require students to have the latest state-of-the-art browser software available. These documents are static. Expository mathematical documents use the Web and its facilities to explain mathematics. This can be done in various ways: using an applet to illustrate a construction in geometry, manipulating a mathematical equation with a computer algebra package, etc. This is often done as a smaller part within the context of a static mathematical document. Authors of dynamic documents are aware of the state-of-the-art technology, focussing on functionality, they use it, and they are prepared to invest time in this.

Let us survey some current practices next. For conventional research articles and classroom documents, the situation with respect to authoring tools is quite similar, because the mathematics that is being displayed is static. Practically only one familiar editor is used, which is a form of TeX (often LaTeX) in most cases or sometimes MSWord with an equation editor. Documents encoded in TeX are ASCII-files, which are compiled to DVI, PDF, PS or HTML files for presentation on the Web. For the HTML case one extracts an HTML file with GIF-images, using a LaTex2HTML-like package, or an HTML + MathML file, using newer extraction packages like TtM. Finally, Techexplorer allows a presentation of TeX-encoded formulas on the fly. But this technique did not find as many users as were expected from the developers. Documents written using MSWord with the Equation Editor are published on the Web as PDF files or as DOC files. Another option is to save the document (within MSWord) as a Web page, generating HTML + GIF or HTML + MathML.

For expository mathematical documents, one wants to use dynamic authoring tools. This occurs in various ways. From some computer algebra packages like Maple or Mathematica, one can save the document as an HTML+MathML file, which allows to export MathML expressions by cut and paste from the document to the computer algebra package. In the near future, computer algebra systems probably will be directly accessible through Web services, allowing to connect to the system via commands in the HTML page with the aim to insert output from computer algebra systems. Dynamic mathematics that does not involve computation can be added to a document by providing a script code, e.g. using MathML actions. A common aspect of these features is that everything is still in an experimental state and only available for the technically knowledgeable. Most potential authors, editors and publishers have no precise ideas about these techniques.

An important problem for the preparation of Web documents in mathematics is the development of an appropriate input procedure, taking the given publication techniques into account and remaining open for enhancements. For instance, for common papers in mathematics a very easy procedure should be available for authors to transform their document into the MOWGLI format. Authors and editors certainly cannot be requested to edit XML files directly without suitable tools, and even these tools should be very easy to handle to be accepted at least by authors, who have a strong motivation for electronic enhancements of publications. Documents working with formalised mathematics in Coq, want to have an automatic translation from Coq to MOWGLI, which can be edited for better presentation afterwards. Tools for authors to deal with such problems are still under development. Here developments at the partner MPG with the Living Reviews as a test bed will contribute to this. But the investigations also should include journals from core mathematics.

Finally some first results of MOWGLI dealing with the problems introduced above should be summarised. There is a clear decision about the conversions to be arranged between the different offers. OMDoc is taken as an intermediate format suitable for HTML/MathML rendering. There is a conversion of the Coq proof assistant into OMDoc. General users of MOWGLI will go to the interrogation interface of HELM/Coq and the internal interface for Coq users via the OMDoc-search interface. Simultaneously efforts are undertaken to make Coq more XML compliant. These important interactions have been established in the first year of MOWGLI. As a case study, applications of Coq for theorem proving in elementary geometry and plane incidence geometry are available. A first release of the MOWGLI metadata scheme is available from DFKI. This is compliant with other metadata recommendations like those for EULER. Content dictionaries are available already, though controlled vocabularies are still missing. They will be developed for sub-domains like those handled by MPI at first, in order to provide a background for the content mark up of the Living Reviews.

References

1. The Dublin Core Metadata Inititiative. http://purl.org/dc/
2. Living Reviews in Relativity. http://www.livingreviews.org
3. Mathematical Markup Language (MathML) 2.0 W3C Recommendation, 21 February 2001. http://www.w3.org/TR/MathML2/
4. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
5. Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation 27 March 2000. http://www.w3.org/TR/rdf-schema/
6. SOAP Version 1.2 Part 0: Primer. W3C Working Draft 17 December 2001. http://www.w3.org/TR/2001/WD-soap12-part0-20011217
7. Extensible Markup Language (XML) Specification. Version 1.0. W3C Recommendation, 10 February 1998. http://www.w3.org/TR/REC-xml
8. XSL Transformations (XSLT). Version 1.0, W3C Recommendation, 16 November 1999. http://www.w3.org/TR/xslt
9. Asperti, A., Padovani, L., Sacerdoti Coen C., Schena, I.: Formal Mathematics in MathML. Proceedings of the First International Conference on MathML and Math on the Web, October 20-21 2000, University of Illinois at Urbana-Champaign
10. Asperti, A., Padovani, L., Sacerdoti Coen, C., Schena, I.: Formal Mathematics on the Web. Proceedings of the Eighth International Conference on Libraries and

Associations in the Transient World: New Technologies and New Forms of Cooperation, June 9-17, 2001, Sudak, Autonomous Republic of Crimea, Ukraine

11. Asperti, A., Padovani, L., Sacerdoti Coen, C., Schena, I.: XML, Stylesheets and the re-mathematization of Formal Content. Proceedings of Extreme Markup Languages 2001 Conference, August 12-17, 2001, Montréal, Canada
12. Asperti, A., Padovani, L., Sacerdoti Coen, C., Schena, I.: HELM and the semantic Math-Web. Proceedings of the 14th International Conference on Theorem Proving in Higher Order Logics (TPHOLS 2001), 3-6 September 2001, Edinburgh, Scotland
13. Barras, B. et al.: The Coq Proof Assistant Reference Manual, version 6.3.1, http://pauillac.inria.fr/coq
14. Tim Berner's Lee: The Semantic Web. W3C Architecture Note, 1998
15. Huet, G., Plotkin, G. (eds): Logical Frameworks. Cambridge University Press 1991
16. Huet, G., Plotkin, G. (eds): Logical Environments. Cambridge University Press 1993
17. Kohlase, M.: OMDoc: Towards an Internet Standard for the Administration, Distribution and Teaching of mathematical Knowledge. Proceedings of Artificial Intelligence and Symbolic Computation, Springer LNAI, 2000
18. Kohlase, M.: OMDoc: An Infrastructure for OpenMath Content Dictionary Information. Bulletin of the ACM Special Interest Group for Algorithmic Mathematics SIGSAM, 2000

SINM: The Italian National Information System for Mathematics

Virginia Valzano[1] and Maria Carmela Catamo[2]

[1] SINM national Coordinator and SIBA general Coordinator
University of Lecce, Coordinamento SIBA
Via per Monteroni, Edificio La Stecca, 73100 Lecce (Italy)
siba@siba2.unile.it
[2] Technical collaborator SIBA and SINM
University of Lecce, Coordinamento SIBA
Via per Monteroni, Edificio La Stecca, 73100 Lecce (Italy)
catamo@siba2.unile.it

SINM (Sistema Informativo Nazionale per la Matematica) is the Italian National Information System for Mathematics.

It enables the Italian mathematical community to have easy access to a coordinated system of bibliographical, documentary, full-text and multimedia resources.

It aims at the development, diffusion and sharing of electronic information resources regarding mathematics, at a less possible waste of technical and financial resources.

The SINM Web site is accessible at the URL http://siba2.unile.it/sinm .

Since 1991 SINM is coordinated by the University of Lecce, in particular by SIBA Coordination (http://siba2.unile.it), in agreement with the Italian mathematical University and research organizations libraries.

Within this Information System, SIBA Coordination has developed many national projects for the cataloguing, digitalization and fruition of bibliographical and documentary material, for the publication and consultation of electronic journals and for the conversion in electronic format of the back volumes.

To cite some of them:

- the *National Journals Catalogue of Mathematical, Physical, Computer and Technological Sciences*;
- the *OldenMath (Olden Mathematical books and documents)* System;
- *SINM-MPRESS (Mathematical Preprints Search System)*;
- *REIM (Riviste Elettroniche Italiane di Matematica)* Project.

The **National Journals Catalogue of Mathematical, Physical, Computer and Technological Sciences** contains the bibliographical descriptions and holdings of the journals of the Italian scientific libraries belonging to SINM as well as the registry data of the same libraries; moreover, it contains the bibliographical descriptions of the journals available in electronic format and links to the relevant Web sites. It is integrated with the *Archive of Indexes* of journals of major interest for mathematicians (http://siba2.unile.it/cgi-bin/waisidx), with

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 24–28, 2003.

the *Electronic Journals Catalogue Directory* (http://siba2.unile.it/ej-catalogue) – both realized by the same SIBA Coordination – and with other network resources (databases and electronic journals full-text).

The Catalogue, accessible by Web at the URL http://siba2.unile.it/archives/bibsearch.html, allows the user to find the desired information in an extremely simple way, to request automatically copies of journal articles by e-mail (clicking on the library of interest and the relevant e-mail address), exclusively for scientific purposes, or to access directly the electronic version available both on the Web servers of the publishers and on the *ScienceDirect OnSite Server* of CASPUR, Rome (http://periodici.caspur.it)

The more, dynamic links enable to extend the searching on the databases shared on the *ERL-WebSpirs* server (http://siba2.unile.it:8590), on the *Electronic Journals Catalogue Directory* (http://siba2.unile.it/ej-catalogue) and on the *ZMATH* (http://siba-sinmdb.unile.it/ZMATH) and the *MATHDI* (http://siba-sinmdb.unile.it/MATHDI) Databases, to the Catalogue, in order to locate journals and to find documents.

The **OldenMath** System (http://siba3.unile.it/archives/omsearch.html) intends to catalogue rare and valuable editions in the mathematical area held by the Universities and research organizations belonging to SINM and to digitize the same editions, partially or completely, by means of the SIBA Coordination System for digital acquisition and processing of images (http://siba2.unile.it/sedi/labim.html).

The software for the on-line management and consultation, developed ad hoc by SIBA Coordination, is based on the CDS/ISIS System (environment MS-DOS and UNIX).

The Web interface being realized for an easy access to the catalogue uses the search engine WAIS-ISIS.

The search template allows the location of a specific work within the catalogue, searching for title, author and/or editor, publisher and/or typographer, publication place and date.

It allows moreover to navigate on authors, publishers, typographers, publication place and date, series and/or collective titles.

By a common browser (such as Internet Explorer 4.0 or equivalent) the user can access bibliographical descriptions of documents and the relative digital images. Moreover, it can browse entirely digitized documents, get information about the libraries to which they belong and access their relative Web sites.

The current OldenMath Catalogue (http://siba3.unile.it/archives/omsearch.html), realized within the homonymous Project and the *I17 Initiative* of the *Coordinated Project of the Universities of Catania and Lecce*, contains bibliographical descriptions of olden, rare and valuable editions, held by mathematic libraries of the University of Pisa, Bologna, Milano and Padova.

Bibliographical descriptions follow the ISBD(A) standard.

Images refer to the partial or entire reproduction of some editions and to the more significant pages of the other ones (title page, incipit, colophon, etc.)

The resolution of the images stored (in JPEG format) in the on-line Catalogue, accessible by Internet, is of about 550x850 pixels; the resolution of images stored in the historical archive, for a more in-depth study by scholars, is of 2000x3000 pixels.

On each image accessible by Internet a logo has been applied, indicating the name of the University of Lecce and of SIBA Coordination (who realized the images), in order to prevent somehow any embezzlement.

To cite some volumes present in the OldenMath Catalogue:

- the II Tome of the *Elements d'algèbre* of Euler, *De l'Analyse Indéterminée*, published at Lyon in 1774, held by the library of the Department of Mathematics, Computer Sciences and Physics of the Pisa University; the volume has been entirely digitized and it is possible to browse its pages one by one; to proceed skipping 20 pages at a time or to go directly to the desired image;
- a work of Plücker, *Analytisch-geometrische Entwicklungen*, also held by the Department of Mathematics, Computer Sciences and Physics of the Pisa University and entirely digitized;
- a volume of the *Geometria* by Cartesio, held by the library of the Department of Mathematics of the Milano University, from which currently only the more significant pages have been digitized;
- two entirely digitized specimens of the library of the Mathematical Seminar of the Padova University: *Optice* by Isaac Newton and *Mathematicae Collectiones* by Pappus Alexandrinus;
- the *De planis triangulis* by Magini, held by the Mathematics Library of the Bologna University;
- the 2nd edition of *Algebra* by Raffaele Bombelli, published at Bologna in 1579, volume that Riccardi defined as "rare and valuable". It is a very important work for the history of mathematics. The volume belongs to the "Bortolotti Fund" that takes its name from the Bolognese mathematician and historian who discovered, in the Library of the "Archiginnasio", a manuscript containing the IV and V volumes of the work, and he published it.

The **SINM-MPRESS** (http://siba-sinm.unile.it/mpress) is the Italian National Index of Mathematical Preprints; it collects and indexes the preprints of Italian mathematicians, and is included in the international system MPRESS/ MathNet.preprints (Mathematics Preprint Search System, http://euler.zblmath.fiz-karlsruhe.de/MPRESS).

It is based on Harvest, a highly flexible software that works in a distributed way, giving the possibility to index both preprints entirely hosted on the SINM-MPRESS server and those stored on other Web sites.

Moreover, SINM-MPRESS is interoperable with the ETRDL System (http://www.iei.pi.cnr.it/DELOS/EDL/edl.htm) of Computer Sciences and Applied Mathematics preprints of CNR and ERCIM (European Research Consortium for Informatics and Mathematics).

REIM (Riviste Elettroniche Italiane di Matematica = Italian Electronic Mathematical Journals) is a project aiming at the coordination and the development of the National System for the Web publication and consultation of the Italian electronic mathematical journals.

The project aims also at the conversion in electronic format of the back volumes of the journals, based on the system developed by EMIS (European Mathematical Information Service, http://siba-sinmemis.unile.it) within the ERAM project (Electronic Research Archive for Mathematics, http://www.emis.de/projects/JFM).

The REIM System for the management and consultation of electronic journals has been implemented by SIBA Coordination in January 2000 with the publication of the electronic version of the Journal “Note di Matematica” (http://siba2.unile.it/notemat) by the University of Lecce.

Like the ESE System (http://siba2.unile.it/ese), implemented by the same SIBA Coordination within the *ESE* project (Electronic Scientific Publishing of University of Lecce), REIM is based on standard and open technologies (SQL, PHP), on the use of standard formats for the access and electronic distribution of documents (PDF, PostScript, TeX) and on the use of standard communication protocols (HTTP).

REIM enables the editorial management of electronic journals, the loading of full-text documents and the management of the relative metadata by means of a specific Web interface (http://siba2.unile.it/sinm/reim).

The more, it allows to consult journals by means of a sole Web interface (http://siba2.unile.it/sinm/reim/search). It enables to search by title, author, abstract, keywords, MSC classification (Mathematical Subject Classification) and DOI code (Digital Object Identifier) contemporaneously on one or more journals.

Each indexed article in the REIM System has a DOI code (assigned by SIBA Coordination) automatically generated and registered by the same system in the international index (DOI Directory).

The DOI coding is universally acknowledged and enables the unequivocal and permanent identification of each document in electronic format through the assignment of an alphanumeric code to that document by the publisher.

The REIM System enables to control the access to full-text documents through the authentication of the user by IP address or by password. The authorized user may access the full-text document directly on the REIM System.

The REIM System enables also to consult, through the same Web interface, the electronic journals of other publishers: the System indeed allows also the indexing of the metadata concerning external electronic documents. In this case the System refers to the full-text on the publisher Web site.

SIBA Coordination furthermore has achieved numerous consortium agreements with producers and distributors of electronic resources aiming at an easy access to electronic journals and to the major databases of the mathematical area. Among these databases we can mention *MathSci* and, in particular, *ZMATH* (http://siba-sinmdb.unile.it/ZMATH) and MATHDI (http://siba-sinmdb.unile.it/MATHDI), whose Italian national mirrors are managed by SIBA Coordination. These databases, as mentioned above, allow also dynamically linking to the *National Journals Catalogue* and to the other network resources accessible through the SINM Web server.

SINM is connected with other European Information Systems, with the main national and international mathematical Organizations and Associations, in particular with the *Italian Mathematical Union* and with the *European Mathematical Society.*

Together with the latter one and with other European partners, SIBA Coordination cooperates to the *LIMES* (Large Infrastructure Mathematics-Enhanced Services) project, financed by the European Union for the scientific and technological development of *ZMATH Database* (http://www.emis.de/projects/LIMES/). Within the LIMES project SIBA Coordination is the *Italian Editorial Unit* of ZMATH (http://siba-sinmlimes.unile.it/editZMATH/), it attends the reviewing of the Italian journals, monographs and proceedings and is actively involved in the development and the updating of the same database.

Management of Informal Mathematical Knowledge – Lessons Learned from the Trial-Solution Project

Ingo Dahn

University Koblenz-Landau

Abstract. Most mathematical knowledge is not formalized. It exists in the form of textbooks, articles and lecture scripts. We discuss a way to make this knowledge nevertheless usable for automated knowledge management services. Our approach is based on Slicing Book Technology. We describe this approach and summarize the experience gained during its implementation in the European project Trial-Solution.

1 Introduction

It seems generally agreed that the ability to manage knowledge in an efficient way is of crucial importance for the information society. However, there is much less agreement on what this means in practice. Knowledge management is interpreted by many as the art of convincing people in an institution to share their knowledge. Others interpret it as a way to retrieve pieces of text – called knowledge objects – from a database.

We shall not dive into this discussion, but will rather consider what are the requirements to mathematical knowledge management from the point of view of the intended end user and how fulfilling these requirements may be supported by automated means. The end users we have in mind are people who need mathematical knowledge for certain purposes. Examples of such purposes are

- solving new mathematical problems
- applying mathematical knowledge to solve problems in other areas like engineering, economics or natural sciences
- passing an exam in Mathematics or in another science which uses Mathematics as a tool.
- teaching Mathematics

We shall not discuss how to manage knowledge in a library or in a lecture, though these problems are obviously related.

In the following we shall point out that the new means of the information society can be better exploited using a new approach to the storage, management and reuse of information. We illustrate the benefits that can be achieved with some examples and sketch the infrastructure that is needed to realize these services. Electronic online publications have posed many questions related to copyrights

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 29–43, 2003.

in a new framework – we shall provide some thoughts on these issues. In the final section we describe a vision towards a new quality of knowledge management support.

The paper refers explicitly to the management of informal mathematical knowledge, i.e. knowledge that is not described in a formalized, machine readable format with a precise semantics. We acknowledge that such formalized knowledge exists and offers quite new possibilities for correctness and for the retrieval of mathematical information. However, the vast bulk of mathematical knowledge exists in an informal version only, mostly in the form of textbooks and articles in mathematical journals. To these every-day forms of mathematical knowledge, the tools available for formalized knowledge cannot be applied. Moreover, formalizing informal mathematics requires still a major intellectual effort and cannot be expected to stay inline with the rapid development of mathematical knowledge.

When we confine ourselves subsequently to Mathematics, this is done since we have most experience in this field – in fact, more than in other fields – mathematical papers tend to have a clear and precise structure with very fine grained and clearly identifiable knowledge objects. However, it will become clear in the following that the major ingredience of our approach is, to make use of the structure of the document. But clearly structured documents occur in many fields – we mention textbooks in other fields, legal documents and technical documentations as the most obvious examples – and our results can be transferred to these fields too.

Our presentation will be to a large extent guided by the experience gained in the project Trial-Solution which was funded in part by the European Commission. The author would like to thank all members of the Trial-Solution Consortium that helped to gain practical experience with a new way to manage mathematical knowledge.

2 Knowledge Management Vs. Document Management

Usage of documents has a long tradition in human history. Documents are the standard form in which knowledge is communicated among humans. When knowledge is requested, it is everyday practice that humans go to a library or a web search service and retrieve a document – and only then they start retrieving the knowledge they need from this document. Humans have developed beautiful designs and sophisticated techniques like typesetting to ease working with documents.

Computers can proceed differently. They do not have a sense for the beauty of a book. Simple and clear structures are much more helpful for them, even if this leads to a much larger mass of data which they have to process. They can retrieve the knowledge they need from databases or from raw experimental data. They can also quickly scan thousands of web pages and extract information from these pages – provided this information has been encoded in a computer friendly form!

It is only over the last five years – in connection with the hype of XML as a meta-lingua franca for the exchange of data – that it became generally acknowledged that a clear separation of content and presentation helps to serve humans as well as computers in making use of the available information. We hasten to mention that this is not quite correct. Mathematicians have partly separated content and presentation already since around 1990 when they started to use broadly LaTeX to write their journal papers and books. This was possible due to a high degree of standardization of the language of mathematical formulas. Thus mathematicians and publishers of mathematical texts have gained a lot of experience in managing their knowledge in a way that is equally friendly to computers and humans. This emphasizes that studying the experience of mathematical knowledge management may be of special interest also for other fields.

A LaTeX document consists of LaTeX source code which describes what is to be presented. The presentation itself is automatically computed, based on directives that are collected in central style files. These style files are often made up by experienced designers. Many publishers maintain their own style files. These style files support the authors in getting their documents nicely formatted, according to the directives of the publisher.

As a side effect, LaTeX has introduced standard ways of managing structural and referential information – tables of content, key phrases, references, citations etc.. All these are important helpers for managing mathematical knowledge and mathematical documents. Other knowledge presentation systems like word processors or web browsers still lack similar tools. Also the typesetting quality of LaTeX systems is unsurpassed by these other wide spread systems. Only specialized electronic publishing systems can produce better layouts than LaTeX more easily – however mostly at the cost of freezing the content in the form of a print page, which makes it hard to reuse the content any more.

In fact, LaTeX describes the content only in as much as it is necessary to calculate a human-friendly *presentation*. For example it is not (necessarily) encoded in a LaTeX document whether a multiplication symbol denotes a (commutative) multiplication between numbers or a (non-commutative) multiplication between matrices. This is sufficient since the author can assume that the (human) reader will infer this from the context. In some cases this disambiguating context will be given explicitly in the document, in other cases the author will rely on the reader's knowledge of the "standard mathematical usage", for example to infer that a small greek letter epsilon denotes an arbitrarily small real number.

When computers are to access this mathematical content correctly, the disambiguating context must be given in a computer readable form too. This is a major part of the formalization of mathematical knowledge. To provide this is tedious for authors and not required by readers who do not want to manipulate the mathematical objects occurring in the document with a computer.

Thus we are facing the situation that most mathematical documents are currently available in the quasi standard format of LaTeX that permits an easy adaptation to different presentation styles and has moreover standard ways to encode some relations between documents (citations) and within documents (references).

It will turn out to be even more important that LaTeX provides standard ways to encode the structure of documents and the specific constructs that occur in mathematical documents. As in popular word processors, the author has a lot of freedom to influence the structure and the layout of the documents. However, unlike these word processors, LaTeX discourages deviations from the quasi standard. Moreover, in LaTeX the easiest way for the author to introduce his/her own constructs is, to do it in a central style file where it can be easily found and adapted later if the need occurs. This leads to the fact that mathematical documents are better structured and easier to adapt to different presentation needs than documents from many other fields. However, note that all these features apply at least to the same extent to XML documents from any field, which however unfortunately are still rare.

Having documents with a machine-recognizable structure and with machine-readable meta information opens up the possibility to access directly the knowledge contained in these documents and reuse it outside of the context it was written for. This is the essence of Slicing Book Technology which was introduced by the author in 1999. It consists of the following steps [Da01a]

1. Decompose existing books into semantic units
2. Add a knowledge base of meta data
3. Design an intelligent advisory system that uses these meta data
4. Compose personalized documents tailored to the learner‘s current needs – on the fly

Slicing Book Technology was presented at the LearnTec Conference 2000 [Da00a] and at the AERA Conference 2000 [Da01]. The first book using this technology [WD00] appeared in 2000 using the SIT-Reader – a server software from Slicing Information Technology (http://www.slicing-infotech.de).

The European project Trial-Solution was launched in February 2000. It investigates the potential of Slicing Book Technology for the combination of semantic units from different sources. The following discussion in this paper will summarize some of the experience gained within this project.

Within the project tools for Slicing Book Technology are developed and a library of textbooks on undergraduate mathematics are reengineered. The library of semantic units built in the project now comprises more than 25,000 units extracted from 5,000 pages of text. The semantic units are reusable parts which can be as small as an exercise, an example or a theorem or they can be as large as a proof. We mention in passing that the decision on the size of these slices is to a certain extent dictated by the intended reuse and by economic constraints of the possible reengineering efforts.

In the sequel we shall not discuss technical details of Slicing Book Technology. Instead we shall concentrate on some experiences which appear to be interesting from the point of view of knowledge management.

Slicing Book Technology enables a series of new services for the reuse of knowledge. These services will be described in the next section. Before giving this, it is, however, worth noting that Slicing Book Technology introduces a new approach to the communication of knowledge which goes beyond what most people ever have experienced.

The SIT-Reader gives the user the possibility to compose his/her own documents on the fly with qualified automated assistance to ensure that nothing important is lost. This requires the user to go through several steps.

- locate the information that is of primary interest
- eventually ask the system to complete the selection
- eventually iterate the previous two steps
- when the selection is complete, request the generation of the document
- when it turns out that something is missing or that the document contains superfluous parts, go back an modify the selection, eventually with automated support.

This procedure is much more complex than the usual single-click by which users normally get ready-made static documents from the web. In reaction to user feedback the SIT-Reader was augmented with a possibility to obtain complete sections, paragraphs etc. with a single click on their headline. In these cases the system would only perform very basic completions of the document. This gives the reader rapidly what he/she expects from his/her experience from other systems but it completely misses the new flexibility that is available. Needless to say that it is described in the online help system of the SIT-Reader what this flexibility provides for the reader and how it can be used; needless to say also that this online help is hardly consulted when the reader easily gets what he/she expects.

Interrogations of undergraduate students, that have been using the system, revealed that their basic approach to managing their knowledge is very much document centric. They frequently expect (and believe they have the right to expect) from their professors a single document that contains exactly what they have to learn for this particular lecture. This view changes only slowly when they have to *apply* knowledge to solve problems for which the solution is not prefabricated in their textbooks.

We note that this document centric approach to knowledge management is by no means restricted to undergraduate students! Even experienced librarians and publishers normally believe they have done their job well, if they have delivered to their clients a pile of documents which contain the requested knowledge – even if it is buried in the documents somewhere between a mass of material not related to the client's needs. And in fact with static documents they had no choice – with Slicing Book Technology they have.

In order to guide the reader to apply the new tools, the SIT-Reader has been equipped with a wizard. This wizard guides the reader step-by-step from a search for some interesting topics to the generation of a personalized document that leads the reader from parts that he/she apparently knows directly to the learning objectives without detours.

3 Knowledge Management Services

Having accepted that knowledge management is different from document management we gain the opportunity to provide documents from knowledge stored in a database. These documents can be built for different purposes in a different way.

The most simple way is to collect slices by search functions. Full text search or key phrase search are the most common search methods. When slices are classified according to their types, search can be restricted to retrieve slices of a particular kind only. For example it becomes possible to collect exercises on a particular topic only. This restriction of search procedures by types is a standard feature of XML documents. We observe that it can be realized without translating existing content into XML. The only prerequisite requirement is a classification of slices according to their type.

Mathematical knowledge is organized in a systematic way. New knowledge is derived from other knowledge. Different pieces of knowledge – theorems, definitions examples etc. – serve different purposes. Thus mathematical knowledge is organized in a semantic network of interrelated objects of different types. When we understand this network and the way in which it is used in practice, we can design principles for the generation of documents for a particular purpose for a particular user with specific knowledge.

This means, that knowledge objects cannot be considered in isolation. Not each collection of such objects will be meaningful or useful. The first step to support the generation of personalized documents that suit the reader's needs is, to explicate the semantic network of knowledge objects. How this is done in the Trial-Solution project will be described in the next section. For now we note that the essence of this explication is, to define conditions for the reusability of knowledge objects in certain situations.

This can imply, that a certain slice can be used only in conjunction with another slice.

Let us consider for example the decision of a lecturer to include material from a textbook into her lecture. Initially she faces a large number of knowledge objects from a book or even from a library. She will first select the material that is most important for her to teach. This can be a set of theorems, definitions or examples. In order to include these into her lecture, she will have to include other knowledge objects which are necessary for understanding them. These new objects will again build on other objects etc. At a certain point, our lecturer will decide that some of these prerequisites are already known to her audience or that some of the proofs are too detailed. Moreover she may feel the need to augment the already found slices by slices from other books which, perhaps, provide examples from a specifically interesting domain.

Thus in the course of preparing her lecture, our lecturer has made quite a number of knowledge management decisions. These should be supported by automated tools.

In another situation, in order to motivate a learner, it can be useful to present a collection of examples that apply the knowledge to be learned. Also this can be done automatically.

To prepare an exam, our lecturer may search for problems on certain topics. The students may be aware of the topics to be learned and may also search for such problems – perhaps only for those, for which also solutions are available, and they may request automatically the knowledge that is required to solve these problems.

Knowledge management at universities has an important human component. Our lecturer may select the knowledge to be learned for the next seminar. Then each student may augment this selection automatically by other knowledge that corresponds best to his interests and pre-existing knowledge.

Knowledge management may also support cooperative work. In this scenario a group of people comes new into a project. Each participant has a particular expertise but has to acquire specific additional knowledge in order to play the intended part in the project and to communicate with the other members of the team.

The team leader can prepare for each new member a particular script for reading that avoids what the person knows already and leads to what is specifically needed to be known. It may be necessary to collect slices from various books to present all required information and these books may have overlapping content. In this scenario the rules for the generation of documents should be set up such that the scripts for all members of the team use the same slices when they have to provide the same information in order to facilitate communication between the team members during the work, but to lead to the knowledge specifically required.

4 Metadata

If the content of a slice is not formalized, computers cannot use it. However, since the knowledge to be managed is in the content, computerized knowledge management services require descriptions of the content itself and on the conditions for its reuse that can be easily processed by computers but avoiding the high costs of content formalization. This information on the data (being contained in the content) is called metadata.

We can distinguish four kinds of metadata.

1. Key phrases
2. Titles
3. Attributes
4. Relations

Key phrases are the most important tool to describe the content of a knowledge object. Key phrases are widely used. Already authors assign frequently key phrases in their books. Key phrases can be more than just strings. Sometimes they are organized in hierarchies like

Number
- Natural
    - o Even
    - o odd
- real
- complex

Generally, this hierarchy need not be a tree, a key phrase can be subordinate to several key phrases. For example “Incompleteness of second order logic” can be subordinate to “Incompleteness” and “Second order logic”. However it seems feasible to assume that this relation of subordinateness does not contain loops. Key phrases can be related with each other, for example “Multiplication” is related with “Division”. Key phrases can be synonyms of each other.

The structure in which key phrases are organized is called a thesaurus. Thesauri can be considered as advanced forms of ontologies. It can be observed that key phrases frequently describe topics that are discussed in the slices to which they are assigned. Thus a thesaurus provides the set of topics with a structure. Electronic knowledge management tools can use these structures in various ways. For example related topics can be included automatically in a search, synonyms can be taken automatically into account.

The importance of the thesaurus structure for the management of the knowledge objects suggests to keep the key phrases in a structure separate from the book to which the key phrases are assigned. Then the book itself will not contain the key phrases but only pointers to the key phrases in the thesaurus. This use of structured thesauri was a main reason for the Trial-Solution project to deviate from the metadata schemas that have been elsewhere designed for learning objects.

Having the thesaurus separated from the book simplifies thesaurus maintenance considerably. All changes on the key phrase system have to be made only in one place – the thesaurus. For example, augmenting key phrases in the thesaurus with their translations in another language such that all these variants reside in the thesaurus under the same identifier, will make all books using this thesaurus searchable for users of that other language without further effort.

Assigning key phrases in a good way is not an easy task. Though there are norms for building key phrases, these norms are frequently not respected by authors.

Even if they would have been respected, this would have little impact: Electronic tools can handle large sets of key phrases, much larger than a human can browse. For example thesauri for books in the Trial-Solution project have usually 500–1,000 key phrases per book. Nevertheless humans can make use of large thesauri through appropriate tools. For example the user can search for a key phrase and get pointed to the related key phrases. But the user will not pick a key phrase from a list of all key phrases. Instead he may enter only an initial part, a substring or a synonym of a key phrase and the tool will suggest a number of key phrases that match the query. Thus the user is not forced to enter a query in accordance with the norm.

When books are sliced in a fine-grained way, such large thesauri are necessary to describe small slices correctly. They must be assigned by competent personnel who knows the mathematical subject well. Existing key phrase systems are much too rough to be sufficient for detailed descriptions of mathematical contents. For example, the well known MSCS classification system only differentiates between different topics of mathematics, but all knowledge objects that are handled in a book will usually fall into the same MSCS category.

Therefore it is necessary to have the possibility to introduce new key phrases. Though the language of mathematics is to a large extent standardized, it seems inevitable that different people will chose different versions of a key phrase (for example singular vs. plural) for the same thing. Also people assigning key phrases may not be aware that this key phrase has been already assigned for another book.

This may not be a problem as long as only one book is concerned and key phrases assigned to other books do not matter. However, key phrases are a powerful tool to establish relations between knowledge objects from different sources.

For example, suppose that a user is studying a certain slice and wants to know about related material. Suppose moreover, that this slice has assigned a pointer to the key phrase “Logarithm”. Then it is possible to search other books for slices which have the same pointer assigned. This will no longer be possible, if the author of the second book has reinvented the key phrase “Logarithm”, not knowing about the first book, since his reinvention will be kept under a different pointer. Also if the second book uses “Logarithms” instead of “Logarithm”, . the thesaurus will keep a different pointer.

Therefore it is necessary to incorporate new key phrases into an existing standardized thesaurus. This incorporation should not only standardize key phrases but it should also create directives to replace non-standard key phrase assignments automatically by standardized ones. This standardization requires a tool for thesaurus maintenance that can generate such replacement directives. It has to find simple replacements (like duplicate key phrases or synonyms used as separate key phrases) automatically and it has to allow an interactive work for the more complex tasks. Within the Trial-Solution project FIZ Karlsruhe in cooperation with Trinity College Dublin has implemented such a tool – the Trial Solution Key Phrase Server. The Mathematical Societies or institutions on behalf of these, like FIZ, are the natural places to maintain a central thesaurus for their fields in their relevant languages.

Titles describe the content of a slice. These titles are a help for the human to navigate. They are human readable and usually in the language of the slice. Chapters and sections usually have titles assigned by authors, but when the slicing is more fine-grained, similar titles have to be assigned automatically or manually.

Like titles, *attributes* are assigned to individual slices or groups of slices. However, unlike titles, attributes are taken from a controlled vocabulary. This makes them suited for automated processing. A very important attribute is the type of a slice. The Trial-Solution project uses a set of 18 types like “Text”, “Theorem”, “Proof” or “Example”. These types are very useful to decide, whether a certain slice is presented to a particular user. Thus a user not studying mathematics will frequently be interested in Examples but not in Proofs. Also the user can exploit these types. For instance in the SIT-Reader it is possible to select with a single step all exercises found in a search for a particular key phrase – say for the preparation of an exam.

Other schemas, like IEEE LOM, use more attributes to characterize knowledge objects, for example a characterization of the difficulty of an object. Such attributes have not been selected for the Trial-Solution project since they seemed too subjective to assign in order to serve as the basis of an automated service.

Relations are the third kind of metadata. They link different knowledge objects with each other. Though theoretically relations are possible which link any number of objects with each other, we found that binary relations are sufficient for our purposes. In order to understand the choice of relations made in the project it is useful to recall some of its particular aspects.

Relations have been assigned for use by an automated system, more precisely by an automated inference system. Therefore, there was no need to introduce relations that can be defined from others. For example there is no need to assign the inverse of a relation. Also if to understand slice A one has to understand slice B and to understand slice B one has to understand slice C, then there is no need to state explicitly that one has to understand slice C in order to understand slice A since this can be inferred. This leads to a considerable reduction of the amount of relations that have to be assigned.

Actually the Trial-Solution Project uses four different relations between slices.

1. To understand A one has to understand B (internally: A *refers* B). This is the most important relation. It is used for example to add prerequisites in documents, to complete the model of the user’s knowledge by automated inference and to find applications of knowledge objects. Explicit references in mathematical documents often indicate such a relation.
2. A does not make sense without B (internally A *requires* B). This models phrases like “[A] Modify the function in formula [B] (1) such that ...”. In such a case B will be included in any document if A is. Note that B can require other slices which then, of course, have to be included too.
3. A explains B (internally: B *isExplainedBy* A). This is used when it cannot be modeled as “A is of an explanatory type and refers to B”.
4. To understand A one has to understand some part of B (internally A *isKnown Blocker* B). This is used if B contains several ideas which cannot be separated into different slices, perhaps since they occur in a single sentence. This is essentially like the first relation, except that from the fact that A is known one cannot infer that B is known too. That explains the strange name the relation has received in the project.

Since the project has to handle a large number of knowledge objects it was important to assign these meta data in an efficient way. For example if a section builds on all parts of another section, all slices from the first section would be related to all slices from the second section, leading to a large number of relations to be assigned.

To avoid this, the project introduced *meta data inheritance*. This means that meta data assigned to some groups of slices, like a chapter, a section or a group of examples, can be inherited by all slices in this group. In a similar way, key phrases can be assigned to a group of slices and inherited downwards to all slices.

Beside the meta data mentioned so far there are other meta data which are more designed for human use. For example the author and the copyright situation as well as the language(s) of the knowledge object have to be mentioned somewhere. It would clearly be a waste of resources to repeat this information for each slice. Instead, this information is assigned only to the top node of a sliced document and is inherited downward the document hierarchy.

Inheritance is based on the hierarchical structure of the documents. Fortunately, this is always possible. Relying on the hierarchical document structure was also uncritical for the project since this structure was already of central importance for the basic service – the generation of the personalized documents to be delivered to the reader.

We mentioned already that it usually does not make sense to extract a piece of knowledge from a document and consider it in isolation. Now, the same holds technically. An item of a list cannot be isolated from the list environment, i. e. from the information that it is part of a list, where the list begins and where it ends. To deal with such information, the author introduced the concept of a sliced document as a *hierarchical file tree*.

Conceptually, a sliced document can thus be visualized as a tree of directories in a file system, where each of these directories contains a start file and an end file where the start file contains the begin of an environment and the end file contains the end of that environment if needed. Moreover, the directory can contain subdirectories of the same kind. Those directories which do not contain subdirectories have files that contain the proper content.

For example, a directory representing a chapter will contain in its start file the headline of the chapter. Start or end files may be empty in which case they can be omitted.

Now, when a document is to be composed from slices, the set of slices that should go into this document is considered. This set is completed by all nodes in the file tree that lay on a path from one of these slices to the root of the file tree. This subtree of the tree representing the complete document is traversed in a depth-first manner. Whenever a node is entered, the start file found there is included and when it is left, the end file is added. Content files are added when they are found. In this way, using the hierarchical structure of the original document, a technically correct document is built. Note that, when chapter headings are contained in start files as described above, the chapter heading will occur in the document whenever content from this chapter is included. The hierarchical file tree model opens up more possibilities for the management of rights as we will see in the next section.

The meta data assigned to knowledge objects must be combined with each other and with information about the user in order provide the services described in the last section. This combination is governed by general rules that are processed by an automated knowledge management system. An example of such a rule is

If the user is a learner and looks for an exercise on topic T then from each group of such exercises two exercises are included into the document. Moreover all prerequisite units for these exercises are included.

In addition there may be other rules required in order to complete the description, for example:

If n units are to be selected and there are more than n units available for selection, then select those units which are closer to the beginning of the book but are not yet known to the reader.

The knowledge management system used in the Trial-Solution project has been developed by P. Baumgartner. It is described in [BF02].

5 Copyright Issues

Making material available online immediately provokes the question of whether this supports infringements of copyrights. Since it is relatively easy to copy and modify electronic material, some publishers and authors are concerned that their material may be stolen and their revenues affected, though this is not supported by actual experience – quite contrary, in many cases putting material online increases selling of the printed product ([JE01]).

Nevertheless, it is necessary for a knowledge management system that uses protected material to support the protection of this material. But not only copyrights have to be respected. It is of equal importance to protect the moral rights of the authors.

The Trial-Solution project delivers knowledge to the reader only in the form of pdf documents, the actual LaTeX sources from which these documents are generated, are kept on the server. Adobe provides methods to restrict copying or printing of these documents that could be applied to further extend protection of the document. It would be also possible to include in each pdf document the login name of the user for whom it was generated so that in case of unlawful use he could be excluded from the further use of the server, though this is not applied.

To support the moral rights of the authors, but also of the copyright owners, the project has decided that each delivered slice from a document should be accompanied by the information from which document it came (the witness document) and what was its position in this document. A theoretical discussion of this concept was given in [LW01].

When documents are generated, the witness document and who is the author is given in the top line of each delivered page. For the information on the position in this document, the hierarchic file tree structure, described in the last section, is used. The root of the tree representing a sliced witness document is denoted by a unique identifier. Then each node in a tree has assigned a number, called *sourcereference*, that shows its position among the nodes having the same father node. Numbering of such nodes starts with 0 or 1 in order to be compatible with the eventual numbering in the witness document.

For example, if chapter 3 of a book starts with an introductory text, the introductory text gets sourcereference 0, section 1 gets source reference 1, section 2 gets sourcereference 2 etc.. In order to characterize the position of a slice in a witness document, the sourcereferences on the path from the root to this slice are collected in a string where they are separated by a / symbol. This string is prefixed by the identifier of the book and another /. The result of this is called the *logical identifier* of the slice. In our example, the second section of the third chapter of a book with the identifier mybook has the logical identifier *mybook/3/2*. Note that logical identifiers give a much more precise way to cite eventually small parts of documents than just mentioning the witness document only.

In order to include the logical identifier in the pdf documents generated for the reader, the project has developed two methods. The first method prints aside of all slices marginalia with their logical identifier. This method has the advantage that the logical identifiers remain visible also when the document is printed. It has the disadvantage that in some situations (for example for floating objects like figures or when the author has used the margin already) the automated typesetting system may have problems to find a good position for putting the marginalia. Though methods have been implemented to detect and handle such situations, a satisfactory solution may not always be possible.

Then a second more reliable method is used. When personalized pdf documents are created, they are equipped with pdf bookmarks. Each bookmark has a title for navigation, preceded by the source reference of the node in the hierarchic file tree represented by this bookmark. Thus, in the above example, the bookmark for the third chapter would consist of the title of the chapter with a prefixed 3 while the title of the second section has a prefixed 2. Then the logical identifier – up to the identifier of the book which however is replaced by the top line of each page – can be easily reconstructed from these sourcereferences.

In order to support assignment of appropriate credits, the Trial-Solution project uses the following *path delivery principle* whenever documents are to be generated:

> *Whenever a slice from a document is put in a composed sliced document, all nodes on the path in the hierarchical file tree from the root down to this slice together with their meta data must be included too.*

This makes sure that inherited meta data are always retained. Thus also information about the author, the publisher or rights of use attached only to the root node will always be included.

Though not supported within the Trial-Solution project, Slicing Book Technology provides the possibility to compose new personalized documents not from witness documents only, but also from documents that have been composed from slices of other documents (higher order composition). The project has agreed on the position that the compilation of a book from slices is an intellectual effort which deserves to be mentioned. The path delivery principle has been designed in support of this.

So, when some slices are taken out of two or more sliced documents, these slices are exported with the metadata on their path, including the identifier for the book and the information about authors and rights. The paths taken from each of these sliced books form a number of tree. Then the sliced book to be generated from the selected slices will be represented by a tree that is obtained from the exported trees, joint all immediately below a new root node. This new root node then can carry metadata denoting the title and the author of the compilation and eventual additional regulations for the rights of reuse. However, for each slice the information closer to the slice itself – especially that coming from the original author – will take precedence over the newly attached regulations.

6 Infrastructure

The Trial-Solution project has developed a series of tools to enable the delivery of the aforementioned services. These tools are

1. a tool for automated slicing of documents and assigning basic metadata, developed by Slicing Information Technology Berlin;
2. a tool for automated key phrase assignment, developed by CWI Amsterdam;
3. a web based tool for the manual revision of the results of the previous tools, developed by the University Koblenz-Landau and by the University of Nice Sophia Antipolis;
4. web based tools for unification and maintenance of thesauri developed by FIZ Karlsruhe and Trinity College Dublin;
5. a web based tool for generating personalized documents as described above. This tool was developed by the University of Koblenz-Landau and Slicing Information Technology Berlin, based in part on the aforementioned SIT-Reader. It contains an automated theorem prover developed at the University Koblenz-Landau for automated knowledge management.

These tools have been applied on a number of books from Springer-Verlag Heidelberg, from Verlag Harri Deutsch, from the Technical University Chemnitz, from the University Koblenz-Landau, from the University at Cologne but also from Teubner-Verlag and Heldermann-Verlag which are not members of the project. The main manual work on metadata assignment was performed at the Technical University Chemnitz. The Open University UK designed the evaluation procedures and the Heidelberg Academy of Sciences backed the aforementioned discussion on copyright issues.

For the future we expect a number of extensions of the knowledge management services described above. For example, the personalized books may be sent to a print on demand service. In fact, this has been already implemented in the project and is tested in cooperation with a *print on demand service* at the Technical University Chemnitz.

One deficiency of the technology described here is, that information on the knowledge of the user can be only obtained directly from the user. Due to the restriction of the source material – preexisting documents – an automated or semiautomated system to assess the knowledge of the user would have to be implemented separately. This goes beyond the project, however, the Delivery Tool of the project will be equipped with an open interface that permits external assessment systems, after passing a security check, to modify the user models. It is planned to test this interface by implementing a combination with the assessment system of the e-learning platform WebCT.

For a wide deployment, the current architecture will have to be extended. Instead of a central database of materials we envisage a distributed system where several sites offer their material, negotiate automatically about conditions of delivery within a framework of contracts between the rights owners (including the possibility of free delivery) and where the knowledge management systems of each site are specialized in certain fields and communicate with each other to find the most appropriate knowledge for their human clients.

References

[BF02] Baumgartner, P., Furbach, U.: Automated Deduction Techniques for the Management of Personalized Documents. Ann. Math. Art. Int. – Special Issue on Mathematical Knowledge Management, Kluwer 2002 (to appear)

[Da00] Dahn, I.: Symbiose von Buch und Internet. Proc. Learntec 2000, Karlsruhe 2000 551–558

[Da01] Dahn, I.: Automatic Texbook Construction and Web Delivery in the 21st Century. J. of Structural Learning and Intelligent Systems **14**(4) (2001) 401–413

[Da01a] Dahn, I.: Using Networks for Advanced Personalisation of Documents. Proc. SSGRR 2001, CD-ROM Edition, L'Aquila 2001

[JE01] Jensen, M.: Academic Press Gives Away Its Secret of Success. The Chronicle of Higher Education, Sept. 14, 2001

[LW01] Metadata for Advanced Structures of Learning Objects in Mathematics – An Approach for Trial-Solution. Trial-Solution internal Report 2001

[WD00] Wolter, H., Dahn, I.: Analysis Individuell. Springer-Verlag, Heidelberg 2000

RusDML – A Russian-German Project for Establishing a Digital Archive of the Russian Mathematical Publications

Galina A. Evstigneeva and Andrei I. Zemskov

Russian National Public Library for Science and Technology
12 Kuznetski Most, Moscow, Russia
fo2.@gpntb.ru, zemskov@gpntb.ru

Abstract. The article describes a project with the aim to develop the core for a distributed accessible digital archive for all Russian publications in mathematics. Russian mathematics had obtained a high world-wide reputation for quite a long period. Having the Russian mathematical publications available in digital form will provide an important contribution to the world wide initiative to establish the Digital Mathematics Library DML. The article addresses the main features of the project like the professional structure for the digitization process, collaborative structures for caring about the preparation of the bilingual metadata, international cooperation for caring about the input and providing a system of mirrors of the archive with several access possibilities. The description goes along a combined project proposal submitted to funding agencies in Russia and Germany. Many tools developed for the project will play a pioneering role and can be used for digitization activities in other fields

1 Objectives and Basic Requirements

The goal of RusDML is to digitize a core collection of Russian journals in mathematics, which so far available in printed form only and, by making them accessible in the web, to facilitate the world wide access to them. Having succeeded with this a further activity may go for comprehensiveness, i.e., to perform the digitiziation of all Russian mathematical publications, including monographs, series of collections of papers, encyclopaedias, handbooks, proceedings volumes, deposited articles etc.

