TEX user habits versus publisher requirements

Lolita Tolenė

Abstract

Typesetters always balance on the thin line between unlimited author creativity and strict publisher requirements to produce full-text XML. In this paper we present both sides.

TEX is designed in a way that offers wide capabilities to achieve the desired goal in many different ways. Therefore a huge collection of TEX packages has been created over the years. Even more local macros are used every day. We present which TEX packages are commonly used for the scientific content and what proportion of them comes from the standard sources, such as CTAN and TEX Live. We give some insights into authors' habits using TEX for writing scientific content. Also keeping XML in mind, we discuss how and why these habits are important for typesetters while preparing papers for publishing.

1 Introduction

Scientific content preparation for publication is a substantive task, where a typesetter must balance the researcher's vision of how the content is best presented for the scientific community with the publisher's requirements for the journal style and XML[1] content. The XML format has become a standard in storage and making the electronic documents available. Therefore almost all publishing houses we have encountered provide a DTD[2] for XML production, which defines structure directed not only to appearance, but very often to the meaning of the content.

In our workflow, PDF and XML are produced from LATEX documents. TEX, the formatting engine of LATEX 2ε, is highly portable and free. Therefore the system runs on almost any hardware platform available [2]. So TEX has become the standard text processing system in many academic high-level scientific and research institutions.

In processing the incoming LATEX[3] manuscripts, typesetters strongly depend on the stability of TEX distributions and source file contents. TEX is designed in a way that offers wide capabilities to achieve the desired goal in many different ways. Therefore a huge collection of TEX packages has been created over the years. Even more local macros are used every day. This is all very attractive for the authors, but with the XML format in mind, it often becomes a burden.

In the following we discuss difficulties related to manuscripts becoming printed copy, while meeting publisher requirements (Section 2). In Section 3 we provide a statistical overview of about 90 000 STM (scientific, technical and medical) LATEX papers prepared for publishing in the last 7 years. In Section 4 some final remarks will be given.

2 From manuscript to printed copy

A manuscript prepared for publication eventually becomes printed copy, meeting journal style requirements. It also contains enriched structure, which is converted into XML, valid for a publisher-specific DTD. Requirements directed to the meaning of content are the most difficult to fulfil and we always search for some ways to ease this process. In the following we discuss most common obstacles for producing a valid XML structure from LATEX documents.

LATEX is used to display the content in the desired way, very often forgetting about the meaning and consistency. Broken math formulas (see the upper part of Fig. 1) or a single phrase split across several cells in a table (see LATEX code and its output in the lower part of Fig. 1) are good examples.

LATEX can be used to change the appearance of some content or to create some symbols in many different ways, but often such code has no equivalent in the XML (see examples in Fig. 2). Such structures have to be replaced with a Unicode symbol or converted to pictures.

Publisher requirements for XML usually state to use MathML[4] for mathematical content. A graphic made from a formula is not very pleasant to the reader's eye; it does not scale so smoothly as a MathML object and most importantly it contains no constituent information, and is not editable.

Some content enrichment is done having in mind world wide databases, identifying authors by ORCID (Open Researcher and Contributor ID) or another code and connecting them to their papers, counting paper citations and determining journal impact factors, connecting research sponsors to grant numbers. Therefore frontmatter and backmatter are crucial parts of the paper. Depending on the publisher, requirements for, e.g., the bibliography references and their citation links are very strict and structure-difficult. One of the ways to fulfil these requirements is to create a hooked version of some standard TEX distribution package (such as `natbib` or `hyperref`), which generates necessary additional XML-oriented

[1] e**X**tensible **M**arkup **L**anguage.

[2] **D**ocument **T**ype **D**efinition.

[3] We rarely encounter manuscripts written in plain TEX, so we use LATEX concepts throughout this article.

[4] **Math**ematical **M**arkup **L**anguage.

```
'$R=\{x|x$ is real $\}$' instead of '$R=\{x|x \mbox{ is real } \}$'

                                    ***

\begin{tabular}{ccccc}
\hline
Sample & Depth (cm) & Weight of Sample  & CRS & Pb-210 age \\
Number &  & Counted (g) & sediment accumulation & (year AD)\\
&&& rate (g/cm$^{2}/$yr)$^{\mathrm{a}}$\\
\hline
...
\end{tabular}
```

Sample Number	Depth (cm)	Weight of Sample Counted (g)	CRS sediment accumulation rate (g/cm^2/yr)[a]	Pb-210 age (year AD)

...

