Twenty centuries of mathematics:

Digitizing and disseminating the past mathematical literature

John Ewing, Executive Director, American Mathematical Society

"If you have built castles in the air, your work need not be lost; that is where they should be. Now put the foundations under them."
-Henry David Thoreau, Walden, Chap.18

Mathematicians have talked quietly for some time about the need to digitize the past mathematical literature. During 2001, the conversations became more intense as several new digitizing projects were announced. Should we coordinate those projects? Could we integrate the recent literature that is already in digital form? How could we digitize far greater amounts of older material? The goal was to create a virtual library containing much of the past literature--a library that could eventually grow into a "World Mathematics Library"

In a conversation in mid-2001, Philippe Tondeur (the Director of the Division of Mathematical Sciences at NSF) outlined his vision for such a library. While I was sympathetic, I pointed out that one needed a plan, or at least an outline, and that even with a plan there were many obstacles. Philippe persuaded me to write this "concept paper" based on our conversation, and consequently turned my pessimism into a proposal.

Since that time, a group headed by Cornell University was awarded a planning grant to consider the next steps in carrying out a massive digitizing project. Mathematicians and agencies from other countries have expressed interest in an international effort. And the impossible sums of money needed for funding seem almost possible (even if most of the other obstacles remain).

The opinions expressed in this paper are the author's, and do not necessarily represent opinions of the American Mathematical Society.

Mathematics has always relied on its scholarly literature. From the time of Euclid's Elements[1] (about 300 BC), mathematics thrived because key literature was passed from generation to generation. In modern times, the process accelerated, changing the way mathematicians carry out research. Because it is impossible to study and digest all relevant literature in a broad area, mathematicians find themselves navigating the literature--moving from one paper or book to another, perusing results and proofs, and relying on references in order to link to the next item. The linking process has become

[1] "The *Elements* form, next to the Bible, probably the most reproduced and studied book in the history of the Western World. More than a thousand editions have appeared since the invention of printing, and before that time manuscript copies dominated much of the teaching of geometry." [Struik, Dirk J. *A Concise History of Mathematics*, 4th ed, Dover, New York, 1987, p. 49.]

more important as the literature has grown, and it is one of the reasons electronic publication has great potential benefit for mathematical research.

Reliance on past literature is common to all disciplines, but time scales differ. In some areas of science, literature more than a few years old has value mainly for historical reference. For mathematicians, work from ten, twenty, or even one hundred years ago is relevant and useful in research. Like all scientists, working mathematicians will use and reference more recent work the most, but having the ability to access the older literature is of essential value to research mathematicians. Even when only a small fraction of the references come from literature in the distant past, those references may be the key to successful research.

As the scholarly community moves forward into the digital age, more and more of the current (and recent) literature will be available in electronic form online. The more that is available, the more the community will derive value from the ability to navigate easily from item to item.[2] But for mathematics, navigation will have limited value as long as the bulk of the *past* literature is accessible only in paper form. In mathematics, making the past 20 centuries of scholarly literature available online can have a profound effect on research, both now and in the future.

This concept paper outlines a possible mechanism for making much of the past mathematical literature available online for everyone. Such a large project has a number of potential difficulties. But in many respects it is a tractable project with a well-defined goal and clear benefits to the research community. On the one hand, it is the sort of effort that might be undertaken in any discipline. On the other hand, mathematics is an ideal discipline in which to test such a project, both because it is relatively modest in size and because the need for digitizing the past literature is so clearly understood. The international mathematical community understands that need, which makes this suited for international cooperation as well.

For mathematics, this is a project that ties the past to the future in a way that is consistent with the present transition in scholarly publishing. All mathematicians will benefit.

Overview

There are three goals for this project: (i) digitize a preponderance of scholarly mathematical literature that is not already in digital form, (ii) set technical standards for making digital mathematical literature accessible online, (iii) negotiate a protocol for making future digital mathematical literature available in the future. While many people will view the first goal as the essence of the project, achieving the other two goals is essential to make the project worthwhile.

[2] Linking was a persistent theme at the Second UCSU-UNESCO International Conference on Electronic Publishing in Science, which took place in February 2001. The Proceedings can be found at http://associnst.ox.ac.uk/~icsuinfo/.