Mathematics is a science where the availability of electronic publications and retro-digitised documents lead to considerable improvements 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, – mathematical results do not have a date of expiration. This gives evidence for the need and the actual demand to install RusDML.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 44–51, 2003.

The idea to pursue the project is in accordance with the world-wide tendencies of establishing digital collections of scientific publications (in particular publications on mathematics) providing Internet access to distributed archives. The idea of the global DML has been addressed in a first comprehensive way in the White Paper by J. Ewing ([JE]). In addition to better access, the digital mathematical archive would serve as an electronic repository and would provide a preservation facility for the printed collections. In this line one could see also a rational exploitation of library premises, providing better conditions for preservation especially for acid-paper publications, which are endangered by deterioration.

In accordance with these objectives the basic requirements for RusDML should be the following: The archive should be open and accessible world-wide. Distributed copies should guarantee the safety of the data and facilitate the access from different parts of the user community. The RusDLM should be part of the global network providing access to digital publications in mathematics. The development of RusDLM should integrate international co-operations. Links with other offers will be highly desirable in a later phase.

2 Ongoing Digitization Activities

As a consequence of the huge amount of publications to be retrodigitized for the final version of RusDLM several steps will be needed to reach the final goal. Even before starting the project it will make sense to check what already is available from ongoing projects and can be used for RusDML.

Scientific and experimental studies of the problems for the creation of digitized full text thematic collections of the scientific literature have been performed during several years. Some important existing repositories covering parts of mathematics are the following: ERAM (see [HB], [BW]), standing for Electronic Research Archive in Mathematics, and NUMDAM (see [NUM]) on the academic level, having digitized a total of 1.2 million pages so far and integrating publications of publishers like Springer among others; JSTOR, a multidisciplinary non-for-profit offer, located in the USA and providing charged access to about 30 mathematical journals; "Backfiles for the Future", developed and offered by the publisher Elsevier on the commercial level; DIEPER as a European project to investigate cooperation for digitization and development of appropriate tools and system on the multidisciplinary level, etc.

R&D activities on digital mathematical collections are pretty wide. There are several national digitization projects for mathematics in the application phase, and it is likely that they can rely on some seeding money from the European Union. The American side hopes to get some similar support from the American NSF. At Tsinghua University in Beijing they succeeded to digitize the old Chinese mathematics within a national project, but also countries like Columbia already undertook some steps to digitize their cultural heritage in mathematics. Comprehensive information on these activities could be obtained from the DML-homepage ([DML]).

Like the other offers RusDML will present a part of the global system, and total compliance with international trends, standards and protocols is taken care about through international cooperation. Thus world-wide accessibility will be guaranteed. More specifically the digital content will be accessible through both, direct web access and document ordering systems, delivering a requested issue or article at the user's office or at home with no extra difficulties. Therefore, the professional work using mathematics will be made more efficient. Furthermore, the retrospectively digitized content can be matched with the offer of born digital material, enlarging existing collections of full text electronic documents.

As a result of the DIEPER project a first sample issue for RusDML even will be available in advance to the project itself. The contents of the most traditional Russian journal in mathematics, Mat. Sbornik, had been scanned by the DIEPER partner in Helsinki, and after some additional work on the access data this journal will be available as a RusDML prototype due to kind agreements with the Helsinki University Library for using their files and the Russian Academy of Science (RAS), Moscow Branch, to make the digitized articles freely accessible through the web. Even more, the editors and scientists at RAS will have the possibility to enrich the information on their journal by comments, historical remarks or any kind of addition, which seems to be of interest in relation to the scientific merits of the corresponding article. This will be an added value for the journal, and it only can be obtained in a convenient way, after having the journal digitized, and equally important, after having provided a structure where useful additional information could be handled in a searchable way.

3 The Co-operational Network

As a key issue the Russian-German cooperation between several partners in both countries will be the organizational base of the project. Scientifically this is a consequence of the traditional good cooperation between Russian and German mathematicians for three centuries. As everybody knows, this cooperation survived some political catastrophes. But even now, when we have a period where Russian mathematicians partially try to publish in other languages, there is a comparatively high demand of Russian publications in Germany. It is no question that for Russian mathematicians the digital offer to be installed with RusDML will be a highly desirable improvement of their literature supply. But the two libraries involved on the German side still have the image to be reliable reference sites for this, and providing the content of RusDML will make them unique sites for users who are not likely to go to a provider in Russia. As a consequence, a bilingual access structure with enhanced facilities for those with weak Russian reading capabilities will be one of the most important requirements for RusDML. This feature also will be a pioneering work for other digital offers of Russian publications.

Going back to the details, the main Russian partner will be the Russian National Public library for Science and Technology (GPNTB) scientifically backed up by the Mathematics Division of RAS. Interests of other Russian libraries will be respected in bilateral agreements, because RusDML may establish GPNTB as one center of excellence for digital offers of Russian mathematics, but there is no aim to interfere with the interests of other mathematical libraries in Russia. Hence their aims and mission will be respected and taken into accordance, when delivery of documents, linking of offers and mirroring services should be taken into account.

The same philosophy applies to the German partners. These are the State and University library in Goettingen (SUB), the Technical Information Library in Hannover (TIB), representing the contributions from Zentralblatt MATH through its editor-in-chief, Professor Bernd Wegner, the Technical University Berlin (TUB). The different roles of these partners will be explained below. But as an essential common facility it had been agreed, that all three libraries, GPNTB, SUB and TIB, have the option and almost the obligation to install a full copy of RusDML. All project participants are well prepared for the collaboration, because they have pretty good collections in mathematics and they have long experience with the handling and administration of electronic offers.

GPNTB is a major library in Science and technology for the Russian Federation. Its collection comprises 8 million items of national and foreign publications. The library provides comprehensive approach to Russian collections in its role as a State Depositary and recipient of obligatory free of charge copies (mandatory copies) of all publications in their domain. All journals selected for RusDML are available at the collection of GPNTB, starting just from the first issue of the journal in its first publication year to the current production. Moreover, the library is experienced in library automation and information technologies. The OPAC contains more than 300.000 records since 1990. The Union Catalogue of publications in science and technology for Russia and the other New Independent States (NIS) (which describes the former USSR countries) contains more than 500.000 records since 1987.

There are 250.000 registered library users and annual circulation is 3.2 million documents. From a total of 350 PCs 135 are dedicated for the convenience of the users. The GPNTB web server registers about 20.000 visits daily. The library has good experience in cooperation with library networking projects like LIBNET, LIBWEB etc. The specialists at GPNTB have developed the electronic automated library administration and access system IRBIS, which had been installed and is used at more than 170 libraries in Russia and the CIS-countries. Evaluations by German specialists proved good compatibility of the IRBIS system with foreign library automation systems and therefore this system can be considered as professional and competitive on the international level. For details the web server of GPNTB may be consulted ([GPNTB]).

The Russian Academy of Sciences organizes born electronic offers of their journals and provides them freely for Russian users through their IZIR system. Like with Mat. Sbornik they will consider the digitization of their journals as an added value, and customize everything in a convenient way for their users. In this sense RAS clearly supports the RusDML initiative and this will be an extremely helpful assistance for the negotiations with other Russian publishers and editors for getting the license to digitize their materials. Concerning the scientific exploitation of the RusDML, RAS may play a leading role to explain the many advantages of the enhanced digital offers.

As mentioned above the Technical University Berlin (TUB) will play a special role through the editor-in-chief of Zentralblatt MATH, Professor Bernd Wegner. This is not a primary journal, its main role is to provide a comprehensive reference data base in mathematics. It provides bibliographical data, indexing information and reviews or abstracts in English. Hence the core metadata for RusDML will be available there, because all journals in RusDML are evaluated by Zentralblatt, and employing the linking facilities from the database to full text offers, it can be used as a simple access tool to the holdings of RusDML. Integrating the reviews as a special addition into the metadata, also users with low reading ability in Russian can decide, if they really want to go into the details of an article or not. Without adding the reviews the same role will be played by Zentralblatt for the other offers of DML.

The State and University Library in Goettingen (SUB) runs a professional digitization center with good technical equipment. They have developed a customized workflow combining the preparation of the documents for digitization, the scanning of the articles and the development of the access data in an integrated procedure. They have long-term experience in digitization projects. In particular, SUB participated in the DIEPER (Digitized European Periodicals) project and is providing the archive in the ERAM project. They are supported by DFG to provide a complete collection of publications on pure mathematics, which enables them to serve as an ideal background for digitization projects in this subject. The digitized collections will enable SUB to offer a better and quicker service to their users. They have a good experience in the installation and provision of access to digital collections.

One of the main objectives of the Technical Information Library in Hannover (TIB) is to develop and provide complete collections of publications in the areas of applied mathematics and engineering sciences. They have good experience with literature in Slavic languages and comprehensive collections of documents in these languages. Until recently they even prepared translations of Russian publications on particular request. They will make good contributions to help with the arrangement of the bilingual metadata, though for the Russian side this also will be a main task of GPNTB. The digitization of their good holdings of Russian mathematical literature will to add value to their library services, as it will be the case for all other participating libraries.

4 Content, Responsibilities of the Partners, and Workflow for RusdML

As mentioned above, there are some basic requirements for the project. Most importantly the digital archive should be easily accessible world-wide. All three participating libraries should spend combined efforts to take care of the long-term preservation and readability of the digital collections. Later upgrades of the offers can be imagined leading to more convenient access to electronic archive and to improved search facilities. For example, one important item is the linking from the references to their web offer. To achieve these goals all partners supposed to share their efforts and results and as a consequence they should serve as mutual mirror sites for the complete archive of RusDML.

The participating libraries will support international standards for RusDML as recommended by the DML project and others. The project is open for cooperation with other initiatives, should be able for further expansion to cover the Russian publications more completely. There will be additional digital collections provided by another institutions, which are not in the core list of documents recommended for RusDML, for example. As a first added value for the service provided full text search may be provided and links to and from reference databases are highly desirable.

A big variety of Russian publications in mathematics is available. Hence several stages of the project have to be considered. On the first stage RusDML will start with processing journals from a core list of about 120 titles, which are covered by bibliographical databases of Zentralblatt MATH and the Jahrbuch database. This core list should be corrected and extended. For instance, in the collection of GPNTB some quite interesting mathematical journals can be found which are not listed in the Zentralblatt database. For this a list of additions has been developed, possibly containing as a journal, which has been classified as a series of collections of articles by Zentralblatt. A registry of Russian publications in mathematics will be developed and extended during the project period. The following factors will be taken into account: There is an ambiguity in the interpretation of the document and publication type. For example, is it a journal or not. There are journals, which were published by USSR publishers and after the change, for the time being are published by a publisher in the New Independent States. There are titles, for which the publication has been cancelled, or it has been temporarily interrupted and then taken up later again, possibly under a new name.

Coming to the distribution of obligations and total amount of work the following figures have to be considered. Starting with the approximately 120 Russian journals of the first list, which have been processed by Zentralblatt and the Jahrbuch, joint estimates by GPNTB and SUB came to a figure of about 2 million pages. Using the existing digitization infrastructure at both sides it is agreed that the handling of the structural metadata, which are necessary for controlling the page numbers and the scanning are shared at equal parts between GPNTB and SUB. This delegates about 1 million pages to GPNTB. This will be done on the basis of uniform formats and common protocols with respect to technical issues.

For providing search and access facilities the articles have to be provided with metadata. Here GPNTB will be responsible for the Russian text of the metadata and SUB and TIB will care jointly about the English version of the metadata. To this purpose all three participating libraries should agree on the specification of the metadata format and structure. Generation of a first core set of English metadata is the task of the Technical University Berlin. The participating libraries will get these on the basis of a special permission from the editor-in-Chief of Zentralblatt and the Jahrbuch, both being the same person. They may be supplemented by the review/abstract article by article for the convenience of users with low Russian reading capabilities. These data are basic for the whole project and should be organized journal by journal, volume by volume and article by article in advance to the scanning. The provision of these data will initiate the digitization procedure. Simultaneously they serve as a reference to organize links to and from the databases. This will provide the user with a first access structure to the digitized articles even before the complete set of metadata will have been developed in both languages.

It is supposed to migrate from the first list of journals to a more complete one by 2003. The later list will be compiled after negotiations with publishers and editorial boards on for getting the approval for the creation of a digital version of the corresponding journal. There will be workshops and meetings between the participants of RusDML with the aim to determine the list of contributions of each participating libraries more precisely, taking into account special abilities of the partners, to exchange information on formats, tools and holdings and to improve the ability of the staff to handle new developments within RusDML.

A digitization plan will be elaborated for the period of the project. Preparatory evaluations of the compatibility of the software at the partners will be made. Technological details for the development of a uniform processing technology, the formation of the electronic archive, the installation of interfaces and user services, linking tools to Zentralblatt and the Jahrbuch data bases will be elaborated. Quick provision of easy access to the full text information will be the main goal of these efforts. For different tools test phases involving users may be considered.

Having in mind that prototypes of tools will be developed which may be portable to other digitization projects, good information dissemination on the results obtained in RusDML will be a key issue. This will be in accordance with the pioneering role of the project.

References

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

[DML] URL: http://www.library/cornell.edu/dmlib/

[GA] Evstigneeva, Galina A., Wegner, Bernd: O proekte sozdanija elektronnogo archiva russkich publikatsii po matematiki. Proceedings of LIBCOM 2002, Yershovo, November 2002 (to appear)

[JE] Ewing, John: Twenty Centuries of Mathematics: Digitizing and disseminating the past mathematical literature.
http://www.ams.org/ewing/Twenty_centuries.pdf

[GPNTB] URL: http://www.gpntb.ru

[NUM] URL: http://www-mathdoc.ujf-grenoble.fr/NUMDAM

[BW] Wegner, Bernd: 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.

Predicting the Future of Scholarly Publishing[⋆]

John Ewing

> I believe that the motion picture is destined to revolutionize our educational system and that in a few years it will supplant largely, if not entirely, the use of textbooks. *Thomas Edison, 1922*

> It is probably that television drama of high caliber and produced by first-rate artists will materially raise the level of dramatic taste of the nation. *David Sarnoff,1939*

When Orville Wright flew his airplane over a small stretch of rolling grassland in 1903, the managing editor of Scientific American[1] predicted that thousands of planes would soon fly over every city, delivering patrons to theaters. On the eve of the First World War, two famous British aviators[2] argued that planes would prevent wars in the future (because they brought people together). Scientists, engineers, and futurists have *always* conjectured the consequences of technology. In the case of planes, the experts were right in recognizing that they would profoundly affect our lives in the coming century... but they were certainly wrong in foretelling what that effect would be.

Once again the experts are predicting the future. The *digerati*[3] tell us that the Internet has changed everything, that technology will revolutionize the way we do business, and that nothing will again be the same. Maybe. But the experts provide few facts to back their predictions, and they preach a digital future as an act of faith rather than a reasoned conclusion. It's hard to tell hype from reality when someone promotes technology with religious zeal.

What about scholarly publishing? Here, a *special* group of experts is predicting (and promoting) the future. The experts foretell the imminent collapse of scholarly journals and some advocate revolutionary replacements – refereed postings, e-prints, and overlays. In many countries, government agencies have embraced these predictions, providing support for alternatives – PubMed Central, the Public Library of Science, the arXiv. And experts offer miraculous solutions to previously intractable problems, describing a revolution in scholarly publishing that will provide universal free access to scholarship... at no cost to anyone. The "free" alternatives seem to be enticing solutions to our present, very real problems.

[⋆] Based on a talk given at the Conference on Electronic Information and Communication, Tsinghua University, China, August 29-31, 2002.

[1] Waldemar Kaempfert, 1913

[2] Claude Graham-White and Harry Harper, 1914

[3] A term used in *Digital Mythologies,* Thomas Valovic, Rutgers University Press, New Brunswick, 2000.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 52–58, 2003.

How can we *predict* the future rather than merely *wish* for it? Good predictions are difficult without facts – facts about the past and about the present. This sounds obvious but, amazingly, experts and enthusiasts often dismiss past experience. They argue that since everything will soon change, experience is not relevant. This kind of sophism is especially prevalent in discussions about the Internet, where experts tell us the old rules no longer apply.

But they are wrong: Making predictions without facts is mysticism, not science.

1 Facts

Here are some facts about the current environment in scholarly publishing.

Alternatives to journals have been widely publicized, and some of these are remarkably successful. The best known in mathematics are the arXiv (http://www.arxiv.org) and MPRESS (http://mathnet.preprints.org). The former is a repository of papers, and the latter is a distributed system with links to repositories, including the arXiv itself.

- As of mid-2002, the mathematics sections of the arXiv hold approximately 20,000 papers, with about 15,000 of those contributed by individuals and the remainder migrated from previously existing preprint servers.[4] The total number of papers, including those migrated from other preprint servers, can be determined using the total number of papers given at http://front.math.ucdavis.edu/ .
- MPRESS has links to about 25,000 papers (including those in the arXiv).
- Since 1998, mathematicians contributed 12,618 papers to the arXiv (through mid-2002). During this same time, Math Reviews indexed more than 280,000 journal articles.

Alternatives to journals are more popular in some fields than in others, but they have played a prominent role in discussions about electronic publishing. In 2001, the Association of Learned and Professional Society Publishers conducted a survey of scholars in many disciplines[5]. One part of the survey considered preprint servers.

- When asked whether preprint servers were important in their work, about one-third (32%) said Yes. (Among physicists, 55% answered Yes.)
- When asked whether they *used* preprint servers, 12% of the respondents said they did. (Among physicists, 32% did.)

[4] The number of papers contributed by individuals can be determined by counting submissions for each year at http://arxiv.org/archive/math

[5] *The ALPSP research study on authors' and readers' views of electronic research communication*, Alma Swan&Sheridan Brown, Key Perspectives Ltd, ISBN 090734123-3. The survey was sent to approximately 14,000 authors of scientific papers across many fields. The response rate was about 9%.

Most scholars don't understand the scholarly literature – not its contents but rather its extent and complexity. When mathematicians think about "journals", they think about the best known and most visible – the ones they scan on the new-journals-shelf in the library. But the mathematical literature is far more complex and diverse. MR divides all journals into two classes: Those from which *every* article is either indexed or reviewed (the "cover-to-cover" journals) and those from which articles are selected for inclusion (the "others").

- In 2001, MR indexed or reviewed 51,721 journal articles[6].
- Those articles came from 1,172 distinct journals.
- In 2001, 591 (50%) of the journals were "cover-to-cover".
- That left 581 (50%) classified as "other".
- And 30,924 (60%) of the articles were in "cover-to-cover" journals.
- Leaving 20,797 (40%) articles in the "other" journals.
- This means that 40% of the journal literature is outside the "mainstream" mathematics journals!

Almost all discussions about scholarly communication focus on electronic publishing. There is a recognition that the transition from paper to electronic is proceeding more slowly than first imagined, but almost no one understands how slowly.

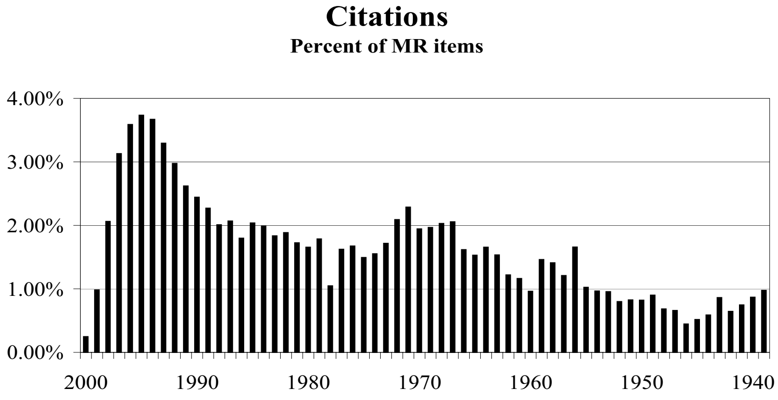
- In 2001, only 46 (4%) of the journals covered by MR were primarily-electronic[7].
- Only 1,272 (2.5%) of the articles were in primarily-electronic journals.
- On the other hand, approximately 34,000 (67%) of all articles had links, meaning that at least that many are available in electronic form.

Mathematicians have always known that past literature is important. Because MR recently added reference lists for articles for some journals, it is now possible to make the dependence more precise. The reference lists currently cover journals from 1998 to the present and include 336,201 citations to journal articles.

- Of all references, 53% were to articles published prior to 1990.
- More than 28% were to articles published prior to 1980.
- This is especially striking because the number of journal articles increased over time. Examining the number of papers covered annually by MR from 1950-1990, the percentage of MR items cited in recent papers varies between one and two percent for almost every year during the entire period.

[6] For present purposes, books, proceedings, and all items other than journal articles are not counted.

[7] The term "primarily-electronic" is not precise, but indicates journals that are either electronic only or that have a subsidiary paper copy added to the electronic version, which is viewed as primary.



Many scholars have commented about the high cost of commercial journals, but few have noted their number and size. They are gradually dominating the scholarly literature.

- In 2001, only 349(30%) of the journals were commercial.
- Yet they published 25,008 (48%) of the articles!
- Moreover, looking back ten years, we see a clear trend. In 1991, only 24% of the journals were commercial, publishing 38% of the articles.

What drives the expansion of commercial journals? While most scholars concentrate on the *costs* of journals, *revenues* are the crucial figures in understanding journal economics.

- A rough estimate[8] suggests that the revenue from each article in commercial journals generates about $4,000 in revenue (which may be off by a factor of 2.) Therefore, the 25,000 mathematics articles in commercial journals in 2001 generated about $100 million in revenue for commercial publishers.
- An even rougher estimate suggests that for the non-commercial journals, each article generates about half as much revenue. Even for these, therefore, the total revenue was about $50 million in 2001. (Again, this may vary by a factor of 2.)
- And it's important to remember that mathematics is only a small fraction of all scholarly publishing. There are about 25,000 journals in science, technology, and medicine alone[9]. Just one commercial publisher, Elsevier, derives more than a billion dollars in revenue from its science journals.

[8] Competition and cooperation: Libraries and publishers in the transition to electronic scholarly journals (see §2), A. M. Odlyzko. *Journal of Electronic Publishing* 4(4) (June 1999), www.press.umich.edu/jep/ , in the online collection *The Transition from Paper: Where are we Going and how will we get there?*, R. S. Berry and A. S. Moffatt, eds., American Academy of Arts & Sciences, www.amacad.org/publications/trans.htm, and in *J. Scholarly Publishing* 30(4) (July 1999), pp. 163-185.

[9] This figure is often quoted by the Association for Research Libraries, although it is hard to determine its precise source.

These are the facts: Many scholars (although not most) promote alternatives to journals, but many fewer actually use them. Journals continue to dominate the scholarly literature in mathematics. Almost all journals are in both paper and electronic format, and almost none are electronic-only. The journal literature is highly dispersed, contained in many journals, including those that cover disciplines outside mathematics. The older literature is extremely important for current research. And finally, commercial journals are taking over an ever-larger fraction of the literature, with enormous financial incentives driving the trend.

What should we conclude from these facts? Here are two alternative predictions.

Prediction 1: The alternative models expand and pressure journals. The independents, with only scant operating margins, diminish further. The commercial journals, with deep pockets, continue to expand and add features. Commercial publishers consolidate and eventually dominate the scholarly literature.

Prediction 2: The alternative models expand and pressure journals, driving out the independent journals. The alternative models solve all their problems – financing, covering the disperse literature, archiving, etc. The commercial publishers close down their journals and walk away with their enormous profits.

Which of these predictions is correct? Many scholars hope for the second; only the first is supported by the facts.

2 Ecology

What's bad about promoting technology rather than predicting its consequences? We discovered the answer recently when we examined what a century of technological progress had wrought. The answer is *ecology.*

We normally think of ecology in terms our natural environment, but ecology can refer to any system and its relationship to the surrounding environment. The ecology of scholarly publishing includes many things – a system of refereeing and reviewing, the use of publications in hiring and promotion, the way in which scholars view their legacy of research. Most experts on electronic publishing dismiss these things as unimportant; it's why so many get predictions wrong.

Should we worry about wrong predictions based on ignorance? Of course we should. If we fail to recognize that 40% of mathematical scholarship is published in multidisciplinary journals, we will design alternatives that ignore almost half the literature. If we believe that e-only journals are growing in number, when the number is shrinking, we may invest in the wrong trend. And if we ignore the fact that commercial journals take up not just more dollars but more shelf space as well, we risk sitting by like Nero while scholarly publishing is destroyed. Ignorance about the past and present of scholarly publishing is more than careless exuberance about the future; it means we can neither predict that future nor understand how to shape it. Remember – while some ecological disasters are caused by greed or malevolence, most catastrophes occur because well-intentioned people did not foresee the consequences of new technology.

The ecology of scholarly publishing is embedded in the far larger ecology of publishing, which currently has many forces driving change, and few of those forces have anything to do with scholarship or the academy. Ecological disaster for scholarly publishing would be swift and (largely) unnoticed by anyone outside academic life.

Many of the alternatives to journals may temporarily solve the problem of costs and speed of publication. For those who believe scholarly journals are merely a way for publishers to sell research back to the scholars who created it, this may seem like a fine solution. But people with publishing experience do *not* subscribe to this reductionist view. Journals are not just a way to distribute words on pieces of paper or screens; they are complicated institutions, involving authors, editors, libraries, researchers, publishers, professional societies, and administrators. Each has a role to play, and each has interests represented by the institution.

The institution of journals exists because scholarly publishing is not meant only for today's scholars but for future scholars as well – for our children and our children's children. Scholarly communication is more than sending papers to one's colleagues. Validation? Archiving? Financial incentives? These are all about sustaining scholarship for the future, not about exchanging papers in the present. Who will watch over collections when enthusiastic volunteers move on? Who will pay the costs of ever-changing servers and software to keep papers accessible? Who will provide the huge sums for archiving – not only saving the bits but updating the format of millions of papers? Surely we should not rely on government agencies, which have an increasingly short-term view in all their activities.

Many of the experts on electronic publishing assure us that these questions have easy answers. But we need to remember the lessons of the past: Predicting the consequences of technology is an uncertain business. Can we solve the problem of archiving in the future? To wave our hands with the assurance that technology will find solutions is like waving our hands for nuclear waste or carbon dioxide or fluorocarbons. We need to worry about the future because no one else will worry about something as fragile as scholarship.

3 Conclusions

What should we conclude? Should we steadfastly maintain the status quo? Do we avoid technology altogether? Of course not. *We should experiment; we should try out new things; we should tinker with technology and find better ways to communicate.*

But in carrying out our experiments, we need to be cautious and we need to be humble. We should remember that in the past smart people were unable to predict the effects of technology. There is no reason to believe that today's smart people are any better at predicting than yesterday's. Trying out anything that comes to mind without understanding the effect on the entire system of scholarly communication may be exciting, but surely it is not wise.

We also need to be forward-looking. The *essence* of scholarship is what we leave for future generations, not what we produce for today's. Scholarly communication is not about us – it's about the future of our discipline. Many enthusiasts who promote new projects ignore this principle. Make changes now, they argue and worry about whether they are sustainable later. But if scholars themselves don't worry about their future, who will?

What about the experts? Treat them with skepticism. More information is better? Maybe. But nearly everyone is experiencing information overload today; perhaps the quality of information is more important than the quantity. Faster is always better? Maybe. But the bottleneck on the Internet is the person receiving the information, who often is not able to process what is already provided. The Internet will solve the problems of scholarly communication? Maybe. But scholarship and the Internet are different in an essential way: The nature of scholarship is long-term; that of the Internet is transitory. Finally, be especially skeptical of the experts who demand that you are either with them or against them. Subscribe to their vision of the future or be branded a Luddite. This is a false dichotomy – resist it.

Responsible caution is not the same as mindless obstinacy. It is possible to promote electronic publishing without promoting the dissolution of institutions that have served us well. It is possible to cultivate and shape those institutions without ripping out their roots. It is possible to have a revolution without renaming the months.

If we have learned anything in the past century, it is that even the most useful technology can destroy those things we value most.

An Information System for Retrieving and Reasoning about Xml-based Mathematical Knowledge

Bernd D. Heumesser[1], Dietmar A. Seipel[2], Ralf-Dieter Schimkat[1], and Ulrich Güntzer[1]

[1] University of Tübingen, Wilhelm-Schickard Institute for Computer Science
Sand 13, D – 72076 Tübingen, Germany
{heumesser, schimkat, guentzer}@informatik.uni-tuebingen.de
[2] University of Würzburg, Department of Computer Science
Am Hubland, D – 97074 Würzburg, Germany
seipel@informatik.uni-wuerzburg.de

Abstract. Xml has become a key language for information interchange and integration over the World Wide Web. Representing and storing mathematical achievements and knowledge in a self-describing, extensible and open manner facilitates Web information systems which simplify co-operation among mathematicians world wide.
In this paper, we describe a *knowledge base management system* for Xml-based mathematical knowledge. Our information system provides different types of information retrieval techniques. Digitally stored and edited mathematical knowledge can be accessed by applying a well-defined and structured access to the types and elements of Xml-documents.
The information system transforms Xml-based mathematical knowledge into a complex Prolog-structure called *field notation*, which serves as the basis for building a digital library of fine-grained mathematical objects. Based on the field notation we provide a powerful and flexible *declarative query language* in a logic programming environment.

1 Introduction

Xml [12,14] is a widespread W3C-standard, which paves the ground for various new kinds of information systems accessible over the Web. For exchanging mathematical knowledge over the Web the Xml-based languages are a key technology. Two Xml-based languages dealing with mathematical knowledge are MathML [16] and OpenMath [3]. In this paper we focus on MathML, but to a great extent the techniques shown here are also applicable to OpenMath. Representing and storing mathematical achievements and knowledge in a self-describing, extensible and open manner facilitates Web information systems which simplify co-operation among mathematicians world wide. This knowledge, which is distributed over the Web, can only be handled if there is a possibility to query it based on the underlying mathematical structure. For instance the Mowgli-project [1] works on a development of a technological infrastructure for

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 59–71, 2003.

the creation and the maintenance of a virtual, distributed, hypertextual library of mathematical knowledge based on a content description of the information.

Using logic programming techniques for dealing with mathematical knowledge is an approach that has also been taken by Dalmas, Gaëtano and Huchet [4], who present a deductive database MFD2 for mathematical formulas. The MBASE system of Kohlhase and Franke [9] provides access to different theorem proving systems based on logic programming components. These two systems put the focus on using deduction for theorem proving.

In this paper, we describe a Web information system for XML-based mathematical knowledge. In particular, we are encoding the collection of ordinary differential equations given by Kamke [8] in the content markup of MATHML. The described information system provides different kinds of information retrieval techniques: The digitally stored mathematical knowledge can be accessed by applying a well-defined and structured language to address the types and elements of MATHML-documents. In addition, we can perform *reasoning tasks* about the knowledge using a rule-based approach in logic programming. From an architectural point of view our information system transforms a MATHML-document into an equivalent PROLOG-structure called *field notation*, which serves as our *Document Object Model* (DOM) for building a digital library of fine-grained mathematical objects. Based on our field notation we provide a powerful and flexible *query language* in a logic programming environment, which bridges the gap between XML-based digital libraries and deductive (inference-based) databases. From a general point of view our system may be regarded as an *intelligent search engine* that is tailored to the application domain of MATHML-documents because it is capable of querying mathematical formulas based on their structure.

The rest of the paper is organised as follows: In Section 2 we present the techniques available for managing mathematical knowledge represented in MATHML. In Section 3 we introduce a data structure for the representation of XML-documents in PROLOG. A PROLOG-library for dealing with this data structure is presented in Section 4. Then we apply this library for the classification of ordinary differential equations and for reasoning tasks on mathematical knowledge. Finally, in the conclusions we give a short outlook to our further research.

2 Managing XML-based Mathematical Knowledge

The eXtensible Markup Language XML is a document language; it is a subset of SGML [7]. Specific XML-languages, so-called XML-applications, for special application domains can be defined.

The main objective of XML is the separation of *content* and *presentation*. The user specifies only the structure of the data. For presenting the data a special presentation schema is used, such as, Cascading Style Sheets or the eXtensible Style Sheet Language XSL [15]. XML is a term-structured language, since XML-documents are labelled trees. XML is well-suited – and widely used – for representing semi-structured data.

Many tools and software systems are capable of dealing with XML. Most systems support XML by enabling operations on the general grammatical structure of XML. They use a Document Type Definition (DTD) or an XML-Schema to support XML-applications of a given domain; but in general they only provide generic operations on the XML-based knowledge. In order to take into account the special semantics of an XML-application it is necessary to define *sophisticated functionality* on top of these *generic operations*.

2.1 The Markup Language MathML

MathML is an XML-language for describing various types of mathematical knowledge. It is also a W3C-standard, which can be incorporated into other XML-documents, such as XHTML [17].

```
<apply><eq/>
  <apply><plus/>
    <apply><diff/>
      <bvar>
        <ci>x</ci>
        <degree><cn type="integer">1</cn></degree>
      </bvar>
      <apply><ci type="fn">y</ci><ci>x</ci></apply>
    </apply>
    <apply><times/>
      <ci>a</ci><apply><ci type="fn">y</ci><ci>x</ci></apply>
    </apply>
  </apply>
  <apply><times/>
    <ci>b</ci>
    <apply><sin/>
       <apply><times/><ci>c</ci><ci>x</ci></apply>
    </apply>
  </apply>
</apply>
```

Fig. 1. The Ordinary Differential Equation $y' + a \cdot y = b \cdot \sin(c \cdot x)$ in MathML.

Like XML, MathML has been designed for serving, receiving and processing content on the Web. The presentation markup may differ from the mathematical structure, because for presentation purposes it is not necessary to specify the exact mathematical structure. The content markup is a language in which mathematical structures can be defined in an exact way. For retrieving and reasoning about mathematical data we use only the content markup of MathML. For example, Figure 1 shows an *ordinary differential equation* (ODE) in MathML-notation. The operator following `apply` is `diff`, which stands for derivation. The first operand consists of a bound variable, for which the derivation is computed, and a degree, which indicates how often to derive.

Some browsers are capable of presenting MathML, e.g., the new Mozilla software, and a lot of other mathematical software can also work with with MathML.

2.2 A Knowledge Base for MathML-documents

We try to answer the question which operations are available for the rather large mathematical knowledge base of MathML-documents. How can retrieval technology be applied to MathML? Which methods for reasoning are available? A potential user, like an engineer, may want to find a solution or a link to a publication for a given ODE. A mathematician may want to do some reasoning on the data such as finding out whether there is an ODE with certain properties in the knowledge base.

We use concepts from various areas for handling Xml-based knowledge, cf. Figure 2: XQuery [18,19] is a general Xml-query language, which is under the process of standardization by the W3C. With XQuery an Xml-document can be queried and transformed. The retrieval facilities of XQuery are not specific to mathematical knowledge, where some additional features are required. Matching a term with another term w.r.t. mathematical structures such as variables, constants and term structure is done by *term rewriting technology*. However, for some of the tasks the technology is not sufficient. *Reasoning tasks* and the option to use a programming environment to extend the given features are required. This can be done in the programming language Prolog [10].

	Xml	Retrieval	Matching	Reasoning	Progr.
XQuery	⋆	⋆			
Term Rewriting			⋆		
Prolog			⋆	⋆	⋆
FnQuery	⋆	⋆	⋆	⋆	⋆

Fig. 2. Available Technology for dealing with Mathematic Knowledge and Xml.

Our approach combines all these methods in an Xml-enabled Prolog-environment called FnQuery. In our system Xml-documents can be represented in a natural way, and it can be queried easily. *Term rewriting* may be incorporated and the embedding in a Prolog–engine supplies the required reasoning and programming facilities. By combining existing techniques in a *logic programming environment* we have built a powerful and flexible system that meets the demands of the new formats of mathematical knowledge.

3 A Document Object Model for Xml-documents

We have built a library in Swi-Prolog [13] featuring complex reasoning tasks about the mathematical knowledge. Prolog has been used as a programming language; in addition the Prolog-engine has been applied for rule-based deduction. We use a Prolog-Dom for Xml-documents with operators for accessing compenents of documents that has originally been introduced in [11].

When working with Prolog, the following *characteristics* of the language are important. Prolog is a logical programming language using Sld-resolution (Sld: Selection rule driven Linear resolution for Definite clauses) [10] as its basic inference mechanism. Further concepts are *backtracking* to find all answers for a query and *unification* for computing variable bindings; this supports relational programming. Furthermore, various libraries are available for Prolog: a library for accessing Sgml- (and hence Xml-) documents, a library for accessing and handling documents on the Web, and a library for building graphical user interfaces.

Prolog can handle term and tree structures nicely. Since Xml-documents are tree-structured, Prolog is expected to be able to handle Xml-documents well. However, there are some problems: Typically, an attribute value of a predicate is accessed by selecting the corresponding position: for example, for the schema ode_classification(Number, Degree, Type, Class) the type T of an ODE with the degree 2 can be selected by ?- ode_classification(_, 2 , T, _). Note that variables in Prolog always start with a capital letter. Assignments in Prolog are always non-destructive: during computations it is not possible to change components of a structure without creating a new structure representing the updated structure.

3.1 Representation of Complex Objects in Field Notation

A complex object O with attribute/value-pairs $\mathsf{a_i : v_i}$ can be represented as an *association list*

$$\mathsf{O} = [\mathsf{a_1 : v_1}, \ldots, \mathsf{a_n : v_n}].$$

If the values are complex objects themselves, then they can be represented as association lists as well; a complex structure can be represented by nested association lists.

This notation leads to certain advantages: The ordering of the attribute/value-pairs is arbitrary. The values are accessed by their names rather than by argument positions. The schema of the data is not fixed and may be changed at run time. This gives extra flexibility for operating on documents. Null values can be omitted and need not be represented – on the other hand we can add new values at run time.

We have extended this formalism for representing Xml-documents in Prolog. This approach is sufficient, but not limited to Xml-documents. It can even be used for structures which are more complex than Xml. For representing Xml-documents in Prolog a new data structure called *field notation* has been introduced, which allows for accessing the components of an XML-document very easily. The field notation serves as our PROLOG-DOM for XML-documents. XML-documents can have attributes and sub-elements. Thus, an XML-document with the tag name “T” can be represented as a triple $\mathsf{T} : \mathsf{As} : \mathsf{C}$, where “As” is an association list for its attribute/value-pairs and “C” represents its content, i.e., its sub-elements. We call the notation

$$\mathsf{O} = [\mathsf{a_1} : \mathsf{v_1} : \mathsf{w_1}, \ldots, \mathsf{a_n} : \mathsf{v_n} : \mathsf{w_n}] \tag{1}$$

for a list of XML-documents *field notation.* Here $\mathsf{a_i}$ is a sub-element of O with the list $\mathsf{v_i}$ of attributes and the content $\mathsf{w_i}$. An XML-document can be represented by a list $[\mathsf{T} : \mathsf{As} : \mathsf{C}]$ containing one element. For example, the following triple represents a part of the XML-document given in Figure 1:

```
apply:[ ]:[
   ci:[type:fn]:[y],
   ci:[ ]:[x] ]
```

Thus, semi-structured data, like XML-documents, can be represented nicely in field notation.

3.2 Access to Components in Field Notation

Techniques for accessing, querying, and updating semi-structured data elegantly are also available. In the following we give some examples for the usage of the field notation and the FNQUERY-language. A more detailed description can be found in [11].

For an object `O` of the form (1) in field notation we can select the list `X = w`$_i$ of sub-elements using the statement `X := O^a`$_i$, and we can select the list `Y = v`$_i$ of attribute/value-pairs using the statement `Y := O@a`$_i$. Here “`:=`” is a binary infix-predicate symbol, which evaluates the second argument, i.e., the terms `O^a`$_i$ and `O@a`$_i$, and assigns the result to the first argument, i.e., the variables `X` and `Y`, respectively.

```
?- O = [ apply:[ ]:[ ci:[type:fn]:[y], ci:[ ]:[x] ] ],
   X := O^apply, Y := O@apply.
X = [ ci:[type:fn]:[y], ci:[ ]:[x] ], Y = [ ]
```

This reminds of the evaluation of an arithmetic expression “X is $3 * (4 + 5)$” in PROLOG where “is” is a binary infix-predicate symbol, which evaluates the arithmetic term “$3 * (4 + 5)$” and assigns the result to the first argument X.

It is possible to have *complex paths expressions* for selecting sub-elements:

```
?- O = [ apply:[ ]:[ ci:[type:fn]:[y], ci:[ ]:[x] ] ],
   X := O^apply@ci^type.
X = fn
```

In pure PROLOG the previous statement would look much more complicated.

A query selecting with multiple path expressions returns a list of objects, namely one object for each selector:

```
?- O = [ apply:[ ]:[ ci:[type:fn]:[y], ci:[ ]:[x] ] ],
   X := O^[apply, apply@ci].

X = [ [ ci:[type:fn]:[y], ci:[ ]:[x] ], [type:fn] ]
```

If a path expression contains variable symbols, then all ground instances of the path expression which are an allowed path in the queried object can be generated on backtracking:

```
?- O = [ apply:[ ]:[ ci:[type:fn]:[y], ci:[ ]:[x] ] ],
   X := O^apply^Path.

X = [y], Path = ci ;
X = [x], Path = ci
```

Finally, using the operator "*", we can *assign new values* to attributes or elements as follows:

```
?- O = [ apply:[ ]:[ ci:[type:fn]:[y], ci:[ ]:[x] ] ],
   X := O*[^apply@ci^type:real].

X = [ apply:[ ]:[ ci:[type:real]:[y], ci:[type:real]:[x] ] ]
```

Notice, that the attribute `type` was changed from `fn` to `real` in the first `ci`-element, while it was set in the second `ci`-element, since no such attribute existed before.

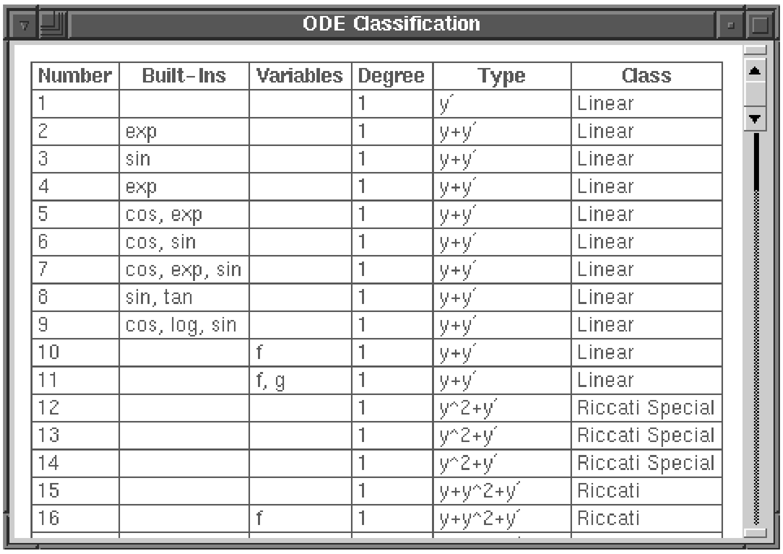
4 Retrieval of Ordinary Differential Equations

In this section we show how the field notation and the FnQuery-language are applied to MathML-documents. Since we are working on a Web-based expert system for ordinary differential equations, some results from this application domain are presented here.

In our current system we can search for a given equation w.r.t. variables and variable function symbols based on unification in Prolog. We are going to incorporate a term rewriting library to allow for enhanced unification possibilities. But before we try to unify the terms we classify and filter the equations to reduce the number of candidates. This kind of filtering is done with FnQuery.

4.1 Classification of Ordinary Differential Equations

Some *properties* are computed for each given ordinary differential equation in the knowledge base. We locate function symbols, which are defined in MathML, e.g., `cos` or `exp`. We build a set of variable function symbols (identifiers with the value `fn` in the `type`-attribute) used in the equation. The system computes the degrees of the derivations; since it is possible for a function to appear in different degrees of derivation, this is a set, too. Then we generate what we call the *type* of an equation, which indicates how the function and its derivatives appear in the equation. Based on these properties we classify a given equation. Our classification predicates assume by default that x is the argument of the derivated function; this can be changed by specifying other values.



ODE Classification

Number	Built-Ins	Variables	Degree	Type	Class
1			1	y´	Linear
2	exp		1	y+y´	Linear
3	sin		1	y+y´	Linear
4	exp		1	y+y´	Linear
5	cos, exp		1	y+y´	Linear
6	cos, sin		1	y+y´	Linear
7	cos, exp, sin		1	y+y´	Linear
8	sin, tan		1	y+y´	Linear
9	cos, log, sin		1	y+y´	Linear
10		f	1	y+y´	Linear
11		f, g	1	y+y´	Linear
12			1	y^2+y´	Riccati Special
13			1	y^2+y´	Riccati Special
14			1	y^2+y´	Riccati Special
15			1	y+y^2+y´	Riccati
16		f	1	y+y^2+y´	Riccati

Fig. 3. Classification of Ordinary Differential Equations.

The classification of some ordinary differential equations is given in Figure 3. For example, the classification of the equation

$$y' + y^2 + (x \cdot y - 1) \cdot f(x) = 0$$

can be found with number 16. The number of the equation is shown in the first column. The second column contains the appearing constant function symbols. The third column indicates variable function symbols. The type of the equation and the class are found in the last two columns.

4.2 Classification Predicates Using FnQuery

The following predicate uses FnQueryfor computing the list `Degrees` of the degrees of an ordinary differential equation `Equation` that is given in field notation; the degrees are computed w.r.t. a given function `Function`:

```
mathml_to_derivation_degree(Function, Equation, Degrees) :-
   findall( Degree,
      ( ( Xs := Equation^_^apply
        ; Xs := Equation^apply ),
        _ := Xs^diff,
        Degree := Xs^bvar^degree^cn,
        [Function] := Xs^ci ),
      Degrees ).
```

The predicate `findall/3` uses backtracking for computing all solutions for the goal given by its second argument. We take the equation and search for an `apply`-tag in an arbitrary position within the term. Then we check if the `apply`-element contains the operator `diff`. Next we extract the degree by accessing the path. Finally, we check if the operator `diff` is applied to the function `Function`. Since the content of an Xml-element is a list, we have to enclose the function in list brackets when we check if this is the content of the `ci`-tag.

By decomposing the equation into sub-terms we get small terms like the one shown in the example below. If the terms are simple enough, their type can easily be determined. By passing the results backwards with respect to the operators passed the way down, we can assemble the type of the equation. The following rule computes the type of `Formula` w.r.t. the function `Function`; it covers only one case – the other cases are handled by alternative rules:

```
mathml_to_term_type(Function, Formula, Term_Type) :-
   _ := Formula^diff, [Function] := Formula^ci,
   !,
   [Degree] := Formula^bvar^degree^cn,
   Term_Type = [[(Degree, 1)]].
```

The predicate `mathml_to_term_type` is called within a top-down analysis of the operator tree of the equation. First we check if the formula contains the `diff`-operator, then we check if it is applied to the function `Function`. We have to enclose the variable for the function in list brackets, since the selected content is a list. Then we check whether the `ci`-tag contains only the given name of the function. Here is an example for determining the term type of the formula y'':

```
?- mathml_to_term_type(y, [
      diff:[]:[],
      bvar:[]:[ci:[]:[x], degree:[]:[cn:[type:integer]:[2]]],
      apply:[]:[ci:[type:fn]:[y], ci:[]:[x]] ],
      Term_Type ).

Term_Type = [[(2, 1)]]
```

Note that the second argument is the content of an apply-element. With respect to the function y given by the first argument, this term has the type (2,1) indicating that it contains a second derivative of y in the first power.

