H. Stamerjohanns

TUG, August 2021
1 Motivation
   - \LaTeX{} and XML
   - Why convert to XML?
   - Previous projects
   - texmlbus: change of focus
2 Website
   - Build system
3 Summary
Just use LaTeX

LaTeX is the format to write math

- millions of scientific publications have been written using LaTeX
- best way to produce high quality math typesetting

Drawbacks

- mixes form and content
- no real semantics
- style files change over time
- no formal validation
- long term preservation?
XML

- not something you want to directly edit
- document can be validated
  ⇒ possible archive format
- JATS *Journal Article Text Suite*
- MathML, XHTML
  ⇒ render document directly in web browser
- easier for searching and indexing tools, screenreaders
Project based on...

arxivml build system

- written at Jacobs University Bremen
- use \texttt{LaTeXML} to create XML
- mass conversion to XHTML
  \[ \approx 500,000 \] documents converted
- create real-world MathML
  \[ \Rightarrow \text{improve LaTeXML} \] \(^1\)

\[ ^1 \text{B. Miller and D. Ginev, https://dlmf.nist.gov/LaTeXML/} \]
Build system

- open source (MIT licence)
- implemented in scripting language (here \texttt{php})
- uses \texttt{SQL} database to store state
- distributes jobs on several hosts
- sets timeout for each job
- analyzes conversion process
  - checks files
  - parses the result files (\texttt{stderr.log})
  - classifies results
- stores information in DB
texmlbus: change of focus

- easy installation
- more interactivity
- other targets than XHTML
- create same target using different systems
texmlbus: change of focus

- easy installation
  ⇒ use Docker images
- more interactivity

- other targets than XHTML
- create same target using different systems
textmlbus: change of focus

- easy installation
  - use Docker images

- more interactivity
  - upload files via browser
  - import files directly from Overleaf

- other targets than XHTML

- create same target using different systems
texmlbus: change of focus

- easy installation
  - use Docker images

- more interactivity
  - upload files via browser
  - import files directly from Overleaf
  - schedule jobs via browser

- other targets than XHTML

- create same target using different systems
texmlbus: change of focus

- easy installation
  - use Docker images

- more interactivity
  - upload files via browser
  - import files directly from Overleaf
  - schedule jobs via browser

- other targets than XHTML
  - result table for each target

- create same target using different systems
texmlbus: change of focus

- easy installation
  - use Docker images

- more interactivity
  - upload files via browser
  - import files directly from Overleaf
  - schedule jobs via browser

- other targets than XHTML
  - result table for each target

- create same target using different systems
  - introduce stages (target combined with image)
  - needs subdirectories for each stage
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files (each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

SQL database
- workqueue
- statistics

manages entries in workqueue
- state and priority

reads workqueue

invokes make

stores analyzed log data

reads log files

default:
create result files and log files

1
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

SQL database
- workqueue
- statistics

manages entries in workqueue
- state and priority

reads workqueue

invokes make

stores analyzed log data

reads log files

default:
create result files and log files
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

default:
create result files and log files

Worker containers
- handle conversion
- call \texttt{latexml} and \texttt{latexmlpost} or conversion commands

Reads workqueue

Reads log files

Stores analyzed log data

Manages entries in workqueue
- state and priority

SQL database
- workqueue
- statistics

Invokes make

3 Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

SQL database
- workqueue
- statistics

manages entries in workqueue
- state and priority

reads workqueue
stores analyzed log data
invokes make
reads log files
default:
create result files and log files

H. Stamerjohanns
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

SQL database
- workqueue
- statistics

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

default:
create result files and log files

reads log files
invokes make
stores analyzed log data
reads workqueue
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

SQL database
- workqueue
- statistics

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

default:
create result files and log files

reads log files

invokes make

stores analyzed log data

reads workqueue

manages entries in workqueue
- state and priority
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

default: create result files and log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