The entire mathematical literature consists of approximately 50 million pages contained in books, journals, and various other publications.[3] There are many ways to digitize the past literature (that is, literature that is not already in digital form), but the only cost effective way[4] is to combine scanning with partial optical character recognition, creating a combination of scanned page image and associated text file (for searching). There is more to the process, of course. Relevant bibliographic data about each item must be captured (usually by keyboarding); items have to be studied and categorized to understand the various parts (articles, chapters, etc.); proofreading of critical data has to be carried out. Estimates for the cost of carrying out these steps in a large scale operation vary, but a rough approximation is $2 per page[5], making the total cost to digitize 50 million pages about $100 million.

At the moment, many projects are underway to digitize past scholarly literature. One of the first of these is JSTOR[6], which provides complete runs of a collection of journals (including about two dozen in the mathematical sciences) to institutions as a package. Several other groups are formulating projects to scan entire collections of journals[7]. Individuals are encouraged to scan and to make available their own papers and books.[8] All this coincides with the explosion of *recent* mathematical literature that has gone online in a great variety of digital forms (and which will become *past* literature in the near future). Many different groups, with many different formats, with many different interfaces. Almost all have the same goal--to make the mathematical literature accessible to mathematicians--but without coordination and standards the effort will founder. Creating a basic set of standards for digital mathematical literature is essential in order to keep all these efforts from merely producing a Tower of Babel.[9]

The call for standards in electronic publishing is not new, and there have been many attempts to set standards for large communities of scholars.[10] An attempt to negotiate standards in this project must necessarily take into account the work that has gone before, which has not always led to wide adoption. In this case, however, it is much more likely that adoption will spread throughout the community. The standards are aimed at a single discipline, and the project will focus attention on the need for standards.

---

[3] This estimate has been made by Keith Dennis, based on past bibliographic studies. The phrase "mathematical literature" is not defined precisely here, which is the first difficulty mentioned below.
[4] The term "cost effective" is relative, of course, but the alternative of keyboarding material would likely increase costs by a factor of 5, taking into account the basic bibliographic work that would still be necessary.
[5] Other estimates have been made that are far lower. See [Odlyzko, Andrew. "The economics of electronic journals", *First Monday* 2(8), August 1997, http://firstmonday.org/, and *Journal of Electronic Publishing* 4(1), September 1998, http://www.press.umich.edu/jep/].
[6] http://www.jstor.org/about/
[7] The latest is the Electronic Mathematics Archiving Network Initiative (EMANI) involving a consortium of libraries and the publisher Springer-Verlag, A number of other efforts are underway in Europe, all with suitable acronyms such as BNF, DIEPER, and NUMDAM. Individual publishers (for example, Elsevier) are already committed to creating their own collections of past literature in digital form.
[8] A recent call to authors, endorsed by the Executive Committee of the International Mathematical Union urges all mathematicians to create their own "collected works"; see http://www.mathunion.org.
[9] Genesis 11:1-9.
[10] For one of the best known, see http://www.openarchives.org.

Creating a collection of past literature requires that one update the collection in the future. Because this means dealing with individual publishers and organizations who disseminate the literature initially, and because the mathematical literature is especially diffuse, it is essential to outline a protocol for updating the collection over time. This will likely be different for books than for journals, and it may be only an ideal rather than an enforceable protocol.[11] It is essential to attempt such negotiation, however.

One important aspect of the digitizing project is *missing* from this description--distribution of the material after the project is completed. Its absence is deliberate, and in fact, it is a key ingredient for the success of the project. While it is possible in principle to create complicated distribution arrangements that involve collecting fees, distributing these to publishers or authors will almost surely burden the project with huge overhead costs. Negotiating these arrangements and maintaining them will consume much energy, which otherwise could be directed at carrying out the project itself.

Rather than complicated distribution arrangements negotiated by the project, the free market can provide ample distribution. The underlying philosophy of this project is to make the raw material available to the entire community, and then to encourage organizations (publishers, scientific societies, libraries, and other groups) to create a variety of mechanisms to access the material along with auxiliary indexing and organization. The raw material (bibliographic data, scanned images, associated text files, and other digital material) will be largely unstructured. Providing useful access to that material will require considerable effort, and neither grants nor a single organization can sustain that effort over long periods of time. But *many* organizations can sustain the effort indefinitely. Some will find ways to distribute the material as a service to the community; others will find ways to add value by indexing or adding other features, and they may charge for the service. All providers will promote their services, making access for the community easier and better suited to individual needs. The market approach guarantees that the material will be available in many ways, in many places, for many years. It also provides a robust mechanism for archiving, similar to the mechanism that has worked well in the past.

