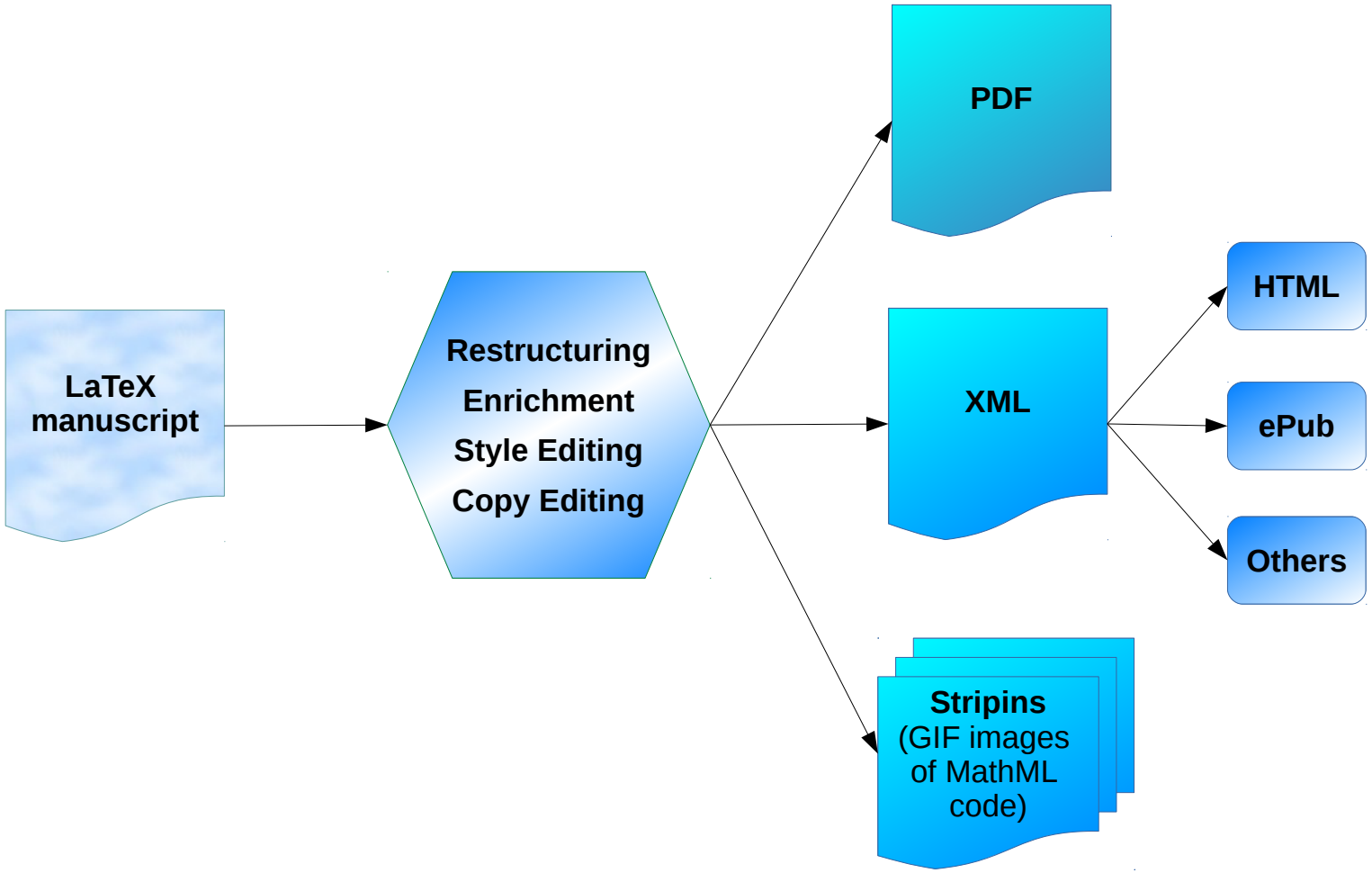
TEX users habits versus publishers requirements