Figure 1: LaTeX code examples of a broken math formula and table headings with phrases split into separate cells.

```
\raisebox{.2em}{$n$}\big/\raisebox{-.2em}{$m$} ⟹ n/m

                                    ***

${\longrightarrow\hspace*{-3.1ex}{\circ}\hspace*{1.9ex}}$ ⟹ ⊸→

                                    ***

$1\!\!1$ ⟹ 𝟙 instead of \usepackage{dsfont}$\mathds{1}$ ⟹ 𝟙

                                    ***

 \newcommand{\forkindep}[1][]{%
     \mathrel{\mathop{\vcenter{\hbox{\oalign{
         \noalign{\kern-.3ex}
         \hfil$\vert$\hfil\cr\noalign{\kern-.7ex}$\smile$\cr  ⟹ ⫰
         \noalign{\kern-.3ex}}
         }}}\displaylimits_{#1}}
     }
```

Figure 2: Examples of symbols created using LaTeX: on the left-hand side is the source code, on the right-hand side, its output.

content without changing the user output. Such actions are extremely dependent on stability of packages and TeX distributions.

On our side, as typesetters, there are few LaTeX to XML converters being used (like TeX4ht or the one described in [1]). Also there are some thoughts to explore LuaTeX-based converter possibilities. Each of them, theoretical or practical, have flaws different than others and one thing they all have in common — in order to produce an XML valid for a publisher-provided DTD, the source content has to be prepared, either changing the TeX source directly or using available TeX distribution tools.

Author creativity can often make the manuscript processing a very hard task. Looking at LaTeX document content, from a typesetter's point of view there are a few important highlights: document class and style packages declaring the formatting of the paper and locally defined macros. Manuscripts constructed with a *heavy* and deep understanding of TeX structures require special accuracy — in order to create an XML which meets publisher requirements some of these structures are dismantled and replaced throughout a corpus (in other words, expanded), and others are converted into pictures while producing an XML. On the other hand, manuscripts using only *light* macros, such as defining repeatedly appearing phrases, measurement units etc., and packages found in one of the main TeX distributions (such as TeX Live or MiKTeX) or CTAN,[5] require very little intervention, mainly oriented toward contex-

[5] **C**omprehensive **T**eX **A**rchive **N**etwork.

tual enrichment for XML production (e.g., author information, funding related information connection to appropriate databases).

If the author uses a publisher-provided template, few changes are noticeable, whereas author-created formatting usually results in a completely different layout from the prepared printed copy, which makes it difficult to notice some unintended mismatch to original output content. Also the author, having put so much work into creating the desired layout, often feels disappointed by the outcome.

Manuscripts written using mostly unstructured plain text usually do not change much from the layout point of view, but the corpus must be given a contextual meaning. Also, strange combinations of primitive TEX command sequences, where usually some widely-known standard coding should have been applied, typically need to be replaced with the standard coding, but a human has to decide whether this is the case. Such manuscripts require reading the author's mind to some extent, e.g., where the theorem or its proof ends — especially difficult if these structures are nested, i.e., a proof contains theorems and proofs of its own, whether the two letters combined into a single glyph should be replaced with an appropriate LATEX command sequence, or if this is some field-related denotation and should be left untouched (for XML a picture should be generated from this symbol), etc.

3 Manuscript content analysis

Data description This article provides a statistical overview of about 90 000 STM (scientific, technical and medical) LATEX papers which have been prepared for 252 journals of well-known publishing houses such as Elsevier, Springer, Mattson Publishing Services, BioMed Central, IOS Press, International Press, and others. The data covers the years 2010 to 2016. Manuscripts have not been sorted in any way, therefore they include random nationalities, institutions, science fields, etc.

The provided results were gathered by analyzing manuscript source files (TEX), which show what researchers use for writing STM content. In order to see what is used overall, `.fls` and `.log` files have been analyzed. Only a small number of manuscripts are sent with compilation output files attached. In the current set `.log` files were found for 6% of manuscripts. Therefore compilation output files must have been produced by recompiling gathered manuscripts. For this purpose three distributions of TEX Live, released in 2010, 2014 and 2016, were at our disposal. The following engines have been used for successful compilation of about 90% of manuscripts: pdfTEX (`latex`, `pdftex`, `pdflatex`), LuaLATEX (`lualatex`), XETEX (`xelatex`). In order to generate the `.fls` files, the option `-recorder` was used.

