Electronic posting and archiving of TeX-composed scientific documents

Laurent SIEBENMANN

Université de Paris-Sud
Matématiques, Bâtiment 425
F-91405, Orsay, France
Email: `lcs@topo.matups.fr`

Abstract

This note is a quick introduction to two proposals that would make TeX's original ".dvi" output format more useful for electronic posting and archiving of scientific documents. The first is compact auxiliary 'atomic' fonts to make Knuth's CM fonts able to serve all (non-Cyrillic) European languages, by using composite characters. The second is the notion of a 'multi-standard' graphics object. This was initially designed to solve the nagging problem of portable previewing of EPS (Encapsulated PostScript) graphics. But it promises much more. A single ".dvi"-based posting will be able to serve all the multifarious graphics abilities and debilities of TeX viewers and printers.

1 Introduction

The race is on to establish data formats for publishing the world's scientific literature in purely electronic form. It is widely acknowledged that the internet now has the capacity necessary to deliver scientific articles in just a few seconds to any computer on the internet (via www = World Wide Web, and also older systems like ftp). Hopefully, the costs for internet transmission will remain quite negligible in spite of commercialization. Rapid access (hard disc) storage cost per page in an electronic library is already negligible. (Provided efficient storage formats are used, the marginal cost is now less than one cent per page for the 4 year life of a hard disk — and it is still falling.)

The leading candidates among electronic formats for the mathematical sciences seem to be:

— ".tex" = TEX's ASCII typescript input format defined by Knuth and elaborated by 'TEX markup formats' such as Plain, LATEX and others. This format usually must be enriched by `\special`-ly included graphics objects. Hence it should perhaps be called the "tex_etc" format.

— ".dvi" = TEX's binary device-independent output format. Similarly, we should talk of the "dvi_etc" format.

— ".ps" = the PostScript page description language of Adobe Inc.

— ".pdf" = Adobe's format for its freeware viewer Acrobat, nourished by its commercial ".ps" to ".pdf" converter Distiller.

A candidate you might have expected is SGML, but SGML does not yet provide adequately for mathematics. Nevertheless Jonathan Fine's SIMIM, a new TEX format, hopefully combines SGML with TEX's mathematics in an interesting way.

How do these formats measure up to expectations? For me, the major fault today of all of these formats (judged along with the software and hardware that support them) is 'low bandwidth' for browsing. Indeed, one assimilates far less of interest, in an hour of electronic browsing, than in a well-run physical library.

I believe that this is a sign that browsers (also called viewers or previewers or screen drivers) are in their infancy. Their speed should improve with increased microprocessor power. The 10-fold increase in processor clock speed from 8 to 80 mhz in the past 10 years augurs well. As does the doubling of bus width from 16 to 32 bits. There is also an excellent prospect that public fascination with bulky imagery available on the internet will push data transport speeds and prices on the internet *ahead of and beyond* the needs of the browsing scientist. Of course if I am wrong, there will be crippling traffic jams. One *bad* omen is that the resolution of images will lag behind — if the progress from 72 dpi to 104 dpi in 10 years is any indication.

Equally worrying is the widely heard prediction that internet services will be too expensive for third world countries and also for any scientists without robust financial support. It is based on the observation that the internet is being inexorably commercialized.

One can hope that a combination of growth and competition will make the the necessary services affordable. However, there is a grave problem casting its shadow on this optimism — it can be stated as follows. Given that prices tend to be as high as a market can bear[1], and that science teachers and researchers have relatively little financial clout (and students still less) when playing in the same league as business and professional groups, we would be wise to ask whether scientists are playing a losing game on the internet.

My answer is that, to assure excellent service from the internet in spite of lacking financial clout, scientists should adopt their own particularly *efficient* methods of electronic publication. So efficient that even third world scientists will have adequate access to the the electronic literature.

In this struggle for efficient communication, the scientists' secret weapon is TEX. Secret because I believe that TEX's inherent complexity will restrict its use to a small circle surrounding the scientific community in which it was developed. Donald Knuth paid extraordinary attention to efficient operation of TEX; he designed the input ".tex" typescripts and the binary output ".dvi" page description files to have information density comparable to that of ordinary ASCII prose (to within a factor rarely worse than two or three). And like prose they are compressible to well under 50%.

To some of you, the idea that TEX should be done 'on the cheap' is repulsive. I claim that flexibility of design can make TEX able to do 'high tech' publishing with the same software and often the same macros — for example including color MPEG animations. Furthermore, I want the same posting to serve both poor and rich scientists. A poorly equipped reader might see a still, black-and-white image instead of the color animation, but he should be able browse the same posting.