Lolita Tolenė
lolita.tolene@vtex.lt

May 2, 2017

VTeX

LaTeX first workflow



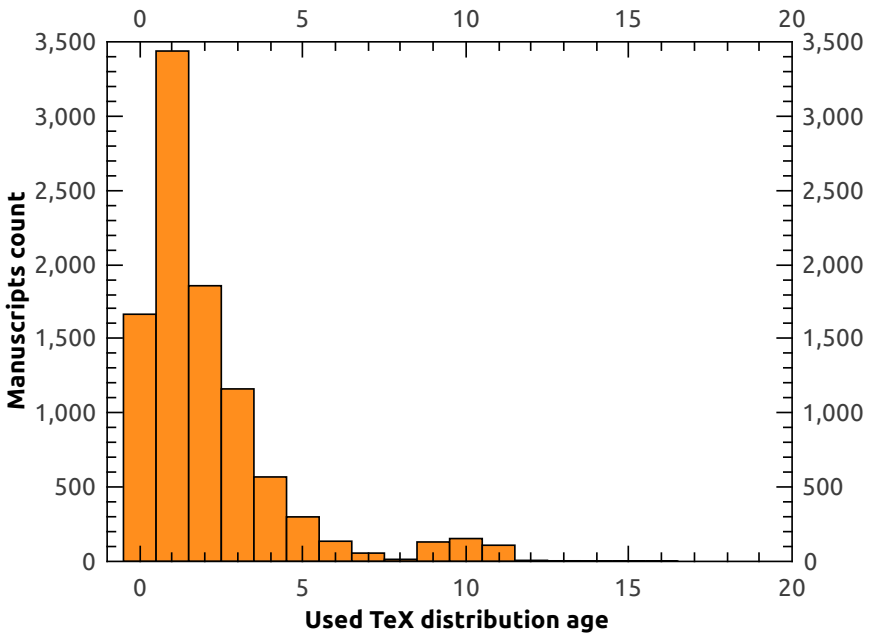
About LaTeX manuscripts

- $\sim$ 90000 LaTeX manuscripts published during 2010–2016 in 252 STM journals of publishing houses:
  - *Elsevier*
  - *Springer*
  - *Mattson Publishing Services*
  - *BioMed Central*
  - *IOS Press*
  - *International Press*

About LaTeX manuscripts

- ∼ 90000 LaTeX manuscripts published during 2010–2016 in 252 STM journals of publishing houses:
  - *Elsevier*
  - *Springer*
  - *Mattson Publishing Services*
  - *BioMed Central*
  - *IOS Press*
  - *International Press*

- LOG files attached in 6% of cases
- matching PDF files attached in +20% of cases



About LaTeX manuscripts

- $\sim$ 90000 LaTeX manuscripts published during 2010–2016 in 252 STM journals of publishing houses:
  - *Elsevier*
  - *Springer*
  - *Mattson Publishing Services*
  - *BioMed Central*
  - *IOS Press*
  - *International Press*

For-mat	Manuscripts 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

Analyzing LaTeX manuscripts

Extracted $\sim$ 90000 LaTeX manuscripts covering 2010 to 2016

Recompiled with TeX Live 2010, 2014 or 2016

- 90% successful compilation
- used pdfTeX (`latex`, `pdftex`, `pdflatex`), LuaTeX (`lualatex`), XeTeX (`xetex`)
- Generated FLS files by passing `--recorder` option

Extracted data from TEX frontmatters

- 85% created XML structure for exact analysis
- The rest analyzed using `RegExp`

Determined the main TEX file of a bundle by the criteria

- TEX file is standalone in structure
- TEX file is not used for other TEX file compilation as input
- TEX file is not a *style* or *class* in structure