Organization and timing

Administration of such a project requires more than volunteers and committees--it requires a small staff with central control of the many groups working on the project, perhaps distributed throughout the world. That staff may be under the administrative control of one or more existing organizations (to minimize overhead), but it needs to be dedicated solely to carrying out the project. While details are hard to specify in advance, there needs to be a director, administrative assistants, technical advisors, and legal consultants (see below).[12]

[11] Currently, a window of 5 years has been proposed for journal articles; that is, publishers release their material to such projects after 5 years. For books, the time limit is much more difficult, and many publishers view books that are even 20 or more years old as valuable intellectual property.

[12] Budget estimates are difficult to make at this level of detail, but a rough estimate is that total administrative cost will be approximately 20% of the total project cost.

The job of the central staff is to administer and coordinate digitizing projects (either its own or those carried out by other groups), to oversee the work of various advisory committees, and to negotiate about permission to digitize and disseminate the final work. Carrying out this work will require a director with full responsibility for all aspects of the project, advised by committees but with considerable authority to act and to make independent decisions.

During the first phase of the project (likely 1-2 years), three committees will need to be established--content, technical, and advisory. The first will have responsibility to decide which material is to be included in the project. Its work will be ongoing throughout the duration of the project. The second will make decisions about technical standards both for the bulk of the project's work and for the community at large. Its work will be ongoing as well and will be closely connected with archiving, mentioned below. The third (smaller) committee should represent the mathematics community, providing overall advice on major decisions for the project. For example, this committee will have responsibility for establishing protocols for adding material to the collection in the future.

Work on digitizing older literature will continue for approximately 8 years following the initial 2-year period. During this time, material from the project will be made available to the various organizations disseminating it to the community, with the understanding that it will be added to their collections as soon as possible. Because several different groups may be involved in both funding and carrying out the work, quality control on the additional material will be coordinated by a central body under the authority of the central staff. When digital material is available from more than one source, the advisory committee will make decisions based on recommendations of the staff, as well as other considerations.

As the main phase of the project continues, agreements about future additions to the project will be negotiated. Protocols for adding material will be adopted. A process for specifying and modifying standards will be put in place. The aim is to establish a system for ongoing oversight of the project by one or more organizations, with independent financial support for that oversight.

The overall goal of this project is to create a collection of material that represents "past" mathematical literature along with a mechanism for sustaining that collection and keeping it current. At the end of the ten-year period, this should be a system that is sustained by many organizations around the world, each with individual interests but with a common interest to foster mathematical research. Adding material to the collection will become a normal part of the publication process, made cost effective by standardization. Administering the collection will be small scale, and (one hopes) taken on by a small group of organizations.[13]

[13] Such administration can be patterned on the administrative efforts of other standards setting groups, such as the World Wide Web consortium (http://www.w3c.org). These function by soliciting modest donations from supporting organizations along with volunteer help.

Major problems

There are four major problems in carrying out such a project and sustaining it once it is complete. Solving these will not be easy, but finding solutions will be essential to success. These four problems ought to be the central focus of initial planning.

(1) ***Content***. People involved in indexing mathematical literature (like the staff at *Mathematical Reviews* or *Zentralblatt*) recognize the difficulty in selecting what should be included in such a collection. At Mathematical Reviews, approximately 110,000 items are considered for inclusion each year; only about 75,000 are actually added to the database. Deciding which to include is agonizingly difficult. The mathematical literature is far more diffuse than most people realize.[14] Not only are there hundreds of current journals, but many journals publish mathematics mixed with economics, psychology, physics, etc.. Deciding to include only full runs of journals means either that a large amount of the mathematical literature will be missed or that a large amount of the added material is not mathematics (in any sense). Deciding to include selections of articles from journals adds enormous editorial costs to the project.

The situation for books is even more complicated. Should one include textbooks? What level is appropriate? What about books that are at the boundary of mathematics and another area? Again, making individual decisions is costly.

And for both kinds of material, making decisions is a highly charged, often political process (as any reviewing and indexing journal can attest.) What languages should be included? What if an item is known to have major errors? How are multiple editions handled? Are unpublished works included (and what is meant by "published work")? Deciding the content is far more complicated than asking a committee to decide which journals or publishers should be included--it is a process that requires careful thought in advance, and careful administration later in order to avoid massive additional costs.

(2) *Copyright*. This is often misunderstood and underestimated by people thinking about such projects. When undertaking to digitize runs of journals from specific publishers, obtaining permission to digitize the work merely requires obtaining a handful of signed agreements from publishers (who are known in advance). In seeking to digitize an entire field, dealing with copyright issues requires understanding complicated legal issues, often with international copyright law, which is notoriously complex. It means dealing with hundreds of publishers, many of whom are not easily identifiable or who are no longer in business. It means dealing with thousands of authors or their heirs for the rights to reproduce books, which in many instances include material (for example, photos) with uncertain copyright status. This adds an enormous administrative cost to the project.