SQL database
- workqueue
- statistics

reads workqueue

stores analyzed log data

reads log files

H. Stamerjohanns
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

SQL database
- workqueue
- statistics

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call \texttt{latexml} and \texttt{latexmlpost} or conversion commands

default:
create result files and log files

reads workqueue

stores analyzed log data

invokes make
**texmlbus build system**

**Interactive usage via browser**
- add and remove documents
- queue conversion jobs
- manages document files

**SQL database**
- workqueue
- statistics

**Build Manager**
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

**Worker containers**
- handle conversion
- call `latexml` and `latexmlpost` or conversion commands

**Document files** (each article in subdirectory)

**Build Manager** reads log files

**Worker containers** invokes make

**SQL database** stores analyzed log data

**Build Manager**

**Worker containers** create result files and log files

Manages entries in workqueue
- state and priority

**Interactive usage via browser**
- add and remove documents
- queue conversion jobs
- manages document files
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

SQL database
- workqueue
- statistics

reads log files
invokes make
stores analyzed log data

manages entries in workqueue
- state and priority

reads workqueue
default:
create result files and log files

Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files (each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

SQL database
- workqueue
- statistics

reads log files

create result files and log files

invokes make

stores analyzed log data

reads workqueue

manages entries in workqueue
- state and priority
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Worker containers
- handle conversion
- call \texttt{latexml} and \texttt{latexmlpost} or conversion commands

SQL database
- workqueue
- statistics

manages entries in workqueue
- state and priority

reads workqueue

reads log files

default:
create result files and log files

invokes make

stores analyzed log data
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Document files
(each article in subdirectory)

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

default:
create result files
and log files

reads log files

reads workqueue

stores analyzed log data

Worker containers
- handle conversion
- call latexml and latexmlpost or conversion commands

invokes make

SQL database
- workqueue
- statistics

管理工作队列文件
- 管理工作队列和优先级

H. Stamerjohanns
TUG, August 2021
Interactive usage via browser
- add and remove documents
- queue conversion jobs
- manages document files

Shared Volume
Document repository

MySQL DB
- statistics

Build Manager
- operates on workqueue
- schedules make jobs on worker containers
- analyzes log files

Webserver / PHP
- invokes make
- stores analyzed log data

Worker containers
- handles conversion commands
- call conversion commands

default:
create result files and log files

reads log files
reads workqueue
stores analyzed log data
reads log files
Statistics for samples

<table>
<thead>
<tr>
<th>result</th>
<th>count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>exception</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>notice</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>error</td>
<td>1</td>
<td>20.00</td>
</tr>
<tr>
<td>success</td>
<td>4</td>
<td>80.00</td>
</tr>
</tbody>
</table>

Detailed return values

<table>
<thead>
<tr>
<th>return value</th>
<th>count</th>
<th>%</th>
<th>marked for rerun</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing_macros</td>
<td>1</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td>warning</td>
<td>1</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td>no_problems</td>
<td>3</td>
<td>60.00</td>
<td></td>
</tr>
</tbody>
</table>

Top Fatal Errors for (xml)

Top Errors for (xml)
Upload and import articles

- Add files...
- Start upload
- Cancel upload
- Delete selected

Please specify a set when you import

Drop files here
Import articles from Overleaf

Set to import to: Please specify a set when you import

ProjectId: Id of project

Name: The project name you would like to use

Username: Your username

Import project
### overleaf

#### Alphabetic list of documents

<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Directory</th>
<th>pdf</th>
<th>xml</th>
<th>xhtml</th>
<th>jets</th>
<th>pagelimit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2021-08-03</td>
<td>overleaf/presentation</td>
<td>no problems</td>
<td>missing_macros</td>
<td>no_problems</td>
<td>queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td>18:34:38</td>
<td></td>
<td>EnFile DestFile</td>
<td>DestFile</td>
<td>EnFile DestFile</td>
<td>queue</td>
<td></td>
</tr>
</tbody>
</table>