Analyzing LaTeX manuscripts

Extracted $\sim$ 90000 LaTeX manuscripts covering 2010 to 2016

Recompiled with TeX Live 2010, 2014 or 2016

- 90% successful compilation
- used pdfTeX (`latex`, `pdftex`, `pdflatex`), LuaTeX (`lualatex`), XeTeX (`xetex`)
- Generated FLS files by passing `--recorder` option

Extracted data from TEX frontmatters

- 85% created XML structure for exact analysis
- The rest analyzed using RegExp

Determined the main TEX file of a bundle by the criteria

- TEX file is standalone in structure
- TEX file is not used for other TEX file compilation as input
- TEX file is not a *style* or *class* in structure

Analyzing LaTeX manuscripts

Extracted $\sim$ 90000 LaTeX manuscripts covering 2010 to 2016

Recompiled with TeX Live 2010, 2014 or 2016

- 90% successful compilation
- used pdfTeX (`latex`, `pdftex`, `pdflatex`), LuaTeX (`lualatex`), XeTeX (`xetex`)
- Generated FLS files by passing `--recorder` option

Extracted data from TeX frontmatters

- 85% created XML structure for exact analysis
- The rest analyzed using `RegExp`

Determined the main TeX file of a bundle by the criteria

- TeX file is standalone in structure
- TeX file is not used for other TeX file compilation as input
- TeX file is not a *style* or *class* in structure

Analyzing LaTeX manuscripts

Extracted $\sim$ 90000 LaTeX manuscripts covering 2010 to 2016

Recompiled with TeX Live 2010, 2014 or 2016

- 90% successful compilation
- used pdfTeX (`latex`, `pdftex`, `pdflatex`), LuaTeX (`lualatex`), XeTeX (`xetex`)
- Generated FLS files by passing `--recorder` option

Extracted data from TEX frontmatters

- 85% created XML structure for exact analysis
- The rest analyzed using `RegExp`

Determined the main TEX file of a bundle by the criteria

- TEX file is standalone in structure
- TEX file is not used for other TEX file compilation as input
- TEX file is not a *style* or *class* in structure

Document classes & packages

	Unique classes	Unique packages
Total	365	1845
Since 2015	143	1023
In TeX Live 2010–2016	55	996
In TeX Live 2016	48	970
In CTAN	2	66

Class	Last known source	Use per year
article	TL2016	
elsarticle	TL2016	
amsart	TL2016	
svjour3	Springer	
revtex4	TL2016	

Document classes & packages

	Unique classes	Unique packages
Total	365	1845
Since 2015	143	1023
In TeX Live 2010–2016	55	996
In TeX Live 2016	48	970
In CTAN	2	66

Class	Last known source	Use per year
`article`	TL2016	
`elsarticle`	TL2016	
`amsart`	TL2016	
`svjour3`	Springer	
`revtex4`	TL2016	

Document classes & packages

	Unique classes	Unique packages
Total	365	1845
Since 2015	143	1023
In TeX Live 2010–2016	55	996
In TeX Live 2016	48	970
In CTAN	2	66

Class	Last known source	Use per year
`article`	TL2016	
`elsarticle`	TL2016	
`amsart`	TL2016	
`svjour3`	Springer	
`revtex4`	TL2016	

Document classes & publishers

Class	Last known source	BMC	DUP	Elsevier	International Press	IOS Press	Mattson	Springer
`aastex`	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`	BMC	10%						
`bmcart`	BMC	15%						
`elsarticle`	TL2016	3%		36%				1%
`imsart`	IMS				1%		46%	
`ios-book-article`	IOS Press					4%		
`iosart2c`	IOS Press					28%		
`ipart`	Int. Press				78%			
`jaise2e`	IOS Press					9%		
`svjour3`	Springer	4%						34%
Total		96%	93%	84%	96%	75%	94%	77%