The following predicate mathml_to_xpce_table uses the previously described predicates for computing the table given in Figure 3:

```
mathml_to_xpce_table(Function, Equations) :-
   findall( [Fs_1, Fs_2, Degree, Type, Class],
      ( Eq := Equations^apply,
        _ := Eq^eq,
        mathml_to_function_symbols(Eq, Fs_1, Fs_2),
        mathml_to_derivation_degree(Function, Eq, Degree),
        mathml_to_term_type(Function, Eq, Type, Class) ),
      Rows ),
   xpce_display_table( 'ODE Classification',
      ['Number', 'Built-Ins', 'Variables',
       'Degree', 'Type', 'Class'],
      Rows ).
```

It is assumed that the root element of each equation has the tag apply and that this root element contains a sub-element with the tag eq. The objects extracted by backtracking are all equations. These equations are now processed in the following predicate calls. One of them is the predicate shown above which generates a list of numbers of degrees that occur in derivation.

4.3 Search for an Ordinary Differential Equation

As already mentioned, the term type of an ordinary differential equation is used to shrink the number of candidates for the search for the equation. For example, the term type of the following equation is $y' + y$:

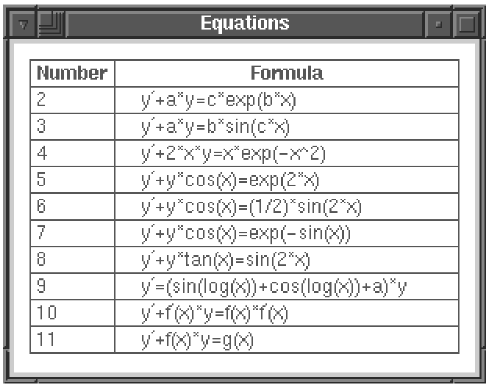
$$y' + y \cdot f(x) = a \cdot \sin(2 \cdot x) \tag{2}$$

The predicate mathml_term_type_to_equations selects all equations with a given term type from the list Equations of all ODEs in field notation:

```
Equations = [ ...,
   3:[ ]:[
      formula:[ ]:[[apply:[ ]:[eq:[ ]:[ ], ...]]],
      built_ins:[ ]:[sin],
      variables:[ ]:[ ],
      degree:[ ]:[1],
      type:[ ]:[y+y'],
      class:[ ]:[Linear] ], ... ]
```

Equations contains one element for each ODE: the tag is the number of the ODE and the content represents the properties. The formula-element contains the equation in field notation.

Figure 4 shows the result of the selection of the equation with the term type $y' + y$. The first column contains the number of the equation. The second column shows a term representing the equation. The further properties of the equations can be obtained by joining the tables of Figure 4 and Figure 3 on the column for Number.



Equations

Number	Formula
2	y´+a*y=c*exp(b*x)
3	y´+a*y=b*sin(c*x)
4	y´+2*x*y=x*exp(-x^2)
5	y´+y*cos(x)=exp(2*x)
6	y´+y*cos(x)=(1/2)*sin(2*x)
7	y´+y*cos(x)=exp(-sin(x))
8	y´+y*tan(x)=sin(2*x)
9	y´=(sin(log(x))+cos(log(x))+a)*y
10	y´+f(x)*y=f(x)*f(x)
11	y´+f(x)*y=g(x)

Fig. 4. Ordinary Differential Equations with the Term Type $y' + y$.

Given the list `Equations` of ODEs and an ODE `Query` in field notation, the following predicate `mathml_query` determines the matching equations and returns a list `Substitutions` of substitutions:

```
mathml_query(Function, Equations, Query, Substitutions) :-
   mathml_formula_to_term_type(Function, Query, Type),
   mathml_term_type_to_equations(Equations, Type, Es),
   findall( Number:Substitution,
      ( Formula := Es^Number^formula),
        mathml_unify(Formula, Query, Substitution) ),
      Substitutions ).
```

Firstly, the type `Type` of the query equation `Query` is determined; then this type is used for selecting elements of `Equations`. The resulting list `Es` is accessed within the predicate `findall`: The expression `Es^Number^formula` extracts the formula of the equations. Note that the variable `Number` is unified with the number of the equation. After that, the predicate `mathml_unify` is called to check if it is possible to unify the selected formula with the formula of the query equation. If this call succeeds, then the substitutions for the two formulas are returned and the pairs `Number:Substitution` are returned in the list `Substitutions`. For example, if `Query` represents the ODE $y' + y \cdot f(x) = a \cdot \sin(2 \cdot x)$ given in Equation (2) in field notation and `Equations` represents the knowledge base of all ODEs, then we get the following:

```
?- Equations = ..., Query = ...,
   mathml_query(y, Equations, Query, Substitutions).
Substitutions = [
   3:[ apply:[ ]:[ci:[type:fn]:[f], ci:[ ]:[x]] - ci:[ ]:[a],
       ci:[ ]:[a] - ci:[ ]:[b],
       cn:[type:integer]:[2] - ci:[ ]:[c] ],
   6:[ apply:[ ]:[ ci:[type:fn]:[f], ci:[ ]:[x]] -
          apply:[ ]:[cos:[ ]:[ ], ci:[ ]:[x]],
       ci:[ ]:[a] -
          apply:[ ]:[ divide:[ ]:[ ],
             cn:[type:integer]:[1], cn:[type:integer]:[2]] ],
   8:[ apply:[ ]:[ci:[type:fn]:[f], ci:[ ]:[x]] -
          apply:[ ]:[tan:[ ]:[ ], ci:[ ]:[x]],
       ci:[ ]:[a] - cn:[type:integer]:[1] ] ]
```

Three matching equations were found: 3, 6, and 8. E.g., the first substitution says that $f(x)$ is replaced by a (specialisation), a by b, and 2 by c (generalisation). If we apply this substitutions to 2, then this equation is identical (modulo commutativity of "$\cdot$") to equation 3 of Figure 4, i.e., $y' + y \cdot a = b \cdot \sin(c \cdot x)$.

5 Conclusions

We presented a system for the retrieval and the reasoning about XML-based mathematical knowledge. The introduced field notation is well-suited for representing and processing XML-documents in PROLOG. Embedded into this logic programming environment we have built a flexible, declarative query language FNQUERY for XML-documents that is based on *path expressions* and *structural recursion*.

We have applied our system to the domain of ordinary differential equations that are represented in MATHML. We can do exact match search (retrieval) as well as classification into Kamke's slots by PROLOG-reasoning for a substantial part of ODEs.

Our further work will go into two different directions. We are currently working on more sophisticated techniques for classifying ordinary differential equations. To enhance the facility for searching a given equation we are going to incorporate a term rewriting library. We are building a graphical user interface with a more readable external language, which is transformed to FNQUERY.

The Semantic Web Initiative [2] is working on the problem of adding *a logic component for reasoning* about documents on the Web. We apply our results towards building a sophisticated search engine for MATHML-documents. As shown in [6] it is possible to extend our techniques to other XML-based languages for representing mathematical knowledge such as OPENMATH.

Acknowledgement. The authors would like to thank R. Loos for drawing their attention to the problem of classifying Kamke's list of differential equations.

References

1. Asperti, A., Wegner, B.: Mowgli – A New Approach for the Content Description in Digital Documents. Proceedings of the 9th Intl. Conference on Electronic Resources and the Social Role of Libraries in the Future, Section 4, Volume 1, 2002
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, May 2001
3. Caprotti, O., Carlisle, D. P., Cohen, A. M.: The OpenMath Standard. The OpenMath Esprit Consortium, February 2000
4. Dalmas, S., Gaëtano, M., Huchet, C.: A Deductive Database for Mathematical Formulas. In: J. Calmet, C. Limongelli (Eds.): Design and Implementation of Symbolic Computation Systems. Springer LNCS, 1996
5. Heumesser, B., Schimkat, R.: Deduction on Xml Documents: A Case Study. Proceedings of the 14th International Conference on Applications of Prolog INAP'2001 – Stream Content Management, 2001
6. Heumesser, B., Seipel, D., Güntzer, U.: An Expert System for the Flexible Processing of Xml-Based Mathematical Knowledge in a Prolog-Environment. Proceedings of the Second Internation Conference on Mathematical Knowledge Management MKM'2003, Springer LNCS (to appear), 2003
7. International Standards Organization: Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML). ISO 8879, October 1986
8. Kamke, E.: Differentialgleichungen – Lösungsmethoden und Lösungen. Akademische Verlagsgesellschaft, 8th Edition, 1967
9. Kohlhase, M., Franke, A.: MBase: Representing Knowledge and Context for the Integration of Mathematical Software Systems. Journal of Symbolic Computation **32**(4) (2001)
10. Lloyd, J.: Foundations of Logic Programming. Spinger, 2nd Edition, 1987
11. Seipel, D.: Processing Xml-Documents in Prolog. Workshop on Logic Programming WLP'2002
12. Suciu, D., Abiteboul, S., Bunemann, P.: Data on the Web – From Relations to Semi-Structured Data and Xml. Morgan Kaufmann, 2000
13. Wielenmaker, J.: Swi-Prolog 5.0 Reference Manual. `http://www.swi-prolog.org/`.
14. Extensible Markup Language (Xml) 1.0. World Wide Web Consortium, October 2000. `http://www.w3.org/TR/2000/REC-xml-20001006`
15. Extensible Stylesheet Language (Xsl). World Wide Web Consortium, October 2001, `http://www.w3.org/TR/xsl/`
16. Mathematical Markup Language (MathML) Version 2.0. World Wide Web Consortium, February 2001, `http://www.w3.org/TR/MathML2/`
17. Xhtml 2.0. World Wide Web Consortium, August 2002, `http://www.w3.org/TR/2002/WD-xhtml2-20020805/`
18. Xml Path Language (XPath). World Wide Web Consortium, November 1999, `http://www.w3.org/TR/xpath`
19. XQuery 1.0: An Xml Query Language. World Wide Web Consortium, August 2002, `http://www.w3.org/TR/xquery`

International Copyright and Mathematics

Wilfrid Hodges[*]

School of Mathematical Sciences Queen Mary, University of London
w.hodges@qmul.ac.uk

1 Mathematicians and Copyright: A Case of Mutual Indifference

Nobody should suppose that intellectual property law was created for the benefit of mathematicians. The chief direct effect of intellectual property law on mathematics is to restrict the copying of mathematical publications. In a world where (we are told) the average mathematical paper is read by 1.6 people, most mathematicians pray for anything that will give their writings a wider circulation.

Nor should mathematicians imagine that the rest of the world cares what the mathematical community thinks about intellectual property law. Mathematicians form a small part of the academic world, and academia as a whole has only a minor voice in the discussions that shape intellectual property law today, either at international or at national level. For example, who spoke at the WIPO International Symposium on the Effects of Technology on Copyright and Related Rights, 2000 [15]?

> Speakers included representatives of the American publishing, motion picture, software and database, recording and photographic industries.

Who are the members of the UK Intellectual Property Advisory Council [21]?

- A company chief executive
- Two directors of university centres for entrepreneurship or IP research (presumably included for their expertise rather than as representatives of universities)
- Two attorneys and two solicitors
- Two inventors who are in business
- An IP manager in a commercial company
- An economist
- The director of a bioethics council.

[*] My warm appreciation to my former colleagues on the IMU Committee on Electronic Information and Communication (CEIC) and to my London colleagues Gary Lea and Chris Reed, for introducing me to many of the issues discussed in this paper. The opinions expressed are my own. My thanks also to Bernd Wegner and the organisers of the ICM 2002 satellite conference on Electronic Information and Communication in Mathematics.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 72–86, 2003.

As these lists illustrate, the pressures that shape intellectual property law come mainly from *the commercial or financial exploitation* of intellectual property.

A few years ago my colleague Bruce Watson and I sought permission from the widow of Kurt Gödel to publish a translation of a paper of his (in which he proved the consistency of elementary arithmetic). In her kind reply she wrote that her attorney had advised her to find out from us the commercial value of the translation. We told her it was nil, and she gave permission. I sometimes wonder how we might have made money out of the translation. Setting it to music, perhaps. But the moral is that most published mathematical papers earn very little money for the mathematician, though they do usually help to provide income for the publisher. From books the mathematician can earn more, but generally not much unless they are elementary textbooks.

So there is usually not much point in taking a mathematician to court on an intellectual property issue. The financial pickings are too small. In the few instances reported below where a mathematician was involved in a legal action, the mathematician always had a contract with a commercial company that was joined in the action.

The last forty years have seen some important changes in Western thinking about the purpose of university research. Most universities now have their technology transfer offices, set up to guide research towards commercial exploitation, always in hopes of bringing money into the university. This trend has ignited some fierce battles over the ownership of research, and sometimes over its secrecy too. As yet, mathematicians are not too visible in these battles. For example one of the protagonists in the current debate at Cambridge University discussed below is a well-known mathematician [14], but his contribution makes no reference to mathematics.

After my talk at the CEIC afternoon in the Beijing 2002 Congress, a member of the audience suggested that since nobody has much interest in taking mathematicians to court, mathematicians should do what suits them in intellectual property matters and defy anybody to stop them.

There is a sound point here: if we mathematicians don't fight for our interests, nobody else will. But I wouldn't recommend anarchy. Whether or not it was put there for our benefit, intellectual property law is a fact. Its chief role is to allow people to make binding agreements with each other about the commercial use of intellectual property, and we do ourselves a service if we use this law wisely. Wisdom consists of knowing what one can and can't do, and what things other people are likely to want to do. For this reason the CEIC published a copyright checklist 'What do you want from your publisher?' [10].

In case I am being too pessimistic about the legal punch of mathematicians, I should mention that Pamela Samuelson, a leading authority on academic copyright, has several times said that she foresees academics becoming more effective lobbyists on copyright issues. For example she recently suggested that [19]

> U.S.-based cryptographers may need to become active legislatively ... to help Congress understand why certain changes need to be made to the [Digital Millennium Copyright Act] ...

And of course the future could be different. Mathematicians might regularly become rich through exploiting the technological applications of their work. (Already quite a few trained mathematicians gather respectable salaries as patent attorneys.) We can dream.

2 Patents

There are several types of intellectual property law. Probably the simplest to explain is the law of patents. If you invent a method for doing or making something useful, and then patent it, the patent gives you a monopoly on industrial or commercial use of the method for twenty years. It's clearly against the public interest to restrict the use of a good idea to a single person, so there have to be compensating benefits to justify granting patents. One is that the patent holder can collect the financial reward for having discovered the method. A second is that the patent holder, in exchange for getting the patent, is obliged to disclose what the method is, so that anybody can use it when the patent expires.

There is a steady trickle of mathematical patents. Most of them are patents of algorithms, for example [23]

> United States Patent 5,373,560
> Schlafly December 13, 1994
>
> ---
>
> Partial modular reduction method
> Abstract
> A method is given for the modular reduction of cryptographic variables,
> a component of many public key cryptosystems. . . .

This was the patent that led to the comical claim that Roger Schlafly had patented two prime numbers. You can't patent a prime number, but you can patent a way of using it.

Another well-known example is the Unisys patent on the Lempel-Ziv-Welch decompression algorithm as used with GIFs. It expires on 19 June 2003. This patent put many software developers at risk of prosecution; a number of websites tell the story. One might reckon that twenty years is almost infinity in software development; a patent holder can make a rich killing very quickly, and then for ten years or more the patent is simply a barrier to progress. So we might hope for a trend to reduce the length of patents. Sadly the trend of recent international agreements in IP law has been mainly to extend the time period.

Most mathematicians never see a patent. Two other forms of intellectual property that play an infinitesimal role in mathematics are trademarks and the law of confidence. I did have a brush with trademark law a year or so ago. On my website I had the slides of a talk I gave to a student society under the title 'The philosophy of mathematics—a bluffer's guide'. On 18 January 2002 I received an e-mail from an agent of the London-based company Oval Projects Ltd, who politely asked me to desist from using the phrase 'bluffer's guide' since it is a registered trademark of Oval Projects Ltd.

Justice MacKinnon once described UK trademark law as being of ‘fuliginous obscurity’. Nevertheless it seems clear that I was under no obligation to do as Oval Projects asked. You can only infringe a trademark by using it yourself in the course of trade, and even then the courts have found that you must be using it *as a trademark* (rather than, say, a description) in order to infringe. My website wasn’t offering to trade anything. Nevertheless a courteous request invites a courteous answer and I removed the item from my website. Had Oval Projects threatened to sue, the position would have been very different; in the UK it’s an offence to make groundless threats to prosecute for trademark infringement.

I know of no cases where mathematicians brushed with the law of confidence. For most mathematicians, having their results kept confidential is the last thing they want. But it could happen. For example editors and referees have a clear duty of confidence towards the material submitted to them. One does see some lax practice. At a talk during an International Congress in the 1970s I heard an interesting new result, and I asked the author about it when I got back to London. He said he had submitted it in a paper but not mentioned it to anybody. So presumably the speaker was a referee breaking confidence.

Secrecy of academic research was one of the issues in the *Pelletier v. Agouron* case ([17] Chapter 5). But here the research was not mathematical; it had medical implications and the defendant was a commercial company.

3 Plagiarism

Copyright, like patents, is a form of intellectual property.

Many people see copyright law as a protection against plagiarism. Some mathematical publishers echo this view. They argue that (i) one needs to be the copyright holder in order to be able to take action against plagiarists in the courts and (ii) publishers are much better able to bear the costs of legal action than authors are. From this they infer that authors should hand over copyright to publishers.

The facts are not so straightforward.

Most mathematicians like to see their results used by other people. They like it even better when they are named as the discoverers of these results. They are angry when they see their best ideas attributed to or claimed by other people. They are irritated when they are given credit for things that seem to them trivial or silly.

The right of an author or artist to have her creations attributed to herself without any devaluing misrepresentation is a central part of what the Berne Copyright Convention calls *moral right.* It plays an important role in French law (*droit moral*) and German law (*Urheberpersönlichkeitsrecht*). The name ‘moral’ means that the right is tied to the author or artist herself, as distinct from economic rights that can be transferred to other people. (One says that moral rights are *inalienable.*)

The vast majority of the world’s trading nations have formally associated themselves with the Berne Convention. But each country decides how to incorporate the Convention into its own legislation, and there is plenty of room for slippages. US copyright law has never explicitly recognised moral rights, except in certain specialised areas such as architecture. (Nevertheless there are other ways of achieving the same effect in US law, for example by claiming defamation.) UK law incorporated moral rights of authors as recently as 1988, and computer programs were excluded. Chinese copyright law in 1990 ascribed moral rights to the 'author', but it redefined the author as the person under whose supervision and responsibility the work was created, which rather misses the point [6].

There are several features of moral rights that make them not terribly helpful for mathematicians under any legal jurisdiction.

(a) *Copyright protects expression but not ideas.* With a few exceptions, the works that carry moral rights are those that carry copyright in general. By the WIPO Copyright Treaty 1996 (an international agreement which extends the Berne Convention and incorporates its main provisions), [26] Article 2:

> Copyright protection extends to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such.

Actually the reference to mathematics is a bit of a red herring here. The point is that legislators have rightly been reluctant to put any longterm restrictions on the spread of useful information, and so (as the slogan goes) copyright applies not to ideas but to the expression of ideas. So your favourite theorems carry no protection at all from copyright law. However, the text of your papers and books counts as expression, and so this text is protected. Also the order in which you put the ideas could count as expression; but somebody who copied your order of exposition could often claim in defence that mathematical arguments impose their own logical arrangement.

(b) *Only the author can take action on moral rights.* Going through the courts is expensive, as publishers remind us. For this very reason you would be a fool to rely on your publisher being prepared to incur the expenditure. But in any case only the holder of the moral rights can take them to court, and we saw that the moral rights can never pass from author to publisher. (There are always some small exceptions: your moral rights might pass to your publisher after your death, depending in part on what you said in your will.)

In fact there are very few recorded instances where accusations of mathematical plagiarism went to court. In 1997 it was reported that Sir Roger Penrose and Pentaplex Ltd (the company which manufactured 'Perplexing Poultry', a jigsaw game based on Penrose's two-tile nonrepeating tilings of the plane [9] p. 531ff) sued the Kimberly Clark Corporation for unauthorised use of Penrose tilings in Kleenex Quilted toilet paper. At that date the Kleenex range had already passed to another company, so presumably the aim was to recover damages. (There is a brief account in [18]. My information on the legal details is secondhand; I mention them only for the principles involved.)

Penrose's tiles are four-sided figures with certain angles at the vertices and certain distances between the vertices. The angles and distances are chosen for the proof of aperiodicity, so presumably they are a consequence of the 'mathematical concept', and this makes them exempt from copyright. But the shape of the line between two adjacent vertices is not determined, and you can use your artistic imagination (as Penrose did) to choose these lines so that the tiles look like chicken. If you do, the result is an artistic creation and hence is copyrightable. I haven't seen any of the legal documents, but from this analysis it does look as if Penrose's case must have been artistic rather than mathematical. He could have argued that Kimberly Clark had violated his moral right not to have his designs besmirched; Pentaplex could have been a party to this complaint if the Pentaplex designers contributed to the shape of the tiles. But it is also possible that the chief complaint was about commercial copying of the design, which (as we shall see) violates copyright though not moral rights in particular.

(c) *Naming and shaming is cheaper and just as effective.* A good example is the article of Allyn Jackson [12] in the American Mathematical Society Notices, which reports the unauthorised and unacknowledged use by John L. Casti of various other people's work in his book 'Mathematical Mountaintops'. In fact the publisher had withdrawn the book and Casti had apologised to the aggrieved authors before Jackson's article appeared. But Jackson set Casti in the stocks before the mathematical community. Plagiarists are put on notice that they should expect this treatment.

Another naming and shaming case was that of a certain C. V. Papadopoulos, who is alleged to have submitted, either under his own name or under a pseudonym, at least thirteen published papers of other authors to various computer science conferences, changing the titles but not the contents. You can read all about him at [7]. This case illustrates how web search engines can pick up plagiarism with devastating efficiency. I've used them regularly on students' essays and picked up quite a few bandits.

(d) *Mathematical common property doesn't have sharp boundaries.* A mathematician who works within an established paradigm is apt to think of the paradigm itself as public property that doesn't need to be attributed. A more senior mathematician who regards himself as the source of the paradigm may take the opposite view. This can happen at all levels. Van Dalen [5] documents one leading researcher's indignation at the way 'his' paradigm is credited. But the same problem arises when not very bright research students think they have added more to their supervisor's suggestions than in fact they have. (Once as a PhD examiner I had to deal with a glaring example. The university in question sensibly recommended that the examiners should treat it as academic incompetence rather than plagiarism.)

The Ethical Guidelines of the American Mathematical Society [2] contain some well-placed comments on moral rights, but they seem to assume that the problem just mentioned never occurs. Thus for example

> The correct attribution of mathematical results is essential.

This is hugely unhelpful. It's of no importance for a modern mathematician to attribute the binomial theorem correctly, or indeed most of the contents of today's textbooks. The problem is to draw a correct line between such cases and the contemporary research results to which the AMS comment certainly does apply. On this the Ethical Guidelines offer no help at all. (Not that one should blame the AMS. It's an inherent property of ethical guidelines drawn up by committees that they fudge all the difficult issues.)

In fact the mathematical community has its own more or less shared notions of what counts as common property and what needs to be credited. Getting a sense of this distinction is one of the things that mathematical researchers have to learn in their training.

Among mathematicians I sense a widespread feeling that individuals who protest too loudly about not being given proper credit are likely to be ungenerous people who themselves give inadequate credit. (In fact I could name one or two examples commonly given.) Perhaps this is a general feeling across the academic world. I don't know whether the facts justify the feeling, but probably it helps keep the peace. A journal or society may be better placed to nail an offender than an individual mathematician is, and also better placed to judge objectively whether there really has been an offence.

4 Restricted Acts

There are certain things that one can do with a literary or artistic work, which are known as *restricted acts* within the context of copyright. (This is British terminology, but it's convenient for general use.) They mostly have to do with copying the work—hence the name *copyright.* Copyright law doesn't directly control these restricted acts; instead it controls who can control them. So copyright law is largely about *powers* in the sense of Hohfeld, i.e. abilities to alter the legal rights and duties of other people. We can separate two questions: What are the restricted acts, and Who can control their performance? This section looks at the first question; we move to the other question in the next section.

One can get a first view of what acts are restricted by looking at the copyright statements in the front matter of books and the covers of journals. Here is an excellent example from the American Mathematical Society, [27] p. iv:

> **Copying and reprinting**. Material in this book may be reproduced by any means for educational and scientific purposes without fee or permission with the exception of reproduction by services that collect fees for delivery of documents and provided that the customary acknowledgment of the source is given. This consent does not extend to other kinds of copying for general distribution, for advertising or promotional purposes, or for resale. Requests for permission for commercial use of material should be addressed to the Acquisitions Department .... Requests can also be made by e-mail to
> `reprint-permission@ams.org`.
>
> Excluded from these provisions is material in articles for which the author holds copyright. In such cases, requests for permission to use or

> reprint should be addressed directly to the author(s). (Copyright ownership is indicated in the notice in the lower right-hand corner of the first page of each article.)

This notice lists the main restricted acts that are relevant to the mathematical material in the book. The copyright owner has the power to control these acts. The AMS is the copyright owner for most of the papers in the volume, and here it uses this notice to state the terms that it imposes. It also notes that where authors keep their own copyright, potential users will need to refer to these authors for the terms that they impose. (In general these powers need not rest with the copyright holder. We come back to this point in the next section.)

Most jurisdictions allow some leeway for people to copy material without having to seek permission. But the principles behind this leeway are different in different countries. In the US the legislators chose to single out some forms of copying as too valuable to be prevented by law. These forms go by the name of *fair use*. Many people have expressed fears that recent international negotiations on the copyright of electronic material are leading to a serious cutback in what is allowed as fair use.

The UK has a more restricted notion of 'fair dealing', unfortunately without a precise legal definition. In Australia likewise you can engage in 'fair dealing' without infringing copyright. UK law suggests that limited copying for 'research or private study' is fair dealing, while Australian law expands this to 'private use'. (This has caused some complexities; for example when is use of the internet private and when is it public? [4] The answer could be important for electronic journals.) There seems to be nothing corresponding to fair use in German or French law; in these jurisdictions, if you don't have an explicit permission to copy then you can't copy.

Not all publishers are as helpful as the AMS. For example take the following [8] p. iv:

> All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

This was published in the US, so US copyright law applies. I don't known whether the publisher has never heard of fair use, or just wishes it didn't exist.

The next example is a paragraph added onto an otherwise excellent statement [13] p. iv:

> This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, re-sold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.

Apparently the publisher wants you to believe, if you bought the book, that these restrictions apply to you. They don't unless they were part of the contract between you and the bookseller. In most cases the publisher was not a party to this contract; perhaps the publisher bound the bookseller to impose such a restriction in the contract of sale, but I bet the bookseller broke this promise. Contrary to what the publisher wants you to believe, if the cover of the book falls off, there is nothing whatever to prevent you making your own cover and then lending the book to your sister!

We saw that (in general terms) copyright restricts the use of expressions rather than ideas. More precisely, copyright applies to the copying and publication of

- text (provided the content could have been said otherwise);
- an arrangement of mathematical facts (if it is not determined by the logic of the mathematical facts themselves);
- a computer program;
- (very recently) the contents of a database;
- a diagram or computer graphic (provided the facts could have been displayed otherwise).

Consider for example the LMS Journal of Computation and Mathematics, an electronic journal. According to its website [16], this journal publishes papers that incorporate:

- large amounts of data (including mathematical tables) [COPYRIGHT!] or experimental results;
- source code for programs [COPYRIGHT!];
- hypertext;
- material inviting comment [COPYRIGHT!] or update (though not unless likely to be of longterm interest).

All of this is electronic matter over and above the copyright-carrying text of the paper itself.

5 Who Holds the Copyright?

Copyright includes a bundle of powers to control various restricted acts that involve the copyrighted work. These powers in turn include second-order powers to transfer the same powers to third parties by contract. For example if I own the copyright on a book, I can make a contract with a publisher that grants the publisher the right to print and market copies of the book, and also allows the publisher the power to negotiate with university libraries the terms on which they will allow readers to borrow or copy the book. Contracting with the publisher in this way, I am said to give the publisher *licences* to do these things. The licence is *exclusive* if it grants the publisher sole right to do whatever the licence allows. For example if I give the publisher an exclusive right to make electronic copies of the book available to the public, then I bind myself not to put a copy on my own website.

There could in theory be a long chain of contracts transferring the same power from one person to another. The author gives the University an exclusive licence, which the university uses to grant a publisher an exclusive licence, and the publisher in turn gives an exclusive licence to one of its subsidiaries, and so on. The effect is much the same as if the exclusive right was a commodity that could be bought and sold in the open market.

At this point there is a great divide between continental European jurisdictions on the one hand, and English-speaking on the other. There are two main differences.

(a) English-speaking law takes seriously the idea that copyright consists of property that can be bought and sold. In fact if you own copyright, you can give it lock stock and barrel (apart from the moral rights) to someone else; this is called *assigning* the copyright. In Germany you can't do this; the nearest you can do is to give someone else comprehensive exclusive rights. German publishers sometimes refer to this as 'assigning copyright', because they know that this is the only way they can communicate with English-speaking authors. (For example Springer-Verlag commonly marks its books

© Springer-Verlag Berlin Heidelberg

even when it indicates that it holds the publication rights by German Copyright Law.) But it's not the same – for example German authors can claim back their rights from the publisher if the publisher ceases to maintain the publication. If you are an American author who would like your publisher to return the copyright to you if at any time he wants to discontinue your book, you will have to negotiate this with the publisher; you could try writing this into the contract that assigns copyright.

(b) In English-speaking law the first owner of the copyright is, roughly speaking, not the author but the owner of the labour that gave rise to the work. If somebody employed you to write a paper, then copyright in the paper belongs in the first instance to the employer and not to you. Many university lecturers are unaware of this, because their universities have an IP policy that makes a gift of the copyright to the lecturers as an explicit or implied consequence of their contract of employment.

This recently became a hot topic at the University of Cambridge. The IP policy of the University had been that intellectual property generated by externally funded research would be owned by the University, but the University would not claim ownership of copyright in 'normal academic forms of publication such as books, articles, lectures, or other similar works generated in the course of externally funded research unless those works have been specifically commissioned by a sponsor' [24].

But in July 2002 the Council and General Board of the University recommended that as from 1 January 2003, the University should assert ownership of all intellectual property generated by its employees in the normal course of their duties. The case for this change seems to rest chiefly on technology transfer through patented inventions. There has been a lot of opposition from academics in Cambridge University who believe that the new policy will in fact be harmful to technology transfer.

It might seem that the University's mathematicians are threatened with loss of copyright for reasons that have precious little to do with copyright. Opponents of the new policy allege that there is an unpublished reason why the University wants to retain copyright, namely to stop lecturers from using University resources to create fancy teaching materials which they then sell on for a profit [3]. So the University's thinking on copyright may be the result of the huge growth of multimedia technology, and perhaps also the expected growth of distance learning which will rely on this technology. These factors affect all universities, and it will be interesting to see how many of them move in the same direction as Cambridge.

6 Questions of Ownership

The naive view is that everything on earth either belongs to some person or is common property. (The 'person' could be an organisation.) Up to a point the law supports this view in the case of intellectual property. The owner is the copyright holder in English-speaking jurisdictions and the auteur/Urheber on the European continent.

The reservations come fast and thick. First, the legal situation of joint research is so complicated that I avoided mentioning it before, and I say no more now. Many legal questions in this area are unresolved.

Next, intellectual property has the curious feature that it suddenly passes into public ownership after a certain length of time. For new work, copyright expires seventy years after the death of the author; for old work the situation is again complicated and varies from one jurisdiction to another.

Next, we saw that the mathematical community has its own notion of what has become common property – so that anyone can use it without having to get permission or cite the author. A priori there is no reason why the mathematicians' notion should have any connection at all with the legal limit of 70 years from the author's death. Courts are very unpredictable in the extent to which they recognise the assumptions and common practice of particular communities.

Next, as we have seen, one should beware of assuming that ownership implies control. If you own X, then in general you have the power to hand over the control of X to someone else; within copyright this happens all the time.

In short, the law of copyright gives doubtful help to any mathematician who wants to campaign under the slogan 'It's my work; why should the publisher/the university steal it from me?'

Nevertheless it is true that there are some concrete legal decisions facing a mathematician who proposes to publish. Here I assume you are an author and you hold the copyright in your work. If you are submitting the paper to a journal, you probably have little choice but to give the publisher the rights that he asks for.

If you want to retain the right to put your material on your website, or to republish yourself if the publisher loses interest in your book, there is no harm in asking. Anecdotal evidence suggests that nowadays most journal publishers allow you to put the paper on your website if you ask. For example Allyn Jackson [11] recently reported that Elsevier always says yes. If enough people ask, it might soon become a standard feature of journal contracts.

You have more freedom if your work is a book. The CEIC checklist [10] lists a number of things that you might bear in mind if you want to negotiate the terms of the contract.

If you license your publisher to authorise reprinting of your work by other publishers, then people who want to include your chapter in their published collection will need to find your publisher rather than you. Your publisher will probably be easier to find than you. Of course if you assign copyright then you give your publisher this power, but you can equally well transfer it by licence. (You might use the opportunity to ask the publisher to promise his best efforts to let you know when he authorises republication of your work. At a conference bookstall I once opened a volume and found inside it a paper of mine which I didn't know was being reprinted.)

Librarians who have set about digitising their back collections have run into the problem that some of the copyright holders or licensees on old material are now hard to trace. Fifty years ago nobody dreamed of digitisation. Even within the lifetime of your copyrights, there could be further technological changes that send the librarians out chasing permissions for mass publication in some new form.

In terms of what they allow, there is very little difference between the following two situations:

(i) You give your publisher copyright but require your publisher to give some things in return (e.g. to license you to publish electronically, or to give you back the copyright if the work goes out of print).
(ii) You give your publisher a licence which is comprehensive except for some rights that you want to reserve to yourself.

In Europe only the second is possible. In the English-speaking world both are possible.

The chief difference between (i) and (ii) shows up when circumstances arise that the licensee hadn't foreseen. If nothing was said in the licences, the default assumption is that discretion lies with the copyright holder. Let me give two examples. The first is non-mathematical and the publisher (or more strictly the broadcaster) is the vulnerable licensee. In the second example the licensee is the mathematical author.

The first example is a dispute between the British comedy team Monty Python and the US broadcasting company ABC Inc. ABC had contracted with Monty Python to broadcast some of Monty Python's sketches. When ABC came to look closely at the sketches, they found some wording that could have offended American audiences. (For example 'intercourse' and 'pert thighs'. This was 1976.) So they cut sections out; but they had not beeen licensed to do this. Monty Python took them to court and won on appeal. The appeal court determined [1] p. 1095:

> 'Since the scriptwriters' agreement explicitly retains for the [authors] all rights not granted by the contract, omission of any terms concerning alterations ... must be read as reserving to [the authors] exclusive authority for such revisions.

This ruling leaves in the balance what would have happened if the agreement had not been explicit about retaining all unstated rights for the author. But other things being equal, the copyright holder will be in the stronger position.

We turn to the mathematical example. Starting in 1995, Eric Weisstein ran a very successful website 'Eric's Treasure Trove of Mathematics', later called 'MathWorld'. In 1997 he signed a contract with CRC Press LLC for them to publish the contents of the site as a book. In 1999 he joined Wolfram Research, and with their support he continued to develop the website. He moved the site to Wolfram Research's website Mathematica.com, and declared that he held joint copyright with Wolfram Research. In 2000 CRC took action under copyright against Wolfram Research, Stephen Wolfram and Weisstein himself, demanding that the website be closed down and that the defendants should pay damages and CRC's legal costs. The court granted CRC a preliminary injunction. After this a settlement was reached out of court; Weisstein continues to run the website, but an annual payment is made to CRC, who retain some rights over the material in the website.

Weisstein's case is the most substantial mathematical case to have gone through the courts in recent years. His own account of the matter is at [25]. It seems that he never foresaw how weak his legal position would be when he had assigned the copyright to CRC; in particular he never took care to see that the contract spelt out what CRC was licensing him to do with the website. But there is an important extra twist: Weisstein apparently didn't realise the importance of making clear that the copyright that he was assigning to CRC was for the contents of the website at a certain date, rather than the website itself. The contract simply says 'the Work', and the court had to spend some time determining that 'the Work' did include the website [22].

There is a further twist even on this. When CRC had first made demands about how Weisstein should run the website, Weisstein responded as many mathematicians would: he grumbled but did as he was asked. The court used this as evidence that he was controlling the website 'at the direction of CRC', and hence that his intention had been to give CRC rights over the website. It doesn't always pay to be obliging.

The case also illustrates two points that we have already noted. First, mathematicians make themselves more vulnerable when they associate themselves with commercial companies who can afford to pay damages. And second, in US law there is no built-in protection against a publisher who (like CRC according to Weisstein) takes the copyright on your work and then fails to market it effectively. A third point worth making is that this case arose only because of advances in electronic publishing. It's hard to see how a similar situation could have arisen before the web.

So within English-speaking jurisdictions both publishers and authors do have some incentive to keep copyright. Probably the publishers will always win this contest, because they understand publishing and mathematicians don't. Rather than do battle on the details of copyright, mathematicians who feel dissatisfied with the present situation are more likely to put their energy into other forms of publishing, for example open access archives.

But in any case both mathematicians and their publishers have a common interest in making explicit in their contracts what each of them wants to be able to do and what each wants the other to do. In a recent case about rights to electronic publication, the US Supreme Court said wisely [20]:

> [Warnings of "devastating consequences" are unavailing. . . .]
> The Authors and Publishers may enter into an agreement allowing continued electronic reproduction of the Authors' works; they, and if necessary the courts and Congress, may draw on numerous models for distributing copyrighted works and remunerating authors for their distribution.

References

1. Abbott, F., Cottier, T. and Gurry, F.: The International Intellectual Property System: Commentary and Materials. Kluwer Law International, The Hague 1999
2. American Mathematical Society Ethical Guidelines, `www.ams.org/secretary/ethics.html`, 7 January 2003
3. Anderson, R.: Campaign for Cambridge Freedoms: Analysis of the Vice-Chancellor's Proposal. `www.cl.cam.ac.uk/~rja14/expropriation.html`, 9 January 2003
4. Brudenall, P.: The future of fair dealing in Australian copyright law. Journal of Information Law and Technology, `elj.warwick.ac.uk/jilt/copright/97_1brud/` 1997.
5. Van Dalen, D.: Brouwer and Fraenkel on Intuitionism. Bulletin of Symbolic Logic **6** (2000) 284–310
6. Dietz, A.: The new copyright law of the People's Republic of China—an introduction. International Review of Industrial Property and Copyright Law **22** (1991) 441ff
7. Euro-Par '95 website. `www.sics.se/europar95/plagiarism.html`, 8 January 2003
8. Goossens, M., Rahtz, S. and Mittelbach, F.: The LaTeX Graphics Companion. Addison-Wesley, Reading Mass. 1997
9. Grünbaum, G. and Shephard, G.: Tilings and Patterns. W. H. Freeman and Co., New York 1987
10. Hodges, W.: What do you want from your publisher? Mathematische Nachrichten **188** (2001) 21–30 and Notices of American Mathematical Society, November 2001, pp. 1176–1182 and CEIC website `www.ceic.math.ca/`
11. Jackson, A.: From preprints to e-prints. Notices of American Mathematical Society, January 2002, pp. 23–31
12. Jackson, A.: Theft of words, theft of ideas. Notices of American Mathematical Society, June/July 2002, p. 645

13. Kaye, R. and Macpherson, D. eds.: Automorphisms of First-Order Structures. Clarendon Press, Oxford 1994
14. Körner, T.: Untitled address to author-members of the Poldovian Academy of Literature. `www.cl.cam.ac.uk/∼rja14/ccf/korner.html`, 9 January 2003
15. Library of Congress website.
`www.loc.gov/loc/lcib/0101/copyright_symposium.html`, 7 January 2003
16. London Mathematical Society Journal of Computation and Mathematics.
`www.lms.ac.uk/jcm/editorial.html`, 12 January 2003
17. McSherry, C.: Who Owns Academic Work? Battling for Control of Intellectual Property. Harvard University Press, Cambridge Mass. 2001
18. Mirsky, S.: The Emperor's new toilet paper. Scientific American 277 (July 1997) i 24.
19. Samuelson, P.: Towards more sensible anti-circumvention regulations. In: Financial Cryptography, 4th International Conference, FC 2000 Anguilla, ed. Yair Frankel, Lecture Notes in Computer Science **1962**, Springer-Verlag, Berlin 2001, 33–41
20. Supreme Court of the United States. New York Times Co., Inc., et al. v. Tasini et al., June 2001
21. UK government IP website
`www.intellectual-property.gov.uk/ipac/std/members.htm`, 7 January 2003
22. United States District Court, Central District of Illinois, Danvill/Urbana Division, CRC Press, LLC v. Wolfram Research, Inc., Stephen Wolfram, and Eric Weisstein, 00-CV-2262, available at `mathworld.wolfram.com/docs/InjunctionRuling.pdf`
23. United States Patent and Trademark Office. `www.uspto.gov/patft/`, 14 January 2003
24. University of Cambridge Technology Transfer Office, Frequently Asked Questions regarding IPR Policy.
`www.admin.cam.ac.uk/offices/tto/faq/ip.html`, 9 January 2003
25. Weisstein, E.: website of MathWorld,
`mathworld.wolfram.com/erics_commentary.html`, 9 January 2003
26. World Intellectual Property Organization (WIPO). Copyright Treaty 1996
27. Zhang, Yi. ed.: Logic and Algebra. Contemporary Mathematics **302**, American Mathematical Society, Providence RI 2002

EMIS 2001 – A World-Wide Cooperation for Communicating Mathematics Online

Michael Jost[1] and Bernd Wegner[2]

[1] Fachinformationszentrum Karlsruhe – Zentralblatt MATH
Franklinstr. 11, D-10587 Berlin
jo@zblmath.fiz-karlsruhe.de
[2] Mathematisches Institut, Technische Universität Berlin
Strasse des 17. Juni 135, D-10623 Berlin, Germany
wegner@math.tu-berlin.de

Abstract. The European Mathematical Information Service (EMIS) is the information server network of the European Mathematical Society. It is based on the voluntary support of several partners from all over the world. This article gives an overview of concepts behind the service, and its main components: The Electronic Library, the collection of databases, and the projects.

Introduction

Several years ago, the increasing development of electronic devices for the publication of papers and books in mathematics led to a drastic change in the communication process between authors and editors, to new ways of distributing mathematical publications to research mathematicians – like electronic journals, and to an extension of the offers of information on mathematical research to the mathematical community. Today, electronic publishing generally is considered as a must, and the linking facilities offered by electronic versions of mathematical publications in combination with other mathematics offers in the web enable researchers and professionals dealing with mathematics to get the mathematical information and tools they are interested in successfully by browsing and searching. In particular, links going to and provided by literature databases and other qualified indexes help them to find their way through the tremendous bulk of current and previous mathematical research papers and a lot of additional items of interest, like pre-prints, educational material in mathematics, software, graphical material and others.

The aim of this article is to report on those aspects related to methods of electronic publishing and electronic communication by exhibiting offers provided by EMIS (European Mathematical Information Service), calling this accumulation of services and projects a "portal". Clearly, not all features of a comprehensive portal site are offered by EMIS, but the collections provided by the sections dealing with the electronic library, the databases and the projects bundle unique services of high interest for mathematics. Though information on conferences,

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 87–94, 2003.

jobs, society matters etc. can also be obtained from EMIS, the three sections above are the highlights of this service and the subsequent exposition will concentrate on them.

1 The General Concept of EMIS

The idea to develop the European Mathematical Information Service EMIS was born at the meeting of the executive committee of the EMS (European Mathematical Society) in Cortona/Italy, October 1994. The installation of the central server for EMIS began in March 1995 in co-operation with FIZ Karlsruhe at the editorial office of Zentralblatt MATH in Berlin. In June 1995 EMIS went online at the URL http://www.emis.de/. This initial installation was extended very soon to the current version of a central server collecting mathematical information and distributing this through a world-wide system of mirror servers.

The World Wide Web access to the contents of EMIS is free for all users, except for the full usage of some databases. In these restricted cases a link leads directly to the corresponding system of database gateways, and the user is subject to the conditions valid for accessing the databases. In any case, users will be able to do searches. But in the case where his institution does not subscribe to the service, only some restricted information will be available from the hit list.

One of the basic ideas for EMIS is distribution through a world-wide system of mirrors where the full content of the service is available at all sites, and updated periodically. This improves the accessibility of EMIS, and it simultaneously is important for the safety of the data and their archiving: if one of the system components fails, it can be regenerated easily from the other components. In principle, every European country has installed one mirror at least, other mirrors have been installed on all continents (except Antarctica).

2 The Electronic Library

The Electronic Library of EMIS (ELibM) aims to present a collection of freely accessible electronic publications which should be as comprehensive as possible. There are four sections: journals, proceedings volumes, monographs, and collected works. In order to guarantee that the electronic publications stored in the Electronic Library meet the quality standards required for articles in traditional print journals, the decision on the inclusion of journals, proceedings or monographs is taken in accordance with the Electronic Publishing Committee of the EMS. Hence, no items will enter the library which have not been evaluated and recommended by a referee within the editorial procedures of the corresponding journal or series. This is in particular important in order to rule out the reservations of many mathematicians who have the opinion that electronic publishing will damage the quality of mathematical publications.

Most of the journals in the Electronic Journals section are completely produced elsewhere, and EMIS only serves as an additional distributor. In some cases, however, the e-journal is produced by EMIS from the original source files provided by the editors. We prefer that the offer of the electronic version is installed at the site of the editors, such that a mirror of the journal can be taken over by EMIS. The organizers of EMIS provide support for this purpose.

The e-journals section contains purely electronic journals as well as electronic versions of print journals (dual journals). Most of the dual journals are published at a low-budget level, and hence the risk of loosing subscribers to the print version due to the free electronic offer currently is considered as low by them. Some of them give the electronic offer with a certain delay to EMIS such that the earlier availability will be considered as an advantage of the print version.

Acknowledging that the electronic versions are becoming increasingly important for the users, this delay period will be reorganized. During the period which is considered by dual journals as the most important one to keep libraries subscribing to them, the access to the electronic version may be provided only to subscribers. To enable the access control for this purpose, the journal will be stored on separate servers, though the metadata should be made freely accessible in ELibM. ELibM will offer links to the complete articles. After a period to be decided by the journals themselves, the full content may be transferred to the system of mirrors of EMIS where it can be read without having a subscription. Such a structure also will support ideas like posting articles "online first", which speeds up the publication procedure considerably.

To get some idea about the journals distributed by ELibM some samples are mentioned below:

- Annales Academiae Scientiarum Fennicae Series A. Mathematica
- Annals of Mathematics
- Archivum Mathematicum (Brno)
- Beiträge zur Algebra und Geometrie
- Commentationes Mathematicae Universitatis Carolinae (Prague)
- DOCUMENTA MATHEMATICA
- The Electronic Journal of Combinatorics
- The Electronic Journal of Differential Equations
- Electronic Research Announcements of the AMS
- Electronic Transactions on Numerical Analysis
- Geometry and Topology
- Journal de Theorie des Nombres de Bordeaux
- Living Reviews in Relativity
- Matematicki Vesnik (Belgrade)
- Mathematical Physics Electronic Journal
- El. J. of the Argentine Society for Informatics and Operations Research
- Seminaire Lotharingien de Combinatoire

More or less all freely available electronic journals in mathematics are mirrored in ELibM. The total number of journals in ELibM is about 60 at present, including three discontinued (older volumes still available), and five to be announced shortly. The journals section is complemented by a collection of about 20 proceedings volumes and collections of articles. The total number of full text articles in these two sections is about 11.000 at present. Additionally, ten electronic monographs are available in ELibM, and two collected works (Riemann and Hamilton).

