Preserving Our History

The use of TEX over the last decade and a half to write papers, lecture notes, and even ephemera has moved from the unusual to the commonplace. Indeed, some of my younger colleagues can't remember using anything else but TEX to write mathematics. For those of us somewhat longer in the tooth, we remember using other software, which was preceded by the little golf balls that allowed typing of mathematical symbols, which in turn was preceded by writing in the mathematics by hand (with the hope that the typesetting would introduce only a few errors).

The utility of this wonderful piece of software has been greatly enhanced by the continuing and massive growth in the capability of the accompanying hardware. Indeed, in the early 1990s when the first implementations of TEX appeared on desktop machines, I wrote a review for the *Notices* with some comparative timings.[1] I had a test document of about 100 pages that contained an appropriate mixture of ordinary (for that time, English, without accents) text and mathematics. It could be run though TEX in times varying between one and five minutes. When I try the same thing on my newest desktop machine, the job takes 0.004 seconds. Viewing it another way, three such jobs can be run in the time it takes to refresh the screen once.

One of the happier results of this migration to TEX and the improvement in hardware has been the ability to put our papers on personal webpages so that anyone with a standard computer configuration can acquire them. This usually means making a PDF or a PostScript file available for download. The infrastructure of the Internet makes this transfer of files easy and transparent, and, as a result, the little postcards that were mailed to request reprints has joined those little golf balls as historical curiosities.

Improvements in hardware have another happy consequence: papers written in the predigital era and consequently not available on the Internet are not beyond redemption. They, too, may be made available for download. Over the past few years there have been significant advances in the hardware and software used to scan paper documents. With the right equipment, scanning several hundred or even a thousand pages is not difficult.

There are two approaches, both of which work well. The first is to use a standalone scanner. Robust models with document sheet feeders are available for under US$1,000. These usually include the software for doing the scanning, and sometimes OCR (optical character recognition) software is also included. There are lots of options when using such software, so here are some suggestions.

When scanning the pages, the software can produce color, grayscale, or black and white files. Unless there is a compelling reason, black and white is usually the best choice for older documents. There is also a choice of resolution: 200, 300, 600, or 1200 dpi (dots per inch). Usually the 600 dpi is the best choice.

There are also several different types of files that can be produced by the scanning software. All of them have some compression: these come in two types: lossless (no data lost during the compression) and lossy (some data irretrievable). A lossless compression is the best. There are also different file formats, the most common being PDF and TIFF. The PDF files are the ones to put on your webpage; they can be read on any modern computer with readily available software. Note that there are two different types of PDF files: ones that are image only and those that are also text searchable. The latter type is preferable, and most scanners can produce them.

There is also a compelling reason to keep lossless TIFF files. Newer and smarter software will emerge that will do things we can't do today. If you keep the TIFF files, there will be no need to rescan since the information is already in an industry-standard format.

There are frequently other options to consider. For example, many scanners allow the page to be autostraightened, that is, pages that are tilted because of inaccurate feeding or printing will be rotated into horizontal and vertical alignment. This is, more often than not, useful. Another option is despeckling, that is, the removal of very small dots on the page on the assumption that they are either background to the page text or errors introduced while scanning. The despeckling may be set to be more or less aggressive. This choice depends on the quality of the material being scanned. If there are handwritten symbols, its probably best to leave the despeckling off. If the document being scanned is printed, despeckling may be useful. For those with an experimental bent, running a few sample pages with different settings can be very helpful.

A second approach is to use a photocopier. Many of them come with scanning software built in: you feed in the pages and the image files are emailed back to you. Usually there are fewer options than with a scanner. The default resolution is usually 200 dpi, so be sure to set it to 600 dpi. Most of the provisos given above for scanners are also valid for photocopiers.

I recently carried out a two-month project that involved scanning of some 53,000 pages. It really wasn't difficult. In fact the hardest part of preserving your mathematical history may be taking the staples out of those old pages.

Our mathematical history is important and worth preserving. A bit of effort by all of us can produce a significant body of mathematical literature. Let's get our history out of the file cabinet and onto the Web!

—Michael Doob
University of Manitoba
mdoob@ccu.umanitoba.ca

A version of this article appeared in IMU–Net*, September 2008.*

[1] TEX *and the Single CPU (I), March 1990, and (II) December 1991.*