Table 1: Formats used by researchers for manuscript compilation.

Format	Manuscript count
pdflatex	2962
latex	2489
platex	54
xelatex	52
tex	11
amstex	4
pdftex	4
platex-sjis	4
lualatex	3
eplatex	1
mpost	1
uplatex	1

Of course, we encounter only a portion of papers produced worldwide. Therefore, in most cases concrete numbers have no meaning here, and only percentages will be provided. All of the statistical data presented here can be accessed at `github.com/vtex-soft/statistics.tex-manuscripts`, and interested readers are encouraged to explore further.

One of the main interests in analyzing this data is to get a picture of which TEX family tools are most popular from a researcher's point of view and how it changes (if it does) over the years.

TEX tools used As noted above, only 6% of manuscripts were provided with their compilation `.log` files. Additionally, 20% of cases had PDF files matching the `.tex` filename. Table 1 shows compilation formats used by researchers, extracted from `.log` file content and PDF metadata. While the most commonly used engine is pdfTEX, 10 252 (48%) manuscripts were compiled to DVI first instead of directly producing a PDF file.

Most researchers use the latest TEX distribution version. But as one can see from Fig. 3, there are a number of authors who compile their manuscripts with TEX distributions up to ten years old.

Further analysis has been split into two main parts, separating the document classes and packages used.

Document classes Throughout the papers, 366 unique document classes were found. Only 15% are in TEX Live distributions since 2010, 2 are in the current CTAN file list (namely `svjour` and `smfart`) and other classes are distributed by publishing houses or created by authors (see Table 2).

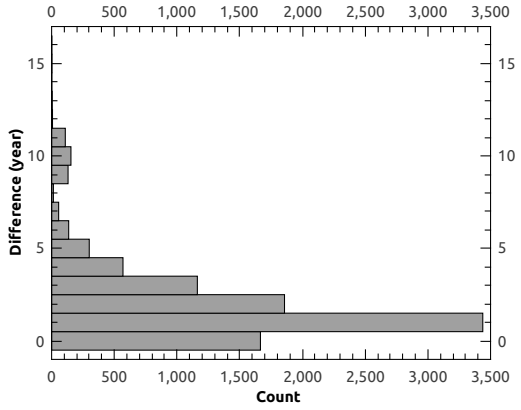

Figure 3: Age frequency of TEX distribution originally used by manuscript authors.

Table 2: Counts of unique document classes and packages in the analyzed manuscripts.

	Classes count	Packages count
Total	366	1847
Since 2015	143	1023
In TEX Live 2010–2016	55	996
In TEX Live 2016	48	970
In CTAN	2	66

Most commonly, manuscripts were provided using the `article`, `elsarticle`, `amsart`, `svjour3` and `revtex4` classes (see Fig. 4). All of these, except for the `svjour3` class, are included in TL2016. The `svjour3` class is provided by *Springer*. Over the years it has replaced `svjour` and `svjour2` classes, but only the first version, `svjour`, is currently found on CTAN.

While the `article` class is used independent of publishing houses, for other classes in Table 3 one can see a relationship with the publishing house by which a manuscript has been accepted. This relationship between class and publisher is natural, because the latter usually promotes certain classes for particular journals or groups of journals, and provides templates for authors to use. Such publisher-oriented manuscripts are more easily transferred into publication-ready papers and require less intervention, and therefore there are fewer typesetting errors and layout changes. Sadly, this is the case for only a relatively small number of papers. The total number of different classes used, shown in Table 3, shows that a substantial number of manuscripts are written using rare classes, or classes normally used for another publisher's journal.

Figure 4: Most common document classes used in the manuscripts. Word size reflects frequency.

LATEX 2_ε was introduced over 20 years ago, but we still encounter substantial use of the outdated LATEX 2.09 version. Over 1000 of the analyzed papers were formatted using `\documentstyle` command (the most often loaded class is `article`).