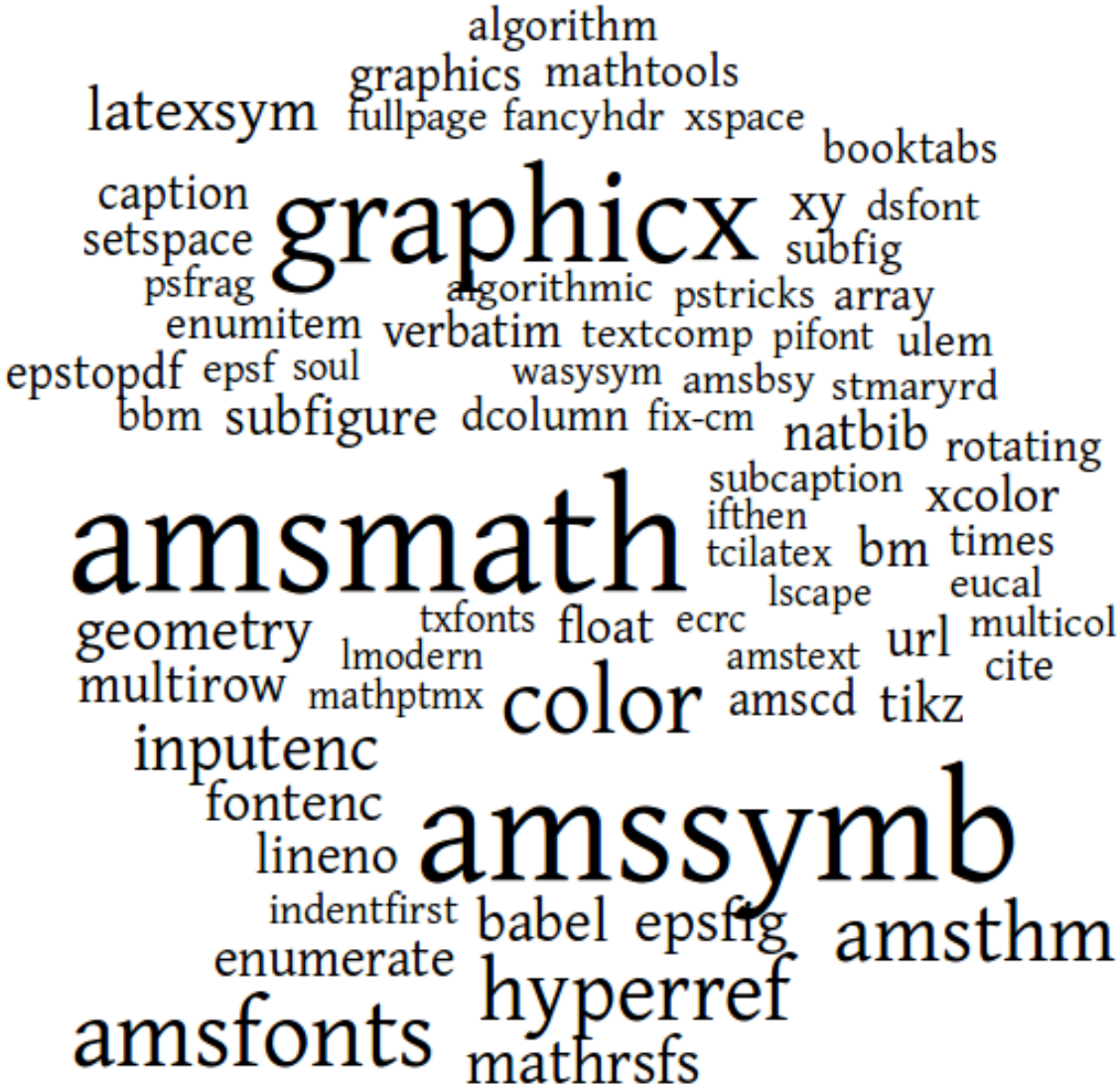
The `article` class is used in 59% of cases

