Handling of LaTeX-ETDs and TeX-Conversion

Günter Törner
University Duisburg-Essen - Location Duisburg-
toerner@math.uni-duisburg.de
Lotharstr. 65, 47057 Duisburg, Germany
www.uni-duisburg.de/FB11/STAFF/Toerner.html
Keywords: LaTeX, TeX, Natural Sciences, Mathematics, MathDiss, archiving quality, Conversion aspects

Abstract

Within the natural sciences, in particular in mathematics, most of the ETDs are based on LaTeX-coded sources whereas the presentation format is more or less PDF. However, in the international library SYSTEMs the LaTeX source files are not stored in general, they are ignored since there is no competence in this format type in the libraries.

On the other hand there exist servers, e.g. the ArXiv-Server or MathDiss International, accepted by the scientists, which trust in LaTeX-files exclusively and which believe in their archiving quality.

So the question arises:

- *How should we handle LaTeX-ETDs in the library SYSTEM?*
- *In the context of ETDs in natural sciences and especially in mathematics we have to discuss further aspects of LaTeX-files:*
- *How to check the inner consistency of delivered LaTeX-documents consisting of various LaTeX-Files?*
- *LaTeX as archiving format?*
- *Conversion aspects (e.g. TeX to XML, in particular TeX to MathML)*

The Philosophy of TeX

Mathematical texts are highly condensed information packages using different symbols and formulas. Thus, comprehension and understanding heavily depends on the visual quality of the typesetting. Thus, far in the past, typesetters always tried to produce a high quality within mathematical texts. As Kuntz (2001) pointed out, Ellen Swanson's book Mathematics into Type is a unique and important contribution to the literature of technical typesetting. It set a standard for how mathematics should be translated from a handwritten manuscript to a printed book or document. At that time, Swanson's book was intended primarily as a resource for technical typesetters.

But time has now changed considerably. With the advent and wide availability of TeX, most mathematicians can take a more active role in producing typeset versions of their work.

What is TeX?

Again we refer to Krantz (2001, p. 13) who gives an excellent introduction into the topic. TeX is a computer package for producing document output of typeset quality. TeX is a computer typesetting language - a high-level computer programming language. It is used by individuals to write letters and articles with a personal printer as well as by technical publishers to produce books on specific typesetting device. TeX is particularly powerful and useful for preparing manuscripts with mathematical formulas and tables. Since being invented and introduced by Donald Knuth in the early 1980s (see Knuth (1980)), it is now almost universally used by mathematician to typeset mathematics, and is also widely used in areas of science and engineering that emphasize mathematics.

We follow again the condensed description in Krantz (2001, p. 14).

Part of the reason for TeX's long life and wide use is that it implements a "markup language" instead of creating output on a computer screen as you enter it. The computer file that you write for a manuscript is a text or ASCII file that contains commands that describe in more-or-less plain English terms how you want the pages formatted and what mathematical symbols you want to use. Viewing output as it might be printed is a separate step. This has the advantage that your computer source file is independent of improvements to your computer hardware. Currently laser printers have around 1200 dots per inch or around 10.200 horizontal dots per page while most current computer monitors go up to at most 1200 or 1600 horizontal dots per screen. So you can see that from the outset, screens display material to a different tolerance than the printer will print it. With the help of its markup commands, TeX positions elements on the printed page to within 10^{-6} of an inch. In effect, the TeX code mandates exactly how you want the document to look. The screen provides an approximation to this truth, usually suitable for accuracy checking. Depending on the quality of your printing device, the printer will give a hard-copy rendition of the ideal document dictated by the TeX code. The superb degree of accuracy of which TeX is capable should protect TeX from obsolescence for many decades to come.

Fortunately, you do not have to decide yourself how to position each character on the printed page to within a millionth of an inch. TeX will make most of these decisions itself. You can use TeX commands to customize these decisions, but this is usually not advisable unless you have a considerable amount of experience with either typesetting or TeX. However […] there are a number of situations where human intervention is necessary for highest-quality output. One of the main purposes of this book is to explain what those situations are and how to deal with them, or else how to communicate

with a copy editor and/or typesetter so that lie or she can deal with them.

Like many languages that have been around for some time. TeX has a number of dialects. The most common dialects are LaTeX, Plain TeX, AMS-TeX, and AMS-LaTeX. These are roughly as close as American, English and British English, with Plain TeX being slightly more divergent. Computer installations that support one of these dialects will generally support all four.

In fact, TeX has been used by law offices and even by the periodical TV Guide. The portability - in the large and in the small - of TeX makes it a powerful and versatile tool.

TeX as a typesetting language within mathematics

As remarked above, TeX and his formats are worldwide accepted by mathematicians. Since the source code is of ASCII type, it is easy to exchange manuscripts and run the file of a colleague on one's own machine. Thus by exchanging small files, it is easy to get virtually access to the full manuscript with its high-quality typesetting format. Furthermore, learned societies of mathematics and the International Mathematical Union (IMU) regard TeX as *'a congenial and accessible way to give documents some structure without adding unreasonable burdens on the author'*.

The deficits of TeX

Above we remarked that TeX is implementing a markup language, however this markup language first of all addresses the layout character of TeX, not the semantic aspects of mathematics. Nevertheless, there are some features implemented in this language which give rise to semantic units, in particular this applies to the LaTeX format. We list without comments some of these elements (definitions, lemmas, theorems, arrays, formula, equation, quotations and many different list environments). Whereas LaTeX is strong in creating environments on a macro level, the specific micro level is not as detailed addressed the macro structure. That is why that there is no satisfying automatic conversion into the XML-world. The main semantic structure is virtually represented on the layout level and thus has to be decoded by the mind of the reader.