The access to the journals in EMIS is organized quite conventionally by clicking through web pages and lists of contents. On the home page of EMIS a list of mirrors is provided where the site with the (probably) best access can be clicked. Then a choice can be made, to enter the Electronic library through the short list of journals without graphics or to use the full display of these items. The first one is preferable, if the choice of journal is clear already and if one wants to avoid the lengthy transfer of the graphical data associated with this journal. The full display contains also background information on the editorial policy of the corresponding journal and instructions how to submit an article. In some cases style files for such a submission can be found on this level.

For all articles DVI- and Postscript-files are available, sometimes also TEX-source codes can be found in addition to that. PDF offers are coming up rapidly and will be obligatory in the near future. By clicking one of these files, the content is transferred to the computer of the user and can be viewed there. Also, printing or storage of these files is possible at the site of the user, but he is requested to respect the copyright policy according to the rules of the corresponding journal. Access to the section of Proceedings Volumes is organized in a similar way.

Admittedly, after seven years of EMIS some of these features will have to be modified and modernized. As already mentioned, the request for PDF-files will become a must, because the pre-installed readers at electronic access facilities in libraries and desktop-computers of research mathematicians already point into that direction. For other offers special measures have to be taken to make the article readable or printable. Decisions about affordable systems of digital object identifiers (beyond DOI) have to be taken to enable a richer linking system between electronic articles. The Zentralblatt MATH accession numbers are one choice for that. These, together with a more standardized set of metadata, will enable the installation of a more professional access structure going beyond just clicking on pages of contents. Such a facility is highly desirable, after having stored such a lot of articles in ELibM, and a first prototype will be available soon.

3 The Databases Section

This section contains four items: MATH – the online version of Zentralblatt MATH, MATHDI – the online version of a similar service for education in mathematics, MPRESS – a global pre-print index, and a database on geometric objects.

To connect EMIS with the databases of **Zentralblatt MATH** and **MATHDI** is one part of the increasing involvement of the European Mathematical Society in the edition of these reviewing services. In contrast to the variety of “databases” offered by commercial publishers now, the word “reviewing” is taken quite seriously and it is not “abstracting” only. A more detailed description of Zentralblatt MATH and MATHDI is given in the article by Olaf Ninnemann in this volume.

In contrast to these two for-pay databases, **MPRESS** is provided as a freely accessible service. It stores combined information on mathematics pre-prints available on the web. The gathering of information is done by robots, which are run by national brokers for harvesting of metadata. This procedure leads to a data structure which only allows for simple search facilities. MPRESS has no ambition to offer a pre-print server itself, only links to full texts are provided. The service is supervised by EMS among others. Countries which support the harvesting are Germany, France, Austria, and Italy. In addition to this some special servers are harvested by MPRESS. Among them are the Topology Atlas and the arXiv.

As a new item, a link to a free offer of high-quality **geometric models** and animations has been arranged. This is a preliminary version, and it has to be investigated how the data of these models could be stored in a convenient way, to make them accessible within the same menu as is provided for searching mathematical articles. But as a first solution the different entries will be reviewed in Zentralblatt MATH, because they consist of fully peer-reviewed articles on their own, though in contrast to conventional mathematical publications they are providing a lot of geometric enhancements.

4 The Projects

EMS is involved in four projects, where three of them are funded by the European Union and one is funded by Deutsche Forschungsgemeinschaft. The first three are LIMES, EULER, and the Reference Levels project, while the other one is the Jahrbuch-Project. The one closely related to Zentralblatt MATH is LIMES (Large Infrastructures in Mathematics - Enhanced Services).

The objective of the **LIMES** project is to upgrade the database Zentralblatt MATH into a European-based world class database for mathematics and its applications by a process of technical improvement and wide Europeanization. Upgrading the existing database, improving the present system and developing a new, distributed system both for the input and output of the data are necessary to allow Zentralblatt MATH to use the latest developments and to anticipate future developments of electronic technologies. Again, a detailed description of the LIMES project is given in the article by Olaf Ninnemann in this volume.

From April 1998 to September 2000 the European Commission has been funding the **EULER** project in the framework of the ‘Telematics for Libraries’ sector from the Telematics Applications programme, and after that in 2002 the EULER-TAKEUP project in the framework of the ‘Information Society Technologies’ programme. The main goal of these projects was to integrate different, electronically available information resources in the field of mathematics.

Today, EULER is a European based world class real virtual library for mathematics with up-to-date technological solutions, a sound sustainable business model, well accepted by users.

In particular, EULER provides a world reference and delivery service, transparent to the end user and offering full coverage of the mathematics literature world-wide, including bibliographic data, peer reviews and/or abstracts, indexing, classification and search, transparent access to library services, co-operating with commercial information providers (publishers, bookstores).

The EULER services provide a gateway to the electronic catalogues and repositories of participating institutions, while the latter retain complete responsibility and control over the creation and maintenance of their data collections as well as the access provisions pertaining to their offerings.

The EULER Consortium is now a registered incorporated society (association), according to German law. It consists of full members and associate members. Only not-for-profit institutions can become full or associate members. In addition the Consortium recognizes another category: commercial parties may join as sponsoring members.

Full members consist of the current EULER participants and new members who are prepared to contribute actively to the Consortium, while associate members participate as information providers only. They provide metadata and regular updates of their data according to Dublin Core based EULER specifications. Commercial partners (sponsoring members) are participants who contribute their metadata according to Dublin Core based EULER specifications and for which a higher financial contribution is required – they cannot be full members.

The Consortium acts through its members to ensure the continuation of the EULER services. Members are required to provide metadata and regular (at least monthly) updates of their databases according to Dublin Core based EULER specifications. The necessary scripts for the conversion are the responsibility of the member. The Consortium can provide initial and basic technical advice. Co-ordinating tasks as well as technical and administrative tasks will be carried out by the Executive Committee. As necessary such tasks can be distributed among the full members. A Scientific Supervisory Board consisting of the European Mathematical Society (EMS), their appointed representatives, and other suitably qualified and recognized members is responsible to ensure the scientific quality of the EULER service, and advises the Executive Committee on scientific matters. The European Mathematical Society (EMS) represents the interests of the mathematical community for which the EULER service is made. The EULER services will remain under control of organizations representing the public interest.

Several initiatives are currently globally underway to establish a comprehensive Digital Mathematics Library, consisting of scanned images of the whole corpus of historical works, and genuine electronic publications in mathematics. The European Mathematical Society has expressed its interest in contributing to such a development, co-ordinate European activities, and liaise with other global partners. Several of the original partners of the EULER initiative are involved in these and similar activities. Concrete examples of initiatives are

a) the ERAM project which builds a collection of scanned material based on the "Jahrbuch über die Fortschritte der Mathematik" (1868-1942, see also below), and
b) the EMANI initiative which includes key players from Europe, the US (Cornell), and China (Tsinghua University).

Results so far are promising, e.g. the ERAM database covers now about 150.000 publications, with a big portion linked online to the archive of scanned works (fulltext archive) at SUB Göttingen. It is already now available in EULER.

Research funding agencies world-wide seem to be interested in these developments, which makes it probable that the idea of a global Digital Mathematics Library will eventually be implemented. (See also the article of the executive director of the American Mathematical Society: John Ewing: Twenty Centuries of Mathematics: Digitizing and Disseminating the Past Mathematical Literature, Notices AMS, Aug 2002, 49(7), 771-777). In such an environment, where several partners would on a global scale work on such a distributed scanning and preservation project, a powerful end-user discovery tool will be needed that works independently from local specialities and formats, is capable of integrated homogeneous retrieval of heterogeneous distributed sources, and is scalable to cope with the amounts of data that are to be expected. EULER has proven that its model is an optimal choice for such a discovery system. Only few adaptations seem to be necessary.

The aim of the **Jahrbuch**-Project, which officially is called ERAM (Electronic Research Archive in Mathematics), is to capture the "Jahrbuch ueber die Fortschritte der Mathematik" as a classical bibliographic service in mathematics in a database and to use this activity to select important publications from the Jahrbuch period (1868-1943) for digitization and storage in a digital archive. The database will not be just a copy of the printed bibliography. It will contain a lot of enhancements like modern subject classifications as far as possible, keywords giving ideas about the contents in modern terms, and comments relating classical results to modern mathematical research areas. These features will remain open for additions within a living project.

The digital archive (built up in connection with the database) covers selected publications as well as whole series going beyond the Jahrbuch period. It will be linked to Zentralblatt MATH and the Jahrbuch database. In the final version all facilities associated with current retrospective digitization projects will be provided. But for the initial period the offer consists of scanned images and metadata for access only. Eventually, the content may be distributed to mirrors and combined with similar archiving activities in mathematics.

Finally, the study "**Reference levels** in School Mathematics Education in Europe at the age of 16", as it was suggested to the European Commission, identifies "Reference Levels" concerning knowledge and competencies in the domain of mathematics that can become common to all countries in the European Union, and perhaps in other countries.

5 Conclusion

The offers in EMIS mentioned above and their distribution through a system of mirrors provide a unique facility for quick and easy access to qualified mathematical publications. These tools can be used by the mathematical community at suitable sites simultaneously for free or at modest rates. This is an advanced service to provide an alternative to offers of commercial publishers which cannot be afforded by the majority of potential users anymore. Mathematicians have to look for their own systems to maintain a reasonable infrastructure for communicating their research achievements. EMIS is one part of this enterprise.

It has to be pointed out that EMIS cannot survive on the current level without the big group of its supporters, who serve as volunteers for maintaining and installing electronic journals, caring about submissions and transfer of content and keeping the mirrors running. Without these activities EMIS would not have been possible. This shows that a viable service can be maintained with the collaboration of several volunteers, in contrast to those who argue that all these activities can only be pursued seriously on a commercial level. At least in mathematics, the publication cycle relies on voluntary services from the mathematical community: publishers do not pay for receiving the articles from the authors, editors of journals and proceedings volumes provide their service for free, referees do not even get the mailing expenses for their contribution to evaluate the papers, and finally readers have to look for public funding to pay for the access to the publications. Hence the question is only where to shift the voluntary support and where to spend the funding.

Math-Net[*] Means Not Just a Page!

Michael Kaplan

Zentrum Mathematik[**] der TU-München[***], D-80290 München
kaplan@ma.tum.de

Abstract. In April 2002 the Executive Committee of the International Mathematical Union (IMU) endorsed a recommendation[1] to install Math-Net Pages. The Math-Net Page[2] for departments or research institutes is a web portal for mathematics departments and institutes that presents information in a standardized, well-structured, and easy-to-use format. The Math-Net Page is only the first step towards building a distributed information system for the international mathematics community. Other Math-Net services like the Navigator[3], Persona Mathematica[4] or math-net.preprints (MPRESS)[5] can use the Math-Net Page as a standardized entry point. All these services together form the nucleus of a global electronic information and communication system for mathematics (which could be used in a very similar way for other sciences too). Since the Math-Net Page for departments and MPRESS are very well accepted in the community the next step could be a world directory of mathematicians driven by the Math-Net community.

1 Introduction

One important step towards a comprehensive information and communication system is the knowledge about colleagues and their professional activities like publications, projects and teaching. As usual in the world wide web there exists an awful lot of unstructured information in individual pages. Besides this individual information some professional societies offer membership lists in different formats.

Some of the homepages are hard to find and some of the membership lists offer only very basic information. Inspired by the success of the mathnet.preprints (MPRESS) preprint search engine the Math-Net initiative in Germany began to build a distributed information system for personal information called Persona Mathematica. The basic idea is the same like for mathnet.preprints:

[*] http://www.mathnet.org/

[**] http://www.ma.tum.de/

[***] http://www.tum.de/

[1] http://www.math-net.org/Math-Net-Recommendation.html

[2] http://www.ma.tum.de/Math-Net/

[3] http://www.math-net.org/navigator/

[4] http://www.mi.uni-koeln.de/Math-Net/persona_mathematica/

[5] http://mathnet.preprints.org/

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 95–99, 2003.

- An easy to use tool produces standardized homepages that contain standardized XML/RDF metadata (nearly the same tool may be used to produce homepages out of existing databases).
- These homepages are gathered and put in a central database.
- A nice search and browse interface gives easy access to the collected data.

This layout has many important advantages:

- If many standardized homepages exist in the community, then every researcher will know their structure soon and therefore find the desired information very fast.
- Since the homepages are produced by the person itself, the information contained is
  - correct
  - up to date
  - and legal!
- XML/RDF metadata are a mighty but quite complicated tool. On one hand they are indispensable for a good search engine, on the other hand they are unreasonable to use for the normal user, who is not specialized in this topic.

A proposal for the layout and content of professional homepages and the advantages of this approach will be discussed more in detail in the following sections.

2 Proposal of a Professional Homepage

2.1 Layout

The main structure of the proposed page consists of two blocks

Basic information (name, affiliation, phone, email, address, ...)
Further information (Research, Collaborations, Teaching, Professional Societies, ...)

The first block provides the basic information that everybody should have on a professional homepage. This is the nucleus of the page, which has to be there in any case. Often this will be the only information one gets.

The second block is mostly optional and provides links to more information in the world wide web. It is similar to the Math-Net Page for departments or research institutes and divided in 6 main information blocks:

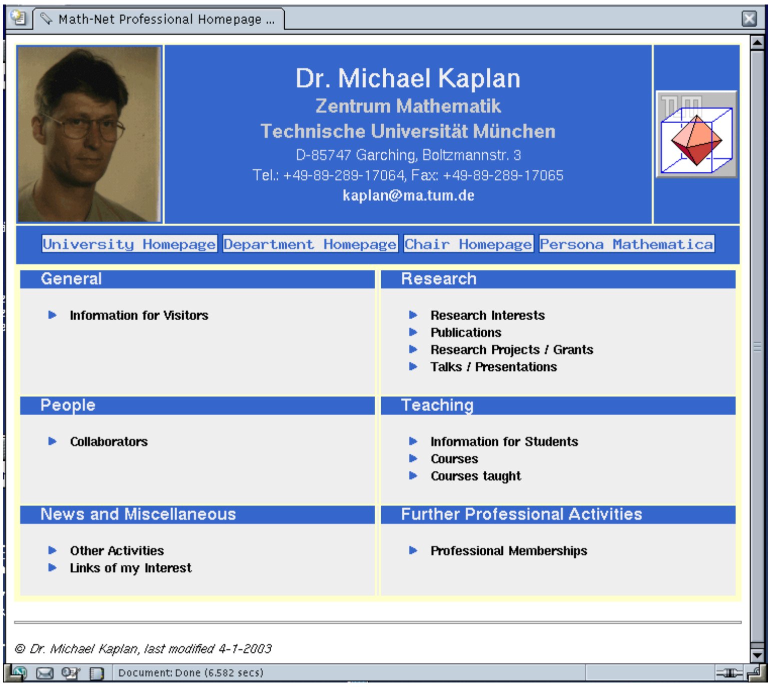
General	Research
Collaborations and Cooperations	Teaching
News and Miscellaneous	Professional Societies and Activities

Finally these main sections include some standardized headings where one can provide links to more material in the world wide web, for example

- Information for Students,
- Courses,
- Courses taught,
- Teaching and Course Materials

in the section 'Teaching'. These headings are still under discussion. They are all optional, but have to be selected from a given list to guarantee a controlled vocabulary.

All together my professional homepage would look like



Math-Net Professional Homepage ...

Dr. Michael Kaplan
Zentrum Mathematik
Technische Universität München
D-85747 Garching, Boltzmannstr. 3
Tel.: +49-89-289-17064, Fax: +49-89-289-17065
kaplan@ma.tum.de

University Homepage | Department Homepage | Chair Homepage | Persona Mathematica

General
- Information for Visitors

Research
- Research Interests
- Publications
- Research Projects / Grants
- Talks / Presentations

People
- Collaborators

Teaching
- Information for Students
- Courses
- Courses taught

News and Miscellaneous
- Other Activities
- Links of my Interest

Further Professional Activities
- Professional Memberships

© Dr. Michael Kaplan, last modified 4-1-2003

Document: Done (6.582 secs)

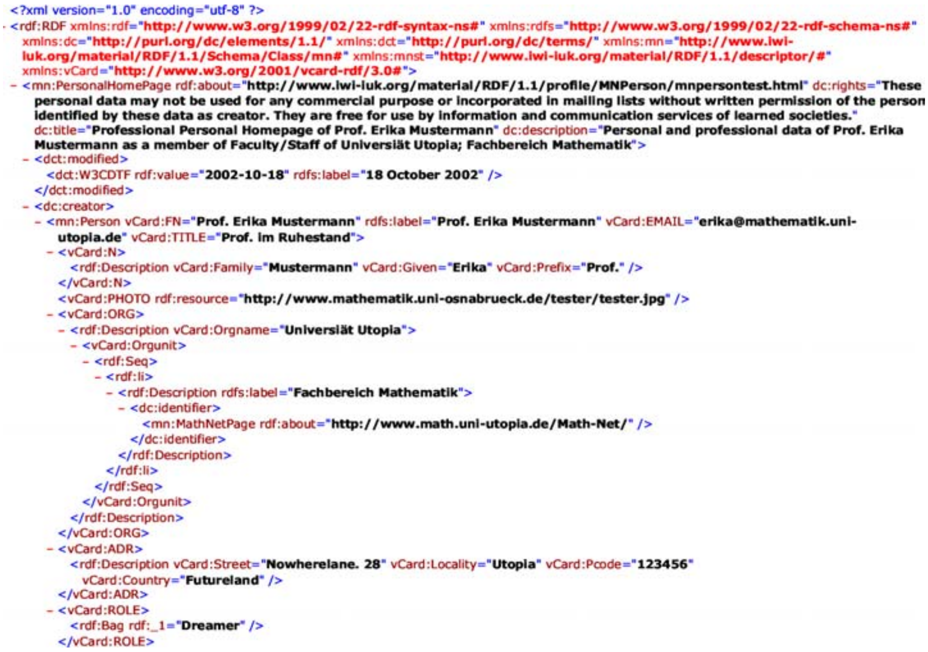
Example layout of a professional homepage

2.2 Internal Value

While the layout of the page is made for human beings and meant to help them finding information on the corresponding person, the internals of the page are especially important for search engines. The page does not only contain the usual XHTML[6]-Markup (HyperText Markup Language) and CSS[7] (Cascading Style Sheets), but also DC[8] (Dublin Core) metadata encoded in RDF[9] (Resource

[6] http://www.w3.org/MarkUp/
[7] http://www.w3.org/Style/CSS/
[8] http://dublincore.org/
[9] http://www.w3.org/RDF/

Description Framework) using XML[10] (Extensible Markup Language) as an interchange syntax. The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. The other mentioned standards were released by the W3C[11] (World Wide Web Consortium) which develops interoperable technologies for the web. Further used standards are vCard[12] by the Internet Mail Consortium and Unicode[13]. By using these standards a very high degree of interoperability is secured.
The TAB[14] (Math-Net Technical Advisory Board) published a so called application profile[15] for professional homepages. This reference description fixes and explains metadata for professional homepages. These metadata provide the basis for a powerful retrieval.



```
 <?xml version="1.0" encoding="utf-8" ?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/1999/02/22-rdf-schema-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dct="http://purl.org/dc/terms/" xmlns:mn="http://www.iwi-
   iuk.org/material/RDF/1.1/Schema/Class/mn#" xmlns:mnst="http://www.iwi-iuk.org/material/RDF/1.1/descriptor/#"
   xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#">
 - <mn:PersonalHomePage rdf:about="http://www.iwi-iuk.org/material/RDF/1.1/profile/MNPerson/mnpersontest.html" dc:rights="These
     personal data may not be used for any commercial purpose or incorporated in mailing lists without written permission of the person
     identified by these data as creator. They are free for use by information and communication services of learned societies."
     dc:title="Professional Personal Homepage of Prof. Erika Mustermann" dc:description="Personal and professional data of Prof. Erika
     Mustermann as a member of Faculty/Staff of Universiät Utopia; Fachbereich Mathematik">
   - <dct:modified>
       <dct:W3CDTF rdf:value="2002-10-18" rdfs:label="18 October 2002" />
     </dct:modified>
   - <dc:creator>
     - <mn:Person vCard:FN="Prof. Erika Mustermann" rdfs:label="Prof. Erika Mustermann" vCard:EMAIL="erika@mathematik.uni-
         utopia.de" vCard:TITLE="Prof. im Ruhestand">
       - <vCard:N>
           <rdf:Description vCard:Family="Mustermann" vCard:Given="Erika" vCard:Prefix="Prof." />
         </vCard:N>
         <vCard:PHOTO rdf:resource="http://www.mathematik.uni-osnabrueck.de/tester/tester.jpg" />
       - <vCard:ORG>
         - <rdf:Description vCard:Orgname="Universiät Utopia">
           - <vCard:Orgunit>
             - <rdf:Seq>
               - <rdf:li>
                 - <rdf:Description rdfs:label="Fachbereich Mathematik">
                   - <dc:identifier>
                       <mn:MathNetPage rdf:about="http://www.math.uni-utopia.de/Math-Net/" />
                     </dc:identifier>
                   </rdf:Description>
                 </rdf:li>
               </rdf:Seq>
             </vCard:Orgunit>
           </rdf:Description>
         </vCard:ORG>
       - <vCard:ADR>
           <rdf:Description vCard:Street="Nowherelane. 28" vCard:Locality="Utopia" vCard:Pcode="123456"
             vCard:Country="Futureland" />
         </vCard:ADR>
       - <vCard:ROLE>
           <rdf:Bag rdf:_1="Dreamer" />
         </vCard:ROLE>
```

Metadata of a professional test homepage (partial)

Since DC, RDF and XML together are quite complicated, it is important that the user gets a tool to produce those metadata.

[10] http://www.w3.org/XML/
[11] http://www.w3.org/
[12] http://www.imc.org/pdi/vcardoverview.html
[13] http://www.unicode.org/
[14] http://www.mathematik.uni-osnabrueck.de/TAB/
[15] http://www.iwi-iuk.org/material/RDF/1.1/profile/MNPerson/

3 Legal Aspects

Personal data are protected all over the world by different laws. Therefore it would be very difficult for a program to follow all rules in all countries. But if somebody uses a tool for the generation of a professional homepage and publishes this page on a webserver then there is no legal problem for the one running the search engine. Moreover our professional homepage will bear a legal notice that the data on the page may be collected for information and communication in the learned societies but may not be included in commercial mailing lists or databases. Apart from this legal notice it has to be stressed that the page will only include professional information and nothing private.
If the owner of a database with personal data publishes it in the web - for example a professional society like the AMS - then the database owner has to have the right to do so. In this case it would also be possible to generate professional homepages out of the database. The main advantage of this proceeding would be the interoperability of the data whereas the usual webpages produced out of the database are not at all interoperable!

4 Further Steps

Of course the exact layout and the headings for the page still need some discussion in the community. But this is just the beginning:

- It is quite easy to build a tool that reads data from a HTML-form and produces a professional homepage from these data. Nearly the same tool could be used as well, to read data from an existing database. It should also have an edit function, such that the information in an existing professional homepage can be changed easily, without having to reproduce the whole page.
- Since the planned system is highly distributed, it needs an alerting mechanism, that tells the search engine where to find newly installed professional homepages.
- If a homepage cannot be found by the search engine then suitable measures have to be taken.
- The user of the system should get an easy to use but powerful interface for the retrieval.
- The data should be crosslinked with other Math-Net services like the navigator.
- The Math-Net page for institutes usually exists in english plus in one or more optional other languages. It has to be discussed, whether this is also necessary for professional homepages.
- With the existing proposal a professional homepage contains up to about 40 links. This poses some problems for the simplicity and usability of the system, that have to be solved.

Please direct comments or suggestions on the proposal for professional homepages to the author.

Mathematics Subject Classification and Related Schemes in the OAI Framework

Antonella De Robbio[1], Dario Maguolo[1], and Alberto Marini[2]

[1] Mathematics Library, University Library System
University of Padova, Italy
[2] Institute for Applied Mathematics and Information Technology
National Research Council (CNR-IMATI), Milano, Italy

Abstract. This paper aims to give a feeling for the roles that discipline–oriented subject classifications can play in the Open Archive movement for the free dissemination of information in research activities. Mathematics, and Mathematics Subject Classification, will be the focuses around which we will move to discover a variety of presentation modes, protocols and tools for human and machine interoperability. The Open Archives Initiative (OAI) is intended to be the effective framework for such interplay. In the first part of this paper, we start by describing the most important subject classification schemes in mathematics and related disciplines. Then we sketch the structure of discipline-oriented schemes in view of browsing and we give an account of different browsing modalities, implemented in the tools we produced and collected in The Scientific Classifications Page. Finally we give an insight on the design, implementation and use of a programming language for the generation of hypertextual presentations of complex structured data. In the second part, we list different strategies for e-print communication in scientific research, up to the basic definitions of the Open Archives Initiative. A review of the functionalities actually implemented in OAI compatible archives managed by the EPrints software will lead us to some working hypotheses about the roles that subject classifications in mathematics and related disciplines can play in the scenarios of the Open Archives movement.

1 Subject Classification Schemes

Subject classification schemes are primary tools for the organization of knowledge and terminology in scientific disciplines. They are produced mainly by professional societies, or academic and research institutions, often to be employed in their own bibliographic databases. Although many of the issuing bodies have national or regional scope, subject classification schemes are generally international in scope, and are intended to be a communication tool for the international scientific community.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 100–111, 2003.

1.1 Schemes for Mathematics

Mathematics Subject Classification (MSC)[3] is developed by the editorial offices of the two world's most important bibliographic databases for mathematical research:

- MathSci, which is produced by the American Mathematical Society, and
- Zentralblatt MATH, which is produced by the European Mathematical Society, the Fachinformationszentrum (FIZ) Karlsruhe, Germany and other Editorial Units all over Europe.

MSC covers all branches of pure and applied mathematics, including probability and statistics, numerical analysis and computing, mathematical physics and economics, systems theory and control, information and communication theory. MSC underwent in time a number of revisions; the latest version came valid in January 2000, so it is called MSC2000.

On the side of mathematics education, the Zentralblatt für Didaktik der Mathematik Classification Scheme[4] is used for the bibliographic database MATHDI, which is edited by the European Mathematical Society, FIZ Karlsruhe, and Zentrum für Didaktik der Mathematik at Karlsruhe University, in cooperation with Math Doc Cell (France)

1.2 Schemes for Computing, Physics, Control and Information Technology

In the field of computing, including hardware, software, networking, theory, methodologies and applications, the most important tool is the Computing Classification System[5]. It is developed by the Association for Computing Machinery (USA) to classify items in the directories Computing Reviews and Guide to Computing Literature, which are edited by the same body. Section 68 Computer Science of MSC was designed in rather tight matching with a great part of CCS.

In the fields of theoretical, experimental and applied physics and astronomy we have the Physics and Astronomy Classification Scheme (PACS).[6] Section 02 Mathematical methods in physics of PACS closely resembles the top level codes for pure mathematics, probability and statistics of MSC. PACS is prepared and revised, at least biennially, by the American Institute of Physics. A version of PACS is established as Section A of INSPEC Classification.[7] INSPEC is a bibliographic information service provided by the Institution of Electrical Engineers (UK). It covers physics, electrical engineering, electronics, communications, control engineering, computers and computing, and information technology. INSPEC Classification has three other major sections:

[3] http://www.ams.org.msc/

[4] http://www.mathematik.uni-osnabrueck.de/projects/zdm/

[5] http://www.acm.org/class/1998/

[6] http://www.aip.org/pubservs/pacs.html

[7] http://www.iee.org.uk/publish/inspec/docs/classif/html

- Section B: Electrical & Electronic Engineering
- Section C: Computer & Control
- Section D: Information Technology

1.3 Schemes for Economics

The fields of economics are increasingly involved in mathematical arguments, both in theoretical and specific topics; and conversely, mathematical problems and theories even more often arise from economic domains. This can be seen by the place mathematical topics take in the Journal of Economic Literature Classification System,[8] developed by the American Economics Association for its indexing journal and for the corresponding EconLit database. Such topics are mostly located in the 62 Statistics, 90 Operations research, mathematical programming, and 91 Game theory, economics, social and behavioral sciences sections of MSC2000.

1.4 Discipline Specific and General Schemes

Besides these, many other subject classification schemes exist for use in any scientific discipline or field of disciplines. Yet other schemes are the general ones, not oriented to specific disciplines, such as Dewey Decimal Classification.[9]

2 Classification Schemes: From Structure to Browsing

2.1 The Common Structure of Subject Classification Schemes

The structure of subject classification schemes, be they discipline specific or general, is essentially the same: a relational system of categories, identified by alphanumerical codes, whose meaning is specified by descriptions or scope notes in some natural language (primarily, for current scientific research, English; translations and multilingual editions are frequently made available). Generally there is one main relation, which in most cases is tree-shaped (monohierarchical, or, simply, hierarchical) and the categories are called nodes. Sometimes, however, the main relation is a more relaxed partial order, allowing nodes to be under more than one node (so the relation is called multihierarchical). Other relations are considered as cross-references, allowing connections between diverging paths of the main relation. Subject classification schemes vary in time through succeeding versions; one version keeps valid for indexing and searching in a bibliographic database for a more or less long period of years. Two subsequent versions can be related by linking categories in the older and the newer version which hold some correspondence in meaning, even if the relation may not be one-one, or structure preserving, due to splits, merges, reorganizations, deaths and births of topics, as represented in the positions of the two versions.

[8] http://www.aeaweb.org/journal/elclasjn.html

[9] http:www.oclc.org/dewey/products/inde.htm

For example, Mathematics Subject Classification has 5531 categories in a three-level hierarchy. The top level counts 63 nodes. Cross-references, often equipped with explanatory text ("For ...") are of the following types: see also - see mainly - see. Some notes for coordinate indexing (and searching) are present. Physics and Astronomy Classification Scheme has a four-level hierarchy. The top level counts 10 nodes, the second level 66 nodes.

2.2 From Structure to Browsing

Due to their structural features, subject classifications are effective tools for browsing and searching in bibliographic databases, catalogs and other kinds of metadata repositories.

Moreover, subject classifications can set up knowledge organization tools for lexical collections extracted from metadata or fulltext databases, for terminologies, glossaries, dictionaries or encyclopedias, surveys, up to distributed libraries of natively digital documents or digitalized paper document. The set of descriptions of a classification scheme is itself a primary terminological resource.

2.3 H-volumes in the Scientific Classifications Page

Different modes in browsing subject classifications can be exploited by hypertextual techniques. We managed to produce various tools to demonstrate some of these modes. The Scientific Classifications Page[10] collects such tools. It is presented both in English and in Italian. It includes the following sections:

- The Mathematics Classification Page
- Mathematics Subject Classification MSC and Dewey Decimal Classification DDC
- Relating Scientific Subject Classification

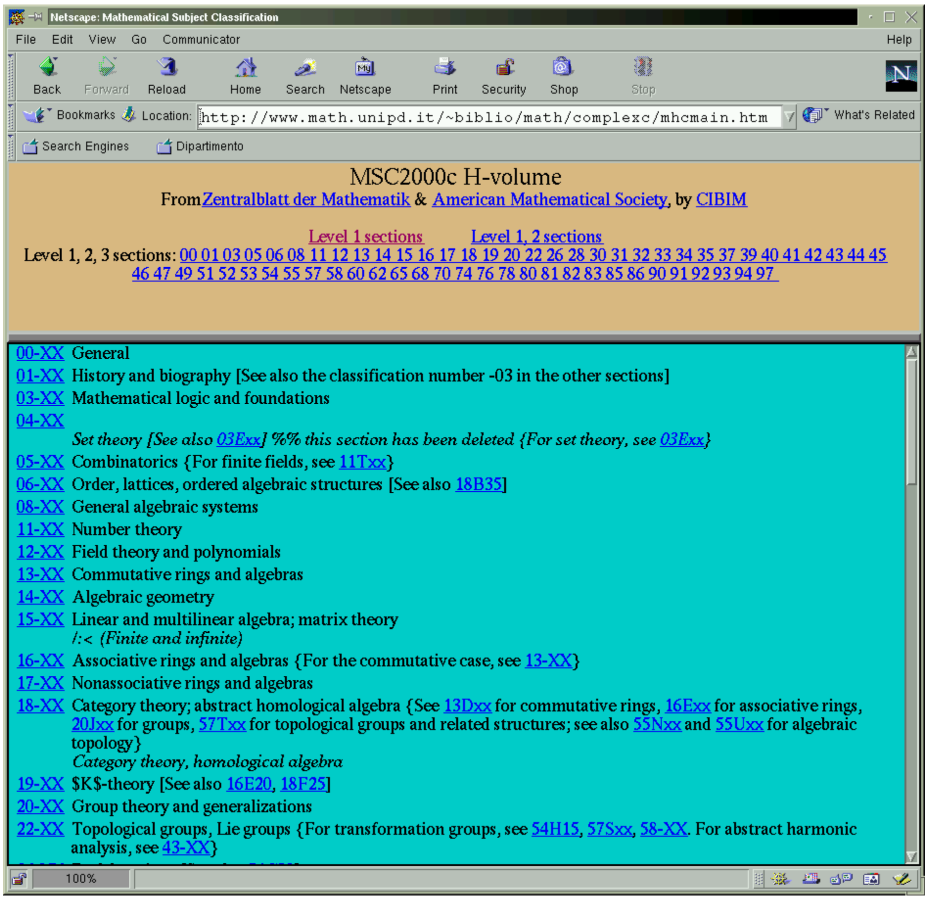
The tools we produced consist of systems of syntactically simple but highly connected and coordinated HTML pages, called H-volumes. H-volumes can amount even to thousands of files, written in plain HTML with simple JavaScript routines; in our working environment they are generated by a pool of standard C programs, starting from ASCII files, which present lists of records without redundancies and glossaries concerning attribute values. H-volumes can be employed to display any kind of structured information set, such as directories, biographical collections, metadata collections, databases, glossaries, dictionaries, encyclopedias, etc.

The actual production of H-volumes starts from ASCII files obtained by manipulating existing data sets and texts, in particular available Web pages. This preparation activity is worked out partly by hand (i.e. using interactively some flexible source editor), partly making use of text processing procedures developed contextually to the development of procedures for HTML page generation.

Let's now turn to see the sections of The Scientific Classifications Page in some detail.

[10] http://www.math.unipd.it/ biblio/math/eng.htm

The Mathematics Classification Page. The Mathematics Classification Page collects six hypertextual frame presentations of the latest version of Mathematics Subject Classification, MSC2000. From an ASCII file containing the whole MSC2000, a simple frame presentation was obtained. From the same file a double view presentation was obtained too. The former process, generating a simple frame presentation, was worked out on a file containing an Italian translation of MSC2000, while, by processing the two files in combination, we obtained a simple frame presentation, which displays interleaved English and Italian texts. From a file resulting from a comparison of MSC2000 with the 1991 version, we obtained a simple frame presentation which includes changes from MSC 1991: Finally, from the combination of the first ASCII file with a file which contained data about subject-specific pages of relevant Websites, we obtained a simple frame presentation, with guide pages linking to those subject-specific pages of Websites.


Netscape: Mathematical Subject Classification
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.math.unipd.it/~biblio/math/complexc/mhcmain.htm
What's Related
Search Engines Dipartimento
100%

MSC2000c H-volume

From Zentralblatt der Mathematik & American Mathematical Society, by CIBIM

Level 1 sections Level 1, 2 sections

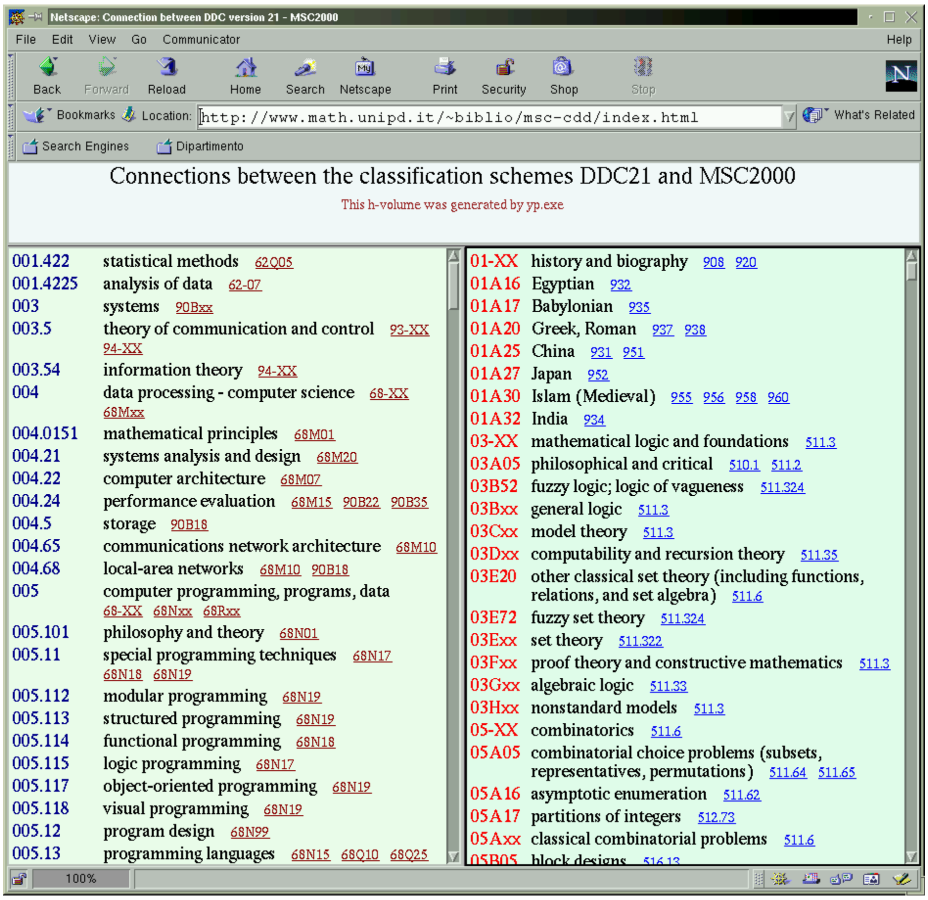
Level 1, 2, 3 sections: 00 01 03 05 06 08 11 12 13 14 15 16 17 18 19 20 22 26 28 30 31 32 33 34 35 37 39 40 41 42 43 44 45 46 47 49 51 52 53 54 55 57 58 60 62 65 68 70 74 76 78 80 81 82 83 85 86 90 91 92 93 94 97

00-XX General
01-XX History and biography [See also the classification number -03 in the other sections]
03-XX Mathematical logic and foundations
04-XX
Set theory [See also 03Exx] %% this section has been deleted {For set theory, see 03Exx}
05-XX Combinatorics {For finite fields, see 11Txx}
06-XX Order, lattices, ordered algebraic structures [See also 18B35]
08-XX General algebraic systems
11-XX Number theory
12-XX Field theory and polynomials
13-XX Commutative rings and algebras
14-XX Algebraic geometry
15-XX Linear and multilinear algebra; matrix theory
/:< (Finite and infinite)
16-XX Associative rings and algebras {For the commutative case, see 13-XX}
17-XX Nonassociative rings and algebras
18-XX Category theory; abstract homological algebra {See 13Dxx for commutative rings, 16Exx for associative rings, 20Jxx for groups, 57Txx for topological groups and related structures; see also 55Nxx and 55Uxx for algebraic topology}
Category theory, homological algebra
19-XX K-theory [See also 16E20, 18F25]
20-XX Group theory and generalizations
22-XX Topological groups, Lie groups {For transformation groups, see 54H15, 57Sxx, 58-XX. For abstract harmonic analysis, see 43-XX}

Fig. 1. Comparison of MSC versions 1991-2000

This is an example of simple frame presentation. The top frame is a sort of Table of Contents, which gives access to different slicings of the scheme: single list presentations of the classification categories at level 1 and 1-2, and an indexed set of list presentations which covers the whole scheme. For the latter, the top frame displays the list of the first 2 digits of the codes of the 63 level 1 categories; each item in the list points to a page which is displayed in the frame below, containing a list presentation of the subtree below the indicated level 1 category. In this way, the long list of all the classification categories is divided into a number of sublists, so you can browse the classification scheme by transferring only files of moderate size.

On the other hand, double or multiple view presentations can be exploited to navigate through transversal links either inside one version of a classification scheme or among more schemes or versions: you can move to and from parallel views of them.



http://www.math.unipd.it/~biblio/msc-cdd/index.html

Connections between the classification schemes DDC21 and MSC2000

This h-volume was generated by yp.exe

001.422	statistical methods	62Q05
001.4225	analysis of data	62-07
003	systems	90Bxx
003.5	theory of communication and control	93-XX 94-XX
003.54	information theory	94-XX
004	data processing - computer science	68-XX 68Mxx
004.0151	mathematical principles	68M01
004.21	systems analysis and design	68M20
004.22	computer architecture	68M07
004.24	performance evaluation	68M15 90B22 90B35
004.5	storage	90B18
004.65	communications network architecture	68M10
004.68	local-area networks	68M10 90B18
005	computer programming, programs, data	68-XX 68Nxx 68Rxx
005.101	philosophy and theory	68N01
005.11	special programming techniques	68N17 68N18 68N19
005.112	modular programming	68N19
005.113	structured programming	68N19
005.114	functional programming	68N18
005.115	logic programming	68N17
005.117	object-oriented programming	68N19
005.118	visual programming	68N19
005.12	program design	68N99
005.13	programming languages	68N15 68Q10 68Q25

01-XX	history and biography	908 920
01A16	Egyptian	932
01A17	Babylonian	935
01A20	Greek, Roman	937 938
01A25	China	931 951
01A27	Japan	952
01A30	Islam (Medieval)	955 956 958 960
01A32	India	934
03-XX	mathematical logic and foundations	511.3
03A05	philosophical and critical	510.1 511.2
03B52	fuzzy logic; logic of vagueness	511.324
03Bxx	general logic	511.3
03Cxx	model theory	511.3
03Dxx	computability and recursion theory	511.35
03E20	other classical set theory (including functions, relations, and set algebra)	511.6
03E72	fuzzy set theory	511.324
03Exx	set theory	511.322
03Fxx	proof theory and constructive mathematics	511.3
03Gxx	algebraic logic	511.33
03Hxx	nonstandard models	511.3
05-XX	combinatorics	511.6
05A05	combinatorial choice problems (subsets, representatives, permutations)	511.64 511.65
05A16	asymptotic enumeration	511.62
05A17	partitions of integers	512.73
05Axx	classical combinatorial problems	511.6
05B05	block designs	516.13

Fig. 2. Navigating transversal links

Here is an example of double view presentation, showing connections between categories from the Dewey Decimal Classification, 21st edition, and MSC2000.

The double view presentation included in The Mathematics Classification Page is actually a duplicated simple frame presentation of the whole tree of MSC2000 in English which allows walking through cross references while keeping vision of the contexts of both endpoints of the selected cross references.

Mathematics Subject Classification MSC and Dewey Decimal Classification DDC. The Mathematics Subject Classification MSC and Dewey Decimal Classification DDC section of The Scientific Classifications Page includes two English language presentations:

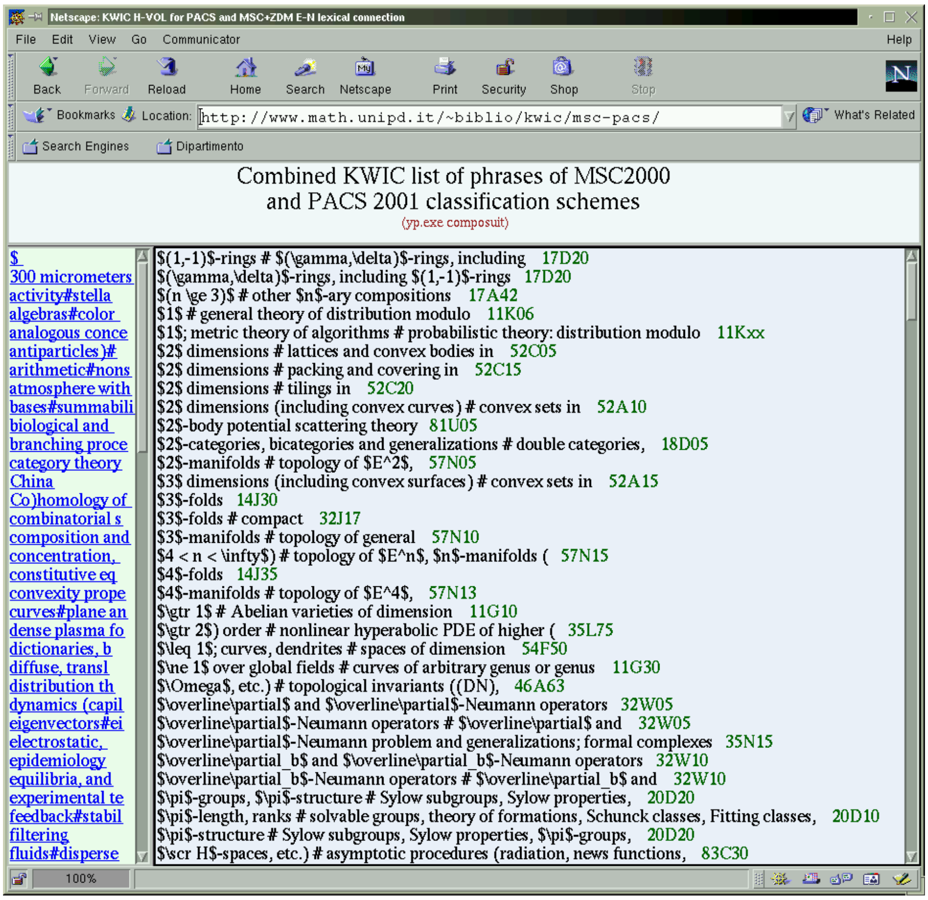
- the just shown page of connections between categories from the Dewey Decimal Classification, 21st edition, and MSC2000
- a KWIC list h-volume for the combined set of descriptions of:
- a revision proposal for the 510 DDC section, Mathematics, presented in January 2001
- MSC2000

The sections E–N of the ZDM classification, encoded as 97E–97N in the MSC style.

KWIC list h-volumes (as in Fig. 3) are devised for discovering textual similarities among subject descriptions in one or more classification schemes or versions, in order to obtain suggestions about possible affinities of contents. A KWIC list (KWIC shortens KeyWords In Context) presents every description through its circular permutations, beginning with a significant word or phrase; the overall list is ordered along the list of significant words. By a method similar to that employed for simple frame presentation, long ordered list, as generally a KWIC list is, can be endowed with some sort of distributor allowing to reach quickly determined points or sections of the long sequence. A distributor can be built with pointers to initial letters, initial words of paged sections, sublists dealing with particular categories of entities. The list of permuted descriptions, subdivided into smaller manageable lists, is displayed on the right, while the distributor appears in the left frame. KWIC lists may not be intended for the end user, rather as a help in establishing structured connections within or among classification schemes. This activity, although can greatly benefit from automated techniques, requires an amount of field specific knowledge which can't be automated, at least with the current technologies. The connections so discovered can be subsequently displayed through multiple view presentations of the involved classification schemes.