In prospective summary, our goal should be to extend the domain of extraordinary efficiency of TEX from the paper publishing world of the 1980's to the electronic publishing world of tomorrow — and without sacrificing high-tech ambitions.

Return to the question: In what format should science be electronically published? In my view, TEX's native formats ".tex" and ".dvi" are strong candidates that seem at least competitive with alternatives: ".ps", ".pdf", ".html", ".rtf" (= RTF of MicroSoft Inc.), etc. which will, I believe, be evaluated in other talks. But it seems unlikely that one format will win on all fronts.

The ".tex" format (in LATEX dialect) is currently the most popular on the pioneering preprint server ftp://xxx.lanl.gov that was initiated by Paul

[1] This occurs particularly with monopolies and oligopolies, which are widespread in the world communications industry.

Ginsparg. My article [1] can be interpreted as an attempt to make Plain TeX as efficient and as archival a markup system as LaTeX.

On the other hand, ".dvi" format, although intended for the job by Don Knuth, seems (as we shall see) to be in a checkmate position and could lose out entirely if not further developed.

I will focus briefly on two innovations that should help the ".dvi" format to avoid checkmate and then go on to surpass all competitors in many tasks. Hopefully, I will have occasion to explain them in greater detail elsewhere. Initially, I explained them on internet discussion lists, and my often eminent interlocutors made significant contributions. Each remains tentative in the sense that it is just a sketched plan for action waiting for support. Developers can immediately examine many details by consulting the list archives indicated below, or by contacting me personally.

2 An 'atomic' complement to CM fonts

One problem with ".dvi" postings is specifically a European one. No single font encoding quite manages to cover European needs. It is well known that the Knuth's CM fonts are not sufficient for the Poles, Czechs etc. Even the EC fonts with numerous accented letters are not sufficient for the Latvians, Welsh and others. My suggestion is to return to Knuth's CM approach in which accents are separate entities and essentially all accented characters are composite. To achieve optimal typographic quality the author will use virtual fonts — but the virtual fonts are to be eliminated in the posted ".dvi" file by use of Breitenlohner's "dvicopy". The most plausible scheme would be to form a set of (small) fonts parallel to CM fonts and containing all the (relatively few) 'atomic' (= non-composite) glyphs missing from CM but necessary for some language of Europe with a Latin-based alphabet. This is compatible with the EC fonts in the sense that virtual EC fonts can be based on CM and this 'atomic' extension.

(A somewhat similar proposal was made for Adobe-distributed Type 1 fonts by Pierre Mackay, and has now been acted on by the LaTeX group for the PSNFSS fonts.)

The window of opportunity for acting on my proposal above is opening up with the approaching completion of the EC fonts — to be reported on in a talk by Jürgen Knappen.

The original discussion of this proposal took place on the math fonts list math-font-discuss@cogs.sussex.ac.uk from March–May 1994; see its archives on ftp cogs.sussex.ac.uk (contact Alan Jeffrey, alanje@cogs.susx.ac.uk).

3 'Multi-standard' graphics integration

The most glaring current defect of ".dvi" format is the lack of a standard `\special` syntax for inclusion of graphics objects — for it tends to make the ".dvi" file non-portable.[2] In addition, to make ".dvi" competitive with other formats such as ".pdf", it will be absolutely necessary to provide for on-screen viewing of the included graphics objects — for the norm of the graphics object is often *not* supported by the viewer application. We shall observe that these intertwined problems happily break up neatly into two disjoint problems.

3.1 Establish a `\special` standard for a single norm

In more detail, standardize a syntax for including one species of graphics object — but a syntax that carries over without essential change to all other graphics norms. This essentially means that the the image rectangle's width and height (on paper or screen) are explicitly or implicitly specified in the `\special`'s argument string viewed in isolation. Reserve an 'extra' version of this syntax for a 'multi-standard'.

An existing syntax having this 'universality' is that used by Rokicki's *dvips* for integration of EPS graphics via his "epsf.tex". It happens to be the syntax most commonly supported by other drivers/viewers.

There is, at long last, a TeX Users Group committee discussing at least the case of this problem for EPS files. And another $\mathcal{AMS}$ committee. Let us wish them the intelligence and stamina and goodwill necessary to deliver the goods. The long term health of Knuth's ".dvi" format seems to require such a standard.

The usual retort to such pronouncements is "don't hold your breath while waiting". Fortunately, thanks to a new ASCII version ".dvt" of ".dvi" format (by Geoffrey Tobin; see CTAN), `\special` syntax in postings can be readily

[2] My "boxedeps.tex" TeX macros (on CTAN) first made ".tex" typescripts using EPS graphics portable in spite of this lack of `\special` standardization. Such standardization remains, I believe, necessary to assure reasonable portability of ".dvi" files.

revised and updated. Consequently there is absolutely no reason to delay. On the basis of Rokicki's most widely implemented current `\special` syntax, all viewers/drivers can act immediately to support 'multi-standard' graphics, as I now indicate.