[1]
<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Directory</th>
<th>pdf</th>
<th>xml</th>
<th>xhtml</th>
<th>jats</th>
<th>pagelimit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>queue</td>
<td>queue set</td>
<td>queue</td>
<td>queue set</td>
<td>queue</td>
</tr>
<tr>
<td>1</td>
<td>2021-08-03</td>
<td>samples/A_quick_guide_to_LaTeX</td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:40:36 queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td>17:40:36</td>
<td></td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:34:36 queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:34:36 queue</td>
<td>queue</td>
</tr>
<tr>
<td>2</td>
<td>2021-08-03</td>
<td>samples/basic</td>
<td>warning</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:34:46 queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td>17:39:48</td>
<td></td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:34:46 queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:34:46 queue</td>
<td>queue</td>
</tr>
<tr>
<td>3</td>
<td>2021-08-03</td>
<td>samples/testmath-sample</td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:35:56 queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td>17:39:48</td>
<td></td>
<td>missing_macros</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:35:56 queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03 17:35:56 queue</td>
<td>queue</td>
</tr>
</tbody>
</table>
This is pdfTeX, Version 3.14159265-2.6-1.40.29 (TeX Live 2019/Alpine Linux) (preloaded format=pdflatex 2021.5.23) 3 AUG 2021 15:40
entering extended mode
restricted \\write18 enabled.
\%\line parsing enabled.
**main.tex
(./main.tex
LaTeX2e <2020-02-02> patch level 5
L3 programming layer <2020-03-06>
\input{texmlbus}
Document Class: article 2019/12/20 v1.4l Standard LaTeX document class
\input{article}
File: size10.clo 2019/12/20 v1.4l Standard LaTeX file (size option)
)
\part{count167}
\section{count168}
\subsection{count169}
\subsubsection{count170}
\paragraph{count171}
\subparagraph{count172}
\figure{count173}
\table{count174}
\captionstyle{skip47}
\belowcaptionskip{skip48}
\bibindent{dimen134}
\input{amsfonts/amsfonts.sty}
Package: amsfonts 2013/01/14 v3.01 AMS font symbols
\input{amsfonts/amsfonts.sty}
Package: amsfonts 2013/01/14 v3.01 Basic AMSFonts support
\emptytoks=toks15
\symAMSa=\mathgroup4
\symAMSb=\mathgroup5
LaTeX Font Info: Redeclaring math symbol \hbar on input line 98.
LaTeX Font Info: Overwriting math alphabet \mathfrak in version \bold
(Font)
U/euf/m/n --> U/euf/b/n on input line 106.
\input{amsmath/amsmath.sty}
Package: amsmath 2020/01/20 v2.17e AMS math features
<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Directory</th>
<th>pdf</th>
<th>xml</th>
<th>xhtml</th>
<th>jats</th>
<th>pagelimit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>queue</td>
<td>queue</td>
<td>queue</td>
<td>queue</td>
<td>queue</td>
</tr>
<tr>
<td>1</td>
<td>2021-08-03</td>
<td>samples/A_quick_guide_to_LaTeX</td>
<td>no_problems</td>
<td>no_problems</td>
<td>no_problems</td>
<td>queue</td>
<td>warning</td>
</tr>
<tr>
<td></td>
<td>17:40:36</td>
<td></td>
<td>EnrFile</td>
<td>DestFile</td>
<td>2021-08-03</td>
<td>17:40:36</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>queue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2021-08-03</td>
<td>samples/basic</td>
<td>warning</td>
<td>no_problems</td>
<td>no_problems</td>
<td>queue</td>
<td>no_problems</td>
</tr>
<tr>
<td></td>
<td>17:39:48</td>
<td></td>
<td>EnrFile</td>
<td>DestFile</td>
<td>2021-08-03</td>
<td>17:34:46</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>queue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2021-08-03</td>
<td>samples/testmath-sample</td>
<td>no_problems</td>
<td>missing_macros</td>
<td>no_problems</td>
<td>queue</td>
<td>warning</td>
</tr>
<tr>
<td></td>
<td>17:39:48</td>
<td></td>
<td>EnrFile</td>
<td>DestFile</td>
<td>2021-08-03</td>
<td>17:35:56</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>queue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<?latex