Relating Scientific Subject Classifications. The Relating Scientific Subject Classifications section of The Scientific Classifications Page contains a set of English language presentations (in one case bilingual):

- a double view presentation, showing connections between categories from the ACM Computing Classification System (1998), and MSC2000



Combined KWIC list of phrases of MSC2000
and PACS 2001 classification schemes
(yp.exe composuit)

```
$
300 micrometers
activity#stella
algebras#color
analogous conce
antiparticles)#
arithmetic#nons
atmosphere with
bases#summabili
biological and
branching proce
category theory
China
Co)homology of
combinatorial s
composition and
concentration,
constitutive eq
convexity prope
curves#plane an
dense plasma fo
dictionaries, b
diffuse, transl
distribution th
dynamics (capil
eigenvectors#ei
electrostatic,
epidemiology
equilibria, and
experimental te
feedback#stabil
filtering
fluids#disperse
```

```
$(1,-1)$-rings # $(\gamma,\delta)$-rings, including   17D20
$(\gamma,\delta)$-rings, including $(1,-1)$-rings   17D20
$(n \ge 3)$ # other $n$-ary compositions   17A42
$1$ # general theory of distribution modulo   11K06
$1$; metric theory of algorithms # probabilistic theory: distribution modulo   11Kxx
$2$ dimensions # lattices and convex bodies in   52C05
$2$ dimensions # packing and covering in   52C15
$2$ dimensions # tilings in   52C20
$2$ dimensions (including convex curves) # convex sets in   52A10
$2$-body potential scattering theory   81U05
$2$-categories, bicategories and generalizations # double categories,   18D05
$2$-manifolds # topology of $E^2$,   57N05
$3$ dimensions (including convex surfaces) # convex sets in   52A15
$3$-folds   14J30
$3$-folds # compact   32J17
$3$-manifolds # topology of general   57N10
$4 < n < \infty$) # topology of $E^n$, $n$-manifolds (   57N15
$4$-folds   14J35
$4$-manifolds # topology of $E^4$,   57N13
$\gtr 1$ # Abelian varieties of dimension   11G10
$\gtr 2$) order # nonlinear hyperbolic PDE of higher (   35L75
$\leq 1$; curves, dendrites # spaces of dimension   54F50
$\ne 1$ over global fields # curves of arbitrary genus or genus   11G30
$\Omega$, etc.) # topological invariants ((DN),   46A63
$\overline\partial$ and $\overline\partial$-Neumann operators   32W05
$\overline\partial$-Neumann operators # $\overline\partial$ and   32W05
$\overline\partial$-Neumann problem and generalizations; formal complexes   35N15
$\overline\partial_b$ and $\overline\partial_b$-Neumann operators   32W10
$\overline\partial_b$-Neumann operators # $\overline\partial_b$ and   32W10
$\pi$-groups, $\pi$-structure # Sylow subgroups, Sylow properties,   20D20
$\pi$-length, ranks # solvable groups, theory of formations, Schunck classes, Fitting classes,   20D10
$\pi$-structure # Sylow subgroups, Sylow properties, $\pi$-groups,   20D20
$\scr H$-spaces, etc.) # asymptotic procedures (radiation, news functions,   83C30
```

Fig. 3. Sections E–N of the ZDM classification in MSC style.

- separate KWIC lists of descriptions of MSC2000, of PACS 2001, of ACM Computing Classification System (1998)
- combined KWIC list of descriptions of MSC2000 and PACS 2001, and of MSC2000 and ACM Computing Classification System (1998).

2.4 Towards a Presentation Generating Language

The H-volumes we produced are not intended to be taken as ultimate references, but as prototypes capable to clarify the real problems to face for the production of more complete and professional h-volumes and to test their effectiveness as documentation tools. In fact, the development of such prototypes brought to the specification of parametrization mechanisms, data structures and processing modes which induced to define a programming language oriented to the manipulation of hypertextual presentations and to displays of mathematical structures.

The definition and the implementation of an experimental language called TAMP (Text Analysis Manipulation and Presentation) was actually started up. TAMP is aimed to the analysis of text files of specified format (TeX, HTML, XML, etc.), the organization of specific knowledge bases endowed with links to other Internet resources and their presentation through HTML pages. The language is implemented by means of a single C program, called YP, reading and generating only plain ASCII files. The first input file, characterized by the extension .ypg, is the source file of the program to execute. Many other specific files pointed out in the program are read and written. Such files contain either data or sources of specific programs, dedicated to generate HTML files or other publishable files (e.g. TeX files), to prepare intermediate files, e.g., lists following defined orderings and collecting items provided by partial unordered files (in particular files extracted from Web pages), or to control manipulations of some types (actually few) of mathematical structures starting from relatively simple expressions of basic ones in order to produce readable presentations of significant structures, possibly in a good consulting context. The implementation is only at a "less than 1 version" and is poor in many respects, but has some peculiarities that allowed the production of practical Web pages and whose developments seem worthy of investigation.

The language can control many data types: the basic ones are integers (but not yet real numbers) and strings; it controls aggregates of basic data as sequences, tables and sequences of sequences. Moreover it's possible to manipulate some specific presentation structures (indexing KWIC lists, glossaries, etc.) and the representations of specific mathematical structures (permutations, partitions, graphs, trees, paths in combinatorial plane, etc.). While a good choice of operators on basic data types and their aggregates is provided, only few operators acting on specific structures are implemented. On the other hand the implementing program YP has good extensibility features: the data types are parameterized, simple schemes allow the introduction of identifiers and general functional characteristics of new operators and their actions can be implemented in routines whose collocation and role are not difficult to tune with the characteristics of existing operators. Among richer data types the language provides some kinds of constructors, composite entities targeted to build presentation structures. A typical example is given by the so called KWIC engine: its definition requires to specify the fields of a flat file, the catalogued routines charged to distinguish and accept these fields, the catalogued routines commissioned to build the different fields of final KWIC items and the parameters required by some routines. Specific statements allow to activate the constructors giving the possibility to choose for them parameters such as schemes controlling files to be generated and prefixes of their names.

An important characteristic of the language is the possibility to define automata at different levels of generality. The automata of the more general type can be defined by a specific rich jargon opening the possibility to determine effective models of acceptors, transducers, text analysers and text generators, typically through successive refinements. Moreover, the translator of the proposed language can be used with a versatile preprocessor allowing substitutions, inclusions, selections and iterations of good reach: its control structures can act on variables concerning strings, integers and files. This preprocessor limits the actual major language drawback, i.e. lack of modularity. A group of statements that would be natural to encapsulate in a module can be recorded in a file endowed with dummy strings: this file can be included in other source files, either in the main one or in a file that can be included similarly.

3 The OAI Framework

3.1 E-print Communication: Tools and Networking Architectures

Scientific research relies heavily on the rapid dissemination of results. So the slow formal process of submitting papers to journals has been augmented by other, more rapid, dissemination methods [1], [2]. Originally dissemination involved printed documents, such as technical reports and informal conference papers. Then researchers started taking advantage of the Internet, putting papers on ftp sites and later on various web sites. But these resources were fragmented. Searching through them resulted to be very difficult, and there was no guarantee that information would be archived at the end of a research project. Different strategies for scientific research communication via e-prints have been developed in time, which involve:

- small specialized archives
- centralized archives such as arXIv[11] for physics and related disciplines, mathematics, nonlinear sciences, computer science; and CogPrints[12] for cognitive science, artificial intelligence, computational linguistics and neuroscience
- single or networked institutional archives, such as NCSTRL[13] and the ERCIM Technical Reference Digital Library[14] for computer science and mathematics
- distributed networks connected by some interoperability protocol, such as RePEc[15] for economics, and DoIS[16] for library and information science
- umbrella servers, such as MPRESS[17] for mathematics
- servers connected to groups of journals or sponsored by commercial publishers, etc.

Web search and cash engines like Researchindex (formerly Citeseer)[18], provide a solution which has been appreciated especially by people in the computing area. E-prints posted in personal homepages without any specific care about metadata are harvested and cashed; the service is comprehensive with reference linking.

[11] http://arXiv.org
[12] http://www.cogprints.soton.ac.uk
[13] http://www.ncstrl.org
[14] http://www.iei.pi.cnr.it/DELOS/EDL/ETRDL_Con/
[15] http://www.repec.org
[16] http://d ois.mimas.ac.uk
[17] http://mathnet.preprints.org
[18] http://citeseer.nj.nec.com

3.2 The Open Archives Initiative

The Open Archives Initiative (OAI)[19] is an international effort to develop interoperability standards for disseminating content over the Web. OAI stresses the separation of being a data provider (i.e., publisher) and being a service provider (i.e., interface for search, browsing, reference linking). On the other hand, nothing prevents the same system to embody and integrate both functions. It is even possible for individual researchers to develop personal open archives, which can be accessed to build tailored personal web sites and other services, as well as harvested into department archives.

The base concept of the OAI is metadata harvesting, which is realized in the OAI Protocol for Metadata Harvesting[20]. So it no longer matters where papers are archived; the papers in all registered OAI-compliant archives can be harvested using the OAI protocol into one global "virtual archive" by Open Archives service providers.

3.3 OAI Compatible Refereed Self-archives: The EPrints 2 Software

EPrints[21] is a free (General Public License) software for managing e-prints archives, developed at the Electronics and Computer Science Department of the University of Southampton (UK). It is aimed at organizations and communities rather than individuals. It provides an interface for system administrators, for archive editors to process submissions, for authors to deposit papers, and for users to access papers by searching or browsing metadata. The system comes configured to run an institutional pre-prints archive, but can be reconfigured with utterly different metadata fields and content. Any version of EPrints is fully interoperable with the current OAI Protocol for Metadata Harvesting.

4 Conclusions

Our work has been directed to the definition of text processing methodologies for the development of hypertextual presentations of complex documentation structures. Such presentation modalities can enrich the browsing functionalities of archives and service providers in the OAI framework, allowing a full network of bridges among specific subject areas to guide advanced research communication activities.

In particular, we are investigating the possibility of providing the EPrints software with tools modeled on the experimental ones we produced for the Scientific Classification Page. Centering with Mathematics Subject Classification, bridges can be launched and passed through inside mathematics and among the disciplines that live and develop with mathematics. This is equivalent to say that bridges can be launched all over the world of scientific and technological

[19] http://www.openarchives.org

[20] http:www.openarchives.org/OAI/openarchivesprotocol.htm

[21] http://www.eprints.org

knowledge, if we are aware of the dynamics that mathematical disciplines are ever more moving in modeling and computing activities for every field of human knowledge.

References

1. De Robbio, Antonella, Maguolo, Dario, Marini, Alberto: Scientific and General Subject Classifications in the Digital World. High Energy Physics Libraries Webzine, Issue 5, November 2001 http://doc.cern.ch/heplw/5/papers/4/
2. Marini, Alberto: Text Processing for Presentation and Manipulation of Mathematical Resources. Paper presented at the Workshop “Electronic Media in Mathematics”, Coimbra (Portugal), September 13–15 2001 http://www.mat.uc.pt/EMM/index.html

Metadata Models – International Developments and Implementation

Heike Neuroth and Margo Bargheer

Göttingen State and University Library (SUB), Germany

1 Introduction into Metadata

Definitions of metadata often describe them as "data about other data", sometimes refined through the expression "structured data about data". This definition over-simplifies the facts, that metadata on one hand have been in use long before the digital age, be it in library catalogues or on inventory cards of museums, and that on the other hand the entity they represent does not necessarily need to be in form of bits and bytes. A highly generic definition for metadata therefore could be "(structured) information about (digital) objects"[1]. Transferred to the digital world "structured information" stands for "structured data".

The term metadata does not solely refer to the representation of existing parts of reality such as data streams or objects, but is applicable as well to describe descriptions – metadata collections like subject gateways or digital libraries can be described in a metadata set as well. The unambiguous use of the term metadata therefore requires to regard the context as well as the levels of complexity to be described and the receivers of the information supplied through metadata. Metadata usually carry structured information like "Author", "Title", "Subject" etc, bits of information, that are semantically interconnected.

Instead of a fixed term or a standardised format metadata therefore should be understood as a form of language to exchange information on objects like books or digital resources as well as for purposes like archiving. Tom Baker describes Dublin Core[2], one of the most common metadata formats, as "pidgin for Digital Tourists"[3]. Baker stresses the point that the structure of metadata forms a grammar, whose "basic pattern is easily grasped" and which "is well-suited to serve as an auxiliary language for digital libraries" (ibid.).

Metadata thus offer an elementary level of understanding, either between machines, human beings or in between the two of them.

[1] A more abstract definition could be "structured information on specific and describable sectors of reality". "Specific and describable sectors of reality" could mean "a certain stream of data" as well as "an object", "a document" or "a human being". It should be emphasised, that metadata usually draw their meaning from the unambiguous relation to the reality they stand for. If there is no proof for the existence of the described, the description might be meaningless in the field of digital data.

[2] http://www.dublincore.org/

[3] http://dlib.org/dlib/october00/baker/10baker.html

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 112–121, 2003.

1.1 Types and Function of Metadata

Who benefits from metadata, assumed they have been stringently applied? For the creators of digital objects (resources, documents, collections), be it a producer or a scientist, metadata offer the possibility to prepare a formalised description of their work and thus exercise control over it. For cataloguers metadata form an established tool for the description of particular types of resources. They enable implementors to set up reliable schema and information providers to share such reliable schema through cross-walks and harmonising processes in order to establish core sets, which are mentioned below in further detail. A major benefit of metadata is given to end users performing information retrieval, who can identify, locate and compare relevant resources and recognise the respective functional requirements.

According to their different statements on the nature of the described digital objects metadata can be divided into the following types:

- Content metadata/descriptive metadata
  refer to description of content or further bibliographic information, can be specified according to document type and/or subject. Technical details of real-life objects such as museum artefacts or books belong to this metadata type as well.
- Administrative metadata
  carry information for the distributed administration and the maintenance of archiving systems such as versioning of the metadata set or date stamps and signatures regarding metadata modifications.
- Structural Metadata
  allow the navigation in archiving systems by offering information on hierarchical levels (journal → article, monograph → chapter, artefact → detail).
- Technical Metadata
  carry information on the digital nature of the described resource/document such as size, format, resolution or colour.
- Preservation metadata
  inform about the durable preservation of digital objects and the storage/presentation format regarding migration and emulation.
- Terms and condition metadata
  allow the disclosure of copyrights, intellectual property and retrieval conditions such as payments or registration.
- Metadata about metadata collections
  allow users to cross-search distributed archiving systems (e.g. RSLP collection description[4]).
- Metadata about the use of metadata
  for example project/domain specific application profiles such as Renardus[5], DC Libraries[6], etc. Metadata of this type are usually stored in registries

[4] http://www.ukoln.ac.uk/metadata/rslp/isadg/

[5] http://renardus.sub.uni-goettingen.de/renap/format.html

[6] http://www.dublincore.org/documents/2002/09/24/library-application-profile/

carrying information on application profiles and element sets of the respective co-operation partners in distributed archiving systems.

Metadata feature certain aspects of functionality. Their service function allows structured access to different resources by offering possibilities like cross-searching, cross-browsing, result display, result ranking or result sorting. Their technical and administration functions allow long-term maintenance, metadata exchange and sharing as well as reliable archiving.

2 Application Profiles, Namespaces, and Registries

2.1 Application Profiles

To fully meet the potential of metadata as an elementary language for interoperability it is necessary to establish application profiles. Application profiles usually consist of several element sets[7], but at least one of them. Application profiles can be used "as a way of making sense of differing relationship that implementors and namespace managers have towards metadata schema, and the different ways the use and develop schema." (Heery/Patel 2000). Application profiles describe elements necessary for a certain implementation by specification of obligations such as "mandatory", "strongly recommended" or "optional" in order to ensure that certain information such as "Title" and "Creator" will be supported by a specific project or domain specific implementation. Application profiles disclose which organising body or institution maintains which element set and provide guidelines and best practice for each element. Application profiles offer the possibility to shape domain-specific variations. Some examples:

- DC-Education[8]
  is based on the Dublin Core element set and includes elements from the IEEE Learning Object Metadata (LOM) element set. Target groups of education material for example thus can be specified on the metadata level already.
- DC-Government[9]
  includes all fifteen Dublin Core elements, supplemented with domain-specific elements refining information on rights (security classifications such as "top secret"), dates (reliable information on publishing dates), subjects, relation.
- DC-Libraries
  is based on the Dublin Core elements and includes elements of MODS (Metadata Description Object Schema), that allow the expression of roles for Creator/Contributor or the differentiation according to genre. DC Libraries support "Library of Congress Subject Headings" as an encoding scheme for Temporal Coverage.

[7] The term "Element set" should be used instead of the formerly term "namespace", as "namespace" prevails as an XML-specific term defined and maintained by W3C.

[8] see DCMI Education Working Group,
http://www.dublincore.org/groups/education/

[9] see DCMI Education Working Group,
http://www.dublincore.org/groups/government/

- EULER AP[10]
  as a mathematics-specific application profile extends the DC elements by differentiating the role of the creator according to scientific publishing conditions or extending DC source type lists with EULER specific types.
- Renardus AP[11]
  Renardus as an interdisciplinary broker service offering sophisticated cross-browsing extends the DC elements by supporting several encoding schemas.

2.2 Element Sets (Formerly Called Namespaces)

Element sets describe a well defined set of metadata elements, according to domain or subject-specific requirements and to different implementations. Besides the pure graduation of information, element sets define semantics and syntax for each metadata element, serving as "a vocabulary that has been formally published, usually on the Web; it describes elements and qualifiers with natural language labels, definitions, and other relevant documentation." (Baker 2000) Metadata generation should follow clear cataloguing rules. With Baker's approach, understanding metadata element sets as a vocabulary, it is thus the underlying grammar to the vocabulary that conditions those cataloguing rules into a consistent semantics and syntax. This consistency is reached, when each element is treated as "a unique identifier formed by a name (e.g., Title)" (ibid.). Examples for element sets and their abbreviations are listed below:

- Dublin Core Metadata Element Set, Version 1.1[12]
  dc
  Dublin Core is one of the most common metadata formats maintained by the Dublin Core Metadata Initiative and it's bodies, the respective working groups such as e.g. DCMI Libraries. The metadata element set can be refined in certain fields with Dublin Core Qualifiers (DCMI Metadata Terms[13]), abbreviated by "dc terms".
- Agricultural Metadata Framework[14]
  fao
  stems from a common initiative by the FAO (Food and Agriculture Organisation of the United Nations) and OneWorld Europe[15], which is based on Dublin Core, but specified according to the requirements of the agricultural field.
- IEEE Learning Object Metadata[16]
  ieee-lom

[10] http://www.emis.de/projects/EULER/metadata.html
[11] http://renardus.sub.uni-goettingen.de/renap/renap.html
[12] http://www.dublincore.org/documents/dces/
[13] http://www.dublincore.org/documents/dcmi-terms/
[14] http://www.fao.org/agris/MagazineArchive/MetaData/TaskForceonDCMI.htm
[15] http://www.oneworld.org
[16] IEEE (Institute of Electrical and Electronics Engineers, Inc.) Learning Technology Standards Committee (LTSC). Originating URL: http://ltsc.ieee.org/wg12/

is a subject specific element set for the field of education. Target groups of education material for example thus can be specified on the metadata level already.
- RSLP[17] Collection Description Metadata
  rslpcld
  refers to the special requirements the description of large collections pose.
- Metadata Object Description Schema (MODS)[18]
  mods
  "The Library of Congress' Network Development and MARC Standards Office, with interested experts, has developed a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. As an XML schema, MODS is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. A mapping from Dublin Core metadata set to MODS schema is now available[19].

2.3 Metadata Registry

Metadata element sets are mainly used for the mixing and matching of data. To ensure their reliability, interoperability and long-term maintenance, projects for setting up metadata registries have been started. They arose from the "recognition of the benefits of shared data dictionaries leading to the specification of a formal registration process in the standard ISO/IEC 11179.[20]" Metadata registries serve as reference tools for a wide range of complex data sets by promoting the re-use of already defined elements, disclosing data definitions on element sets used in local or subject-specific implementations, thus ensuring their authoritativeness. Other objectives are registries of controlled vocabularies within particular domains or developments of domain specific application profiles. SCHEMA and MetaForm are non-authoritative registries, which do not involve in the launching of definitions and standards:

- SCHEMAS[21] project, funded by the European Commission under the Fifth Framework Programme, "has provided a forum for metadata schema designers involved in projects under the IST Programme and national initiatives in Europe" (SCHEMAS Homepage). The CORES registry[22] will be developed to register application profiles and metadata element sets.
- MetaForm[23] (SUB), part of the German Meta-Lib[24] project, is a database with a special focus on the Dublin Core and its manifestations in various

[17] Research Support Libraries Programme, http://www.rslp.ac.uk/AboutUs/
[18] http://www.loc.gov/standards/mods/
[19] http://www.loc.gov/standards/mods/dcsimple-mods.html
[20] DELOS White Paper, 2003: Principles of Metadata Registries. Still to be published.
[21] http://www.schemas-forum.org/
[22] http://www.cores-eu.net/
[23] http://www2.sub.uni-goettingen.de/metaform/index.html
[24] Metadata Initiative of German Libraries,
http://www.dbi-berlin.de/projekte/einzproj/meta/meta00.htm

implementations. MetaForm supports mapping processes between DC applications and other formats, crosswalks between the Dublin Core Metadata Element Set and its various dialects as well as crosscuts of a particular Dublin Core element (e.g DC.Creator) through different formats.
- The Dublin Core Metadata Initiative is the main authoritative body for defining and maintaining the Dublin Core element set. Terms and definitions, encoding schemes and qualifiers as well as controlled vocabulary systems used within the Dublin Core community need to be acknowledged by the usage board[25] to ensure Dublin Core's consistency and further development and are registered in a specific Dublin Core registry[26].

Summarising the preceding paragraphs, the concept of application profiles and element sets, put down in metadata registries, follows certain objectives. They prevent the implementation of metadata schema based on in-house solutions, who are time-consuming and double effort and might work only locally. In this sense they increase the interoperability among different implementors dramatically and encourage cooperation between several partners resp. projects. The workability of metadata schema is ensured through the definition of unique identifiers for single elements as well as particular element sets or controlled vocabularies systems. These concepts ensure that bodies like management authorities feel responsible for the maintenance of metadata elements. Only durable and maintained concepts ensure interoperability and can serve as reliable interchange formats for cross-searching distributed services and heterogeneous environments.

3 Metadata Implementation

3.1 Core Set of Metadata

The concept of metadata core sets is a prerequisite for an advanced service such as cross-searching. Cross-searching implies on the one hand, that searching over distributed and sometimes heterogeneous metadata collections via one single user-interface is technically supported, on the other hand, that searching results bring up comparable resources meeting the user's needs. To meet this objective, besides the technical development such as software architecture thorough preparatory work on the participating partner's metadata element sets is needed. This includes the following:

- analysis of semantics and syntax of each element
- investigation of all qualifiers in use. This refers to refinements as well as encoding schemes such as country-lists, classification systems or keywords, thesauri or controlled vocabularies

[25] The DC usage board consists of 7-11 members, who are knowledgeable about DC and actively work in the metadata community. Originating URL: http://dublincore.org/usage/

[26] see ongoing developments in the DCMI Registry Working Group, http://www.dublincore.org/groups/registry/

- analysis and harmonisation of cataloguing rules. This applies to core elements like “Title” or “Creator” as well as to “Description” and qualifiers such as keywords
- equalisation of rules on repeatability of each element
- analysis and harmonisation on obligation rules such as “mandatory”, “strongly recommended” and “optional”
- analysis of language qualifiers in use. It is highly recommended to investigate and harmonise language qualifiers for the fields “Title”, “Description” and “Subject”, especially in multilingual services or international co-operations that allow cross-searching over language boundaries

This preparatory work is usually conducted by detailed questionnaires and following discussions between the partners involved. In the end it should lead to the determination of a core set of metadata through identifying the minimum set of metadata elements that are needed to reasonably run the service and the maximum set of elements that each partner is able to support sufficiently. The definition of each element entering the core set is based on a “Format of Entry”.

3.2 Renardus Application Profile

Renardus aims to “provide a trusted source of selected, high quality Internet resources for those teaching, learning and researching in higher education in Europe. Renardus provides integrated search and browse access to records from individual participating subject gateway services across Europe.”[27] It will serve as an example to illustrate the concept of application profiles already mentioned. The application profile of Renardus is based on five element sets, encoded in XML/RDF.

- Dublin Core Metadata Element Set,
  Version 1.1: Reference Description [dc1.1]
- Dublin Core Qualifiers Element Set, which includes other DCMIS Elements and Dublin Core Qualifiers [dcterms]
- Dublin Core Type Vocabulary [dcmitype]
- Renardus Metadata Element Set [rmes]
- Renardus Metadata Element Set Qualifiers [rmesq]

The following examples show how different element sets have been implemented into the application profile. Cataloguing rules thus can be deducted from the application profile.

- Title and Title.Alternative
  title: dc1.1 (mandatory, not repeatable, language tag)
  title.alternative: dcterms (optional, repeatable, language tag)
- Creator
  dc1.1 (strongly recommended, repeatable)
  LastName, FirstName: rmesq (strongly recommended, repeatable)

[27] http://www.renardus.org/

- Description
  dc1.1 (mandatory in text version, repeatable, language tag)
- Country
  rmesq (strongly recommended, repeatable)

4 Long-Term Preservation

4.1 Preservation Metadata

Preservation Metadata inform about the durable preservation of digital objects and the storage/presentation format regarding migration and emulation. This includes the description of the complete "life cycle" of a digital object, especially when digital objects are not digital-born but results of transformation processes such as digitisation:

- provenance information like original format of the book
- date of digitisation
- technical information of the digitisation process
- presentation and storage format (supporting migration and emulation)
- RMS (rights-management-system), supplying information on terms and conditions of access, copyright and intellectual property
- etc.

There are a lot of projects and initiatives which are working on preservation metadata. The following list is not complete but provides an overview:

- Preservation Metadata for Digital Objects: A Review of the State of the Art, A White Paper by the OCLC/RLG Working Group on Preservation Metadata (January 31, 2001),
  http://www.oclc.org/research/pmwg/presmeta_wp.pdf
- Preservation Metadata and the OAIS Information Model,
  A Metadata Framework to Support the Preservation of Digital Objects, A Report by The OCLC/RLG Working Group on Preservation Metadata (June 2002),
  http://www.oclc.org/research/pmwg/pm_framework.pdf
- A Recommendation for Preservation Description Information, A Report by The OCLC/RLG Working Group on Preservation Metadata (April 2002)
  http://www.oclc.org/research/pmwg/pres_desc_info.pdf
- A Recommendation for Content Information, A Report by The OCLC/RLG Working Group on Preservation Metadata (October 2001),
  http://www.oclc.org/research/pmwg/contentinformation.pdf
- Metadata Encoding and Transmission Standard (METS),
  http://www.loc.gov/standards/mets/
- National Library of Australia – Preservation Metadata for Digital Collections, Exposure Draft,
  http://www.nla.gov.au/preserve/pmeta.html

- Metadata for Preservation – CEDARS Project Document AIW01 (August 1998),
  http://www.ukoln.ac.uk/metadata/cedars/AIW01.html
- Cedars Guide To : Preservation Metadata (March 2002),
  http://www.leeds.ac.uk/cedars/guideto/metadata/
- NEDLIB: Metadata for long term-preservation (July 2000),
  http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm
- Rebecca Guenther: The joint work of the OCLC-RLG Preservation Metadata Working Group, RLG Open Forum at ALA June 16, 2002,
  http://www.rlg.org/longterm/forum02/guenther.html

4.2 OAIS (Open Archival Information System Reference Model)[28]

The OAIS is a conceptual framework for an archival system important for the long-term preservation of digital objects, stemming from the work of the space data community. Since it's beginnings in 1997 the framework has gained international recognition due to the common effort of RLG[29], OCLC, and many members of their respective organisations in shaping the reference model and adapting it for the use in libraries, archives and research repositories. Especially archive designers and maintainers can benefit from a reliable framework such as OAIS, which provides common concepts and terminology. The OAIS has been implemented already into the NEDLIB project[30], run by the Koninklijke Bibliotheek, Den Haag. The participation of international organisations ensures the maintenance and further development of the framework.

4.3 Open Issues

Although OAIS provides a sophisticated framework, several issues still need to be developed in the future. This especially refers to distributed archiving systems such as EMANI, the electronic mathematical archiving network initiative[31]. Projects like EMANI work on concepts, how to handle the granularity of digital objects (such as journals/articel or monograph/chapter) in archiving systems. Other questions to be solved refer to the development of a minimum set of preservation metadata in distributed services. Automatic generation of metadata and preservation metadata standards, best practices and clear guidelines display fields for further engagement as well.

5 Outlook

The questions thrown open by this article refer to certain related topics. Further developments from there thus promise to be fruitful for those questions. Schema

[28] http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pd
[29] http://www.rlg.org/rlg.html
[30] http://www.kb.nl/coop/nedlib/
[31] http://www.emani.org/

issues for example will be touched by progresses in XML/RDF research, architecture issues will benefit from progresses on the field of de-central discovery systems such as Z39.50[32], LDAP[33] or OAI[34]. Architectural issues will influence at least the administrative metadata already mentioned. Research on formats will refer to presentation and storage format, which is important for migration and emulation of data. It may be expected, that additional metadata are needed to cover the aspect of format sufficiently. Business models, which concern metadata for rights, permission of use, payments and registrations are important as well. Trends and developments in this field should be observed carefully, as they might call for modifications in the respective metadata elements. Metadata are indispensable for the efficient search across multiple collections by supporting interoperability and crosswalks. A prerequisite for such services are reliable registries. Thorough application and use of metadata supports the documentation and maintenance of interrelationships within repositories. With descriptive and technical metadata for example, digitisation processes can be traced. For the dissemination of digitised contents, metadata are a major tool for resource discovery. They allow the documentation of multiple versions of digital objects, be it updated versions, different formats or translations. Metadata not only describe these versions but allow connections and links between the respective objects. Rights and reproduction information are stored safely in metadata. Ambitious applications in the Web such as the large-scale preservation of the Cultural Heritage or developments of the Semantic Web would be unthinkable without metadata.

References

1. Baker, Thomas: A Grammar of Dublin Core. D-Lib Magazine October 2000. Originating URL: http://dlib.org/dlib/october00/baker/10baker.html
2. Heery, Rachel, Patel, Manjula: Application profiles: mixing and matching metadata schemas. Ariadne Issue 25, 24-Sep-2000. Originating URL: http://www.ariadne.ac.uk/issue25/app-profiles/intro.html

[32] http://www.loc.gov/z3950/agency/

[33] see e.g. http://www.openldap.org/

[34] http://www.openarchives.org/

LIMES – An Infrastructure for the Benefit of Mathematicians in the Information Society

Olaf Ninnemann

Zentralblatt MATH, FIZ Karlsruhe

Abstract. Zentralblatt MATH is the most comprehensive and traditional database providing information on publications in mathematics. With a variety of user-friendly facilities, which are associated with the search menu of the web version, it provides a large infrastructure for research and education in mathematics and its applications. Combined access to Zentralblatt MATH and related services is given through the European Mathematical Information Service EMIS, which is a co-operation between several partners worldwide under the umbrella of the European Mathematical Society EMS. Supporting the installation of a more European platform for Zentralblatt MATH the project LIMES deals with the enhancement of Zentralblatt MATH to a European database in mathematics and with a the prototype for a framework to distribute the editorial work to a European network. The aim of this article is to describe the offers of Zentralblatt and their integration in EMIS. It will give a status report on LIMES and an idea of future actions for the implementation of the LIMES results in a European network.

Keywords: mathematics portal, literature databases, mathematics on the web, integrated access, education in mathematics, digital archives.

1 Introduction

Databases with information on scientific literature emerged from the printed reviewing services with the rapid development of electronic devices for the publication of papers. Already the printed services were necessary to get an overview on the increasing production of scientific publications. But with the extended search facilities and the development of the web, providing convenient access by a variety of possibilities for links, these databases became even more important as large infrastructures for research and education in science.

In order to support the role of the reviewing service Zentralblatt MATH as a large infrastructure the project LIMES has been installed. It is a project, which is guided by the EMS and funded within the Fifth Framework Programme (FP5) of the European Community. The acronym LIMES stands for "Large Infrastructure in Mathematics – Enhanced Services". The project fits into the horizontal programme "Improving human research potential and the socio-economic knowledge base" of the FP5, providing and improving access to research infrastructures. The duration of the project will be from April 2000 to March 2004.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 122–131, 2003.

The partners of the LIMES-project are: FIZ – Fachinformationszentrum Karlsruhe (Germany), UJF – Cellule de Coordination Documentaire Nationale pour les Mathématiques (France), Eidetica (The Netherlands), SIBA – University of Lecce (Italy), DTV – Technical Knowledge Centre & Library of Denmark (Denmark), USDC – University of Santiago de Compostela (Spain), HMS – Hellenic Mathematical Society (Greece), TUB – Technical University of Berlin (Germany), and as the supervising society the EMS. The project co-ordinator is FIZ. The contracts for the funding of the projects have been signed in March 2000. The works started in time at April 1, 2000.

The reviewing service Zentralblatt MATH had been founded in 1931 and provided the first reference database for mathematics, starting in 1978 as part of a network called STN. When mathematicians became more and more attracted by the Internet for their mutual communication and information, Zentralblatt also installed a web service. This service has been enhanced permanently by links to other electronic offers and facilities to use it for integrating information from Zentralblatt in primary publications. In principle mathematicians and other researchers can access from their desktop everything needed from mathematics for their daily work. To improve the access facilities even classical mathematical literature, which had been printed only initially and which remains of permanent interest in mathematics, has been digitised gradually through projects like ERAM (see [1] or [6]) and other digitisation projects like JSTOR and NUMDAM and made accessible on-line.

As an extension of these offers, a lot of additional documents of interest for mathematicians appear in the web already like pre-prints, educational material in mathematics, software, graphical material and others. Information databases provide a qualified access to some of these offers, but for the whole set a uniform access facility with a high degree of de-duplication will be highly desirable. As an extension of Zentralblatt MATH the gateway provided by the EULER service via EMIS may provide one tool for this.

2 The Database Zentralblatt MATH

The connection to the database of Zentralblatt MATH via EMIS (see the URL http://www.emis.de/ZMATH) is the result of the increasing involvement of the European Mathematical Society EMS in the edition of this reviewing service. In contrast to the variety of "databases" offered by commercial publishers now, the word "reviewing" is taken quite seriously and it is not "abstracting" only. Qualified and easily accessible evaluated information are the main aspects of Zentralblatt. The precision of the hit lists has to be considerably higher than what can be obtained by the common search engines in the web.

A lot of reviewers are employed for the reports on the mathematical literature. All reviewers work for Zentralblatt MATH more or less as volunteers. There are additional important aspects of quality which should be mentioned only briefly: comprehensiveness, precision of data, competent indexing, easy search menus, convenient linking with full texts and document delivery systems etc. All these aspects are important features for a reliable service, where quick and dirty offers cannot compete with, – though some of them pretend to have this quality, when their promotion is considered. Mathematics as a science, where the truths detected by mathematicians do not loose their validity with time passing by, needs such a precision to maintain control over the achievements in the past and to get a reliable information on what is new.

The title Zentralblatt MATH stands for several offers: the conventional printed reviewing service as well as for the two electronic offers, one on CD-ROM for off-line use and one as the database with online WWW-access. EMIS provides one link to the WWW-access, but for the world-wide distribution additional access has been arranged through mirrors of MATH in Strasbourg (IRMA), Ithaca (Cornell University), Berkeley (MSRI), Mexico City (CINVESTAV), Rio de Janeiro (IMPA), Lecce (SIBA), Moscow (RFBR), Athens (HMS), Santiago de Compostela (USC/CESGA), Edmonton (University of Alberta) and Beijing (Tsinghua University). To serve for different mathematical communities multilingual search interfaces were installed by the partners and the LIMES team. For example, starting with the ICM 2002 in Beijing an interface in Chinese was available.

As is common with other services, the full use of the database only will be possible for subscribers. But it should be mentioned that there is a free component for using Zentralblatt MATH: Anybody can do searches in MATH. But non-subscribers only will get information on at most three hits from their list. These hits will be taken from the top of the list, where the most recent documents are listed. Hence, if the search leads to a small hit list only, then useful information may be obtained by everybody having access to the Internet. In particular as a new feature a quick search facility is offered as a reference checker, which enables quick and precise referencing to everybody and can be used to provide electronic documents with a freely available permanent document identifier.

Some quick details on Zentralblatt MATH: It provides enhanced bibliographic information on all mathematical literature, starting with 1931. It has been founded by Springer-Verlag (Berlin) as an alternative to the Jahrbuch über die Fortschritte der Mathematik (see below), which had a lot of delays with their reports at that time. Initially it had been edited by the Prussian Academy of Sciences in Berlin, nowadays it is a joint enterprise of Springer-Verlag as the publisher and EMS, FIZ Karlsruhe and the Heidelberg Academy of Sciences as editors. The total number of documents for which information is stored in MATH is more than 1.900.000. This increases by almost 80.000 documents annually at present. The update of the database is made every two weeks, which corresponds to the production of the print version of Zentralblatt. The scope of publications handled by this service includes in addition to mathematics its applications in physics, mechanics, computer science, economics, statistics, biology and other sciences.

Searches can be made using the following list of fields: authors, titles, classification (MSC 2000), basic index, source, language, year of publication etc. A search can be formulated in logical combinations of these terms. For the search a graphic menu is available. The information is given in the AMSTEX source code, but several choices for a convenient formula display are available. Download of the hit list at the users site is possible. Links to the full text of the corresponding article have been installed, if this is available electronically. Other options to get the full text of such an article consist of central document delivery services. Buttons to connect to such services and to see, if the corresponding article is available there, are installed in the menu for the search. Document delivery can be arranged by these services electronically or by sending copies by ordinary mail at reasonable rates.

3 Related Databases in EMIS

A variety of services related to Zentralblatt MATH is offered by EMIS. As a companion database for education in mathematics MATHDI should be mentioned first. It is the online version of the printed service of Zentralblatt für Didaktik der Mathematik. All aspects of quality and search facilities are the same like for Zentralblatt MATH. The editors are the European Mathematical Society and FIZ Karlsruhe. The Editorial Committee is in charge of the supervision. The contents of the database comprise about 95.000 documents. MATHDI can rely on a good co-operation with scientists in several European countries.

In contrast to these two charged databases, the pre-print index MPRESS is provided as a freely accessible service. It stores combined information on mathematics pre-prints available in the web. Pre-prints are an important facility for mathematicians to allow quick posting and communication of their results, in particular if an electronic version is made accessible on a server. Peer reviewing and the publication of the final revised version of a paper are lengthy procedures in mathematics. The gathering of information for MPRESS is done by robots, which are run by national brokers for the harvesting of data. This procedure leads to a data structure, which only allows for simple search facilities.

As a new item the link to a free offer of a collection of high-quality geometric models and animations had been arranged. Each model is accompanied by a detailed explanation of its special features and its role in mathematics. Appropriate software enables the user to get different views and special insights and to interact with this structure. Hence each of the models could be considered as a publication of its own which has no counterpart in the printed media. It is a completely new way to confront the user with mathematical achievements. The current installation is a preliminary version, and it has to be investigated, how the data of these models could be stored in a convenient way, to make them accessible within the same menu as it is provided for searching mathematical articles. This collection demonstrates in a very convincing way the new possibilities emerging from electronic publishing.

4 Related Projects for Accessing Literature in Mathematics and Applications

There are two projects with homepages hosted by EMIS, which are designed to provide services related to Zentralblatt MATH. One of them called EULER had been funded by the European Union and is established as a service supported by a still growing consortium. The other one is the Jahrbuch-Project ERAM funded by Deutsche Forschungsgemeinschaft.

The goal of the EULER project was to develop a prototype of an Internet service, which integrates some of the most relevant publication-related electronic resources in the field of mathematics. This should provide a “one-stop-shopping” in heterogeneous resources for the user like OPACs, databases, pre-print servers, electronic journals and other WWW-catalogues. Tools have been developed facilitating the production of DC-metadata for the different resources. Using the Z39.50 protocol the so-called EULER-engine users of the EULER-system can carry out combined searches for mathematical information in all these resources.

The original EULER-project terminated in September 2000. During the exploitation period a model has been developed how to establish EULER as an Internet service. Within a smaller take-up project the transition of the prototype to a permanent service had been arranged successfully until end of 2002. The main goal is to establish a library and society based web service, which includes a system of resources with a well-distributed European background. EULER stands for “EU”ropean “L”ibraries and “E”lectronic “R”esources in Mathematical Sciences. More details on the project can be found in [5] and [7].

The aim of the Jahrbuch-Project, which officially is called ERAM (Electronic Research Archive in Mathematics), is to capture the Jahrbuch über die Fortschritte der Mathematik as a classical bibliographic service in mathematics in a database and to use this activity to select important publications from the Jahrbuch period for digitisation and storage in a digital archive. Longevity is typical for research achievements in mathematics. Hence to improve the availability of the classical publications in mathematics and to enable to get quick information on these, electronic literature information services and digital archives of the complete texts are needed as important tools for the mathematical research in the future. This is the reason why so many efforts are invested in projects like ERAM.

The Jahrbuch was founded in 1868 as the first systematic and comprehensive documentation service in mathematics, published by de Gruyter (Berlin) and edited by the Prussian Academy of Sciences in Berlin. See [2] for a report on the foundation of the Jahrbuch and the ideas behind its publication during the period at the beginning of the twentieth century. In [4] a short history of the Jahrbuch and Zentralblatt may be found. The publication of the Jahrbuch terminated during the Second World War and never started again. The Jahrbuch-database will not be just a copy of the printed bibliography. It will contain a lot of enhancements.

The digital archive built up in connection with the database will be linked to the database and provide all facilities associated with current digitisation projects. The content will be distributed to mirrors and combined with similar archiving activities in mathematics elsewhere. More detailed reports including also the archiving activities have been given in [6] and [1]. The project is in good progress. All data will have been captured until end of 2003, and more than 30% have been enhanced already with the help of volunteers. The integration of the Jahrbuch data into the Zentralblatt database will happen in the first half of 2003, because the captured data already have reached the year of the foundation of Zentralblatt.

A final remark should be made on the content of the digital archive provided by ERAM. The capacity, which could be covered using the funding, will be 1.2 million pages. Almost a million are in the archive already and accessible through Zentralblatt MATH and the Jahrbuch at the GDZ in Göttingen. Direct access at the GDZ also is possible. Among the publications handled are famous journals like Mathematische Annalen, Mathematische Zeitschrift, Inventiones Mathematicae or the Commentarii Mathematici Helvetici.

5 The Objectives of LIMES

The general objective of LIMES is to upgrade the database Zentralblatt MATH into a European based world-class database for mathematics and its applications by a process of technical improvements and wide Europeanisation. Upgrading the existing database, improving the present system and developing a new, distributed system both for the input and output of the data are necessary to allow Zentralblatt MATH to use the latest developments and to anticipate on future developments of electronic technologies. This will make Zentralblatt MATH a world reference database, offering full coverage of the mathematics literature world-wide, including bibliographic data, reviews and/or abstracts, indexing, classification, excellent search facilities and links to offers of articles, with a European basis.

Improvements will be made in three areas:

1. Improvement of content and retrieval facilities through sophisticated further development of the current data sets and retrieval programs.
2. Broader and improved access to the database via national access nodes and new data distribution methods; in particular, improvement of access for isolated universities in regions with economic difficulties and in associated states of Central and Eastern Europe, where a mathematical tradition of excellence is under economic threat; stimulation of usage for all kinds of research as well as for funding organisations before decision making by initial support from two national test sides.
3. Improved coverage and evaluation of research literature via nationally distributed editorial units, development of technologies for efficient database production (exemplified by two further European member states).

The LIMES project sets out to achieve these goals by stimulation of usage through national access nodes and centres for dissemination including development of adequate licensing models. Links to offers of complete documents will be an important addition. Better coverage and more precise evaluation of literature through improved update procedures (for the data-gathering, but also for data distribution to access points) via a network of distributed Editorial Units will increase the quality of the database. Networked access to the database in these days has to be arranged mainly via rapidly evolving WWW technologies, with high potential for further development of innovative access and retrieval methods, and improvements in the quality of data (e.g. author and source identification through data-mining techniques).

Contrary to the centralised American model, Zentralblatt MATH will be extended to a European research infrastructure with distributed sites (Editorial Offices and Access Nodes) in the member states. The EMS representing the scientific community has promoted these ideas among the national societies and individual members.

Since EMS is member of Zentralblatt MATH's Editorial Board, it is guaranteed that the results of the project will be available for usage in further research infrastructure partnerships. Tools and methods developed during the project will be usable by project partners on a royalty free license basis. Further results, documentation and information will be made publicly available. Commercial exploitation of results beyond European research infrastructure services will be considered.

6 The Role of the LIMES Participants

The work for LIMES is organised in several work-packages. The participants have been chosen such that each package is taken care of by one or several participants according to their special expertise.

FIZ operates the central editorial office for Zentralblatt MATH (http://www.zblmath.fiz-karlsruhe.de/), which has been described above in detail. As a consequence of this role FIZ is engaged in all work packages of the project.

As a French service the Cellule MathDoc within UJF provides mathematics research libraries and mathematics departments with technical assistance on computerised documentation. In 1999, the UJF released edbm/w3, the first module of the European Database Manager for Mathematics software. In LIMES UJF is in charge of the extensive development of this software like integration of new tools for search, identification, display, and update, the development and integration of improved access control management, and it helps national access nodes with the development and installation of improved update procedures.

Eidetica is a spin-off company of CWI and had been founded shortly before LIMES started. It was created by researchers in the Interactive Information Engineering theme, combining expertise in linguistics and mathematical clustering of dependency networks for textual domains. Eidetica will mainly develop new models and tools for identification of author and serial names (two of the most requested features from the user community), based on data-mining techniques, and other improvements of the retrieval process.

DTV was established in 1942 as the central library of the Technical University of Denmark (DTU) and the national library for engineering and applied sciences in Denmark. The role of DTV is twofold: 1. Set up and run an editorial unit. The unit is responsible for covering a number of journals and other sources, which had been defined and agreed by DTU/MATH/DTV and the co-ordinator. The unit engages in experimenting with modern, electronic reviewing methods. 2. Establishing a raw data system for the editorial units. Following the agreement of the publishers, this data will be made available for LIMES in two ways: a) for end-user current awareness purposes through a special LIMES gateway added to the Current Awareness System of the EMS and to the Zentralblatt MATH database, b) for cataloguing reuse by the editorial units through the export of raw cataloguing data to the Zentralblatt editorial system.

SIBA is the second partner to care about a streamlined system for the electronic handling of the reviewing workflow. They are supposed to design this in tight cooperation with DTV and the editorial office of Zentralblatt in Berlin. Nevertheless, having alternatives and additions to the solution developed by DTV available, will be useful and provide the partners in the extended network with the option of a choice according to their special situation. SIBA currently handles more journals than DTV, and most of them are available in printed form only. In contrast to DTV they are confronted with acquisition problems, which are characteristic for some of the Italian journals in mathematics.

The role of USDC is to propagate the use of the database in Spain and to facilitate the access to it. They have installed a national site server and care about regularly updating the database and new software releases. They have set up a task group for stimulating the usage of the database by Spanish universities, research centres, and libraries. A network structure for information in mathematics is in development, integrating Zentralblatt MATH with national information offers. This work is undertaken in co-operation mainly with CESGA (Centro de Supercomputación de Galicia) and RSME (Real Sociedad Matemática Española).

HMS, founded in 1918, has the purpose to encourage and contribute to the study and research of mathematics and its applications, as well as to improve mathematical education in Greece. HMS applies a methodology similar to that of the partner USDC in a different environment. They are quite successful in establishing a Greek access consortium. That the mathematical community in Greece is very well organised is rather helpful for this task.