Document classes & publishers

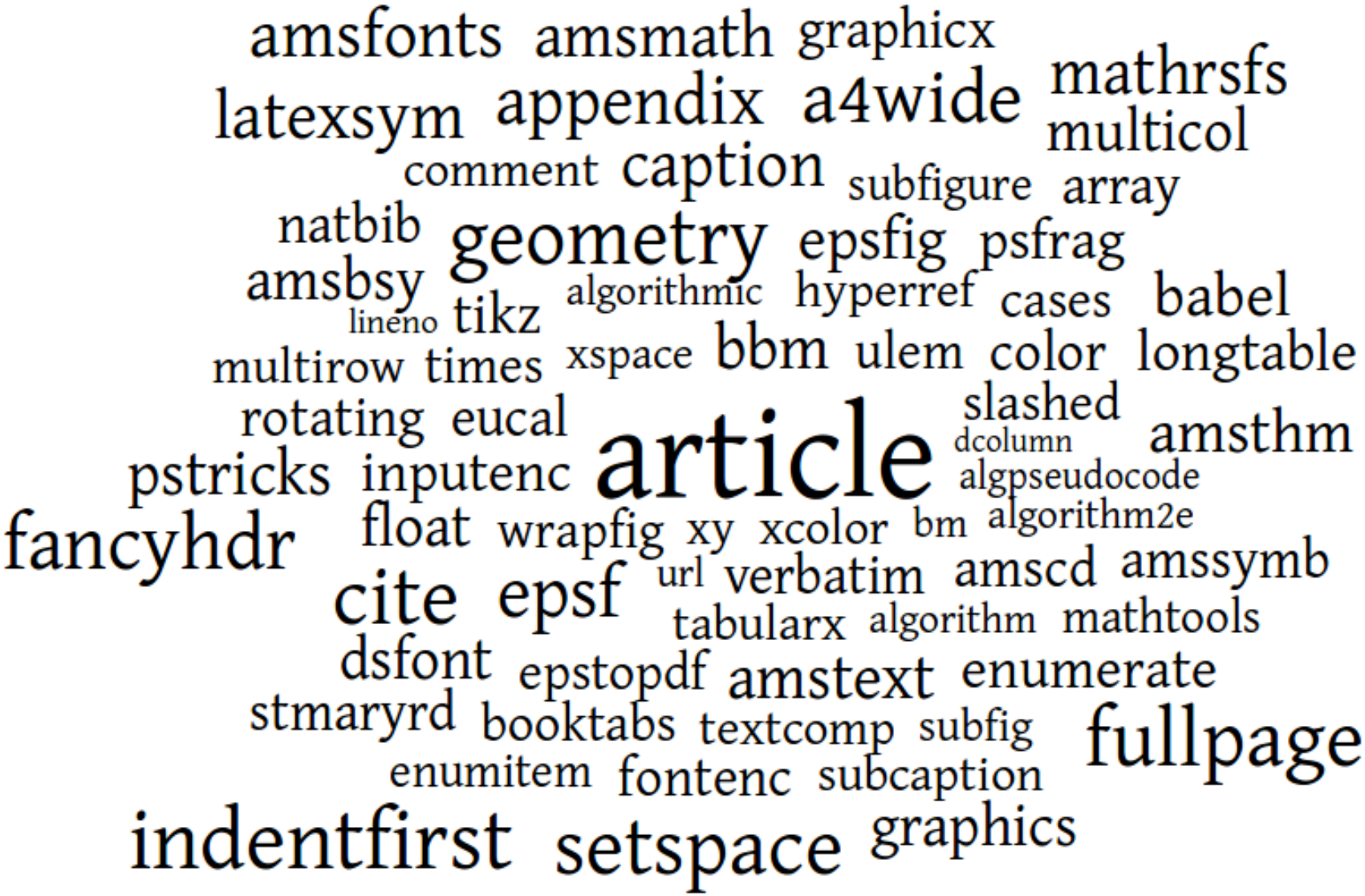
Class	Last known source	BMC	DUP	Elsevier	International Press	IOS Press	Mattson	Springer
aastex	TL2014							6
aicom2e	IOS Press					17%		
amsart	TL2016	22%	74%	19%	2%	1%	11%	8%
article	TL2016	42%	19%	29%	15%	16%	37%	29%
bmc_article	BMC	10%						
bmcart	BMC	15%						
elsarticle	TL2016	3%		36%				1%
imsart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