3.2 Implement a 'multi-standard' approach to graphics integration

A typical example would be the integration of a graphics object "myfig" that is represented as a high resolution color bitmap "myfig.jpg" of JPEG norm for viewers that have such wonderful[3] capabilities, and by a 'simple TIFF' bitmap file[4] "myfig.stf" designed to serve as a lowest-common-denominator format for viewers without JPEG display capabilities. We explain the integration in terms of this example, which is one that no viewer yet handles today.[5] To avoid technicalities, only the article author's intended experience, and that of the browsing reader are described.

The user needs only indicate the name "myfig" on first compiling with TEX. He does so using a macro package for integration of multi-standard graphics; this package could be derived from one of the several integration packages for EPS graphics, by relatively minor generalizations. (The author does not want to have to work with figure dimensions!) The figure preview initially appears crammed into a standard square box; this is because TEX is quite unable to find out the shape or size of the figure since both graphics files mentioned are binary and TEX reads only text files. But, much as with LATEX's cross-referencing, the author is instructed by the log file to preview and then compile a second time. On second and subsequent cycles, the figure appears correctly and can be adjusted using any of the conventional macro commands. If the viewer is able[6] to accept JPEG files, preview is in color, and, if not, the image is the black and white familiar for EPS preview.

[3] Wonderful efficiency and scalability.

[4] Simple TIFF means black-white, 'packbit'-compressed, and with strip structure but no tiles. This is the sort of preview that Adobe Illustrator can include inside EPSF files to permit preview on PCs.

[5] My proposal goes beyond what the pioneering viewers *Textures* (on Macintosh) and *DVIWindo* (on PC) have implemented for EPS graphics preview; if you are not familiar with their remarkable performance, take the time to look. I believe it is fair to say that the 'one-file' EPS-with-preview of the Macintosh blinded most developers to the virtues of the more general 'multi-standard' scheme.

[6] I am hopeful that I will see JPEG color graphics integrated on my home platform (Macintosh) soon. If there is not frontal operating system support for displaying JPEG, there should at least be the possibility to convert to an established local norm (PICT) and use existing abilities of the viewers.

To post the ".dvi" file with its graphics inclusion, the author packages it with the ".jpg" and ".stf" files in a directory — whence the posting format name "dvi_etc". He should ignore an ephemeral ASCII file "myfig.gdf' (say) a temporary 'graphics description file' created by the viewer, which serves to let TEX know the 'bounding box' information on the graphics object that TEX initially could not read.

What happens when a user browses this "dvi_etc" posting with a different viewer supporting the proposed multi-standard? The user will immediately see the best image his viewer can offer. But let us imagine that the user sees only the modest ".stf" and is unhappy because he knows that his viewer can handle good quality ".gif" images (but not ".jpg"). Here the multi-standard again comes to the rescue: it suffices to convert from ".jpg" to ".gif" and the high resolution images appear!

3.3 Popular PICT and WMF scalable graphics become portable

The above JPEG example of multi-standard graphics integration is forward-looking. Here is an application that is useful immediately — and indeed retroactively. One creates a preprint posting using the Macintosh's 'native' vectorized PICT graphics (suffix ".pct") and the multi-standard `\special`. A posting notice recommends that users on other platforms convert[7] the PICT graphics to their favorite local norm, which for PCs means a MicroSoft Windows MetaFile with extension".wmf". The multi-standard graphics integration using a single `\special` will assure that the graphics appears on the end-user's screen as soon as a version in his 'native' norm is available. Successful viewing of vectorized graphics almost always implies optimal printing, but not conversely.

In summary, with multi-standard graphics integration, one can confidently use and post with TEX the 'native' vectorized (scalable) linefigures of Macintosh and Windows for PC. The viewing is then of optimal quality at all scales; this includes the labels if they are based on hinted outline fonts. A high degree of portability will be assured. Currently (without the multi-standard) TEX users convert to ".eps" format for portable printability; but little portable viewability has been achieved.

[7]Powerful conversion utilities are becoming available on most platforms. For the many users who refuse to become involved in conversions, one can hope that www servers will soon automatically supply to any www-connected viewer the graphics norm(s) it requests.

The internet discussion of the above graphics multi-standard proposal can be consulted on www as a hypermail archive:

http://math.albany.edu:8800/hm/emj/

I believe the two standards informally proposed in this article could help make Knuth's troubled ".dvi" format viable, and indeed enviable.

References

[1] L. Siebenmann, Occam's Razor and Macro Management, (these proceedings).