TUB is one of the largest universities in Germany offering curricula for the full scale of engineering sciences. Through the members of its mathematics department, it has a traditional co-operation with Zentralblatt MATH by serving as consultants for the input to the database. In particular, the editor-in-chief of Zentralblatt MATH is member of that department, and he can look back on a 28 years period of experience in this position. TUB will co-operates all partners providing them with continuous advice and caring about the evaluation of results. An important task for TUB is the development of editorial tools.

EMS is a partner in the Editorial Board of Zentralblatt MATH. It has set up a committee whose main task is to define and enact innovative features and procedures in Zentralblatt MATH. The EMS plays two roles in the project: It represents the end-user and gives a pan-European framework to the project as a whole. In both roles, it is part of the Project Management.

7 Some First Achievements

The co-operation of the editorial centre of Zentralblatt in Berlin with their external partners suffers from the restrictions of the current input and administration procedures. These had been designed for a totally centralized workflow. Hence for the work to be done by the external editorial partners some bottlenecks appear, where collaborators from Berlin had to be involved in the workflow permanently.

One of these bottlenecks was the reviewer database providing all information needed to assign papers to them. Access from outside was impossible. This system has been redesigned and installed on a platform, which allows for controlled access from outside now. Clearly, such a database cannot be made freely accessible, but the authorized partners can check the interests and the workload for every reviewer, they can register papers to be assigned to the reviewer, and they can make changes in the reviewer's data in accordance with the centre in Berlin.

A long-standing problem was to increase the flexibility of the input system for the database to be able to handle external input and to insert electronic submissions more systematically. The current system has a highly sophisticated structure with the aim to serve for a lot of different purposes. The input for the mathematics database is one of them, and not of the highest priority. Hence LIMES could not rely on a modification of the current system and had to design a new one, specifically related to Zentralblatt and suitable for distributed users. A prototype for the new input system is available now and ready for being tested in a first release.

The partner UJF has implemented several improvements of the search software. In addition to this they successfully complied with a request, which was not visible at the beginning of LIMES. Links from references of articles to their review in a reference database become a more and more important facility in electronic publications. UJF has developed a look up tool, which will help to produce these links automatically based on the electronic references of an article. The tool has a surprisingly high precision. Also some first tools have been developed allowing an author identification of relatively good reliability. To do the same for journals will be comparatively easy.

The models for the electronic reviewing flow are in good progress. We are just at the borderline to provide an external partner with a copy of the systems, asking him to serve as a test-bed for this structure. The progress at the access nodes had been mentioned above already. Hence looking at the current achievements, the LIMES partners are convinced that they can reach all proposed goals within the project period, which will end with April 2004.

References

1. Becker, Hans, Wegner, Bernd: ERAM – Digitisation of Classical Mathematical Publications. Proc. ECDL 2000, Lecture Notes in Computer Science **1923** (2000) 424–427
2. Lampe, Emil: Das Jahrbuch über die Fortschritte der Mathematik. Rückblick und Ausblick. Atti Congr. Int. di scienze storiche, Rom Vol. XII (1903) 97–104; digital version under http://www.emis.de/projects/JFM/
3. Virgos, Enrique Macias, Wegner, Bernd: Zentralblatt MATH, una gran infraestructura a disposicion de los matematicos espanyoles. La Gaceta de la Real Sociedad Matematica Espanyola Vol. 2, no. 3 (1999) 417–420
4. Wegner, Bernd: Berlin as a Center for Organizing Mathematical Reviewing. http://www.zblmath.fiz-karlsruhe.de/zbl/berlin.pdf
5. Wegner, Bernd: EULER – a DC-based Integrated Access to Library Catalogues and Other Mathematics Information in the Web. Seventh International Conference "Crimea 2000" Libraries and Associations in the Transient World: New Technologies and New Forms of Co-operation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 3–11, 2000, Volume 1: 264–267
6. 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 Co-operation. Conference Proceedings. Sudak, Autonomous Republic of Crimea, Ukraine, June 3–11, 2000, Volume 1: 268–272
7. Wegner, Bernd: EULER – a DC-based Integrated Access to Library Catalogues and Other Mathematics Information in the Web, ECDL 2000. Lecture Notes in Computer Science **1923** (2000) 461–466

Peer Refereeing . . . Will It Be Missed?

Alfred J. van der Poorten

ceNTRe for Number Theory Research
1 Bimbil Place, Killara NSW 2071, Australia
alf@math.mq.edu.au (Alf van der Poorten)

Generally, a referee has to make one of the following recommendations:

1. Publish essentially as is; the only changes necessary are very simple typographical matters which can be changed by the editor.
2. Publish after author's minor revision; the referee suggests points which must be changed before the paper meets the standards for publication.
3. Publish only if the author makes major revisions. (Perhaps the paper is much too long or is badly written. The revised paper will be refereed again.)
4. Reject. (There is nothing salvageable.)

These 'Hints for Referees' [1, p.36] are all very well, but one might wonder just how to distinguish. So Don Knuth reminds referees that to be publishable:

a. The paper should contribute to the state of the art and/or should be a good expository paper. If it is purely expository it should be clearly designated as such.
b. All technical material must be accurate. A referee should check this carefully.
c. The article must be understandable, readable, and written in good English style.
d. The bibliography should be adequate.

1 An Opening Discussion

The issue is that we have somehow to cope with a world in which it is easy and cheap to produce and distribute one's manuscripts. Thus 'publication' is yet more uncontrolled than it was traditionally. In this context, it seems absurd to insist that a paper has not been *published* unless it has suffered the ministration of a peer referee; and this no matter how readily the paper is available nor how elegantly it has been formatted.

Most of us have been brought up to be polite at all times. That upbringing forces us to be yes persons, always acknowledging that our interlocutors might well be in the right. Below I introduce a no person, wantonly disagreeing with common understandings.

Y: Without peer refereeing papers would have errors.
N: Many published papers have errors. Mind you, I don't blame the referees. The author is primarily guilty of those mistakes. In any case, most errors are

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 132–143, 2003.

only relatively minor sloppiness (though occasionally a trivial gap in a proof turns out to be a bottomless chasm).

Y: But if serious papers might have errors how can one rely on the literature?
N: If one uses a published result, let's call that hypothetical result 'ERH', without having the vaguest notion of the principles underlying its proof, one is in effect writing: "*Given* ERH, my claim follows by the following argument." It would make no practical difference to you, as author, if ERH were no more than a conjecture. Besides, the more important a result the more likely the paper has actually been read carefully, for example by an assiduous graduate student or post-doc.

Y: Without peer refereeing many old results would be republished.
N: Many old results *are* republished (and not infrequently the 'new' proof is less elegant than the old ones. Happily, an old result is sometimes published with a new argument that actually explains the result, rather than only proving it).

Y: OK. So refereeing as we know it isn't perfect. That barely matters. Referees mostly work for free. Refereeing doesn't cost anything and sometimes does some good.
N: No! The refereeing process is a *major cost* in the traditional publication process. The process is hugely time consuming both for editors, in finding and corresponding with suitable referees, and in the delays traditional in mathematics in obtaining appropriate reports.

Moreover, this cost — viewed as a cost per published article – is compounded by not only published articles incurring a refereeing cost but also those eventually rejected.

Sure, the traditional process appears to assume that the efforts of referees are of no cost to the mathematical community; that assumption presumes our time has no value (whether for ourselves or to our employers).

Y: But refereeing is vital! Publication in refereed publications is essential for career advancement and recognition.
N: One wonders whether the time might be ripe for a new approach in which career advancement relies on the quality of one's work and on the judgment of peers who have actually read and used it.

Y: But, but, ... You can't just let everything be published!
N: You cannot *stop* anything from being published. Have a look at the web. Besides, the problem (if it is indeed a problem) is much older than that. The photocopier thirty years ago made it easy to promulgate one's preprints.

Once people learned to speak they could announce their ideas, with no editorial control. Worse, once people learned to write they could spread their ideas even further.

Summary. The unthinking view of the person in the street or, for that matter, of the mathematician in the corridor, is that the purpose of the referee is only to certify the correctness and originality of the article's results.

However, no referee (unless, perhaps, assisted by able and energetic graduate students) can possibly guarantee that submissions are error free or new. In practice, it's commonplace for old results to be rehashed — and a good thing too. Well known facts only become known well by repetition; and, in any case, occasionally a new proof of an old fact actually explains it properly. Republishing a result may add to "the state of the art".

Correctness is primarily the province of authors, not of their referees. As referees we are the authors' victims. We are not guilty of authors' crimes.

2 The Exchange Continued

Y: It seems that a referee can only be asked to reject blatant nonsense or material that plainly seems unlikely to be true.
N: Certainly not. Normally, referees should *never* be asked to deal with 'blatant nonsense' or material that is fairly plainly inadequate. That's the editor's task.

Y: But it's the right of an author to have his submitted article refereed formally. It would be improper for an editor to make a unilateral decision.
N: Nonsense. An editor should deal politely with authors ("I regret to have to advise you that your submitted article is not suitable for publication in *My Journal*."); but that's only so as to follow Aristotle's advice to be courteous to the weak. Beyond offering courtesy, anything else done for an author making a plainly inappropriate submission is purely gratuitous.

On the contrary, moreover, I insist that it is improper on an editor's part, and is a gross discourtesy to his referees, to ask them to waste their time formally advising rejection of material that was never, ever, going to be accepted. It's an obligation on an editor to have made such decisions expeditiously.

Y: So referees have no proper task at all. Why then have them?
N: Why, indeed? I'm tempted by your momentary conclusion, but the fact is that referees do play a critical role in the ritual of transforming a manuscript into a refereed publication, namely by making it possible to say of them that the paper has been refereed. Providing an imprimatur ('blessing' the paper) is an important task.

Y: But that's absurd!
N: The value and importance of tradition and ritual should not be underrated.

Y: But surely referees have more than ritual purpose.
N: Of course. Referees perform the invaluable task of advising on whether a conceivably appropriate article is in fact suitable for publication in the journal to which it is submitted.

In the nature of things, a positive recommendation includes the referee implicitly certifying that she has no reason to believe that the major allegations of the paper are unlikely to true and more particularly that she has not noticed any (significant) errors.

Summary. Editors are responsible for more than just the final decision to accept a suitable paper. They must also recognise promptly that certain submissions plainly are unsuitable for their journal and dismiss such papers immediately.

Refereeing does provide an *imprimatur*. More important, though, refereeing is a mechanism for selecting *preferred* papers from potentially suitable papers. Refereeing moderates the *quality* of papers accepted by a given journal.

3 A Digression: On Recognising Blatant Nonsense

When I get a letter purporting to make some great contribution to mathematics I first test its claims by applying several principles.

The first is *the principle of the meatgrinder*. In brief, on my complaining to Kurt Mahler that hard work and apparent ingenuity was making no impact on a problem he had set me, Mahler responded: “Ach, Alf. If you want to get gehacktes Rindfleisch out of a meatgrinder then you must put some *steak* into the meatgrinder.”

The principle of the meatgrinder points out that there’s no gain from busily turning the handle, no matter how energetically or painfully. To get chopped steak out of a meatgrinder you have to put some quality meat into it.

There’s a second important principle for which I don’t as yet have a succinct catchy title. But it it boils down to this:

Mathematics ain’t about getting it right; it’s about not getting it wrong.

My thinking is this: There’s no particular merit in getting an answer to a problem, even a correct answer. After all, applying some algorithm is better done by a machine, and in any case is no more than an organised technique for guessing an answer. An easier way to guess — and therefore a better way — is to look at the answers in the back of the book, or to lean across and look at what the person next to you has written (on the presumption she is either smarter or more energetic or both than you), or to phone a friend, or . . . . Contrary to what we tend to teach, finding an answer isn’t mathematics at all; it’s just guesswork, intelligent guesswork, maybe.

Mathematics begins when one checks and verifies that one’s guess is indeed not necessarily wrong. So it’s a very good idea to illustrate that one’s argument fails when it should, that is, when it has no business in working. And one is doing meaningfully higher mathematics when one dissects one’s argument — not the turgid computations and details, but the underlying logic — into immediately digestible pieces.

What’s the point of this outburst? It’s often quite difficult to prove that an argument is correct. Thus the principal obligation on a mathematician is to struggle to prove that her argument is false. If, and only if, that struggle is properly pursued diligently, but fails, does she have any business purporting that her argument might possibly be right.

Summary. There is no obligation on us as referee or teacher to convince an author (or student) that he is surely in error. It's up to the author to compel us to the view that her claims are not necessarily wrong and may therefore well be correct.

4 Publish or Perish

Y: If a result is new and correct then it warrants publication.
N: Rubbish! Many a new result fills a much-needed gap.

Y: Ho, ho. But smart-alec retorts aside, you'll agree that it would be a loss to mathematics if new results had no avenue of publication.
N: I don't agree. A majority of 'new' results are no more than exercises capable of being carried out by anyone mildly familiar with the results being employed. Not infrequently, they are only cheap corollaries of a ground breaking result. Sometimes they involve considerable and, often, wasted effort in tracking the implications of someone else's slight improvement in understanding; they do not add to our understanding. Worse, many new results are part of a 'cottage industry': a sequence of published minor improvements in understanding forced upon us by a group of authors insufficiently disciplined to begin to digest their ideas before regurgitating them onto the public.

It's not a bad idea, mind you, to write up one's intermediate results and ideas; at the least for one's own use. But that's no excuse for paining an innocent editor or for victimising some referee.

But it's beside the point whether the response above is excessively cynical. It's just not true that a result must be published in a refereed journal lest it be lost forever. Nowadays, for example, one might deposit one's draft in the arXiv.

Y: What's the arXiv?
N: The help files at `http://arxiv.org/help/` answer your first question and the address `http://front.math.ucdavis.edu/` answers the follow-up question you should have asked; namely, "How do I use the arXiv?"

Y: But if people don't publish they will perish!
N: Quite, too true. It's high time that there was some uncoupling of research recognition from formal refereed publication.

Most of us dismiss hope of substantial change in our environment, mumbling sadly to ourselves that "they'll never do it". It might help us to remember that they are us.

Decoupling. It's not obvious just what it might entail to *decouple* recognition and refereed publication.

An extreme interpretation might require one to present one's opus all in preprint or arXiv format, with no explicit hint of where or whether its parts have been formally published. [Sure, that's absurd. But is it any the more silly than the fad of pretending that omitting authors' names from papers somehow leads to more accurate refereeing?]

In contrast to decoupling, imagine a national university funding system[†] which rewards universities for research outputs according to papers published in recognised peer reviewed journals, or in refereed conference proceedings. In such an environment, cottage industries flourish and the notion that refereed publication is at a premium acquires a new meaning.

However, the matter at issue is the coupling of refereed journal publication with promotion and related personal recognition. At the least, decoupling entails giving serious recognition to all work of the candidate rather than confining attention to journal articles. In brief, one argues that the purpose of scholarly publication is scholarly communication. It need not matter whether the item has been refereed. The questions should be whether the item is indeed accessible, and whether peers have chosen to use it. Next one might attempt to judge its quality. The hurdle presented by the refereeing criterion may well be lower. It purports little more than that some soul has looked at the paper and has not found it noticeably weaker than other articles appearing in the journal.

One real issue is the exaggerated and cynical belief that recognition is not at all a function of the quality of one's articles but depends only on their number, or their height and weight. Decoupling will already have occurred if the assessor actually is required to have looked at the candidate's opus.

Summary. One must of course make one's work available to the mathematical community, so that the community may judge it and then use it. It seems reasonable to believe that publication in a generally accessible well-mirrored archive can do that adequately.

5 Some Scandals of Our Present System

Y: Luckily, it's plain that mathematics referees care. They put months of effort into their reports.
N: Goodness me. You did attend on the day the teacher explained irony!

Y: Huh?

I opened with some remarks from [1]. Let me add another, with which I largely agree.

> It is tempting to postpone refereeing tasks by putting the paper aside for a few days. But it takes no longer to do it today than it will in a weeks time.

[†] Amazingly, such a system exists in real life; for details and definitions see `http://www.ro.mq.edu.au/OPUS/guidelines/2.htm`.

It is rarely true that putting a paper aside to allow it to age gracefully improves the quality of one's referee's report.

Of course, I don't do my refereeing within a few days. I find it hard to shake off the ugly tradition in which I was brought up.

Let me mention some other ugly traditions. It is an absurd scandal that respectable mathematics journals can stomach as standard practice a delay time of two years or more between receipt of a manuscript and its publication. One reason for those delays is extensive backlog — over a period the journal has accepted more papers than it is able to publish. But that's a consequence of doubtful refereeing and of editors failing to have the wit and courage to refine their acceptance criteria in the light of the volume and quality of manuscript submitted.

Y: Happily, electronic publication allows journals to catch up on their backlog. With print and postage costs eliminated there need be no restriction on the amount a journal publishes annually. The journal can set its standards and can freely publish all submissions that attain that those standards.

N: Unhappily, not quite. Print and postage expenses are different from other costs only in that they vary according to the number of subscribers and the weight of the publication. Sadly, other costs depend variously on the number of submissions, or the number of papers accepted, or the number of pages that must be edited. Most mathematics publications have so few subscribers that these 'administrative costs' predominate whether or not there is print publication.

Some think that these are 'costs' that need not cost. Administrative tasks can be performed by volunteers. Say to them: "There ain't nothing that comes for free, except perhaps sloppy thinking." Point out that you value your time and that when you volunteer your efforts there is a real opportunity cost, even if it is only the opportunity to give your full attention to the test match rather than just to watch it out of the corner of your eye. In any case, even volunteered time has an upper bound; it is not arbitrarily extensible.

However, now that there is the opportunity to publish without blatantly visible costs it is possible to initiate new 'free' journals, at first totally supported by the enthusiasm of volunteers and employers prepared to support the project. Once existing, such ventures may be maintained by that and the sort of grant support that established projects can attract.

Y: Well, at least electronic communication now speeds up the refereeing process; and that applies also to the traditional journals.

N: Not always 'also'; indeed, in my experience, rarely. One difficulty is that many traditional journals do not conduct their correspondence by e-mail; nor do they ask for electronic submission typed in a suitable flavour of TEX.

You'll hear bleating that so to insist would be unfair to mathematicians with no TEXpertise. And what about mathematicians without access to a computer? Tell those complainants that it's no more unfair than the demands we faced until twenty years ago requiring us to type our manuscripts and then to buy a green pencil with which to insert `\mathfrak` symbols.

In any case, the point is the absence of sensible *default* policies. Just as rules are there to be broken, policies exist so they may be adjusted when appropriate. A policy should not deal with out of the ordinary circumstances, other than to have a rider agreeing to deal with special cases in out of the ordinary ways.

It's ludicrous for an editor not to ask, for preference, that referees respond by e-mail. It's scandalous to be satisfied that a referee agree within a month to post a card promising, eventually, to referee a paper. It's been silly, ever since the universal existence of the photocopier, let alone in the case of manuscripts plainly produced by a computer, to post back copies of a rejected manuscript. It is poor policy or sheer rudeness not to advise a referee of the decision eventually taken by the editor, for preference by sending the referee a copy of the advisory e-mail sent to the author.

Summary. Slow refereeing and severe backlogs in publication should not be accepted as normal. Many journals have editorial policies apparently reformulated in 1953 and unchanged since.

6 Alternative Models for Refereeing

Y: How about making better use of *Mathematical Reviews* and the *Zentralblatt*? If people 'publish' on their personal web page or on the arXiv, then the reviewers can take the place of our present referees. All that needs is for their rewiews also to include a critical component.
N: Manuscript archives, of which the arXiv is mentioned as example, are more than just an efficient way of making one's preprints available. The arXiv fixes versions of papers and provides a well-mirrored searchable source for preprints. Placing a manuscript on one's web page is too impermanent and ill-defined an act to warrant being called 'publication'; so let's deal just with papers placed on manuscript archives of the quality of the arXiv.

First, it is a noticeably greater demand on reviewers to ask that they do more than provide a helpful abstract. Finding reviewers cannot be all that easy. *Math. Reviews* has found it useful (and may have found it necessary) to pay per review by giving its reviewers a discount certificate on AMS books. *Zentralblatt* offers reviewers the author discount on Springer books (and rewarded on a per review basis when it sent an international postage coupon with each review). With more demand made on them, it might well become too hard to attract suitable reviewers/referees.

A second problem is more subtle. It's not so much that the review journals review all research mathematics. Rather, in choosing to review an article the review journals *define* it to be mathematics. Naturally, publication in a well recognised peer refereed mathematics research journal is sufficient to qualify for review. However, a substantial part of reviewed research mathematics appears in learned journals that are not primarily mathematical. It's a nontrivial task to identify those papers, just as it may be difficult to decide whether, say, a paper or book has a meaningful research expository component or is primarily a contribution to mathematical education. Then there's the question of vanity publication . . . .

Not surprisingly, those involved with the review journals are not all that mad keen to begin to think about reviewing mathematics that has not been published nor moderated in the traditional way.

Nonetheless, it does seem an attractive proposition to make broader use of the review journals by way of enhanced MR/Zentralblatt review. Mind you, critical reviewing already occurs implicitly. Compare such positive phrases as: 'This fine paper . . . ' 'These instructive remarks . . . ' with a remark such as: 'The author gives yet another . . . '; or with a review simply consisting of a few words from the abstract. Of course, the quality of reviews varies more with the quality of the reviewer, or the match of reviewer and paper, than according to the quality of the paper reviewed.

Y: Second, there's *refereeing by added comment.* For example, Amazon.com invites customers to critique books and to give them a rating of zero to five stars.

N: But don't average the number of stars awarded! A paper with one serious five star review and one zero star review plainly written by an idiot should not be thought of as a two and a half star paper.

It will rarely be a problem to decide which of the two reviews should be dismissed. In the crunch one might have to appeal to the Botvinnik principle[‡].

It does seem feasible to ask readers of papers on the arXiv to submit reports. One trouble is that there is little tradition of *open* refereeing in mathematics. That need not stop readers from pointing to rare actual errors, but it may inhibit critical comment on whether the paper is 'good' or 'worthwhile', let alone whether it's well written.

Curiously, perhaps, there *is* a long standing tradition of open critical book reviews. Short reviews may be little more than abstracts, but extended reviews sometimes are very instructive. Traditionally, such book reviews first survey the subject and then identify the extent to which the book fits into that survey. Happily, here there is no tradition that it is wrong to be scathing when that is appropriate. Some such reviews are wonderful[§]; ask your friends for their

[‡] On a train journey, grandmaster Botvinnik sees the other passenger in his carriage studying a chess board. The stranger notices Botvinnik's glance of interest and asks him whether he would like a game. Botvinnik begins to say no then shrugs his shoulders and answers that, sure, he'll play; but let's stake a good amount on it, say a hundred roubles.

"A *hundred*!" the passenger cries in horror and astonishment. "But how can you risk such a sum? You don't even know who I am!"

"Exactly", answers Botvinnik.

[§] One of my old favourites includes the line: [The author] seems to use a method of infinite ascent in expounding his proofs, that is, simple ideas are often developed by using more complicated ones.

favourite polemics. Book reviews differ from other refereeing and reviewing tasks also in that the reviewer gets a copy of the book as reward.

Generally, referees are not suitably acknowledged or rewarded. Some change to that is easy. For example, it surely is a simple matter for a journal annually to put on its web site a list of persons who have provided timely referees reports.

The primary purpose of refereeing is to *select* suitable or preferred papers. Nonetheless referees should accept the obligation to assist authors not only in correcting their paper but also in *improving* its organisation, exposition, and results. In extreme cases (*Scientific American*?) staff authors may effectively write the paper. It might be a good thing for editors to notice those cases where a referee has effectively become a co-author and to invite author and referee formally to collaborate.

Summary. Whenever one selects one manuscript but not another, for whatever purpose, a review or moderation process has intervened. Whenever one comments on a manuscript, no matter how incidentally, that remark is a review and it may aid a selection. The notion 'peer refereeing' does need to shed its ritual aspects, particularly if the only purpose of those rituals is to humour tenure and promotion committees.

7 The Future

Y: What do you think the future will bring?
N: It's hard to predict, especially the future.

One guesses that the breakpoint is the ability of libraries to pay for their subscriptions. That ability has been resuscitated by publishers repackaging their offerings and adjusting their prices ingeniously, making different offers according to their view of a library's apparent ability to pay.

But, recall that the mathematical sciences provide only a very small part of the learned publications at issue. Several of my remarks above may not apply beyond mathematics.

While libraries are prepared to pay for it, and while we continue to support it by our preparedness to submit our manuscripts, and continue to offer our contributions as referees and editors, we will continue to have learned journal publication as we know and knew it.

Y: But why do journals still exist? It doesn't make sense for papers first to be published on preprint archives, and then to be republished.
N: It's not that simple.

First, there *is* value added by being allowed to label an article as published in one of the better journals. Second, the editor and referees may have induced actual improvement of the article. Moreover, the journal will likely have warmed up the format and presentation of the manuscript.

Third, learned journals did not all come into existence because publishers and learned societies hoped to make a dollar or two at the expense of university libraries. Many journals are, at any rate in the first instance, a record of research of some society or group, or the minutes of meetings. Consider departmental preprint and technical report series; conference proceedings are a clear case of 'minutes'; note titles such as Séminaire ... , Comptes Rendus de ... , Acta Whateveriensis.

Even though there are now many other opportunities to publish one's work, the motives that once led some groups of mathematicians to display their work are active today and will remain valid tomorrow.

In all this, it's worth noticing that nowadays a 'journal' can be almost virtual, consisting of little more than a `.cls` file. 'Publishing' it need only entail selecting and validating suitable papers, then creating a page of links on the web. Mind you, a 'journal' can be yet more virtual; see Greg Kuperberg's *Open Journal of Mathematics* [2].

The preprint archives also present an extra source for new but traditional publishing opportunities. Could there be a market for honest to goodness books featuring, say, the 'best' papers archived or published over a recent period? In this context also see Jim Pitman's proposal at `http://mathsurvey.org`.

It's well known that lectures by keynote speakers are often published automatically in conference proceedings; now and then, a journal may ask an author to submit a survey article. So, for a book, or a journal, to find its content in the preprint archives is not really a new idea. Mind you, from the quality control point of view it might be an improvement to select articles rather than authors.

Indeed, if journals mined preprint archives an author might get to decide which of several competing journals should get the rights to publish. That'd be a pleasantly positive version of our present quandary as to where to submit our masterwork.

Particularly in the humanities, it is commonplace to find books of readings (consisting of previously published articles linked by wise editorial comment); this is not quite unknown in mathematics but is unusual.

Y: I still miss getting real preprints with nice covers.
N: That's the trouble with home publishing. The output is on A4 paper (or, worse, on quarto if you're a North American) and, no matter how good the LaTeX, or how expensive the fonts, it doesn't look published at all.

Just so, it is a common complaint that it's not easy to browse journals on the web. Searching is much easier, of course, but complainants purport to miss serendipitously noticing the previous article, and the next one.

One expect that all journals will eventually be available electronically, most electronically only (perhaps with printed offprints and an expensive option for print copies for very rich libraries). Publishers might make extra friends if, at least until our generation dies off, they maintained the illusion that their journal is more than just the sum of its articles by including, with each full article, the abstract of the previous and of the next article.

Summary. Journals will remain with us, probably often as no more than virtual versions of their paper and bound originals. Selecting their contents brings up the traditional refereeing issues. There is now, and all the more when almost all journals are electronic there will be renewed opportunities for traditional print books featuring the best of . . . .

Formal peer refereeing differs from other forms of review only in being part of an accepted ritual. It will survive, but only as a diminishing element of informed comment on published work.

References

1. Knuth, Donald E., Larrabee, Tracy, and Roberts, Paul M.: Mathematical Writing. MAA Notes Number **14**, Mathematical Association of America (1989)
2. Kuperberg, Greg: Scholarly mathematical communications at a crossroads. Nieuw Arch. Wisk. (5) **3** (2002), 262–264, arXiv:math.HO/0210144.

Geometry & Topology Publications: A Community Based Publishing Initiative

Colin Rourke and Brian Sanderson

1 Preamble

We are primarily research mathematicians. It's what we do. As research mathematicians we are ideally placed to see the grip that commercial publishers have established on our literature. We'd like to call it the Publisher's Fork in analogy with Morton's Fork.[1] On the one hand, whilst carrying out our research, we need — indeed demand — total access to the literature. If we want to consult a paper, we will do whatever we need to do to find it ... in the old days, we would look in our filing cabinets, then colleagues cabinets, then search libraries or telephone more distant colleagues. These days we are more likely to start at the arXiv or Google before resorting to hard-copy search. The methods may have altered, but the purpose has not. We need to check or use previous work and nothing will stop us until we find it in some form or other. This is the first prong of the fork. The other is that, once we have produced a piece of new research, we want everyone to see it. We don't hide it under a bushel; we want to give it away for all to see. We eagerly talk about it to colleagues and students, freely give lectures on it (in all gory detail if asked) etc. And we do the same for everyone else's research, including anonymous refereeing, editorial advising etc. This is the second prong of the fork.

These two prongs, (a) that we demand access to all other research and (b) that we offer our research and research-related services for free, have given our publishers a huge lever over us. We create a product (published research) give it away free and then buy it back with a totally inelastic demand. Given the current financial ethos under which companies exist only to make money, incidentally providing a service, it is not at all surprising that we are now paying exorbitant fees for access to our research. The usual countervailing mechanism in capitalism — competition — is totally absent in this market because there is *no* competition between journals. Each publishes a different set of papers. The Annals (a well run, academically owned, moderately priced journal) is not in competition with Inventiones (an exorbitantly overpriced Springer journal) although they are of similar standard in a similar subject, because they each carry a different set of papers which we all need to read. We need permanent access to *both* of them.

[1] Cardinal Morton was Lord Chancellor in the reign of King Henry VII. By visiting noblemen of the time he would judge their tax for the coming year. If the hospitality he was given was economical, it was reasoned that his host was saving money and could afford a large gift to the King. If, on the contrary the hospitality was sumptuous, he was evidently wealthy and could afford a large gift to the King. These arguments were the two prongs of the Fork.

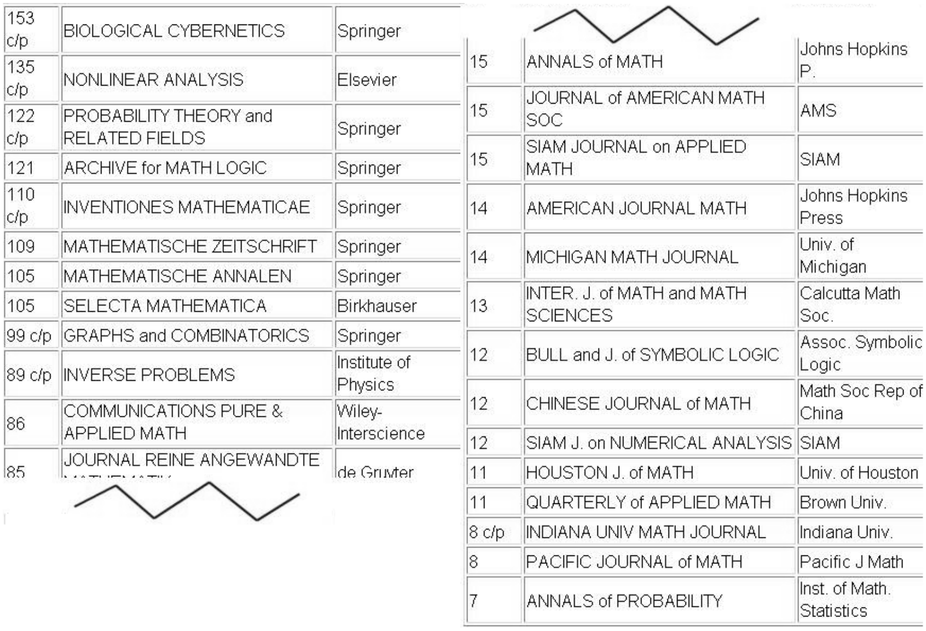
F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 144–156, 2003.

The only journals to which we don't need access are the (very few) journals which publish only low-level papers.

2 Genesis

This situation came home to us with full force about seven years ago when Colin was chair of the Warwick Mathematics Department and Brian was Library Representative and we faced the annual journal cutting exercise. We reasoned: Journals are produced by the community, largely using free labour; mathematicians even do their own typesetting using TeX; the cost of printing in small runs is dropping (it has now dropped to about the same as large runs); journals ought to be getting cheaper; instead they are rising in price far faster than inflation and causing a crisis leading to irreversible damage to libraries. It is interesting to note that a pledge to keep price inflation below 10% per annum was regarded as reasonable by the UK Competition Commission when considering the merger of Reed Elsevier and Harcourt.

Rob Kirby has compiled a relative price list for mathematics journals. Comparing like with like, the price ratio is about 20 to 1 between the lowest priced journals at 7c approx per page and the highest at about $1.50 per page, see Figure 1.



153 c/p	BIOLOGICAL CYBERNETICS	Springer
135 c/p	NONLINEAR ANALYSIS	Elsevier
122 c/p	PROBABILITY THEORY and RELATED FIELDS	Springer
121	ARCHIVE for MATH LOGIC	Springer
110 c/p	INVENTIONES MATHEMATICAE	Springer
109	MATHEMATISCHE ZEITSCHRIFT	Springer
105	MATHEMATISCHE ANNALEN	Springer
105	SELECTA MATHEMATICA	Birkhauser
99 c/p	GRAPHS and COMBINATORICS	Springer
89 c/p	INVERSE PROBLEMS	Institute of Physics
86	COMMUNICATIONS PURE & APPLIED MATH	Wiley-Interscience
85	JOURNAL REINE ANGEWANDTE	de Gruyter

15	ANNALS of MATH	Johns Hopkins P.
15	JOURNAL of AMERICAN MATH SOC	AMS
15	SIAM JOURNAL on APPLIED MATH	SIAM
14	AMERICAN JOURNAL MATH	Johns Hopkins Press
14	MICHIGAN MATH JOURNAL	Univ. of Michigan
13	INTER. J. of MATH and MATH SCIENCES	Calcutta Math Soc.
12	BULL and J. of SYMBOLIC LOGIC	Assoc. Symbolic Logic
12	CHINESE JOURNAL of MATH	Math Soc Rep of China
12	SIAM J. on NUMERICAL ANALYSIS	SIAM
11	HOUSTON J. of MATH	Univ. of Houston
11	QUARTERLY of APPLIED MATH	Brown Univ.
8 c/p	INDIANA UNIV MATH JOURNAL	Indiana Univ.
8	PACIFIC JOURNAL of MATH	Pacific J Math
7	ANNALS of PROBABILITY	Inst. of Math. Statistics

Fig. 1. Kirby letter (extracts) available in full at:
`http://math.berkeley.edu/~kirby/journals.html`

We decided to demonstrate that a high quality journal could be run well and be essentially free and this is why we started *Geometry & Topology*.

3 Quality

Initially the journal was conceived as an electronic journal. At the time electronic publishing was outside the mainstream and had a somewhat tacky image. So our first concern was to ensure real quality. The way we did this was to recruit as high a quality academic editorial board as we could. We were ably assisted in this by John Jones and Rob Kirby. Our editorial board can be seen on our home page, Figure 2. It includes three Fields Medalists and is one of the highest standard boards in Pure Mathematics.

This was the first step. The second step was to involve this board positively in all the decision making. We set up a strong, published procedure slewed towards rejection. The key point is that whilst one editor can reject a paper, it takes three (a proposer and two seconders) to accept one; each of these has her/his name published on the title page of the paper, which militates against casual propositions/secondings. (For more detail, see the section on Refereeing below.)

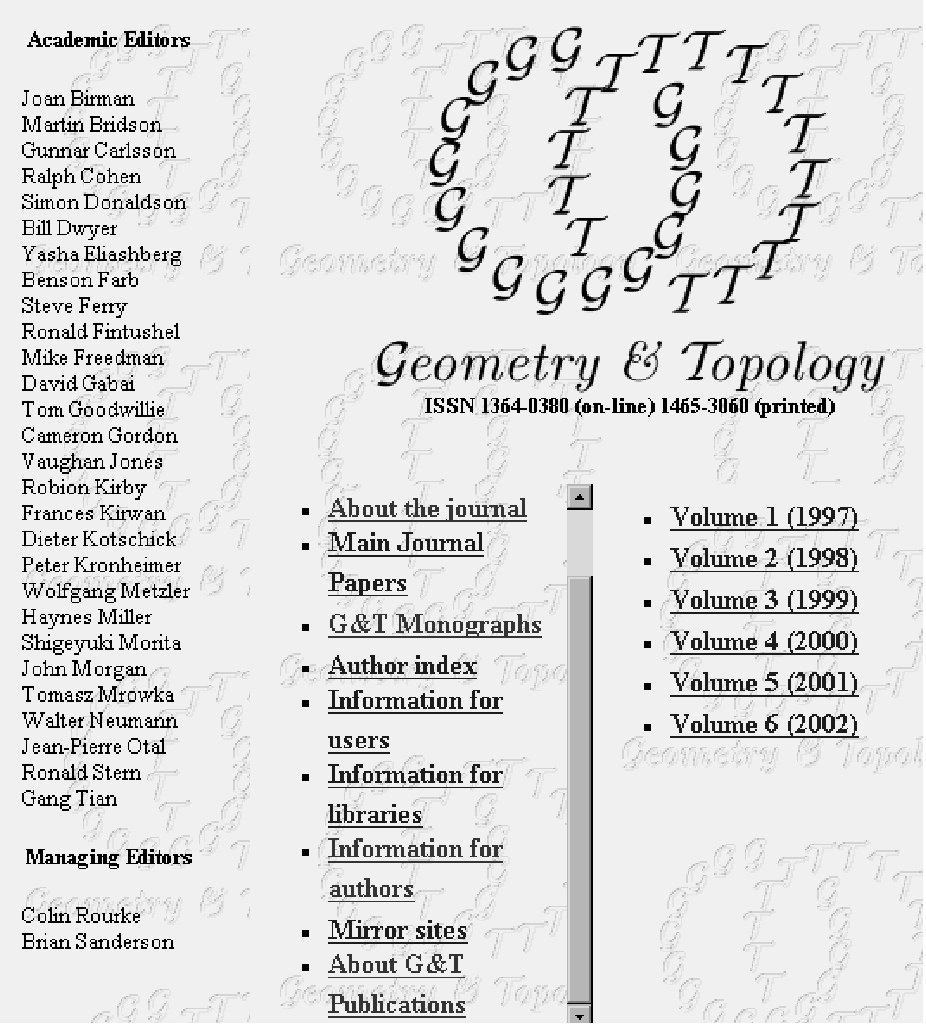
4 Setting Up — Emacs Html and Perl

We used the University of Warwick Mathematics Department Unix computer system. Little more than simple text processing was needed to create the two main ingredients of an electronic journal. These are the html files for the web pages and the perl programs. We did not use any commercial packages. We still use Emacs for text processing and file handling including email. To begin with we knew neither html nor perl but were quite familiar with Emacs having used it to create many TeX files in the past.

Learning html was a relatively simple task. Good guides are available and looking at the source code for existing web pages helped a lot. The only non-text portions of our web pages are still the logos. For these we needed a drawing program. Learning perl was fun but took longer. Our perl programs cut down the time needed to handle submissions and the publishing process.

5 Handling Submissions

To submit a paper an author fills in a web form, Figure 3, which asks for the author's details and details of the paper submitted including an abstract readable in simple text. The paper is attached and the form submitted. At this point we only want a postscript file. If the file has already been deposited on the arXiv we simply want the reference. A perl cgi program on the web server then sends a formatted plain text email to us with the ps file attached. We output the file and run a perl program which performs the following actions:



Academic Editors

Joan Birman
Martin Bridson
Gunnar Carlsson
Ralph Cohen
Simon Donaldson
Bill Dwyer
Yasha Eliashberg
Benson Farb
Steve Ferry
Ronald Fintushel
Mike Freedman
David Gabai
Tom Goodwillie
Cameron Gordon
Vaughan Jones
Robion Kirby
Frances Kirwan
Dieter Kotschick
Peter Kronheimer
Wolfgang Metzler
Haynes Miller
Shigeyuki Morita
John Morgan
Tomasz Mrowka
Walter Neumann
Jean-Pierre Otal
Ronald Stern
Gang Tian

Managing Editors

Colin Rourke
Brian Sanderson

Geometry & Topology

ISSN 1364-0380 (on-line) 1465-3060 (printed)

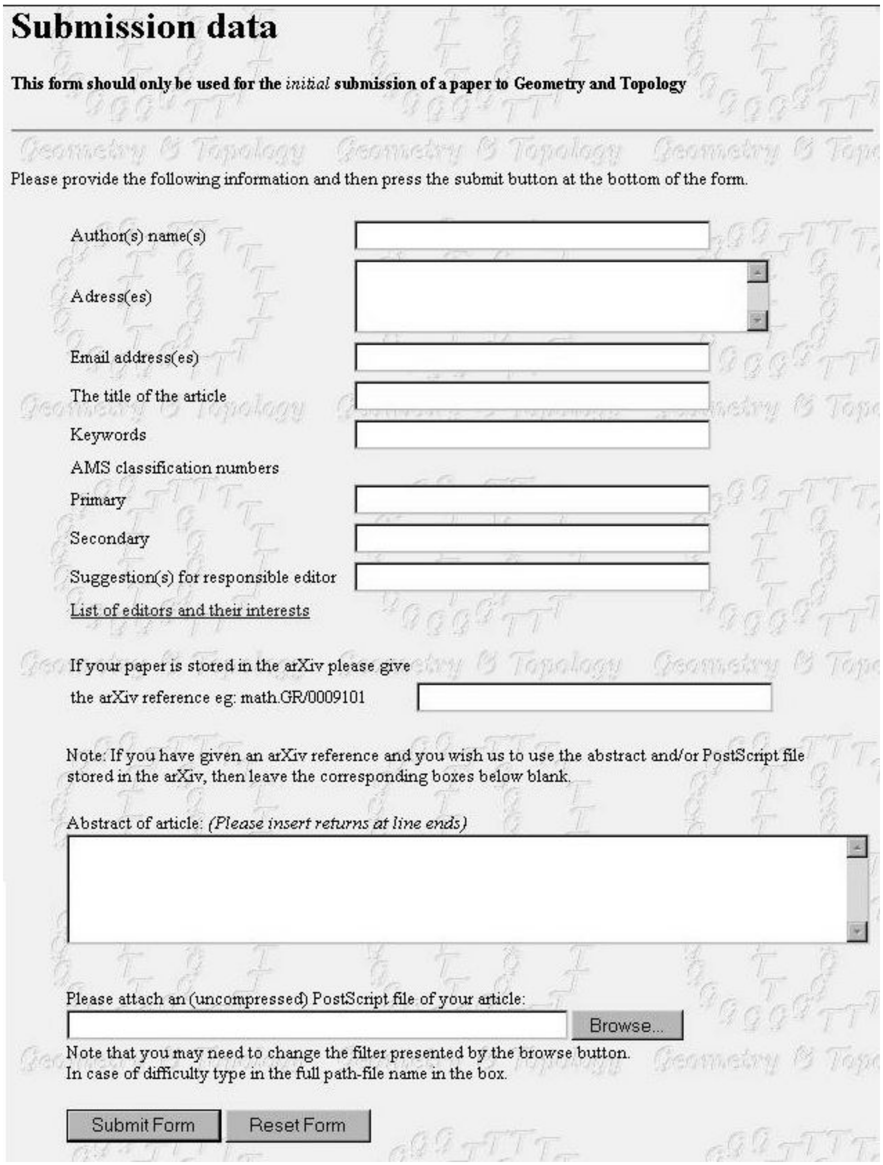
- About the journal
- Main Journal Papers
- G&T Monographs
- Author index
- Information for users
- Information for libraries
- Information for authors
- Mirror sites
- About G&T Publications

- Volume 1 (1997)
- Volume 2 (1998)
- Volume 3 (1999)
- Volume 4 (2000)
- Volume 5 (2001)
- Volume 6 (2002)

Fig. 2. GT home page
`http://www.maths.warwick.ac.uk/gt/index.html`

Load the psfile into a viewer so that it can be checked. Make suitable directories and files. Deposit the psfile. Update the mainlog and our private diary. Email the author to acknowledge receipt in good order. Email the responsible editor with instructions on picking up the ps file. If the author has not suggested a suitable editor we make a choice.

This whole process takes about 10 minutes. From now on we keep an eye on the diary and prompt the editor to nudge the referee when necessary.



Submission data

This form should only be used for the *initial* submission of a paper to Geometry and Topology

Please provide the following information and then press the submit button at the bottom of the form.

Author(s) name(s)

Adress(es)

Email address(es)

The title of the article

Keywords

AMS classification numbers

Primary

Secondary

Suggestion(s) for responsible editor

List of editors and their interests

If your paper is stored in the arXiv please give
the arXiv reference eg: math.GR/0009101

Note: If you have given an arXiv reference and you wish us to use the abstract and/or PostScript file stored in the arXiv, then leave the corresponding boxes below blank.

Abstract of article: *(Please insert returns at line ends)*

Please attach an (uncompressed) PostScript file of your article:

Note that you may need to change the filter presented by the browse button.
In case of difficulty type in the full path-file name in the box.