Packages Only 5% of manuscripts do not contain packages loaded in addition to the document class. Nearly 2000 unique packages were used throughout the manuscripts (see Table 2). More than half of them were found in TEX Live, 66 more are in the current CTAN file list (e.g., `psfig`, `axodraw`, `picins`, etc.; 10 of them are obsolete) and other files are distributed by publishers or created by authors.

The most common packages are shown in Fig. 5 and Table 4. The top of the list is stable throughout the entire time span analyzed: the most commonly used are the American Mathematical Society (AMS) packages, then there are `graphicx`, `color`, `hyperref`, `inputenc`, `mathrsfs` and `epsfig`. Some packages have become more popular in the last two years, notably `hyperref` and `tikz`. It is interesting to note that, while according to CTAN the `graphicx` package is preferred over `epsfig`, use of the latter is decreasing only slowly.

The data shows that a few packages have been used only with certain document class: `fix-cm` with `svjour3` class in 98% of cases, `spr-astr-addons` with `aastex` class in 100% of cases. Packages like the AMS bundle are more likely to be used with any class. Fig. 6 shows how in the last two years used packages are related to the most common `article` class: mostly it is used in combination with styles related to layout formatting, such as `fullpage`, `fancyhdr`, `indentfirst`, etc.

Among the less frequently used packages are a few new styles:[6] `mathpartir` (21 uses, on CTAN

[6] Here a style is called new if it is included in TEX Live 2016, but not in TL2014.

Table 3: The distribution of document classes according to the publishing house to which a manuscript has been submitted, as percentages.

Class	Last known source	BMC	DUP	Elsevier	International Press	IOS Press	Mattson	Springer
aastex[a]	TL2014							6
aicom2e	IOS Press					17		
amsart	TL2016	22	74	19	2	1	11	8
article	TL2016	42	19	29	15	16	37	29
bmc_article[a]	BMC	10						
bmcart[a]	BMC	15						
elsarticle[a]	TL2016	3		36				1
imsart[a]	IMS				1		46	
ios-book-article	IOS Press					4		
iosart2c	IOS Press					28		
ipart[a]	Intl. Press				78			
jaise2e	IOS Press					9		
svjour3	Springer	4						34
	Total (%)	96	93	84	96	75	94	77

[a] Document class uses `article` as parent, therefore the `article` class is used in 59% of all cases.

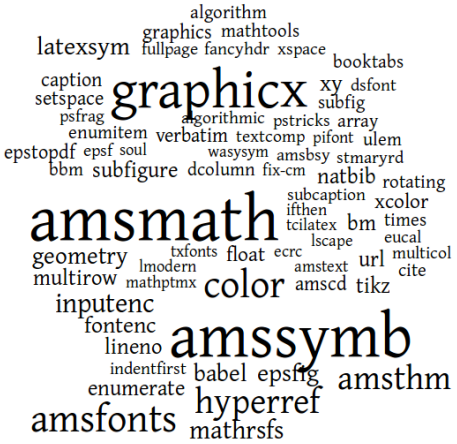

Figure 5: The 70 most common packages, extracted from manuscripts since 2015. Word size reflects frequency.

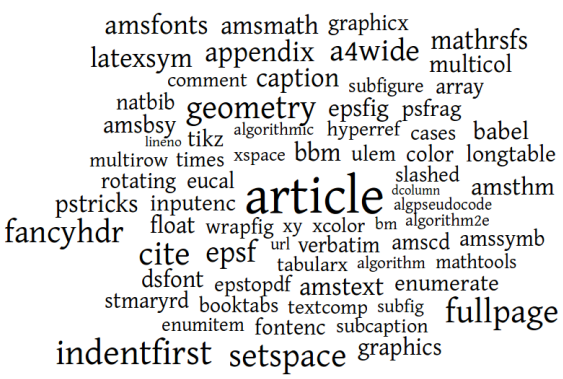

Figure 6: `article` class in relation to the packages shown in Fig. 5. Word size reflects frequency.

since 2016-02-26), `pgfornament` (3 uses, on CTAN since 2016-03-09), `prftree` (1 use, on CTAN since 2014-12-02), and `pstring` (1 use, on CTAN since 2017-01-05). The `todonotes` package appeared on CTAN in 2008-09-06 and its usage in the gathered manuscripts has steadily increased over the years. The package `subcaption` is on CTAN since 2002, but only in 2013 does this package appear among those used, with an increasing frequency since then.