[14] *Mathematical Reviews* corresponds with thousands of sources for the material it reviews, and lists nearly 600 journals that are covered from cover to cover.

All this has been made far more difficult by recent changes in U.S. and international law. The magnitude of the problem is described in an article by Clifford Lynch[15]. In the chapter "Converting older books to digital form," he writes:

> The legalities of such conversions are a much more serious barrier, and one about which the public remains unaware. Roughly speaking, at least in the United States, any book published before the early 1920s is in the public domain (the details of precisely what is in the public domain are very complicated, and aren't crucial here). If you can find a copy, you can scan it, or, if you are willing to pay the labor costs, you can even re-keyboard it with added structural markup into a more sophisticated digital representation. Whether you obtain a new copyright for your converted digital version of the work seems to be legally murky[16], and seems to depend significantly on how much value you add in doing the conversion. This is important because it has implications for the availability of investment capital to convert public domain materials, and for how these materials need to be protected as they are made available, if they need to generate a revenue stream.

For more recent material, Lynch goes on to say in that same article:

> The cost of clearing rights for these works is likely to be hundreds of times greater than the costs of actually digitizing the works.

We can learn a great deal by examining projects that are already in place. JSTOR, for example, has a far easier task of dealing with legal issues because they negotiate with known publishers about complete runs of (usually) several journals at a time. Nonetheless, they expend a large amount of administrative time dealing with legal issues, and employ their own legal staff.

One possible response to the copyright problem is to decide only to include literature that is clearly in the public domain, or for which permission is easily obtained. A rough estimate indicates that more than 90% of the 50 million pages of mathematics remains under copyright. It is likely that half of this requires search and negotiation concerning copyright. Solving the copyright problem by ignoring it therefore requires a major compromise in the original goal of the project--to make a preponderance of the mathematical literature accessible.

(3) *Initial Format*. Of course, setting standards for content that is *already* in digital form is a well known (if not well understood) problem. This will require hard work and substantial negotiation. But even the apparently simple problem of deciding the format of scanned material is extremely difficult. Not long ago, many people would have suggested using some form of compressed TIFF files encapsulated in Adobe PDF format. But, although PDF is widely supported at the moment, support for certain operating systems

[15] Lynch, Clifford. *The Battle to Define the Future of the Book in the Digital World*, http://firstmonday.org/issues/issue6_6/lynch/index.html

[16] For example *The Bridgeman Art Library v Corel Corporation* (97 Civ.6232 (LAK) New York Southern District Court), case, which found that there was no new copyright in images of out-of-copyright artworks.

(Unix) has become problematic. More importantly, there are new, extremely effective formats for scanned images that reduce the size of files by a factor of 3-8 (or more). The most notable of these is DjVu[17], a format developed at AT&T Labs (using wavelets for superior compression and a progressive algorithm for decompressing images, presenting an immediate image that gradually improves). Products implementing DjVu are now owned and sold by LizardTech. Like PDF, DjVu requires special software to view the images within browsers. But the technology is open source and the advantages over more traditional technology are considerable.

Selecting the right initial format--possibly a proprietary format--in an environment that is constantly changing, for a project that lasts over 10 years, is a nearly impossible task. This is closely connected with the next problem, archiving, but it is not the same. (The right initial format for presentation may not be the right format for archiving.)

(4) *Archiving*. This is not so much a problem for the project as it is for those sustaining the collection after the project is complete. Once again, it is a problem that is often misunderstood by people, including experts (precisely because there *are* no real experts in an area like digital archiving, where no one has much experience).[18]

Until recently, there wasn't as much need to consciously archive scholarly journals or books--archiving was (almost) automatic because many copies were distributed to institutions at various locations. One counted on the laws of probability to ensure that at least one copy would be extant years in the future. That one copy could be used to reproduce more copies at a time many years after initial publication.

Two things have changed with electronic publication. First, the copies may not be widely distributed, but rather often reside at one or two sites in electronic form. This is the problem of "robustness", and it's the issue most people think of when discussing archiving. Second, even if a copy of a file is extant many years in the future, it may not be possible to produce copies of the "work", that is, fully functional copies that are identical to those in existence years before. This is because electronic journals and books often consist of files embedded in a larger system that makes use of programs, auxiliary files, and even hardware to render the work. In short, the context in which the work is embedded is often essential to making a faithful copy, and archiving requires being able to reproduce that context. This is often referred to as the problem of "format", but the language makes it sound pedestrian, as if it were merely a problem of presentation. It is, in fact, the central problem of archiving.