Fig. 3. GT submission form
`http://www.maths.warwick.ac.uk/gt/gtsubmit_form.html`

6 Refereeing

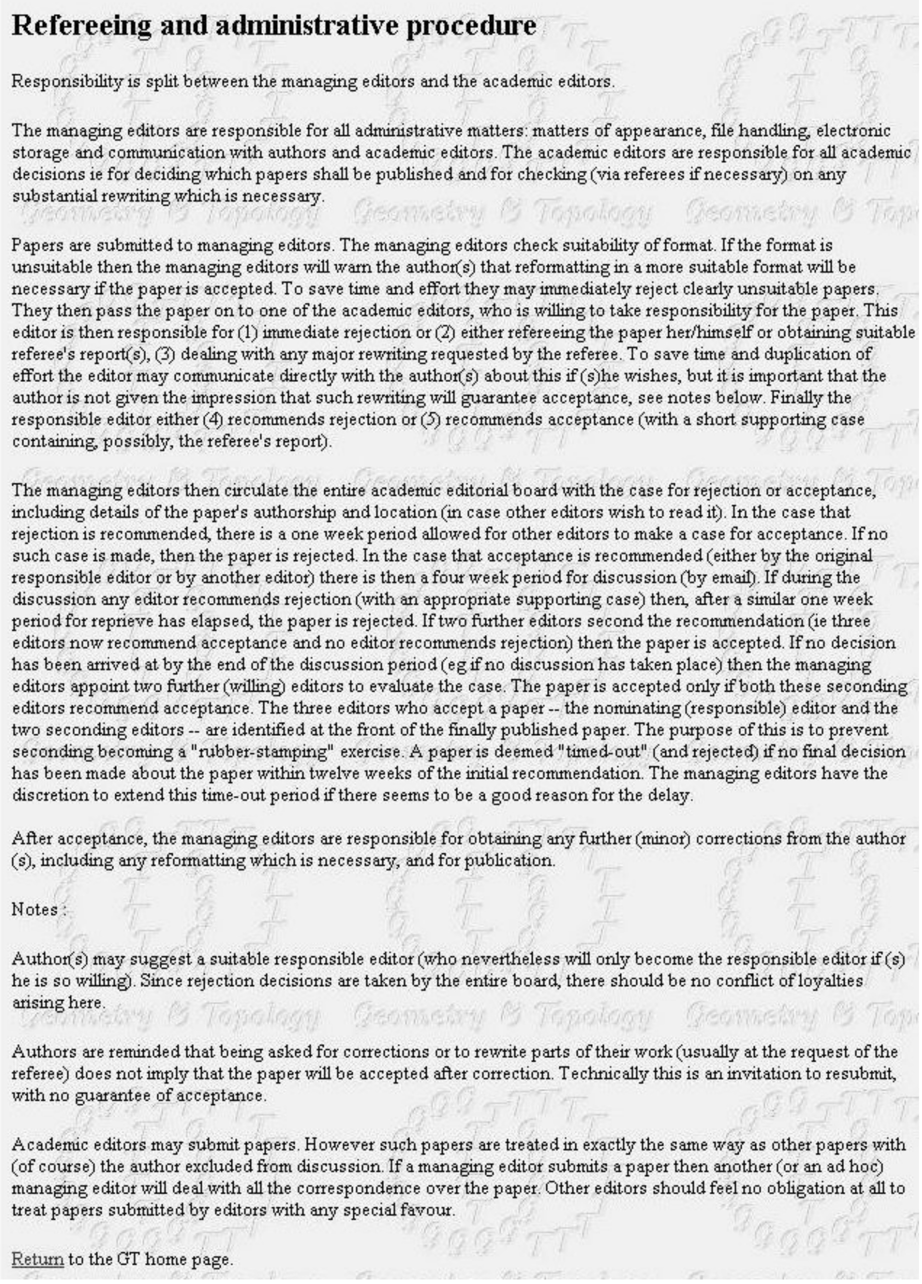
The procedure for handing papers is published on our web pages, Figure 4. The editor is responsible for (1) immediate rejection or (2) either refereeing the paper her/himself or obtaining suitable referee's report(s), (3) dealing with any major rewriting requested by the referee. To save time and duplication of effort the editor may communicate directly with the author(s) about this if (s)he wishes, but it is important that the author is not given the impression that such rewriting will guarantee acceptance. Finally the responsible editor either (4) recommends rejection or (5) recommends acceptance (with a short supporting case containing, possibly, the referee's report). We then circulate the entire academic editorial board with the case for rejection or acceptance, including details of the paper's authorship and location (in case other editors wish to read it). In the case that rejection is recommended, there is a one-week period allowed for other editors to make a case for acceptance. If no such case is made, then the paper is rejected. In the case that acceptance is recommended (either by the original responsible editor or by another editor) there is then a four-week period for discussion (by email). If during the discussion any editor recommends rejection (with an appropriate supporting case) then, after a similar one-week period for reprieve has elapsed, the paper is rejected. If two further editors second the recommendation (ie three editors now recommend acceptance and no editor recommends rejection) then the paper is accepted. If no decision has been arrived at by the end of the discussion period (eg if no discussion has taken place) then we appoint two further (willing) editors to evaluate the case. The paper is accepted only if both these seconding editors recommend acceptance. The three editors who accept a paper – the nominating (responsible) editor and the two seconding editors – are identified at the front of the finally published paper. The purpose of this is to prevent seconding becoming a "rubber-stamping" exercise. A paper is deemed "timed-out" (and rejected) if no final decision has been made about the paper within twelve weeks of the initial recommendation. The managing editors have the discretion to extend this time-out period if there seems to be a good reason for the delay. After acceptance, the managing editors are responsible for obtaining any further (minor) corrections from the author(s), including any reformatting which is necessary, and for publication.

7 Publication

Assuming the paper has been accepted, the publication process begins. Now we ask the author to give us the TeX file in a format which fits our established style. This is detailed in our author instructions of which a fragment is shown in Figure 5. The file may need editing. In particular we will add our standard front page. This is where TeX expertise is required and a badly formatted paper can gobble up our time.

We are usually ready to publish within a day or two of acceptance. The perl publishing program is then run. It performs the following actions:

Get data stored by the submission program. Make pdf and compressed ps files. Modify existing web pages and make new web pages. Move files and create directories. Mail the author and our mailing list to say the paper has been published. Mail the Newsgroup sci.math.research with an announcement.



Refereeing and administrative procedure

Responsibility is split between the managing editors and the academic editors.

The managing editors are responsible for all administrative matters: matters of appearance, file handling, electronic storage and communication with authors and academic editors. The academic editors are responsible for all academic decisions ie for deciding which papers shall be published and for checking (via referees if necessary) on any substantial rewriting which is necessary.

Papers are submitted to managing editors. The managing editors check suitability of format. If the format is unsuitable then the managing editors will warn the author(s) that reformatting in a more suitable format will be necessary if the paper is accepted. To save time and effort they may immediately reject clearly unsuitable papers. They then pass the paper on to one of the academic editors, who is willing to take responsibility for the paper. This editor is then responsible for (1) immediate rejection or (2) either refereeing the paper her/himself or obtaining suitable referee's report(s), (3) dealing with any major rewriting requested by the referee. To save time and duplication of effort the editor may communicate directly with the author(s) about this if (s)he wishes, but it is important that the author is not given the impression that such rewriting will guarantee acceptance, see notes below. Finally the responsible editor either (4) recommends rejection or (5) recommends acceptance (with a short supporting case containing, possibly, the referee's report).

The managing editors then circulate the entire academic editorial board with the case for rejection or acceptance, including details of the paper's authorship and location (in case other editors wish to read it). In the case that rejection is recommended, there is a one week period allowed for other editors to make a case for acceptance. If no such case is made, then the paper is rejected. In the case that acceptance is recommended (either by the original responsible editor or by another editor) there is then a four week period for discussion (by email). If during the discussion any editor recommends rejection (with an appropriate supporting case) then, after a similar one week period for reprieve has elapsed, the paper is rejected. If two further editors second the recommendation (ie three editors now recommend acceptance and no editor recommends rejection) then the paper is accepted. If no decision has been arrived at by the end of the discussion period (eg if no discussion has taken place) then the managing editors appoint two further (willing) editors to evaluate the case. The paper is accepted only if both these seconding editors recommend acceptance. The three editors who accept a paper -- the nominating (responsible) editor and the two seconding editors -- are identified at the front of the finally published paper. The purpose of this is to prevent seconding becoming a "rubber-stamping" exercise. A paper is deemed "timed-out" (and rejected) if no final decision has been made about the paper within twelve weeks of the initial recommendation. The managing editors have the discretion to extend this time-out period if there seems to be a good reason for the delay.

After acceptance, the managing editors are responsible for obtaining any further (minor) corrections from the author (s), including any reformatting which is necessary, and for publication.

Notes :

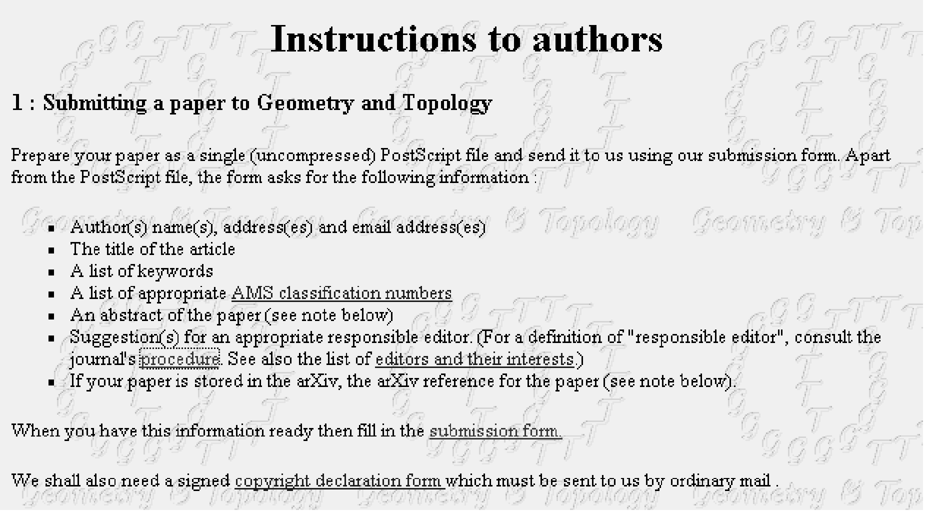
Author(s) may suggest a suitable responsible editor (who nevertheless will only become the responsible editor if (s) he is so willing). Since rejection decisions are taken by the entire board, there should be no conflict of loyalties arising here.

Authors are reminded that being asked for corrections or to rewrite parts of their work (usually at the request of the referee) does not imply that the paper will be accepted after correction. Technically this is an invitation to resubmit, with no guarantee of acceptance.

Academic editors may submit papers. However such papers are treated in exactly the same way as other papers with (of course) the author excluded from discussion. If a managing editor submits a paper then another (or an ad hoc) managing editor will deal with all the correspondence over the paper. Other editors should feel no obligation at all to treat papers submitted by editors with any special favour.

Return to the GT home page.

Fig. 4. Published procedure
`http://www.maths.warwick.ac.uk/gt/procedur.html`



Instructions to authors

1 : Submitting a paper to Geometry and Topology

Prepare your paper as a single (uncompressed) PostScript file and send it to us using our submission form. Apart from the PostScript file, the form asks for the following information :

- Author(s) name(s), address(es) and email address(es)
- The title of the article
- A list of keywords
- A list of appropriate AMS classification numbers
- An abstract of the paper (see note below)
- Suggestion(s) for an appropriate responsible editor. (For a definition of "responsible editor", consult the journal's procedure. See also the list of editors and their interests.)
- If your paper is stored in the arXiv, the arXiv reference for the paper (see note below).

When you have this information ready then fill in the submission form.

We shall also need a signed copyright declaration form which must be sent to us by ordinary mail .

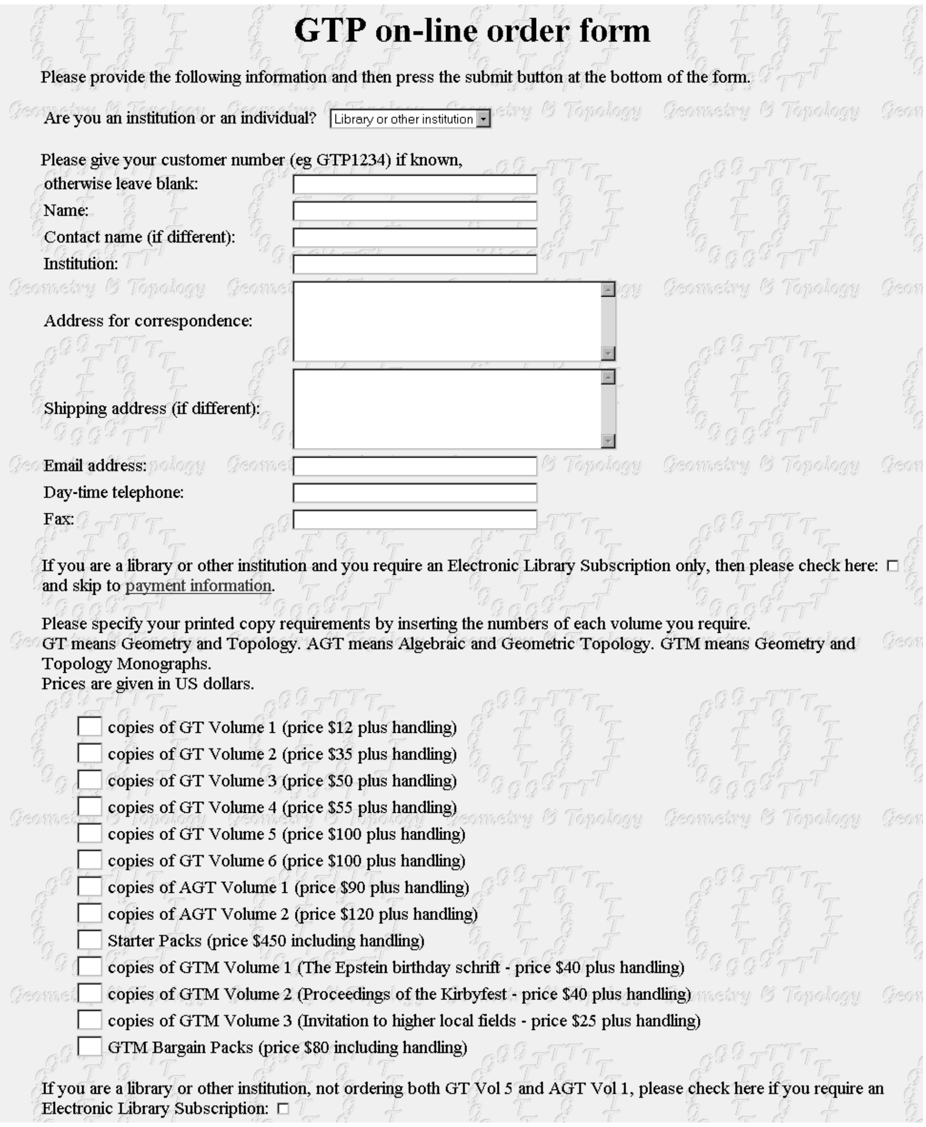
Fig. 5. Author instructions (fragment)
`http://www.maths.warwick.ac.uk/gt/gtauthin.html`

After two weeks we run our arxiv program which submits the paper to the e-Print archive at `http://arxiv.org/`. If it is already there we need to get the author's password in order to upload the new version.

8 Business Model

Initially we wanted to have an open access electronic journal. We fully intend to stick to open access. It soon became apparent that credibility required a print version. At first we used International Press to print our volumes. We have since taken over the printing and use the University of Warwick's print shop. All papers published electronically in a year are printed early in the following year. The price is decided at the time of printing and our target is 10 cents (US) per page. We have both an email and online order form, Figure 6.

An innovation is our electronic subscription (ELS) for libraries. We charge $80 per annum for the right to download and archive the electronic version of our publications and to distribute them though internal or external networks. A library that buys print subscriptions to both our journals is granted an ELS gratis. A library, which subscribes to the print version of just one of the journals, can buy an ELS for half price.



GTP on-line order form

Please provide the following information and then press the submit button at the bottom of the form.

Are you an institution or an individual? Library or other institution ▾

Please give your customer number (eg GTP1234) if known,
otherwise leave blank: □

Name: □

Contact name (if different): □

Institution: □

Address for correspondence: □

Shipping address (if different): □

Email address: □

Day-time telephone: □

Fax: □

If you are a library or other institution and you require an Electronic Library Subscription only, then please check here: □ and skip to payment information.

Please specify your printed copy requirements by inserting the numbers of each volume you require.
GT means Geometry and Topology. AGT means Algebraic and Geometric Topology. GTM means Geometry and Topology Monographs.
Prices are given in US dollars.

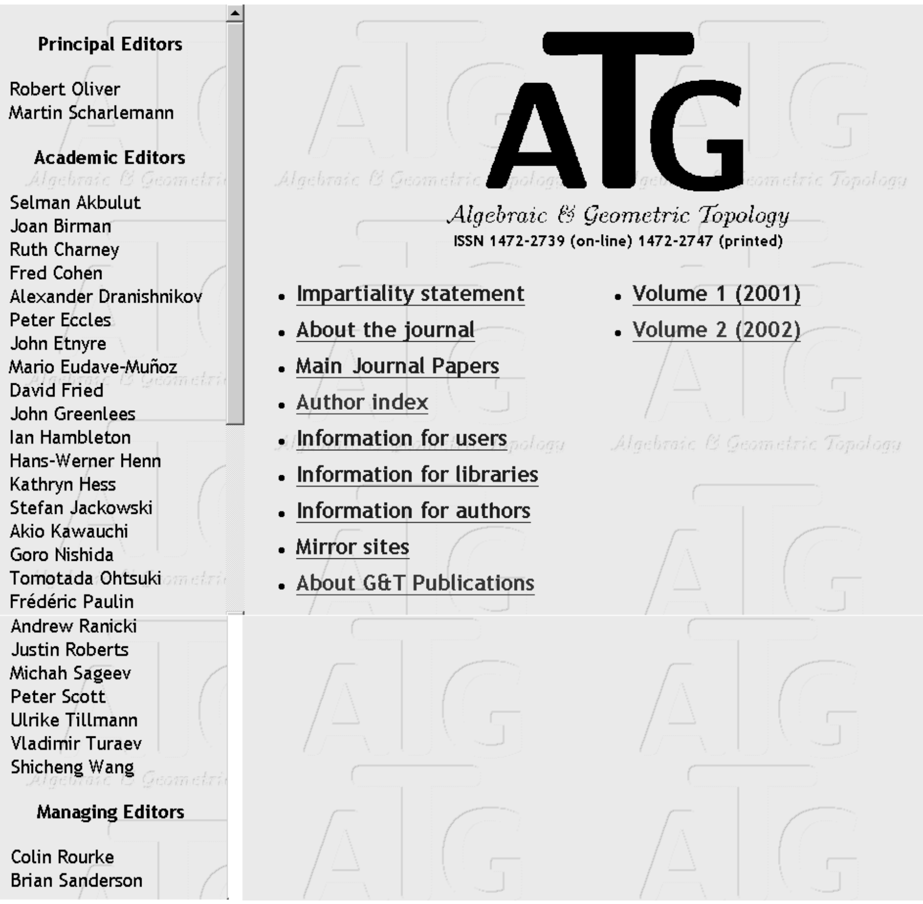
□ copies of GT Volume 1 (price $12 plus handling)
□ copies of GT Volume 2 (price $35 plus handling)
□ copies of GT Volume 3 (price $50 plus handling)
□ copies of GT Volume 4 (price $55 plus handling)
□ copies of GT Volume 5 (price $100 plus handling)
□ copies of GT Volume 6 (price $100 plus handling)
□ copies of AGT Volume 1 (price $90 plus handling)
□ copies of AGT Volume 2 (price $120 plus handling)
□ Starter Packs (price $450 including handling)
□ copies of GTM Volume 1 (The Epstein birthday schrift - price $40 plus handling)
□ copies of GTM Volume 2 (Proceedings of the Kirbyfest - price $40 plus handling)
□ copies of GTM Volume 3 (Invitation to higher local fields - price $25 plus handling)
□ GTM Bargain Packs (price $80 including handling)

If you are a library or other institution, not ordering both GT Vol 5 and AGT Vol 1, please check here if you require an Electronic Library Subscription: □

Fig. 6. GTP on-line order form
`http://www.maths.warwick.ac.uk/gt/gtp-online-order-form.html`

9 Where We Are Now

We currently publish two Journals: *Geometry & Topology* (GT) and *Algebraic & Geometric Topology* (AGT), Figure 7. In addition we have a monograph series *Geometry & Topology Monographs* (GTM), Figure 8, and have published five volumes in this series. The first (printed) volume of GT appeared in 1997 and the first (printed) volume of AGT appeared in 2001. The venture is now called *Geometry & Topology Publications* (GTP), Figure 9.



Principal Editors

Robert Oliver
Martin Scharlemann

Academic Editors

Selman Akbulut
Joan Birman
Ruth Charney
Fred Cohen
Alexander Dranishnikov
Peter Eccles
John Etnyre
Mario Eudave-Muñoz
David Fried
John Greenlees
Ian Hambleton
Hans-Werner Henn
Kathryn Hess
Stefan Jackowski
Akio Kawauchi
Goro Nishida
Tomotada Ohtsuki
Frédéric Paulin
Andrew Ranicki
Justin Roberts
Michah Sageev
Peter Scott
Ulrike Tillmann
Vladimir Turaev
Shicheng Wang

Managing Editors

Colin Rourke
Brian Sanderson

ATG

Algebraic & Geometric Topology
ISSN 1472-2739 (on-line) 1472-2747 (printed)

- Impartiality statement
- About the journal
- Main Journal Papers
- Author index
- Information for users
- Information for libraries
- Information for authors
- Mirror sites
- About G&T Publications

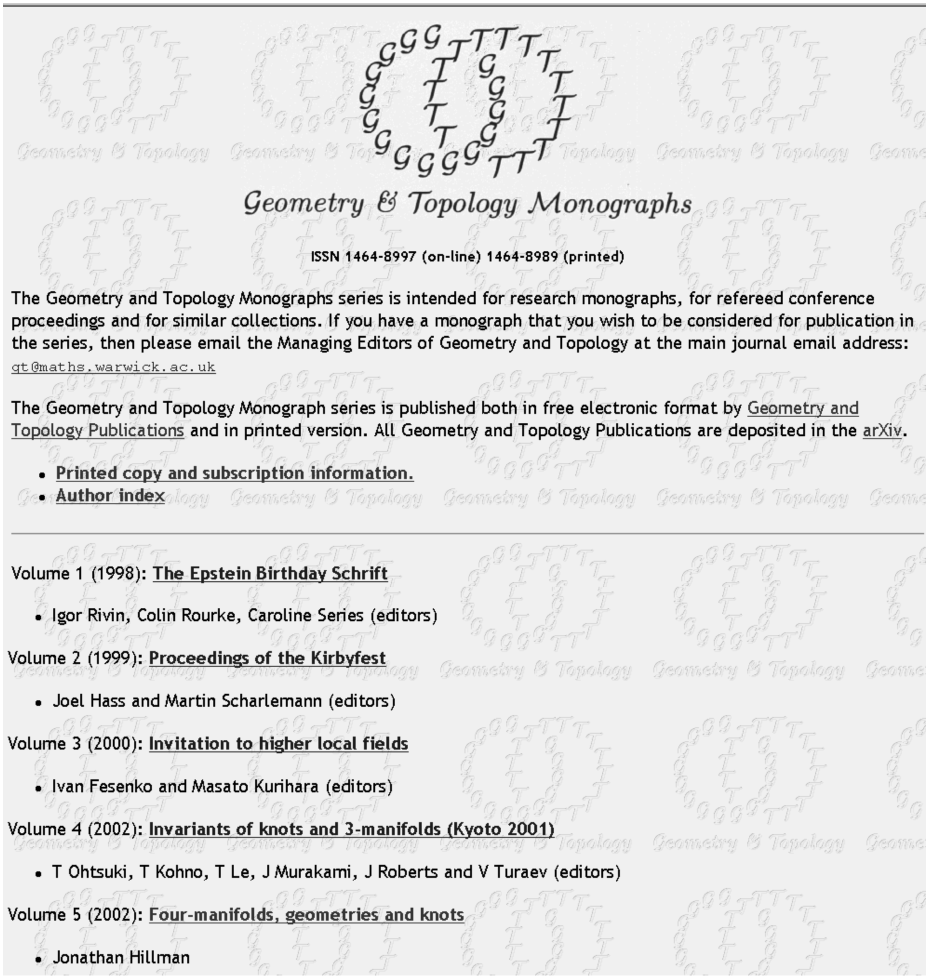
- Volume 1 (2001)
- Volume 2 (2002)

Fig. 7. *Geometry & Topology Publications* home page
`http://www.maths.warwick.ac.uk/agt/index.html`

The model for AGT is different from that of GT. For AGT authors are required to submit their papers directly to an editor. The managing editors of AGT only become involved after a paper has been accepted.

GTP is owned by the University of Warwick and has an executive committee of six members who decide on policy matters.

Financially we are currently at break-even point and on track to produce a surplus which can be used to pay for the routine tasks currently undertaken gratis by the managing editors. Figures 10 and 11 show growth in page numbers and paying subscriptions.



Geometry & Topology Monographs

ISSN 1464-8997 (on-line) 1464-8989 (printed)

The Geometry and Topology Monographs series is intended for research monographs, for refereed conference proceedings and for similar collections. If you have a monograph that you wish to be considered for publication in the series, then please email the Managing Editors of Geometry and Topology at the main journal email address: gt@maths.warwick.ac.uk

The Geometry and Topology Monograph series is published both in free electronic format by Geometry and Topology Publications and in printed version. All Geometry and Topology Publications are deposited in the arXiv.

- **Printed copy and subscription information.**
- **Author index**

Volume 1 (1998): **The Epstein Birthday Schrift**

- Igor Rivin, Colin Rourke, Caroline Series (editors)

Volume 2 (1999): **Proceedings of the Kirbyfest**

- Joel Hass and Martin Scharlemann (editors)

Volume 3 (2000): **Invitation to higher local fields**

- Ivan Fesenko and Masato Kurihara (editors)

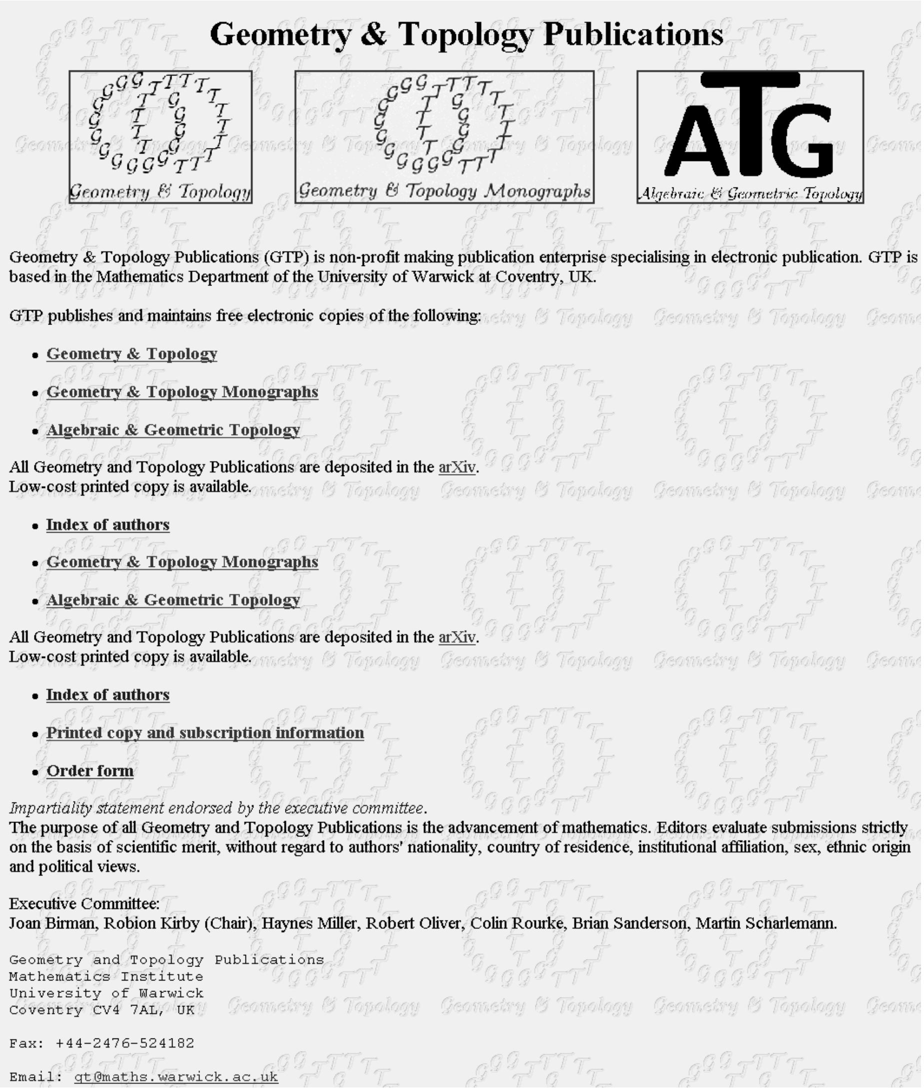
Volume 4 (2002): **Invariants of knots and 3-manifolds (Kyoto 2001)**

- T Ohtsuki, T Kohno, T Le, J Murakami, J Roberts and V Turaev (editors)

Volume 5 (2002): **Four-manifolds, geometries and knots**

- Jonathan Hillman

Fig. 8. *Geometry & Topology Publications* home page
`http://www.maths.warwick.ac.uk/gt/gtmono.html`


Geometry & Topology
Geometry & Topology Monographs
ATG
Algebraic & Geometric Topology

Geometry & Topology Publications

Geometry & Topology Publications (GTP) is non-profit making publication enterprise specialising in electronic publication. GTP is based in the Mathematics Department of the University of Warwick at Coventry, UK.

GTP publishes and maintains free electronic copies of the following:

- **Geometry & Topology**
- **Geometry & Topology Monographs**
- **Algebraic & Geometric Topology**

All Geometry and Topology Publications are deposited in the arXiv.
Low-cost printed copy is available.

- **Index of authors**

- **Geometry & Topology Monographs**
- **Algebraic & Geometric Topology**

All Geometry and Topology Publications are deposited in the arXiv.
Low-cost printed copy is available.

- **Index of authors**

- **Printed copy and subscription information**
- **Order form**

Impartiality statement endorsed by the executive committee.
The purpose of all Geometry and Topology Publications is the advancement of mathematics. Editors evaluate submissions strictly on the basis of scientific merit, without regard to authors' nationality, country of residence, institutional affiliation, sex, ethnic origin and political views.

Executive Committee:
Joan Birman, Robion Kirby (Chair), Haynes Miller, Robert Oliver, Colin Rourke, Brian Sanderson, Martin Scharlemann.

```
Geometry and Topology Publications
Mathematics Institute
University of Warwick
Coventry CV4 7AL, UK

Fax: +44-2476-524182

Email: gt@maths.warwick.ac.uk
```

Fig. 9. *Geometry & Topology Publications* home page
`http://www.maths.warwick.ac.uk/gt/gtp.html`

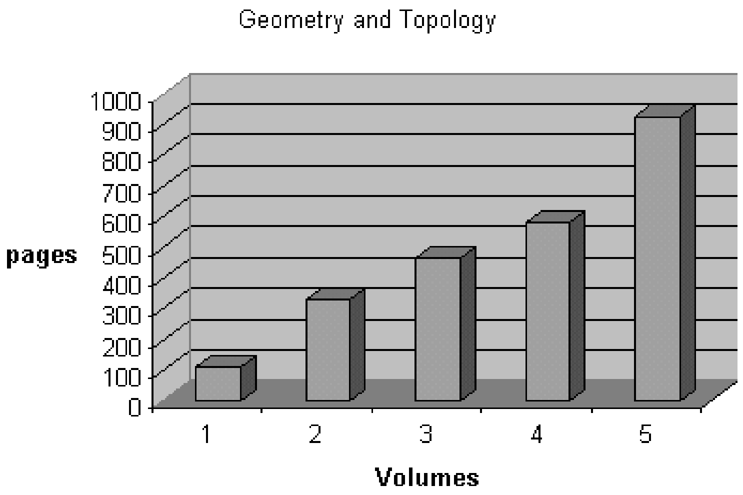


Fig. 10. *Geometry & Topology* Growth by page numbers

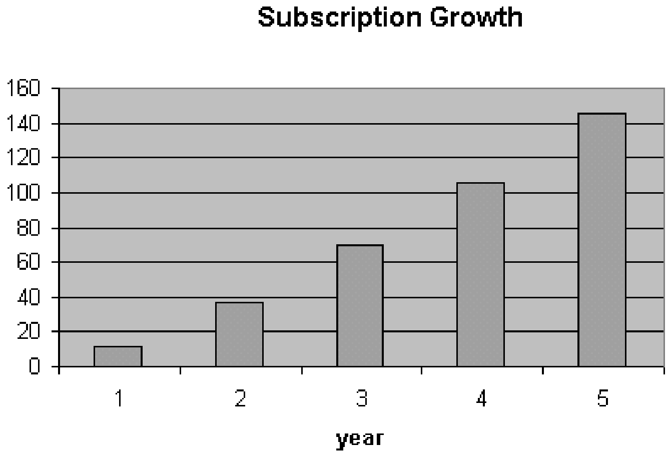


Fig. 11. Growth by subscriptions

We have not needed paid advertising but have relied partly on word-of-mouth and electronic outlets. In particular our journals are endorsed by SPARC (the Scholarly Publishing and Academic Resources Coalition) an initiative of ARL (the Association of Research Libraries).

The Web: Challenge and Opportunity for an Independent Journal

Klaus Kaiser

Houston Journal of Mathematics, University of Houston
Houston, TX 77204-3476
kkaiser@uh.edu

Abstract. To establish a comprehensive Website is for any journal a major task. Most independent journals face the additional obstacle of not having the financial resources for hiring information technology staff. I am going to describe the kind of decisions we had to make to set up a comprehensive Website for the Houston Journal of Mathematics (HJM). The result is a publishing format which may not have all the characteristics of a primarily electronic journal, but it offers more than a paper journal with free abstracts, and file access to its subscribers.
Like some other journals, HJM decided to make electronic editions freely available for registered subscribers of the print edition. Nevertheless, one still needs a "License Agreement Form". This is a legal document that defines the "Scope of the License" and which includes a "Copyright" clause. In its present form, our license excludes electronic files from Inter Library Loans. I will try to explain our opinion on this controversial topic.
All journals, whether commercial or independent, face the problem of "archival quality" of current file formats, primarily of the .pdf format. But there is also the related problem of shared legal responsibilities of possible file maintenance. The paper will present some thoughts on these most pressing issues.

1 What Is an Independent Journal?

Mathematicians like precise definitions. So, how do we define independent journals versus journals that are not independent? Most people classify journals according to their owners, and independent scholarly journals are then more or less journals that are owned by universities. They are perceived as small and cheap. The *Notre Dame Journal of Formal Logic* is a prime example of a small and very inexpensive journal. On the other hand, *Duke Mathematical Journal* is very large and not inexpensive, even on a price per page basis. But both journals are members of *Project Euclid* which is meant to be "A partnership of independent publishers". I guess here the phrase "independent publishers" means publishers that are not part of one of the major publishing giants, for example Bertelsmann, while publishers like the Heldermann Verlag, or International Press, are considered as independent.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 157–168, 2003.

I feel that one should differentiate between independent journals and independent publishers. Certainly, all journals need an *Editorial Board* and a *Publisher. Independent journals are those in which the editors also function as publisher.* Such journals are usually run by university departments. Thus, with respect to fiscal and personnel related matters, they are units of the department and therefore under direct jurisdiction of the chair and the dean of the college. As I see it, the important characteristic of an independent journals is that no relevant decisions are made outside the board of editors and the publishing department.

Some independent journals work with independent publishers to facilitate their shipping and printing tasks. Doing so, these journals still maintain their independence because in such cases the publisher provides only a service. HJM is regularly considering this possibility.

For *commercial* journals, we see a strict separation of the Editorial Board and the Publisher. The editors are responsible for content and academic standards while the publisher takes care of all business related issues, in particular, the publisher determines the price structure of the journal in order to guarantee *profitability. Societal* journals are like the independents *non-profit* but because of their size and financial obligations they are run like commercial journals. Some larger publishing companies, like Cambridge University Press, are affiliated with academic institutions. Most people would consider such *Academic Publishers* as independent because they are different from the publication and media giants. But their journals are in general not independent according to my definition. For example, also here it is the publisher who finally determines the subscription rate. And the editors may have little or no say about the overall design and features of electronic editions.

Electronic journals fall into an interesting category. Most of the new electronic journals can be considered independent. Here, publishing means "Web posting" of accepted articles. Unlike periodical journals, publication is continual. Of course, commercial publishers may also have electronic journals. Most of the independent electronic journals are freely accessible. This is possible because there are only small "real" expenses involved: Computing facilities, office space, technical and secretarial personnel etc. are provided by the publishing institutions. In some cases, libraries have been asked to *sponsor* such journals in order to guarantee free access and to provide funds for periodically published printed editions. Many people feel that independent electronic journals that are freely available, and which provide in regular intervals hard copies (meant primarily as an alternative form of archiving the contents of electronic files) are the wave of the future and may eventually replace traditional journals.

Because of the ongoing electronic revolution, all journals, whether independent or commercial, have to change. I am going to describe part of this process for a typical independent journal, namely the Houston Journal of Mathematics (HJM).

2 Print Editions

As editor of the HJM, I interact with many of our authors, and one thing I can definitely conclude is that mathematicians still want to see their work the old fashioned way: printed on paper. And they want reprints. While authors exchange ideas *via* e-mail and circulate their papers in an electric format, for most mathematicians, a publication isn't a publication unless it is available on paper in a respected journal. This is confirmed in the findings of [4].

Thus, the quality of typesetting and printing still matters. This is an area where independents have a decided advantage. Because most journals use the same LaTeX program for typesetting, journals now look more or less the same. Commercial and societal journals have lost their edge during the electronic revolution because independents can produce journals at a much lower price.

Most journals request that authors send their submissions in the form of a LaTeX file. Here, independent journals, like HJM, may be more persistent than commercial ones. Retypesetting articles is an expensive and time consuming process, besides being wasteful. In order to keep a low overhead, we insist authors to provide us with a useful LaTeX file. In exchange, we go beyond providing authors with the usual tips and helpful tools for accomplishing this. We made an arrangement with our TeX provider, VTeX, to give our authors a steep discount on the same commercial LaTeX implementation that the journal uses.

A major problem for independent journals is finding a reasonable printer. It used to be the case that universities maintained their own printing facilities. Unfortunately, this is no longer the case even for larger public schools, and the number of smaller printing facilities seems to be dwindling. Some major subscription agencies are rushing to fill the gap by going into the "printing plus mailing" business. While this makes a lot of sense, it might create some conflict of interest. This development makes the survival of smaller academic presses even more difficult.

Printing and mailing are the largest expenditures for independent journals. While commercial journals have the advantage of bargaining power, for independents, these costs are nonnegotiable. Although we seem to be stuck with printing and mailing costs that are high and inflationary, I have found that these costs are recoverable through modest increases in subscription rates and through the authors' voluntary contributions. In addition, digital printing technologies already being adopted by printing companies have been reducing the cost of labor. Therefore, I feel that independent journals will be able to provide printed editions at a very reasonable cost for many years to come. In fact, if printed journals should disappear at some point in the future, it will not be because of the high cost of printing, but because the mathematical community has deemed printed journals to be obsolete. No one knows if this is ever going to happen, but right now it certainly doesn't look that way. The predictions of authors like [9, page 2] have not materialized, yet.

3 Electronic Editions

Most journals provide "Electronic Editions" in addition to the traditional print format. Naturally, commercial journals started this first and have to some extent set the pattern for the overall design of a Web presence for scientific journals. Of course, not all features of a conglomerate of dozens, and in some cases of hundreds of different journals apply to a "single journal" site. For example, the practice of providing links to different journals from the same publisher is not an issue if you only have ownership of one journal. As for appearance, there is no universally accepted standard of what a good Website should look like. Like furniture, the architecture of a Website is very much a matter of taste. For independent journals, a clean and simple design is very often a necessity because most independents don't have a Web-trained professional staff to take care of the frequent, regular updates. Our Website is strictly text based, easy to navigate, easy to edit, and it puts content over form.

As I see it, the single most important part of a journal Website are the *Titles and Abstracts*. Here, HJM is somewhat different from most other journal sites. For an author, the paper abstract is an integrated part of the paper. In particular, the paper abstract may refer to items in the literature, and to numbered theorems. And there may be mathematical symbols for which there is no HTML code. Even worse, authors may have used their own TEX macros for text and mathematics. Because of these problems, we ask authors to provide us with a specially prepared Web abstract, preferably in plain English or HTML, but with no TEX jargon, and with all references explicitly stated. Unfortunately, not too many authors are paying attention to our request. Thus we very often have to *detex* the paper abstracts, add the references etc. But the situation is improving. Before we officially post a new issue, authors are now given a secrete URL and they can inspect everything before it becomes official. Of course, this is a luxury probably only independents can afford, namely to personally contact each and every author before posting the papers, send them the final .pdf files and allow for further corrections of the paper, abstract, addresses etc. When an author sees his abstract on the Web, he may realize that it does not express too well what the paper was about. But he has a second chance to e-mail us a rewrite, and if he wishes to do so, he can add a direct link to his homepage with related papers etc. To summarize, the HJM perceives Web abstracts to be more like self-contained author-reviews than as abstracts in the traditional sense.

Now, why is it the case that Web abstracts are important? Before the appearance of the Web, mathematicians had to rely very much on the major reviewing organs in order to learn about the current literature. Nowadays, search engines and pre-print servers can find difficult to get, or still unpublished material. Search engines are timely and stupendously reliable. While an author's abstract cannot replace a carefully written evaluative review, most papers don't get an evaluative review and certainly not what the Reviews call a "featured" review. But through his Web abstract, an author is given the opportunity to describe in plain English and to an open forum, what the merits of his paper are and emphasize what he thinks is new and interesting.

In order to assure a more timely coverage, the major reviewing organs have started to accept electronic files. However, this yields the very important *integrity problem*: Printouts from downloaded files must be identical with the print editions. To ensure this for Print Journals has become a new and major responsibility of the publisher. HJM has addressed this problem by using for print editions exclusively printouts from .pdf files, the same files that have been posted on the Web. Hard copies are not produced from .dvi anymore. This policy can be time consuming. Articles may contain non-standard graphics commands, or obsolete (e.g., old UNIX) usepackages which conversion programs have difficulties to implement in .pdf format. But unless the .pdf file provides exactly the same output as the .dvi file, we do not proceed with posting and publishing an article. Needless to say, we occasionally have to consult professional help from our LaTeX provider.

So, as I'm describing it, the HJM's preparation of Web abstracts goes beyond "copy and paste"-ing abstracts from their source documents as given. Each abstract is individually checked for "Web suitability" and then edited and approved by the author before final posting. A more automated procedure would be to extract abstracts from the .pdf files, and post them together with the list of references. This is what some commercial journals do which offer besides the paper abstracts, also *enhanced* versions for a fee. But this still does not guarantee that the abstracts are self-contained and meaningful. There might be references to numbered theorems, for example.

Another special feature available on our journal's free web version is the index. We have divided the index into five year increments, which are easy to scroll because each page has not more than 250-300 entries, reflecting the 50-60 published papers per year. Freely available, complete Web indexes are one of the nicest and most useful features of journal Websites. Besides helping mathematicians with their research, scrollable indexes make the development of a journal evident and indicate its scope and strengths.

While we considered adding a search facility to the journal's website, we decided it would be redundant because all browsers allow for page search. Actually, now even site search is possible, because of the generosity of Google which provides this service to universities for free. Newer papers in the index are linked to the corresponding Web editions.

Abstracts, Index and Author Information are the basic ingredients of journal Web sites. A comprehensive Website allows also for full length paper access. Before offering paper access, a few decisions have to be made. Most importantly, one has to decide on a file format. HJM decided to use exclusively the .pdf format because it has become the *de facto* standard for electronic postings of more complex and longer documents. Other possible choices are the TeX source file and the .dvi and .ps format. Posting documents in their source form as .tex file is, generally speaking, not very realistic because of the possibility of having to include graphics and other files. Moreover, .tex files can be easily abused by plagiarists who are roaming the Web which is, unfortunately, no longer a hypothetical threat. Like .tex source files, .dvi files need the full TEX program for reading and printing. Of course, TEX is in general not installed on public computers and many mathematicians have access to TEX only on their office machines. If there had been a reader available for TEX , .dvi would have been a viable, *albeit* basic, choice for posting scientific documents on the Web. The .pdf file format has been specifically created for the Web, and TEX files can be compiled quite easily in .pdf, and then be read and printed from a variety of freely available readers. Thus, the .pdf file format has become the most obvious choice for the Web. There is, however, a serious caveat on which I will comment later on.

Before HJM added access to paper files, we offered a complete volume on a CD. We decided to offer the CD because at that time, everybody was asking for it. It came to me as a bad surprise that only a handful of institutions showed interest of buying the CD for the nominal price of $20. We certainly lost money on this experience! Libraries explained their disinterest by saying that patrons were lukewarm about checking out CDs. However, my thinking behind offering a CD had less to do with library patrons. I was interested in giving libraries an alternative means of storing documents, or for archiving their holdings in digitized form. By now, we know that this kind of electronic library never materialized, probably because of copyright concerns. The electronic "holdings" of a library are now typically rights of access to the remote sites of publishers, not the physical ownership of digitized files. It is the publishers who ultimately keep control and ownership of "their" files. So, what we now offer are the electronic editions and we have dropped the CD option.

4 Legal and Business Decisions

4.1 Pricing Electronic Editions

The difficulty with offering Electronic Editions is that nobody knows what kind of overall impact it will have on the subscription base. Electronic access certainly diminishes the need for multiple copies, e.g., one for the main library, and another copy for the department. On the other hand, a comprehensive Website increases visibility and status of the journal which may lead to more, and also to higher quality submissions from prestigious places. For HJM this process has certainly taken place. Since HJM entered the Web, we publish about 15% more pages per annum, and our backlog increased from about two issues to six issues. While we experienced some cancellations, mainly for multiple copies, and from places where several other universities were in the immediate vicinity, our overall subscription base has been stable. We also have more authors who take our plea for voluntary support more seriously and whose contributions have offset to some extent inflationary pressures.

HJM decided to provide electronic access free of charge, but only for subscribers of the print edition. The main advantage of this policy is its simplicity, which minimizes the amount of additional bookkeeping of the subscriber list. Moreover, this way we have made access to electronic files a privilege and not part of the paid subscription. This has some legal ramifications when it comes to assessing responsibilities for file maintenance.

Some independent journals offer electronic access for free. Outside academia, free access has become the norm for daily newspapers but not for magazines. I took a clue from this development and decided that free access would pose too much of a risk of eroding our subscription base. It also looks to me that libraries base their decisions to subscribe to a particular journal not solely on affordability, and on input from faculty, but also whether there is a need to subscribe. Free electronic access is in general no incentive to subscribe to the print edition of a journal. In contrast, I have been told that some libraries are now storing printed copies in book repositories, if electronic access has become available.

Print and electronic editions share the same upfront editorial work which is needed for the preparation of a new issue, but electronic editions require extra labor. Thus, electronic editions are not free *per se*, though there are no costs for mailing and printing involved. Indeed, the additional time and work needed for posting an electronic issue can be quite substantial. Thus, separate pricing of print and electronic editions makes a lot of sense, but in order to avoid additional administrative costs, we decided for one combined price.

Once a library has registered for electronic access, it has access to all electronic issues, even to those for which it didn't have a paid subscription. This is an incentive to subscribe. Especially when (eventually) we will have created an archive of all published articles. On the other hand, if a library cancels a subscription, it has lost a privilege, namely electronic access, even to those issues for which it had a subscription for the hard copies. This serves as a deterrent to the cancellation of subscriptions. Such a policy is justifiable because a subscriber never had to pay for electronic access in the first place. Through many e-mail exchanges, I have the impression that librarians encourage this kind of access policy.

4.2 The Licence Agreement Form and Interlibrary Loans

If a publisher does not provide free electronic access, then in order to obtain electronic access, the university of the subscribing library and the publisher have to agree on a legal document, called the *Licence Agreement Form.* At the minimum, such a document specifies the *Scope of the License*, it contains a *Copyright* clause that limits the usage of files, a clause that relates cancellation of subscription to *Termination of Access* and, finally, it has an *Indemnity and Warranty* clause that protects the publisher from liability claims and, for the benefit of libraries, guarantees the integrity of files. Because we are dealing with regulating the use of new technology, all of these issues enter virgin territory, and, consequently, there is no uniform agreement on any of these clauses. The copyright clause is the most critical of these clauses, and publishers and libraries are changing, or modifying, their positions continually.

When the HJM established full internet access in 1998, we adopted more or less the same positions the AMS held at that time toward copyrights. They have changed their position, but ours continues to read, *"Printing and Downloading of the Electronic Version of HJM articles is permitted solely for Individual use by Authorized Users and only at Permitted Sites."*

This has been correctly interpreted as having the following ramification: Electronic Editions, in any form, are excluded from being accessible through *Inter Library Loans (ILL's).* My explanation for excluding files from ILL's is quite simple: There are no actual loans involved. The "lending" library is not giving up the use of anything (such as a copy of a periodical or a book) because it is on loan. The "borrowing" library, in attempting to exchange files, is in effect using IP numbers outside the scope of the license. Superficial paperwork involving these purported loans does not in and of itself validate them as comparable to traditional loans.

Some publishers do allow limited use of electronic files for the purpose of ILL's, namely for printouts of electronic files, *in lieu* of using photo copies from print editions. This sounds like a good idea. Libraries don't have to search for print editions, which might not even be available at the time of the request, or issues have been misplaced, missing, or mutilated etc.

However, there is another important argument against allowing printouts of electronic files - the possibility of contaminated content. While a publisher has the legal power to restrict or allow printouts of electronic files as ILL's, he has no control over how a library transmits those printouts (the same way he can't tell a library how to do its mail when traditional ILL's are involved).