<resource class="ltx_align_left" src="LaTeXXML.css" type="text/css"/>
<resource class="ltx_align_left" src="ltx-article.css" type="text/css"/>
<title class="ltx_align_left">Quick Guide to LaTeX</title>
</latex>
```
<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Directory</th>
<th>pdf</th>
<th>xml</th>
<th>xhtml</th>
<th>jats</th>
<th>pagelimit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>queue</td>
<td>queue</td>
<td>queue</td>
<td>queue</td>
<td>queue</td>
</tr>
<tr>
<td>1</td>
<td>2021-08-03</td>
<td>samples/A_quick_guide_to_LaTeX</td>
<td>no_problems</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03</td>
<td>17:40:36</td>
</tr>
<tr>
<td></td>
<td>17:40:36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2021-08-03</td>
<td>17:40:36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_problems</td>
<td>no_problems</td>
<td>no_problems</td>
<td>queue</td>
<td>queue</td>
</tr>
<tr>
<td>2</td>
<td>2021-08-03</td>
<td>samples/basic</td>
<td>warning</td>
<td>EnFile</td>
<td>DestFile</td>
<td>2021-08-03</td>
<td>17:34:12</td>
</tr>
<tr>
<td></td>
<td>17:39:48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2021-08-03</td>
<td>17:34:12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_problems</td>
<td>no_problems</td>
<td>no_problems</td>
<td>queue</td>
<td>queue</td>
</tr>
<tr>
<td>3</td>
<td>2021-08-03</td>
<td>samples/testmath-sample</td>
<td>no_problems</td>
<td>missing_macros</td>
<td>no_problems</td>
<td>queue</td>
<td>queue</td>
</tr>
<tr>
<td></td>
<td>17:39:48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2021-08-03</td>
<td>17:35:56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>no_problems</td>
<td>no_problems</td>
<td>no_problems</td>
<td>queue</td>
<td>queue</td>
</tr>
</tbody>
</table>
Quick Guide to LaTeX

A quick guide to LaTeX

What is LaTeX?

LaTeX (usually pronounced “LAY teck,” sometimes “LAH teck,” and never “LAY tex”) is a mathematics typesetting program that is the standard for most professional mathematics typesetting program TeX created by Donald Knuth of Stanford University (his first version appeared in 1978). Leslie Lamport was responsible for creating LaTeX, a more user-friendly version of LaTeX. Programmers created the current version, LaTeX 2ε.

Math vs. text vs. functions

In properly typeset mathematics, variables appear in italics (e.g., \( f(x) = x^2 + 2x - 3 \)). The exception to this rule is predefined functions (e.g., \( \sin(x) \)). Thus it is important to use the correct format for variables and functions correctly. See the difference between \( x \) and \( x_1 \) and \( -1 \), and \( \sin(x) \) and \( \sin (x) \).

There are two ways to present a mathematical expression—inline or as an equation.

Inline mathematical expressions

Inline expressions occur in the middle of a sentence. To produce an inline expression, place the mathematical expression between dollar signs ($). For example, typing \( 90^\circ \) yields 90\(^\circ\) is the same as \( \frac{\pi}{2} \) radians.

Equations

Equations are mathematical expressions that are given their own line and are centered on the page. These are usually used for important equations that deserve to be showcased or equations that cannot fit inline. To produce an inline expression, place the mathematical expression between the symbols \( \text{ and } \). Typing \( \frac{x}{y} \) yields \( \frac{x}{y} \).
The web interface

Result statistics

<table>
<thead>
<tr>
<th>result</th>
<th>count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>exception</td>
<td>1</td>
<td>2.33</td>
</tr>
<tr>
<td>notice</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>error</td>
<td>8</td>
<td>18.60</td>
</tr>
<tr>
<td>success</td>
<td>34</td>
<td>79.07</td>
</tr>
</tbody>
</table>

Detailed return values

<table>
<thead>
<tr>
<th>return value</th>
<th>count</th>
<th>%</th>
<th>marked for rerun</th>
</tr>
</thead>
<tbody>
<tr>
<td>fatal_error</td>
<td>1</td>
<td>2.33</td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>8</td>
<td>18.60</td>
<td></td>
</tr>
<tr>
<td>warning</td>
<td>24</td>
<td>55.81</td>
<td></td>
</tr>
<tr>
<td>no_problems</td>
<td>10</td>
<td>23.26</td>
<td></td>
</tr>
</tbody>
</table>

table as TeX
Summary

- `texmlbus` allows to convert documents and gather statistics about conversions
- especially useful to detect regressions with real-world documents
- `stages` allow to have same targets using different systems
- supports any converter
Things to be done

- add converters more easily
- help to improve \LaTeX{}XML
https://github.com/stamer/texmlbus

Thanks to Overleaf for their support!