DML: Moving Forward

I am troubled. For the past two years, mathematicians have been talking about digitizing the past mathematical literature, a project now called the "Digital Mathematics Library." Mathematics values its past literature more than any other scientific discipline, and putting a substantial portion of that literature online is more than quaint scholarship. The project would have a profound effect on ongoing mathematical research.

The project is ambitious. Digitizing many millions of pages of past literature is costly. Deciding which pages is complicated. Persuading those who own the pages to agree will be hard. But the benefit to mathematics is so great that, when the project was first envisioned, many people agreed we should try to overcome the obstacles, no matter how difficult.

Unfortunately, we are now engaged in a process of creating new obstacles rather than overcoming the old ones. We have made scant progress in addressing substantive issues. And instead of recognizing that we're stuck, we pretend that we are accomplishing something by repeatedly discussing the same side-issues and ongoing projects.

What happened? First, we fell victim to a common temptation (to which academics are especially prone). Faced with a large, complicated set of tasks, we tried to plan them all at the same time, in parallel. That process sounds logical — to divide the tasks and assign each to a committee — but this kind of planning dissipates energy over many different groups and makes it extremely difficult to understand how the various parts of a plan fit together. (And Ewing's universal law of committees states that the likelihood of solving any problem is inversely proportional to the number of committees working on it.) Second, in order to promote the project we dressed the idea in rhetoric ("Digital Mathematics Library"), and over time the façade was confused with the project itself. The original goal (digitize the past literature) morphed into a new goal (create a digital library), with a whole new set of obstacles and problems. New players didn't understand that our flamboyant language was a sales pitch; old players began to believe their own hyperbole.

And so we are now in a holding pattern. A half-dozen groups are supposed to be solving problems, but with no focus, almost no coordination, and little progress. Discussions concentrate on portals, search engines, and administrative structures (all for "the library"). Funding agencies become evermore impatient, proposing to move forward — not by digitizing the literature but by building a software edifice for "the library"!

We have lost our way on what should have been a straight path.

I don't mean it's an easy path: More than others, I believe that the obstacles to digitizing the past literature will be hard to overcome. But we can overcome them only by setting our sites on the goal, analyzing the problems we face, and then planning solutions. A crucial step in planning is setting priorities (which steps do we work on first) and then coordinating solutions (what are the consequences of the solutions to one problem in solving the next).

The goal is to digitize the past mathematical literature; the goal is to digitize the past mathematical literature; the goal is to digitize the past mathematical literature. We need to repeat this as often as possible to keep us headed in the right direction.

Here are three questions that should drive the discussion forward.

1. Because many groups will be carrying out digitization, we need to set *standards* to make the material interoperable. And because digitization is already underway, we need to set those standards immediately. What are the minimum standards to which digitization projects should adhere? This ought to have the highest priority.
2. Because publishers (of various kinds) own most of the material, we need to elicit their cooperation by giving them something in return for the use of their material. Striking that bargain with publishers is essential to the project if we want it to include a substantial fraction of the past literature. What is the precise bargain that should be offered?
3. The publishers must make the bargain with some entity, and the terms of the bargain must be enforceable. One way to accomplish this is to create a central administration, whose only function is to administer the agreements with publishers. Another way is to make standard, but decentralized bargains (for example, with funding agencies). What is the best mechanism for creating (and enforcing) agreements?

With standards, a bargain, and a mechanism for enforcing it, the DML project will begin to function and grow. I *know* that there are many other problems to be solved (updating and archiving, for example), but solving those problems comes next. If there is no material online, we don't have to worry about updating or archiving it.

There will be four types of players —the publishers, the digitizers, the funding agencies, and the mathematicians themselves—and possibly a small central administration. The agencies fund the digitizers. The digitizers process material from the publishers. The publishers provide (at least a major portion of) the digitized material on the web. The administration provides the framework for agreements. And the mathematicians benefit from an ever-increasing supply of mathematical literature on the web. This is a simple, elegant, and effective structure.

Of course, the precise terms of the bargain with the publishers will be important. One suggestion is that publishers can integrate the digitized material into their website provided they make all material older than five years freely available (with varying arrangements for books). It's possible that some other time period will be better accepted; it's possible that multiple forms of the agreement will be necessary.

Not all players will function exactly alike. Some (especially small) publishers will not have the inclination or the means to host their material. Almost certainly, other publishers will step in to host it for them, as they are now doing for the ongoing publication of the journals.

These are details, however, and the heart of any effort to make progress on the digital mathematics library is to focus on a few key planning steps that are essential in the initial stages.

What about libraries, portals, and search engines? They are secondary; think about them after you have solved the initial problem (if at all). The notion that the DML must be a "library" in the sense that it is some well-defined "collection" of accessible material represents our inability to fully accept the nature of the Internet. We don't need to define a collection of material — we need the material itself. A "library" needs to be organized and it needs to be maintained. Who pays for that? And because it is inconceivable that publishers will give up their material to a single collection, the notion is not consonant with any conceivable bargain we might strike. We should move on.

If the material is distributed, do we need a single portal? Of course not. First, there already is a large body of mathematics available on the web without a single portal, and yet extremely valuable. Second, many different portals are likely to be far more useful than one, and those portals already exist (including Mathematical Reviews and Zentralblatt, as well as other nascent projects that list sites). Finally, creating a portal is not a one-time investment but rather a continuing commitment. It requires accurate metadata, robust software that changes over time, and vigilant verification of data and links. Experience shows that this is far more difficult than most novices imagine. Surely that's not a project that fits within the scope of the DML — at least for now. Let portals grow later. Why build the door before we have built the house?

And if not a portal, then shouldn't we have a search engine? We do — it's Google (or whatever comes along that's better). The multibillion-dollar business on the Internet drives innovation in organizing material on the web, and the mathematics community can effectively use the results. Why spend grant money for such efforts when it can be better used for our main goal?

I am troubled, and I know I am not alone because I hear others grumbling. Instead of focusing on those steps that will achieve our intended goal, we are diddling by appointing multiple committees and talking all around the central problems without addressing them. In the meantime, a number of projects are underway, but without any clear standards to guide them. Publishers will soon lose confidence in a project that seems to be floundering. So will funding agencies.

I don't mean to be glib; I understand that this is complex project; I don't know all the answers. But I *do* know that we're asking the wrong questions at the moment. While it sounds trite (indeed, it *is* trite) to say that "a journey of a thousand miles begins with a single step," it's still good advice. We need to take that step ... and a few others as well. And it's clear that before you set out on a journey, it helps to know your destination … and to make sure that the next few steps are headed in its direction.

John Ewing
22 January 2003