The mode of transmission is what should concern us. It is possible that a library decides to digitally transmit documents by scanning a printout, using sophisticated character recognition software. As a possibility, a .pdf file could be created which would then be e-mailed to the "borrowing" library. However, a .pdf file that has been produced this way is almost always of inferior print and display quality compared to the original; but much worse, if the document contains mathematics, the character recognition software may have altered or compromised the content. A publisher cannot permit these compromised files to then get into circulation and be mistaken for originals. It should be clear, then, that if a publisher has made .pdf files available, then only those sanctioned files can be used for (legally limited) e-mail exchanges. I am not sure whether international copyright laws have addressed this point or not.

Some subscribing libraries have provided the HJM only with the IP numbers of their mathematics department members, thus bypassing conflicts they may have honoring ILL requests.

New products and technologies do not always fit seamlessly into old business practices. Liberal distribution of electronic files is not covered by current ILL copyright laws and therefore, unless specifically arranged otherwise with publishers, must be, by default, excluded. Like many other publishers, HJM considers electronic access a bonus for subscribing libraries, one which cannot be shared through ILL's.

Only very few libraries have questioned our journal's ILL policy. My general response to them has been that unlike commercial journals, HJM is so inexpensive that every library with some budget can afford our subscription. In all but one case, every library agreed with me, and then explained their initial opposition to our ILL policy stemmed from the fact that they were not aware that our journal, unlike commercial journals which are highly profitable cash cows for their publishers (see the New York Times [5] on that subject), was not trying to make more money from this ILL policy. In contrast, as a University journal, we only seek to recover our own production costs through modest subscription fees.

It may be the case that eventually State Agencies will negotiate with publishers licences for their whole system of public libraries, thus making electronic exchanges of files between such libraries unnecessary. In the UK, the *British Library* is considering a form of licensing that would allow for controlled access, that is, questioning the idea of unrestricted public access at a public library (cf. Sally Morris [7]. Another possibility would be that a consortium of libraries makes long term subscription committment to selected groups of journals, in exchange for making internet access free or a more liberal policy of ILL's. But before anything like this is going to happen, cheap access for everybody maybe the best solution.

4.3 Implementation and Administration of Restricted Internet Access

Independent journals which are affiliated with major universities generally have access to departmental servers, and this allows them to establish an Internet presence. HJM is very fortunate in that our mathematics department has superb computing facilities and a knowledgable IT staff. Thus we did not have to look outside (e.g. to Project Euclid [3]) for launching a Web presence. The development of the Web has been fast and to some extent pleasantly unpredictable. We only have to think about free global search engines and site searches provided by Google. Because of this, Websites of independent journals can match many features of expensive commercial ones.

While libraries can register directly with us for Internet Access it seems that they prefer to go through their subscription agencies. These agencies now keep a record of subscriptions with internet access and therefore can notify us about changes. Because of this co-operation, the transition from print to *Print plus Electronic Access* has been straightforward and smooth. Our experience very much sustains the claims made in Andrew Knibbe's article [6] on the increasing role of subscription agencies in electronic publishing.

5 Caveat the Adobe Reader

The .pdf format is intimately connected with the Adobe Software Company and two of its products, the *Acrobat* and its free *Acrobat Reader.* However the file format .pdf is open source and there are various products that convert LaTeX files into .pdf. Some are even free, and convert, for example, .ps files into .pdf. The paper by Ockerbloom [8] discusses in more detail the legal relationship between the .pdf file format and Adobe.

The quality of .pdf files is not always high and there are often problems with fonts and graphics; also, screen display and printing can be less than perfect. Unfortunately, no matter by what means LaTeX files have been generated into .pdf files, every new version of AR seems to cause new and unexpected problems for existing .pdf files which originated as LaTeX files. This is an unacceptable situation; the .pdf file format is well understood by a number of professionals who are also familiar with the TeX typesetting language and know about the needs of scientific publishers.

I feel that publishers and libraries should start thinking about the creation of an academic version of something like the Acrobat which would address these issues. According to the information I got from specialists (e.g., from Micropress), the initial development of a high quality reader of .pdf files would be in the neighborhood of $500,000 to $1,000,000, but probably lower. This would include further improvements of the .pdf format, reliable, high quality conversion to .pdf from LaTeX and other essentials. Upgrades of such a product would be much cheaper to produce, but upgrades are necessary to accommodate changing, or new operating systems. Because the .pdf format is open source, any number of developers could be "certified". For paper, government agencies have defined criteria for what constitutes "archival quality." For electronic files, *certification* poses a much more complex problem. It would involve, for example, a certification process for software packages that convert TeX into .pdf files, and another certification process for readers that are supposed to handle (display, print, search etc) such certified files.

While a certification process might be far away, it would be very helpful if the AMS, or other professional organizations, would on a regular basis evaluate and test the various software packages which are currently needed for scientific publishing. It is interesting that about 15 years ago, the AMS did something like that for scientific wordprocessors and printers. I wonder why the AMS is no longer providing this kind of badly needed service to the mathematical community.

I am afraid that the current situation of various well known (free or very inexpensive) patchwork solutions for electronic publishing will not work forever and that finding permanent solutions may need a stronger financial committment from all parties involved.

Nobody knows the future of the .pdf format, and the need for .pdf readers outside the journal publishing community. For the Web, the HTML format is continually improving and various application programs (e.g., for wordprocessing, spreadsheets, data bases etc) may converge to uniformly accepted, but diverse, file formats. Readers for such file formats may become part of Operating Systems, thus making plugins, like the AR, redundant. Thus, even if Adobe survives the current .com shakeup, there is no guarantee that it won't drop the Acrobat at one point in time. However, the .pdf format is a language and as such it will be understood as long as people are willing to learn it. Because publishers depend on the .ps and .pdf format they should make sure that this language stays alive, regardless whether other parties will have interest in it or not.

It has been suggested (e.g. [11, p. 924]) that one should safe-keep electronic documents in their most primitive format. However, it is even now not feasible to recompile periodically LaTeX source files into .pdf in order to make them compatible with the latest version of the AR. This is because LaTeX and some essential components, e.g., enhancements provided by the AMS, are not upward compatible, and older files have to be manually edited in order to compile under such upgrades. A quick calculation shows that this approach is cost prohibitive, even for smaller publishers like HJM. On the other hand, to collect $1,000,000 from publishers on the basis of published pages per year, would amount only to a fraction of their combined printing costs. HJM certainly would support an initiative to establish a monetary pool to be used for the further development and maintenance of the .pdf format, and to create standards that meet the specific needs of scientific publishers. As I see it, the only question is whether publishers and libraries can afford **not** to do so.

Traditionally, librarians have been the custodians of published material. This is still the case for printed matter but does not apply to documents in electronic format. The reason is simple: Libraries do not own files. Libraries may have access to individual files but in most cases, libraries are not allowed to download whole volumes. Thus, their ability to archive electronic material has been curtailed. On the other hand, whether publishers are willing, able, or even legally responsible, to archive and maintain files is untested territory. Publishers and libraries have to come together and find a common ground. Independent publishers may be ready to form partnerships with certain libraries to share the price and burden of file maintenance and archiving. However, it should be clear that the fate of electronic publications should not depend on the good will of a business oriented company, like Adobe, that has no vested interest in academic matters. I feel, more is needed than committees making recommendations. As I see it, the presence is already too worrisome to lose much thought about the future.

Conclusion

The electronic revolution has certainly changed the landscape of the publishing business. It is remarkable that every stage of this still ongoing process seems to have fostered the development of independent journals.

In order to stay competitive, every new technology forces a publisher to rethink its operating procedures. Because of their smaller size, better educated staff, and in many cases the ability to access state of the art computing facilities, many independent journals are well equipped to implement new and more efficient production methods. This explains, at least in part, their much lower journal prices.

For these reasons, I am quite optimistic about the future of independent journals. It will be interesting to see whether some editorial boards of commercial journals decide to go independent, or at least change to an academic publisher, as it has already happened in the well publicized case (cf. [2]) of a prominent journal in computer science.

References

1. Barr, Michael: Where Does the Money Go. Newsletter on Serials Pricing Issues, No. 229 July 13, 1999
2. Birman, Joan S.: Scientific Publishing: A Mathematician's Viewpoint. Notices of the AMS **47**(7) (2000) 770–774
3. Project Euclid. Mission Statement and Project Description. Cornell University. File available at `http://projecteuclid.org`
4. Kiernan, Vincent: Why Do some Electronic-Only journals Struggle While Others Flourish. Chronicle of Higher Education, **1999**, reprint in The Journal of Electronic Publishing June 1999, Vol. 4, Issue 4. File available at `http://www.press.umich.edu/jep/04-04/kiernan.html`
5. Kirkpatrick, David D.: As Publishers Perish, Libraries Feel the Pain, Mergers Keep Pushing Up Journal Costs. New York Times, November 3, 2000.
6. Knibbe, Andrew: A Subscription Agent's Role in Electronic Publishing. The Journal of Electronic Publishing, June 1999, Vol. 4, Issue 4. File available at `http://www.press.umich.edu/jep/04-04/knibbe/html`
7. Morris, Sally: Archiving electronic publications: What are the problems and who should solve them. The Association of Learned and Professional Society Publishers. File available at `http://www.alpsp.org/arcsm00.pdf`
8. Ockerbloom, John Mark: Archiving and Preserving PDF Files. RLG DigiNews, Vol. 5, No. 1, 2001. File available at `http://www.rlg.org/preserv/diginews/diginews5-1.html`
9. O'Donnel, M. J.: Electronic Journals Scholarly Invariants in a Changing Medium. Paper presented at the *Seminars on Academic Computing* 1993. File available at `http://people.cs.uchicago.edu/~odonnell`
10. Strong, William S.: Copyright in the New World of Electronic Publishing. Paper presented at the workshop of Electronic Publishing Issues II at the Association of American University Presses (AAUP) Annual Meeting, June 17, 1994, Washington, D.C. File available at `http://www.press.umich.edu/jep`
11. Committee on Electronic Information Communication of the International Mathematical Union: Best Current Practices: Recommendations on Electronic Information Communication (2002). Notices of the AMS, Vol. 49, No. 8 (2002) 922–925

Math-Net International and the Math-Net Page

Wolfram Sperber

ZIB Berlin
Takustrs.7
D-14195 Berlin
sperber@zib.de

Abstract. Information and communication have become an increasingly important part of mathematical research and teaching in recent years. The electronic medium, especially the Web, has changed the technical base for information and communication dramatically. To use the Web is very advantageous for the mathematical community: mathematicians can publish research results, scripts, software, etc. immediately and without additional costs. For this purpose an organizational infrastructure is needed to coordinate the activities of the mathematical community in the field of electronic information and commmunication so that the mathematical information on the Web is searchable and accessible in an efficient way.
The International Mathematical Union (IMU) has established the Committee on Electronic Information (CEIC) following a resolution of the IMU General Assembly in Dresden 1998. CEIC has worked out recommendations for the use of information and communication. An other activity which is steered by CEIC is Math-Net. Math-Net is a distributed information and communication system designed for the mathematical community. Math-Net and especially the Math-Net Page are described in this contribution in more detail.

1 The Math-Net Project

The Math-Net project in Germany (1997-1999) was the starting point of the Math-Net activities. At this time many mathematical institutions, e.g., departments and institutes, started to build up own Web sites. The most important aims of the Math-Net project were to

- establish high-quality Web sites of mathematical departments and institutes for all kind of mathematically relevant information;
- install a network of persons organizing and maintaining the Web sites of departments and institutes;
- develop methods and tools to standardize the information and the description offered by the mathematical institutions;
- develop services to make the distributed information searchable and accessible in a user-friendly and effective way.

The project resulted in a distributed information and communication system focusing to the Web sites of mathematical departments and institutes in Germany and Austria.

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 169–177, 2003.

2 CEIC and Math-Net

Mathematicians, engineers, economists, students, pupils, and teachers are interested to have access to comprehensive, up-to-date, and cost-saving mathematically relevant information. The information does not only mean mathematical publications and textbooks but also algorithms, mathematical software, e.g., computeralgebra systems, information about projects, persons etc.. And the access to the information must be efficient.

There is a clear need for support and for international coordination of the activities of the mathematical community in this field. Therefore, the International Mathematical Union (IMU) has established the Committee on Electronic Information and Communication (CEIC) in 1998.

CEIC has

- worked out recommendations on electronic information and communication to mathematicians, libraries and publishers, see [1];
- declared the further development and internationalization of the Math-Net activities to be a major aim of its work. The internationalization affects the organizational as well as the conceptual and technical development of Math-Net.

The aims, the principles, and the organization structure of the Math-Net Initiative are formulated in the Math-Net Charter, see [2]. Math-Net bases on voluntary contributions of the Math-Net Members. The use of Math-Net is free. A worldwide mathematical information and communication system has to be distributed and open.

All mathematicians, mathematical institutions, and other information providers, e.g., libraries, publishers or software companies provide mathematically relevant information on their Web sites.

If this is done in a standardized way, Math-Net Services are able to gather, process, and make acessible this information.

The Math-Net Initiative, especially its Technical Advisory Board (TAB), see [3], defines the concepts and develops methods and tools for such a system - on the base of the general developments and trends in the Web. The Web of the future, the Semantic Web, is a semantic-based knowledge system. Standardized semantic metadata are the base for such a system enabling a cross-linking of resources, powerful services and guaranteeing an efficient search.

In the beginning the Math-Net activities are focused on the following topics

- the Math-Net Page, a portal to the core information of mathematical institutions, and the Math-Net Navigator as the corresponding service,
- preprints and the Mathematical PREprint Search System (MPRESS),
- professional homepages for people and PERSONA MATHEMATICA.

The activities cover particularly a standardized metadata description of the resources.

3 Web Resources of Math Departments and Institutes and the Math-Net Page

More than 1,000 mathematical departments and institutes worldwide provide relevant and high-quality information on own Web sites (see the relevant Web Sites in Google [4] or the list of the Pennsylvania State University [5].

The Web sites of the departments cover particularly the information about the institution itself, research, teaching and study opportunities, etc..

Homepages are the portals to Web sites. The homepages of mathematical departments and institutes are often beautifully designed and structured. But they differ dramatically in structure and design, a situation that is rather user unfriendly.

The Math-Net Page is an attempt to improve this situation.

Math-Net Pages should not replace the original homepages of the institutions. Rather they are intended to be as “secondary” homepages for the Web sites of mathematical departments and institutes in addition to the original homepages. As a result the navigation over the Web sites of Math-Net Members is simplified for the human user.

To design the Math-Net Page the existing Web sites of mathematical institutions were analyzed and a schema for the core information was defined. It covers the following six groups:

- General
- People
- News
- Research
- Teaching
- Information Services

Each of these groups has one or more subgroups.

The schema defined by the groups and subgroups was used to develop a proposal for a uniform homepage of mathematical institutions standardized in structure, vocabulary, semantic description, and layout. Fig.1 shows the English Math-Net Page of the TU Munich.

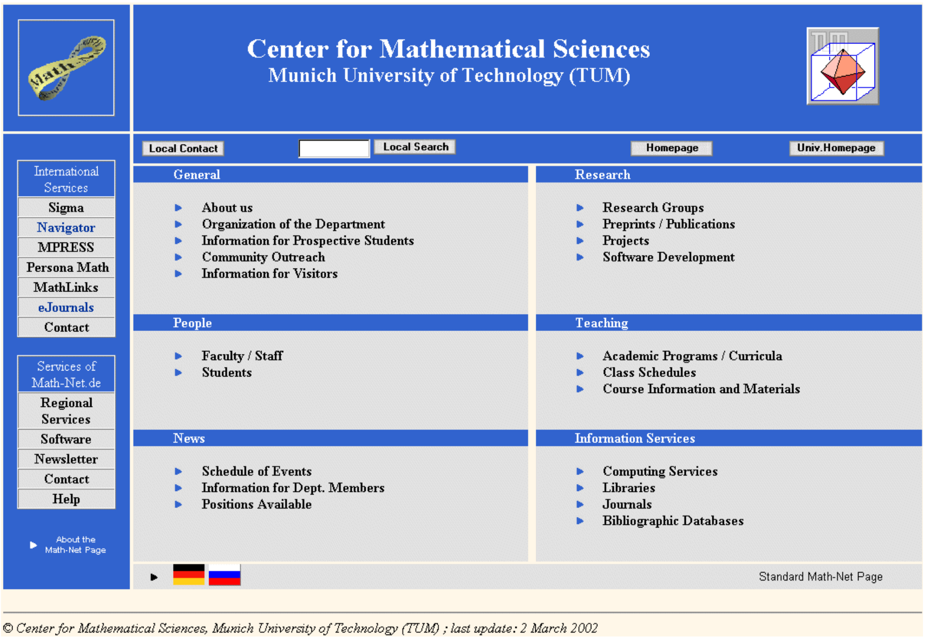
The head of the Mat-Net Page contains the name of the institution, a logo, links to other Webpages, a local search engine if existing, and a contact address.

The main part of the Math-Net Page, the groups and subgroups are arranged on the right hand side, covers links to the information resources of the institution.

On the left side of the Math-Net Page a column provides links to the current Math-Net Services.

3.1 Standard and Alternative Math-Net Pages

The Math-Net Page in Fig. 1 is a so-called “Standard” Math-Net Page. The attribute “Standard” means that there is used a controlled English vocabulary. It was a long-term task to agree on the vocabulary to be used on the Standard Math-Net Page because



Center for Mathematical Sciences
Munich University of Technology (TUM)

© Center for Mathematical Sciences, Munich University of Technology (TUM) ; last update: 2 March 2002

Fig. 1. The Standard Math-Net Page of the Technical University Munich

- the typical names of the groups and subgroups differ between regions and countries;
- the internationalization of Math-Net must admit also Math-Net Pages in languages other than English.

Therefore “Alternative” Math-Net Pages with customized labels were defined besides the “Standard” Math-Net Pages.

3.2 Different Character Encodings

Math-Net Pages in different languages use different character encodings. Worldwide various character encodings are in use, e.g., Latin-1 for West European characters, KOI8-R for Russian cyrillic characters or gb2312 for Chinese character. Unicode is a universal character set which covers most other character encodings (there exist one-to-one mappings between Unicode characters and those in other encodings).
For a broad acceptance the Math-Net Pages should be created in the character encoding that is normally used on the Web sites of the institution.

3.3 Metadata on the Math-Net Page

Math-Net Pages have a uniform layout. They have to allow also a semantic processing of the Math-Net Pages by machines.
Methods and tools for a semantic processing of information, e.g.,

- metadata,
- the Resource Description Framework (RDF),

have been developed in the Web since the mid-nineties. Metadata are in principle not more than "data about data".
Different metadata sets are used for different data. So the DC metadata, see [6], define a set of 15 elements plus qualifiers for a standardized bibliographic description of electronic documents. DC metadata cover, e.g., the title, the creator, or the subject of a resource. vCard, see [7], is a metadata set for the description of information about people.
RDF, see [8], is a general model and syntax to encode metadata. RDF can be presented in Extensible Markup Language (XML).
The Math-Net Page uses different metadata schemas and RDF to present the content of the Math-Net Page in a machine-understandable way.
For more detail about the metadata on the Math-Net Page see [9].

3.4 The Math-Net Page Maker

It is expensive to create a Math-Net Page manually. Misspellings in the XHTML part can produce trouble in the layout of the Math-Net Page, errors in the RDF part produce trouble in the automatic processing of the Math-Net Pages.
For that reason a tool is needed to generate a Math-Net Page in an easy way. The Math-Net Page Maker, see [10], is a form-based tool to generate a Math-Net Page. The user has to insert several data in a cgi-script via a form.
The script generates the complete Math-Net Page.

3.5 The Information Coordinator

An institution taking part in Math-Net has to appoint an information coordinator. The information coordinator is responsible for organizing and maintaining all Math-Net activities of the institution. Especially, the information coordinator should create and update the institution's Math-Net Page(s).

3.6 Acceptance of the Math-Net Page

On 2002-04-12 CEIC and IMU passed a recommendation to all mathematical departments and institutes to create and install Math-Net Pages, see [11]. In the meantime more than 180 departments and institutes from all parts of the world have created and installed Math-Net Pages.

3.7 Prospects

Of course, the concept of the Math-Net Page is not restricted to mathematical departments and institutes. A Math-Net Page for mathematical societies is currently under discussion.
The idea of the Math-Net Page can be used to model portals for other types of Web sites. Especially, the idea can be used to develop a proposal for a standardized professional homepage, see the paper of Kaplan in this volume, or for Web sites of research projects.

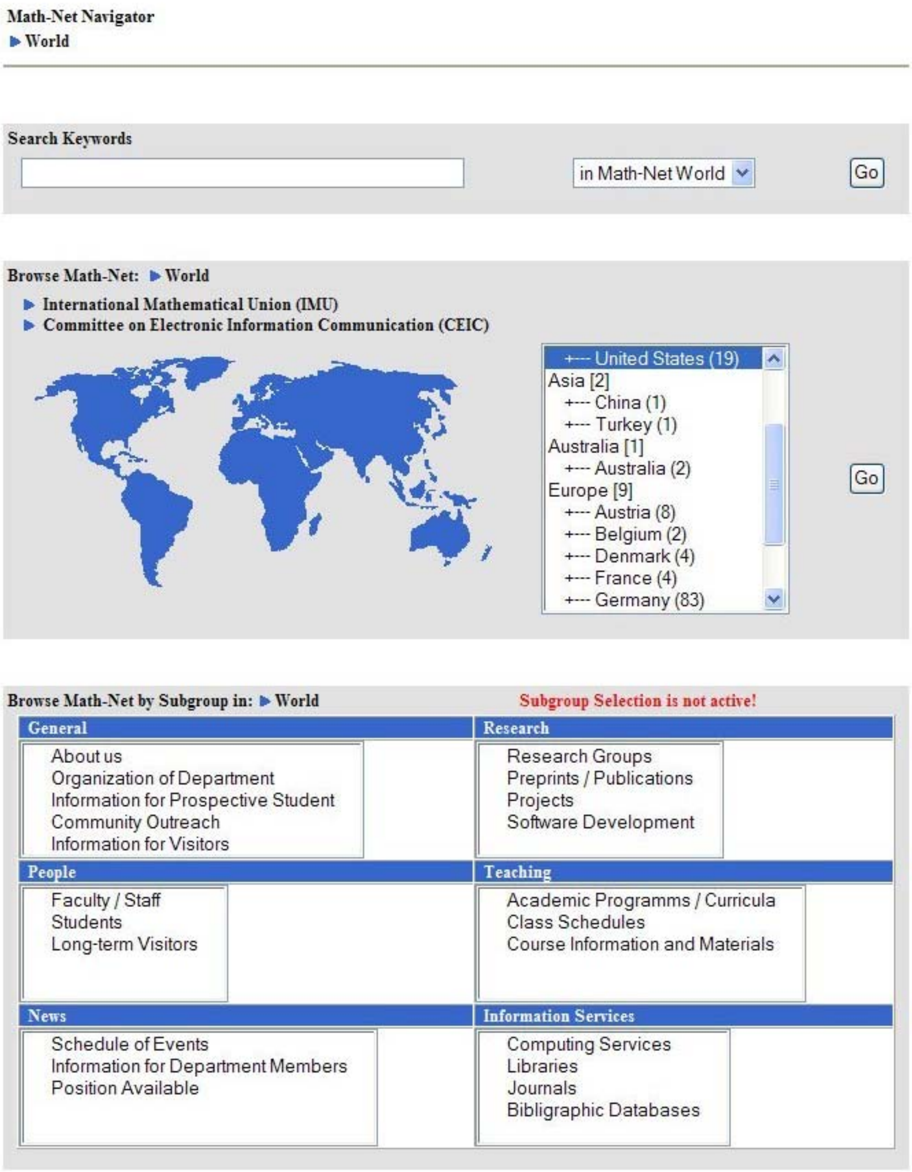
4 Math-Net Services

The institutions taking part in Math-Net make their information resources electronically available in a standardized fashion. The are fully responsible for the quality, accuracy, timeliness, and appropriateness of the data they contribute.
The data of the Math-Net Members are gathered and combined into services. The Harvest software, see [12], is a tool to collect automatically information from the Web and process it.
As a result, the Math-Net Services provide a fast and well-structured access to the mathematical resources within Math-Net.
A high-quality semantic annotation of the objects is necessary to allow high-quality services.
The first Math-Net Service was a preprint index. Today, the service MPRESS [13] is the most comprehensive index of mathematical preprints listing more than 50,000 preprints from 150 different Web sites. PERSONA MATHEMATICA is the Math-Net Services that gathers the data of persons.

4.1 The Math-Net Navigator: Processing of Math-Net Pages

The Math-Net Navigator [14] is the service processing the Math-Net Pages. Therefore, the navigator must be able to process the Math-Net Pages, i.e., the metadata encdcoded in RDF/XML, different character encodings, etc..
The gathered data are stored in a database. The URL of each new Math-Net Page has to be inserted in the database manually. Then the complete Math-Net Pages are daily gathered and processed automatically.
The Math-Net Navigator combines search and navigation at different levels. In more detail, the Math-Net Navigator provides on its top level, see Fig. 2

- a search function (over the servers of all Math-Net Members worldwide);
- geographical-oriented views of the Math-Net Members covering a clickable map and a country list;
- subject-oriented views of the information provided by the Math-Net Members (e.g. lists of all Math-Net Members which provide preprints).



Math-Net Navigator
▶ World

Search Keywords

in Math-Net World ▾ Go

Browse Math-Net: ▶ World

▶ International Mathematical Union (IMU)
▶ Committee on Electronic Information Communication (CEIC)

+--- United States (19)
Asia [2]
+--- China (1)
+--- Turkey (1)
Australia [1]
+--- Australia (2)
Europe [9]
+--- Austria (8)
+--- Belgium (2)
+--- Denmark (4)
+--- France (4)
+--- Germany (83)

Go

Browse Math-Net by Subgroup in: ▶ World Subgroup Selection is not active!

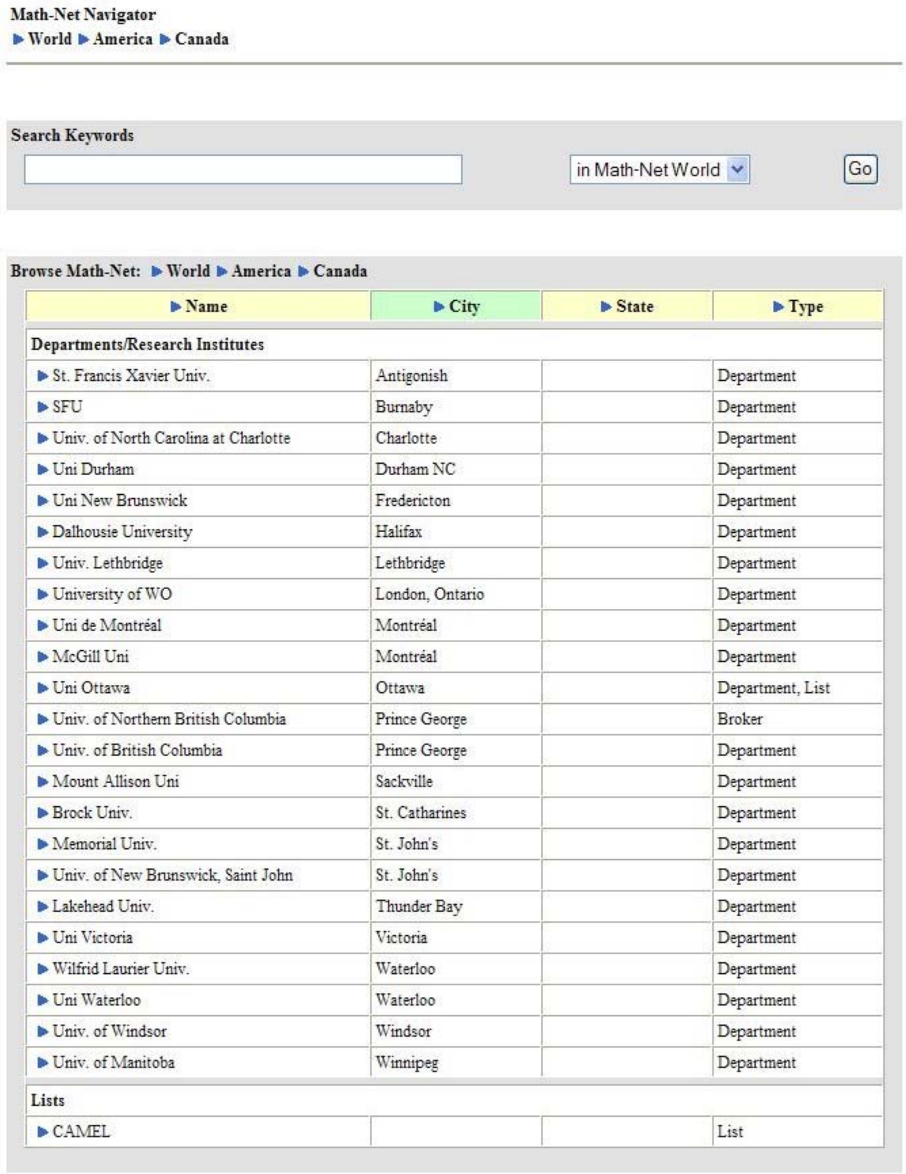
General	Research
About us Organization of Department Information for Prospective Student Community Outreach Information for Visitors	Research Groups Preprints / Publications Projects Software Development
People	**Teaching**
Faculty / Staff Students Long-term Visitors	Academic Programms / Curricula Class Schedules Course Information and Materials
News	**Information Services**
Schedule of Events Information for Department Members Position Available	Computing Services Libraries Journals Bibligraphic Databases

Fig. 2. The top level of the Math-net Navigator

A country (e.g., Canada) view provides the same functionality shown on top level namely

- a search function (over all Canadian Math-Net Members);
- a list of Math-Net Members in Canada;

– subject-oriented views of the information provided by the Math-Net Members (e.g., lists of all Math-Net Members offering preprints). [1]



Math-Net Navigator
▶ World ▶ America ▶ Canada

Search Keywords

Browse Math-Net: ▶ World ▶ America ▶ Canada

▶ Name	▶ City	▶ State	▶ Type
Departments/Research Institutes			
▶ St. Francis Xavier Univ.	Antigonish		Department
▶ SFU	Burnaby		Department
▶ Univ. of North Carolina at Charlotte	Charlotte		Department
▶ Uni Durham	Durham NC		Department
▶ Uni New Brunswick	Fredericton		Department
▶ Dalhousie University	Halifax		Department
▶ Univ. Lethbridge	Lethbridge		Department
▶ University of WO	London, Ontario		Department
▶ Uni de Montréal	Montréal		Department
▶ McGill Uni	Montréal		Department
▶ Uni Ottawa	Ottawa		Department, List
▶ Univ. of Northern British Columbia	Prince George		Broker
▶ Univ. of British Columbia	Prince George		Department
▶ Mount Allison Uni	Sackville		Department
▶ Brock Univ.	St. Catharines		Department
▶ Memorial Univ.	St. John's		Department
▶ Univ. of New Brunswick, Saint John	St. John's		Department
▶ Lakehead Univ.	Thunder Bay		Department
▶ Uni Victoria	Victoria		Department
▶ Wilfrid Laurier Univ.	Waterloo		Department
▶ Uni Waterloo	Waterloo		Department
▶ Univ. of Windsor	Windsor		Department
▶ Univ. of Manitoba	Winnipeg		Department
Lists			
▶ CAMEL			List

Fig. 3. The Math-Net Navigator for Canada

[1] not in the cutout

Of course, the different tools and services of Math-Net provide different aspects of mathematically relevant information. The tools have the same core, the services work together and base on the same design.
All services are available via the central Math-Net portal, see [15].
The present Math-Net, however, is only a first step towards a better information and communication system. Math-Net needs conceptual, organizational, and technical sophistication. And it needs the broad support by the mathematical community. Please take part actively in the Math-Net Initiative!

References

1. The International Mathematical Union, Committee on Electronic Information and Communication: Recommendations on Information and Communication, General Assembly, Shanghai, August 2002, also electronically available at http://www.math-net.org/Math-Net-Recommendation.html
2. http://www.math-net.org/charter
3. http://www.mathematik.uni-osnabrueck.de/TAB
4. http://www.google.org
5. http://www.math.psu.edu/MathLists/
6. http://dublincore.org/documents/dces/
7. http://www.imc.org/pdi
8. http://www.w3.org/RDF/
9. Semantic Annotation in Mathematica and Math-Net, to appear in "Semantic Annotation for the Web", IOS Press
10. http://www.math-net.org/pagemaker
11. http://www.math-net.org/Math-Net-Recommendation.html
12. http://webharvest.sourceforge.net/ng/
13. http://mathnet.preprints.org/
14. http://www.math-net.org/navigator
15. http://www.math-net.org

EMANI – A Project for the Long-Term Preservation of Electronic Publications in Mathematics

Bernd Wegner

Mathematisches Institut, TU Berlin
wegner@math.tu-berlin.de

Abstract. With the rapidly growing activities in electronic publishing ideas came up to install global repositories, which deal with three mainstreams in this enterprise: storing the electronic material currently available, pursuing projects to solve 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. Hence in this domain some pioneering projects have been established trying to tackle the above-mentioned problems.

The article will describe some of these activities and touch the plan to develop a global Digital Mathematical Library (DML). As a special project for mathematics in the archiving area the Electronic Mathematics Archives Network Initiative (EMANI) had been designed. Having in mind that a distributed architecture would be more suitable and reduce the load on the partners for such a project, a network is proposed, which also might be a more open approach for extending the project from a initially restricted solution to a more comprehensive enterprise. For the core of the network, a co-operational system of reference libraries and content providers like publishers and editors has been be set up. On the side of the libraries the following partners have agreed to set up a prototype for the archive: the Tsinghua University Library in Beijing, the Cornell University Library in Ithaca (N.Y.), the French partner Math-Doc and the Lower Saxony State and University Library in Göttingen. The first group of content providers consists of the Springer publishing house, associated publishers and journals posted in the Electronic Library of EMIS. This report refers to the state of EMANI after the fourth EMANI workshop in Paris at the beginning of April 2003.

1 Electronic Offers and Their Providers

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

F. Bai, B. Wegner (Eds.): ICM 2002 Satellite Conference, LNCS 2730, pp. 178–??, 2003.

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. I 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 extend 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. In contrast to the "old world" of printed publications these providers have different aims and it is not always clear for the user what he really could expect from these services, when he is searching for some information or article of his own interest. 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.

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. Even 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. Commonly libraries cover a broad area of subjects and within that they try to be relatively comprehensive. Independent from the frequency of their usage these holdings had been preserved and kept accessible with great care. The objectives of science libraries are user-oriented on one side. On the other side libraries feel obliged to protect the treasure of knowledge they have accumulated in their collection. This makes them also the best choice for solving the problem of the long-term preservation of electronic publications.

Mathematics is a science where the availability of electronic publications and retro-digitised documents lead 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.

2 Citations in Mathematical Articles

This section should provide some arguments why long-term preservation and availability of publications is of particular importance. For non-mathematicians it is not clear at all that mathematics is so much different from other sciences as far as easy availability of older publications will be concerned. For some it is even hard to understand the subjects of mathematical research and the special way how this research is published. For example, extensions and improvements of older results only care about the publication of the additional achievements, and there detailed proofs are essential. Older results may and should be cited, but it is not honest to repeat their proofs in research publications, even if the understanding of these proofs is essential for seeing what the new results are about. Many proofs can be found at one place only. Hence an article is just an addition to a sequence of other articles, more or less tightly interrelated in a structure, which combinatorically is more complex than a tree. It provides another shell to a core of theorems, propositions, examples, models and proofs representing the current knowledge of a subject domain in mathematics. Mathematical research articles commonly are rather efficient in their presentation, and the average publication frequency of a mathematician is quite low compared to other sciences.

Admittedly, parts of research domain in mathematics may be exhibited comprehensively in monographs, but as can be seen by the variety of material in the few research surveys in mathematics (the Itogi Nauki published by VINITI, for example) such monographs with detailed exhibitions of arguments only can cover a part of the domain of reference, giving a motivating introduction with proofs, while the surveys have no space to provide proofs at all, if they really want to be comprehensive. This underlines that references in mathematical papers are not just a matter of honesty, but that at least an essential part of them plays an important role for a complete understanding of the content of an article. The following figures may give a good evidence for the need to have also older mathematical publications available.

We want to consider the case of three journals. The figures are cited from an investigation by Joachim Heinze [6]. The most surprising figures (also surprising to mathematicians) are the numbers of citations before 1992. In the case of the most traditional mathematical journal from Northamerica, 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 cites 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. It would be quite interesting to have a more comprehensive comparison of this type.

3 Current and Future Problems

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 world-wide 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. What everybody experiences with his old releases of word-files, became true meanwhile for the readers of PDF-files, for example. Without conversions, if tools for these exist at all, or simultaneous installation of several versions of the Acrobat Reader the whole range of PDF-files produced in the period, where the Acrobat Reader was offered, is not readable anymore using the last release. 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 there is a variety of plug-ins, which depend on additional software potentially offered with an electronic document. 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 documents in mathematics are considered, because they are most likely to have software depending enhancements. Interactive documents will play an important role in the future. Furthermore, 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.

Hence the best chance to tackle the problem is an open and subject oriented approach as it is described with EMANI in the next section.

4 The EMANI Project

There is a period of approximately 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. As mentioned above, older publications are still of big value for research in mathematics. Hence retrospective digitisation projects increased the current digital content in mathematics considerably. One major of these projects is ERAM (see [3] or [8]).

In the first half of 2001, the Electronic Mathematics Archives Network Initiative (EMANI) had been founded as a special project 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 is proposed, which also might be a more open approach for extending the project from an initially restricted solution to a more comprehensive enterprise. The initiative has been formalised in July 2002 at the EMANI workshop at Cornell University with the partners described in the next section as a first set of members and the author of this article as the co-ordinator of the co-operation.

For the core of the network a co-operational system of reference libraries and content providers like publishers and editors has been set up. In the ideal final version they are supposed to serve for a long list purposes: The basic action will be to store the digital content in mathematics from the content providers at the reference libraries. This will be complemented by retro-digitising all 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 measures can be undertaken to care about the long-term preservation of this content in readable form. First projects for the technical support concerning metadata and access design and the handling of TEX-files have been just initiated.

For example, to have the content stored somewhere will not be sufficient. On the basic level retrospective digitisation will lead to scanned images only, which hopefully can be accessed in some repository. As an important enhancement it will be necessary to improve the usability of the retro-digitised publications by introducing advanced linking and searching facilities and to provide convenient and affordable access to the stored content for mathematicians and professionals using mathematics world-wide.

In the ideal case the reference libraries for EMANI will serve as a backup system for other libraries, which want to store and provide part of the content or refresh their existing offers by updated material. Having in mind the long time scale of the publications to be provided through the network, which go from articles from the 19th century to current publications, a system of distribution agents will be needed. Extensions of the current group may even bring older publications into the system. This may be a good reason to develop new business models for a distribution of mathematical publications in a combined enterprise between reference libraries and content providers.

5 The Initial Phase of EMANI

Such a complicated enterprise like EMANI should start on a smaller well-controllable scale at first only. Once the architecture and the action plan will have been made sufficiently precise, an extension may be considered. The current partners collaborating for the first steps in order to implement the initiative on the side of the libraries are:

- The Cornell University Library, Ithaca, N.Y.: They have a good tradition in retrospective digitisation projects and are involved in the archiving discussion for other sciences also. In particular they are building up an offer of a bundle of electronic journals in mathematics through project Euclid. They serve as a mirror site for EMIS (see [9]).
- The State and University Library Göttingen: Also there some important retrospective digitisation projects like ERAM (see [3] or [8]) and DIEPER are pursued. In addition to this the SUB Göttingen is 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: 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.
- The French partner MathDoC, being a combination of the Orsay Mathematical Library, Paris, and the Cellule MathDoc at UJF in Grenoble: The group in Orsay is co-ordinating a quite comprehensive consortium of French mathematical libraries. The strength of the partner in Grenoble consists in their excellent retro-digitisation project NUMDAM ([4]).

The content providers for the start are Springer-Verlag, Birkhäuser 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 are looking back to a long tradition in publishing mathematics. They are in charge of several of the best journals in mathematics. In contrast to this the ElibM is a co-operation of several journals and editors on a voluntary basis bundling electronic offers in a world-wide system of WWW-servers (see [9]).

First agreements on the architecture of the system have been made. It is common understanding that the storage of the content in a repository will have priority in the near future and that in general copies of the content stored in the system should be deposited at all reference libraries as a matter of safety. Later on also refreshed versions of the content should be exchanged accordingly. It has been also approved that the partners of the initiative will provide their own achievements to support the aims mentioned above as far as possible. But this will become important in a later phase of the project.

An important step in the first phase of the initiative consists of the stepwise transfer of the available electronic content from the content providers to the reference libraries. The libraries will check if the files still can be used for the archiving, adjustments will 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. Hence first tests had been done with files of journal articles.

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. Clearly, display formats like PDF, Postscript, DVI etc. could be reproduced easily from the TEX sources on the fly in principle, where TEX stands for several TEX-dialects. But on the practical side a variety of TEX versions will have to be handled, including different style files, and the production of the journals is based on several companies with presumably different TEX installations. This may lead to different output for the representation format, if the same TEX installation is applied to compile all these files. Obviously the library partners have to develop some expertise to handle these problems. This will be one of the important projects for the next future.

Also new archiving related metadata have to be defined, and an integrated access structure satisfying the needs of all kind of experts who want to work with the archive will be one of the central achievements of the further work in the future. Though links from reference databases could satisfy many of the needs of the mathematicians to get access, the professional handling of the archives will require more than just the metadata used by these services. A lot of information about some technical parameters and the status of the articles will have to be stored.

It has to be taken into account that only some core content of digital publications could be preserved for the future, like it is available from printed versions at present. Preservation of electronic enhancements like links, animations, interactive facilities etc. has to deal with external providers or additional software in the background. This may be impossible in some cases and needs tremendous extra efforts in most cases. Generally it may not be affordable or necessary to keep these additions running for the preservation of the knowledge presented by such articles. For some appropriate replacements may be considered, but then the result should be stored as a new version of the article to comply with the general permanence principle for electronic publications in mathematics. Anyway, it will be worth-wile to check for every article, if problems of this kind may occur, and to decide on a frozen version for long-term archiving if necessary.

The first 18 months of EMANI were guided by 4 workshops in Heidelberg, Ithaca, Göttingen and Paris and by meeting of subgroups at other events. While the first two workshops were important to discuss the general ideas of EMANI and to find appropriate ways of co-operation, the last two dealt with the progress of the work for the different work-packages and several deliverables, which were needed to decide on the further activities. Hence the structure and the work plan for EMANI became quite precise. There is an Executive Board with one representative from each partner and the co-ordinator as chair. An Advisory Board is under construction. The technical and conceptual work will be done by a coordinating compiler group and by several work-packages dealing with 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.

In co-operation with ERAM and NUMDAM considerable progress has been obtained for digitising some of the most important journals in mathematics, which will be accessible from the EMANI partners. Details are described in Section 7.

6 Some Other Archiving Projects

To keep the article short only a rough survey should be given here. Clearly, several approaches to deal with the digital archiving problem are possible, and they have to involve all parties, publishers, libraries and authors. Only a few people still believe, that just copying and storing the articles with their whole web-environment where they are posted is a promising solution. In 2001/2 several of such co-operations started on a more advanced level, trying to tackle the problem with different models. Most of them involve libraries. Here are some examples.

Most prominent is the co-operation between Elsevier and the libraries at Yale University and a Dutch library, caring about the digital preservation of all publications of this big publisher. Harvard University works on the same with Wiley, Blackwell Science and Chicago University Press. LOCKSS is a system of archiving sites co-ordinated by Stanford University. Through the project Harvest Cornell University is involved in the archiving of publications in agriculture. MIT has dedicated some efforts to a special type of electronic publications, the dynamic documents. The New York Public Library is working on the digital preservation of arts journals. 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. An archive for the digital preservation of publications in astronomy and astronomical data has been developed at the Smithsonian Astrophysical Observatory at Harvard. What all of these projects have in common is, that they represent a first approach only and that nobody has a comprehensive solution.

7 Digitisation Projects in Mathematics

Also for older documents searchability will be an important requirement to enable the researcher to find his way in the huge knowledge base of mathematical achievements. Admittedly, no current search engine is able to locate a mathematical result, making its abstract meaning accessible to a user being interested in applying it for his own investigations. Names for some of them will help, and classification codes of special subject areas will restrict the set of documents where to look for the desired information considerably. Hence literature databases for the classical period of mathematics are desirable. They should offer the same facilities like the current literature information services in mathematics, and even more, they should also provide links to the future given by modern mathematics. This was the starting point for the project ERAM, which also will be called the Jahrbuch-project for short.

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). All data from the Jahrbuch will have been keyboarded until end of 2003. They are made accessible in preliminary form in the web, and though for many items the enhancements like English keywords are still missing, the database has found a lot of grateful and satisfied users. The JFM-database will provide access to a digital archive to be built up within the project. The articles are offered in a good presentation format, but facilities for text searches are still missing. Applying OCR to mathematical texts still is confronted with big problems, when formulas should be handled. A first step for a solution of this problem is made by a project based on the co-operation of experts from Japan, Germany and the United States (see [7]).

In co-operation with EMANI ERAM has digitised all the back volumes of Mathematische Annalen, Mathematische Zeitschrift and Inventiones Mathematicae until these journals were posted electronically. The digitisation of the Commentarii Mathematici Helvetice has started. Furthermore, most of the journals, which have installed recent electronic versions in EMIS (European Mathematical Information Service), agreed that all of their print-only back volumes could be digitised and offered within ERAM, and this also has been done. In ERAM, about 950.000 pages have been scanned so far, and the capacity of the project will be sufficient to go for about 1.2 million pages. For more details see the references [3] and [8], or the ERAM-homepage under http://www.emis.de/projects/ clicking on the box for the Jahrbuch.

ERAM could be considered as a part of a global initiative to have all mathematics digitally available. The initiative is called DML (Digital Mathematical Library), and the only current activity is a discussion pursued by a planning group, how such a project could be arranged. DML has a lot of overlap with EMANI, but in contrast to EMANI the global initiative at first will concentrate 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].

At present a lot of efforts are undertaken to get the funding for national digitisation projects for mathematics in Europe. There is the idea to coordinate these activities under a European umbrella, DML-EU. On the world-wide level, the starting point for the discussion of the global DML in 2001 was the White Paper [5] by John Ewing, Executive Director of the American Mathematical Society. The article contained a lot of structural considerations for the DML, and it also addresses the immense problems we will be confronted with when we really want to pursue such a project. As a preliminary project funded by the American NSF a planning group has been formed for the DML and developed work-packages similar to those of EMANI. The planning group will have a final meeting during the funding period in Göttingen at the end of May 2003. The reports requested from the work-packages at the meeting in July 2002 in Washington will be available there. They may serve as a kind of programme how to go on with the DML initiative, though no further funding on a more global level is visible at this moment. Nevertheless the discussions in the planning group have lead to a structure, which is suitable to serve as a common forum for ongoing and future digitisation activities in mathematics.

References

1. Albis, Victor: 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. 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. Ewing, John: Twenty Centuries of Mathematics: Digitizing and disseminating the past mathematical literature. http://www.ams.org/ewing/Twenty_centuries.pdf
6. Heinze, Joachim: 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. Michler, Gerhard: How to build a prototype for a distributed digital mathematics archive library. Proceedings MKM 2001, Linz
   http://www.emis.de/proceedings/MKM2001/
8. 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

9. Wegner, Bernd: 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

Author Index