However, some software exists to convert the layout markup into an XML-based language.

We repeat ourselves when we point out that such a conversion is neglecting the basis semantic elements of the LaTeX source.

Consequences for the libraries

We are confident that in the next years conversion packages will be improved as well XML-based dialects will be developed and context-dependent conversion will decreased the efforts of translation. It is substantial for these software packages that they can work on the basic source file which should be stored.

Nevertheless, there is a further problem which has to be taken into consideration. Usually, a LaTeX based manuscript consists of many files, maybe a root-file, some text files which are organized by the root-file, tables, picture files, data files and so on. Further, the authors will also refer to format packages which are either standardized (see Server) or developed by the authors for the purposes of the manuscript which is produced. It is essential that such a folder with all these files is self-contained and consistent. Normally, it is impossible to check the technical consistency by hand and unfortunately there are nearly no SYSTEMs available which are doing this job automatically. Note that TeX is not a proprietary software! ArXiv and Connell University have developed some devices for running their database.

TeX files as sources for archiving

Officially, TeX is not seen as an archiving format; however, by its open architecture and since the language is ASCII-based, TeX files serves (privately) in many cases as archiving sources. There is no problem to run a well-documented TeX folder from the 1980s. We have information that at the LANL-server, which is a TeX-based server, the more than 200.000 files were updated with a few hours in the last periods with nearly no problems.

The conversion problem - TeX-to-XML

Page Representation Formats

TeX and LaTeX are well suited to producing electronically publishable documents. However, it is important to realize the difference between page layout and functional mark-up. TeX is capable of extremely detailed page layout, specifying precisely where on the page symbols go. HTML is not, because HTML is a functional mark-up language (specifying primarily document structure) not a page layout language. HTML's exact rendering is not specified by the document that is published but is, to some degree, left to the discretion of the browser. This is a deliberate choice. It recognizes that the window size,

resolution, or shape on which a document is viewed will vary from reader to reader, and that therefore layout, font size, and other choices for good readability should be at least partly up to the reader, not the author. The result is that well designed HTML is excellent for browsing, but clumsy for printing.

Most authors are not used to such flexibility, they are used to producing static documents whose appearance is the same for everyone, because, for example, they are copies of a piece of paper. **If you require your readers to see an exact replication of what your document looks like to you, then you cannot use HTML to transmit it, no matter what format it starts in.** That is true not just for translated TeX but also for any authoring tool from which HTML is to be produced. The only way to produce documents whose appearance is completely controlled is to represent them in a page layout language such as such as PDF or Postscript or, for that matter, DVI. These formats are not as good as HTML for browsing, despite substantial hyperlinking ability in PDF, but they are better for transmitting a printable copy. Parentheticaly, word processor formats are less satisfactory for transmitting printable copy, hopeless for browsing, and unreliable for archiving because of the instability of the format.

Mathematics

TeX's excellent mathematical capabilities are absent from HTML and browsers. There are then two main choices for representing equations in HTML: using bit-mapped images, or using browser fonts and tables for layout. The advantage of the bit-mapped approach is that it uses capabilities that are essentially universal to every graphical browser. Its disadvantages are that it requires a separate graphical file for every equation, which becomes very cumbersome and slow to download. Also the alignment and sizing of the graphical equations is uncertain with respect to the rest of the text. The advantages of the font and table approach used by TtH are that one HTML document contains all the information, giving portability and speed of download. The disadvantages are that it depends on having the symbol font accessible on the browser, and that the equation layout is not as compact or elegant as TeX's.

The MathML standard has been developed to represent mathematics in electronic documents. MathML is not HTML. Popular browsers do not currently (Mar 2003) render MathML without additional plugin software or fonts. The standard is in any case that MathML is supported within XML not strictly HTML. What is holding up wider adoption of MathML is not questions of production of MathML, since translators such as TtM are fully up to that job, rather it is the weakness of support in leading browsers. But even when and if MathML is routinely supported by browsers out of the box, documents' appearance will still be in the hands of the browser not the author.

Conclusion

So should I translate to HTML? If you want to provide the easiest browsable format, yes. If you feel it is essential to control the precise layout for aesthetic or other reasons, no. But notice the answer has nothing to do with whether the format starts as TeX.

Aus (Should I translate TeX to HTML or not? Ian Hutchinson; http://hutchinson.belmont.ma.us/tth/shouldi.html)

Bibliography

KNUTH, DONALD , Tau Epsilon Chi (TEX) , a SYSTEM for technical text. Revised version of Stanford Computer Science report number STAN-CS-78-675. American Mathematical Society, Providence, R.I., 1979. 200 pp. \$8.80 (paperbound). ISBN 0-8218-0209-7

KRANTZ, STEVEN G. 1997. A Primer of Mathematical Writing. Providence: American Mathematical Society. 01 TAO 1144. 0 8218 0635 1.

KRANTZ, STEVEN G. 2001. Handbook of Typography for the Mathematical Sciences. Boca Raton: Chapman & Hall / CRC.

SWANSON, E. 1979. Mathematics into Type. Providence : American Mathematical Society.

COMMUNITTE ON ELECTRONIC INFORMATION AND COMMUNICATION OF THE IMUBEST CURRENT PRACTICES : Recommendations on Electronic Information Communication (2002), www.ceic.math.ca/ceic_docs/best_practices/Best-Practices.pdf