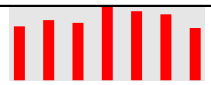
The article class is used in 59% of cases

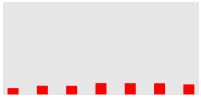
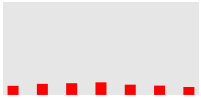
Most commonly used packages

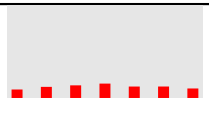
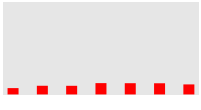
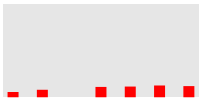
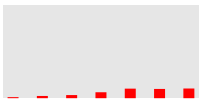


Packages relation to classes



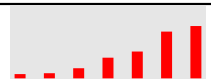
Package	Use count (2010–2014)	Use count (2015–2016)	Use per year
amsmath	52%	59% ↑	
amssymb	51%	56% ↑	
graphicx	51%	46% ↓	
amsfonts	22%	28% ↑	
color	19%	28% ↑	
amsthm	14%	21% ↑	
epsfig	13%	11% ↓	
hyperref	13%	23% ↑	
latexsym	11%	12% ↑	
inputenc	9%	14% ↑	
mathrsfs	8%	12% ↑	
babel	8%	10% ↑	
natbib	8%	8%	

Package	Use count (2010–2014)	Use count (2015–2016)	Use per year
`url`	7%	8% ↑	
`fontenc`	7%	9% ↑	
`subfigure`	7%	7%	
`bm`	6%	7% ↑	
`graphics`	6%	5% ↓	
`multirow`	5%	8% ↑	
`xy`	5%	9% ↑	
`geometry`	5%	10% ↑	
`enumerate`	5%	8% ↑	
`tikz`	3%	8% ↑	
`lineno`	2%	8% ↑	

Package	Use count (2010–2014)	Use count (2015–2016)	Use per year
url	7%	8% ↑	
fontenc	7%	9% ↑	
subfigure	7%	7%	
bm	6%	7% ↑	
graphics	6%	5% ↓	
multirow	5%	8% ↑	
xy	5%	9% ↑	
geometry	5%	10% ↑	
enumerate	5%	8% ↑	
tikz	3%	8% ↑	
lineno	2%	8% ↑	

Package	Use count (2010–2014)	Use count (2015–2016)	Class				Use per year
			amsart	article	elsarticle	svjour3	
mathtools	0.7%	3.2%	21%	21%	40%	5%	
ulem	1.5%	3.1%	9%	25%	32%	8%	
enumitem	0.8%	2.9%	23%	22%	34%	6%	
epsf	2%	1.6%	7%	46%	11%	5%	
microtype	0.5%	1.1%	16%	21%	40%	9%	
tcilatex	1.7%	1%	12%	73%	2%	1%	
todonotes	$< 0.4\%$	1.3%	16%	15%	41%	10%	
subcaption	$< 0.4\%$	2.3%	7%	23%	51%	2%	

- mathpartir – 21 uses, on CTAN since 2016-02-26
- pgfornament – 3 uses, on CTAN since 2016-03-09
- prftree – 1 use, on CTAN since 2014-12-02
- pstring – 1 use, on CTAN since 2017-01-05

Package	Use count (2010–2014)	Use count (2015–2016)	Class amsart	article	elsarticle	svjour3	Use per year
mathtools	0.7%	3.2%	21%	21%	40%	5%	
ulem	1.5%	3.1%	9%	25%	32%	8%	
enumitem	0.8%	2.9%	23%	22%	34%	6%	
epsf	2%	1.6%	7%	46%	11%	5%	
microtype	0.5%	1.1%	16%	21%	40%	9%	
tcilatex	1.7%	1%	12%	73%	2%	1%	
todonotes	$< 0.4\%$	1.3%	16%	15%	41%	10%	
subcaption	$< 0.4\%$	2.3%	7%	23%	51%	2%	