There are several simple schemes for ensuring robustness, including the simple device of replication to create multiple copies (just like paper). Because electronic media may degrade more rapidly than paper, however, there has to be an added step of routine replication to produce fresh copies.[19] Fortunately, making electronic copies is far easier

[17] Extensive information can be found at http://www.djvuzone.org/.
[18] See, for example, http://www.oclc.org/oclc/new/n226/ea.htm .
[19] Recent studies suggest that magnetic media have a lifetime of 10-30 years. Optical media appear to have lifetimes of 100 years or more, but studies are inconclusive.

than making paper copies, which compensates partially for the extra step. Routine replication also addresses the problem of changing media, since a copy can move to whatever medium is currently in use.

One might hope that the format issue can be solved in a similar way--regularly change formats as new come along. There are two reasons this doesn't work. First, "changing formats" is *not* equivalent to making a copy. While making copies is routine and easily done for large volumes of material, changing formats requires special intervention, at least for a fraction of the material. The difficulties depend on the *old* format (something we know in advance) as well as the *new* (something unknown when we create the archive). Even if only a small fraction of the material needs special intervention by technical personnel, this can be enormously costly for a large collection. Those who deal with small personal collections often ignore this point.

There is a second, more subtle reason that changing formats is not a solution to the format problem. The format problem is more than merely preserving the format of a work; it is deciding what information about the environment in which a work is presented should be saved initially and then deciding at each subsequent stage of archiving what information is passed along. It is virtually impossible to save *every* piece of information about the environment. (For example, we likely rely on the ISO standards for recognizing characters and assume conventions about line feeds and returns.[20]) Archiving requires decisions about *which* information will be necessary in the future, and those decisions must be made in the absence of detailed knowledge. Indeed, at the moment, and for some years to come, those decisions must be made without experience as well. There are many, many examples of incorrect decisions made in the past 20 years, resulting in lost work; there is no reason to believe we can avoid incorrect decisions in the future.

To sustain this project, one has to find a way to pay for the potentially large costs to update the format in the future, as well as to make reasonable decisions about what information to pass forward. Maintaining collections at many sites, each with either professional or financial interest in the material, ensures that a large group will want to share those large costs. It will be in everyone's interest to make certain that reliable decisions are made when formats change. Nonetheless, these are issues that extend over long periods of time (often exceeding the careers of individuals involved), and there must be a mechanism to guarantee that archiving issues are dealt with on a continuing basis.

Competition and cooperation

The great advantage of the approach described above is that it effectively balances competition and cooperation. The balance is essential for a project that is international in scope and that spans a decade or more. And the balance is crucial to ensure the effort is sustained once the initial project is complete.

Rather than a few centralized institutions for dissemination of the material, the proposal calls for competition among many organizations to provide access in ways that address a variety of needs. Libraries, societies, universities, commercial publishers can all compete

[20] http://www.iso.ch/iso/en/ISOOnline.frontpage

to add value for the community. This is healthy competition that provides incentives for many people to carry out the work and to sustain it in the future.

On the other hand, there are key areas in which cooperation is essential. Without uniform standards, access to large collections of digital material will be difficult or impossible. Without such standards, the kind of healthy competition above becomes impossible. And without standards, archiving the literature becomes enormously costly, possibly exceeding the resources of even a large group of interested parties.

Cooperation in all phases of this project can be made even more tangible by inviting representatives from many segments of the international mathematical community to serve on the various advisory committees. In addition, many countries have funds available for digitizing collections of scholarly literature. It is possible (and desirable) to divide the job of digitizing the older literature into several large collections, each of which can be done by a separate organization or country. This kind of cooperation, however, requires oversight from a central body, and it will be necessary to coordinate all work using a single body as indicated above.

Initial planning

This document is intended to describe a concept, providing only an outline of the scope of the project, a possible underlying philosophy, and the major issues one must resolve for successful completion. To carry out such a massive project, a small group of interested people (including potential international partners) must engage in far more detailed planning. That planning might be accomplished through a planning grant, administered by a single organization but involving representatives from institutions, libraries, scholarly societies, and publishers.

This project will revolutionize the way in which mathematicians conduct research--it is hard to imagine any single change that will have a greater influence. It remains a dream, of course, but an ideal dream on which to build foundations.