A comparison of the packages loaded directly by the user (extracted from `.tex` files) and those loaded indirectly (extracted from `.fls` files) shows that many packages are bundles of files, and where the user loads only, e.g., the `graphicx` package, by default the `graphics` and `keyval` files are also loaded. It is interesting to note that in 40% of cases the `natbib` and `url` packages are not loaded directly.

AMS packages also dominate when comparing manuscripts with publisher-ready papers. Packages like `url`, `fontenc`, `bm` in the latter set of papers are used with 96%–99% frequency instead of 6%–7%. The `xcolor` package is used 4% more frequently than `color`, but this package overall is used at half the rate than in original manuscripts. Packages rarely used by authors such as `etoolbox` (94%), `ifthen` (18%), `dcolumn` (16%), `array` (16%), `afterpackage` (12%), `atveryend` (7%) are often used in publisher-ready paper preparation.

Package options 85% of the manuscripts' used packages were loaded without additional options. 1453 of the unique packages were never given an

Table 4: Most common packages, split into the time ranges 2010–2014 and 2015–2016.

Package *(2010–2014)*	Usage (%)	Package *(2015–2016)*	Usage (%)
amsmath	52	amsmath	59
amssymb	51	amssymb	56
graphicx	51	graphicx	46
amsfonts	22	amsfonts	28
color	19	color	28
amsthm	14	hyperref	23
epsfig	13	amsthm	21
hyperref	13	inputenc	14
latexsym	11	mathrsfs	12
inputenc	9	epsfig	11
mathrsfs	8	latexsym	11
babel	8	babel	10
natbib	8	geometry	10
url	7	fontenc	9
fontenc	7	xy	9
subfigure	7	enumerate	8
bm	6	url	8
graphics	6	tikz	8
multirow	5	multirow	8
xy	5	lineno	8
geometry	5	natbib	8
enumerate	5	bm	7

option, while 109 of them always had at least one option specified. Among the most frequently used packages, `hyperref` and `inputenc` were given options in 50% and 99% of cases, respectively. The most common options for `hyperref` were: `colorlinks` (23%), `citecolor` (23%), `urlcolor` (19%), `linkcolor` (15%), `breaklinks` (6%), `bookmarks` (6%). The most common options for `inputenc` were `latin1` (40%) and `latin9` (10%). The `graphicx` package was rarely given an option, but the most common was `dvips`.

Fonts While there were some manuscripts compiled with LuaTEX, no OpenType fonts were used. The `fontspec` package was used only two times and `.fls` files show that fonts were loaded using TFM files, Type 1, and virtual fonts. The most common font families are shown in Table 5: mostly the default Computer Modern family is used, unless `amsfonts` package is loaded, from which the `symbols`, `cmextra`, `euler` fonts are used.

4 Reflections

In this article we have briefly presented some statistical data gathered from about 90 000 STM manuscripts. The data shows that researchers most often use stable well-known packages and document classes,

Table 5: Most common font families. Data extracted from `.fls` files.

Font family	Usage (%)
cm	95
amsfonts	92
rsfs	13
symbol	11
zapfding	10
times	10
ec	8
xypic	7
txfonts	6
stmaryrd	3
courier	3
wasy	2
helvetic	1
bbm	1
cm-super	1
doublestroke	1
esint	1
lm	1
bbold	1

while new packages are promoted very slowly. At the same time, authors tend to create and use their own little TEX tools.

Creating non-standard structures or formatting of the look of the manuscript, without consideration of the final structure-based product, creates many obstacles for processing author manuscripts into to-be-published papers. Such actions imply that the knowledge of the good LATEX practices is not spread widely enough, or authors are simply not familiar with publishing-related processes. On the other hand, this creativity shows that TEX and its features have been found useful and popular among the worldwide scientific community.

References

[1] V. Kriaučiukas, L. Razinkovas, and L. Žamoitinaitė. Parsing LATEX for LATEX. In *BachoTEX 2015 proceedings*, pages 31–36. GUST, 2015.

[2] T. Oetiker, H. Partl, I. Hyna, and E. Schlegl. *The Not So Short Introduction to LATEX 2ε*, 2015. `https://ctan.org/pkg/lshort`.

◇ Lolita Tolenė
VTEX
Mokslininkų 2a
Vilnius LT-08412, Lithuania
`lolita.tolene (at) vtex dot lt`