- mathpartir – 21 uses, on CTAN since 2016-02-26
- pgfornament – 3 uses, on CTAN since 2016-03-09
- prftree – 1 use, on CTAN since 2014-12-02
- pstring – 1 use, on CTAN since 2017-01-05

Package options

- 1845 unique packages were loaded in TEX files
- 85% packages used without options
  - 1453 were never given an option
  - 109 were always used with at least one option
- `hyperref` had an option in 50% cases:
  - `colorlinks` – 23%
  - `citecolor` – 23%
  - `urlcolor` – 15%
  - `breaklinks` – 6%
  - `bookmarks` – 6%
- `inputenc` had an option in 99% cases:
  - `latin1` – 40%
  - `latin9` – 10%

Package options

- 1845 unique packages were loaded in TEX files
- 85% packages used without options
  - 1453 were never given an option
  - 109 were always used with at least one option
- `hyperref` had an option in 50% cases:
  - `colorlinks` – 23%
  - `citecolor` – 23%
  - `urlcolor` – 15%
  - `breaklinks` – 6%
  - `bookmarks` – 6%
- `inputenc` had an option in 99% cases:
  - `latin1` – 40%
  - `latin9` – 10%

Going in deeper . . .

- 26.64% manuscripts define new command sequences and registers

Command sequence	Manuscripts		Publish ready files	
	Use count	Files count	Use count	Files count
countdef	2	2		
DeclareMathAlphabet	885	188	446	334
DeclareMathOperator	2087	743	221	28
DeclareMathSymbol	1777	410	1094	944
def	19565	9090	686	263
dimendef	4	4		
edef	432	140	15	11
font	4905	1636	166	81
gdef	334	141	30	20
let	2671	966	1047	747
newboolean	764	587	7	7
newbox	366	234	195	43
newcommand	1173	628	55	28
newcount	607	177	158	21
newcounter	5024	2661	498	273
newdimen	771	386	135	59
newenvironment	19658	10370	553	321
newfam	357	287	5	3
newif	675	464	58	28
newproof	2613	1788		
newsavebox	338	232	57	28
newskip	124	90	6	5
newtoggle	18	13		
newtoks	259	41	2	2
newwrite	6	6		
providecommand	69	25	1	2
Total	65484		5435	

Going in deeper . . .

- 26.64% manuscripts define new command sequences and registers
- Conditionals in 3.16% manuscripts
- Changed category codes in 14.55% manuscripts
  - 35 manuscripts had actual `\catcode` command used 103 times overall
  - `\makeatother` command is often unbalanced with `\makeatletter`
- 26% manuscripts had structures converted into pictures instead of unicode or `MathML` object

XML requirements

Content Model

```
<!ELEMENT  article      %article-full-model;                    >
```

Expanded Content Model

```
(front, body?, back?, floats-group?, (sub-article* | response*))
```

Description

The following, in order:

- <front> Front Matter
- <body> Body of the Document, zero or one
- <back> Back Matter, zero or one
- <floats-group> Floating Element Group, zero or one
- Any one of:
  - <sub-article> Sub-article, zero or more
  - <response> Response, zero or more

Source: https://jats.nlm.nih.gov/publishing/

XML requirements

Content Model

```
<!ELEMENT  contrib-group
                        %contrib-group-model;                        >
```

Expanded Content Model

```
((contrib)+, (address | aff | aff-alternatives | author-comment | bio | email | ext-link | on-behalf-of | role | uri | xref)*)
```

Description

The following, in order:

- <contrib> Contributor, one or more
- Any combination of:
  - <address> Address/Contact Information
  - <aff> Affiliation
  - <aff-alternatives> Affiliation Alternatives
  - <author-comment> Author Comment
  - <bio> Biography
  - Linking Elements
    - <email> Email Address
    - <ext-link> External Link
    - <uri> Uniform Resource Identifier (URI)
  - <on-behalf-of> On Behalf of
  - <role> Role or Function Title of Contributor
  - <xref> X (cross) Reference

This element may be contained in:

<article-meta>, <collab>, <front-stub>, <journal-meta>, <sec-meta>, <supplement>

Source: https://jats.nlm.nih.gov/publishing/

LaTeX into XML: difficulties

Broken math formula

```
$R=\{x|x$ is real $\}$
```

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

Phrases split into separate cells

```
\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)
...				

LaTeX into XML: difficulties

Broken math formula

```
$R=\{x|x$ is real $\}$
```

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

Phrases split into separate cells

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

LaTeX into XML: difficulties

Ignoring standards

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

```
$\nicefrac{n}{m}$
```
$\Longrightarrow$ $\boxed{n/m}$

```
$1\!\!1$
```
$\Longrightarrow$ $\boxed{\mathbb{1}}$

```
\usepackage{dsfont}$\mathds{1}$
```
$\Longrightarrow$ $\boxed{\mathbb{1}}$

Accented letters

```
F$\ddot{\rm o}$rster
```
$\Longrightarrow$ $\boxed{\text{Förster}}$

```
F\"orster
```
$\Longrightarrow$ $\boxed{\text{Förster}}$

LaTeX into XML: difficulties

Ignoring standards

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

```
$\nicefrac{n}{m}$
```
$\Longrightarrow$ $\boxed{n/m}$

```
$1\!\!1$
```
$\Longrightarrow$ $\boxed{\mathbb{1}}$

```
\usepackage{dsfont}$\mathds{1}$
```
$\Longrightarrow$ $\boxed{\mathbb{1}}$

Accented letters

```
F$\ddot{\rm o}$rster
```
$\Longrightarrow$ $\boxed{\text{F\"orster}}$

```
F\"orster
```
$\Longrightarrow$ $\boxed{\text{F\"orster}}$

Refering to unnumbered equation

```
\begin{equation*}
  b+2\tag{*}\label{eq2}
\end{equation*}
```

$\Longrightarrow$ $b+2 \qquad (*)$

```
\eqref{eq2}
```

$\Longrightarrow$ $\boxed{(*)}$

```
\begin{eqnarray}
  d+4 \nonumber \\
  e+5\label{eq4}\nonumber \\
  f+6\label{eq5}\nonumber
\end{eqnarray}
```

$\Longrightarrow$ $d+4$, $e+5$, $f+6$

```
\eqref{eq4} and \eqref{eq5}
```

$\Longrightarrow$ $\boxed{(1) \text{ and } (1)}$

LaTeX into XML: difficulties

Creating new symbols

```
${\longrightarrow\hspace*{-3.1ex}{\circ}\hspace*{1.9ex}}$
```

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

```
\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}}%
}
```

$\forkindep[1][]{}$

Finding new ways to surprise . . .

```
$\mathfrak{G}\hat{\circ}$
$\mathfrak{G}\hat{\color{white}\circ}$
```

$\mathfrak{G}\hat{\circ}$ $\mathfrak{G}\hat{\color{white}\circ}$

LaTeX into XML: difficulties

Creating new symbols

```
${\longrightarrow\hspace*{-3.1ex}{\circ}\hspace*{1.9ex}}$
```
$\longrightarrow\hspace*{-3.1ex}{\circ}\hspace*{1.9ex}$

```
\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}}%
}
```
$\forkindep$

Finding new ways to surprise . . .

```
$\mathfrak{G}\hat{\circ}$
```
$\mathfrak{G}\hat{\circ}$

```
$\mathfrak{G}\hat{\color{white}\circ}$
```
$\mathfrak{G}\hat{\color{white}\circ}$

Thank You

https://github.com/vtex-soft/statistics.tex-manuscripts

lolita.tolene@vtex.lt