texmlbus

H. Stamerjohanns

TUG, August 2021

Outline

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?

Conversion to XML

XML

- not something you want to directly edit
- document can be validated
  $\Rightarrow$ possible archive format
- JATS *Journal Article Text Suite*
- MathML, XHTML
  $\Rightarrow$ 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 `LaTeXML` to create XML
- mass conversion to XHTML

  $\approx$ 500.000 documents converted
- create real-world MathML

  $\Rightarrow$ improve LaTeXML[1]

[1] B. Miller and D. Ginev, https://dlmf.nist.gov/LaTeXML/

Build system

- open source (MIT licence)
- implemented in scripting language (here `php`)
- uses `SQL` database to store state
- distributes jobs on several hosts
- sets timeout for each job
- analyzes conversion process
  - checks files
  - parses the result files (`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

  $\Rightarrow$ use Docker images
- more interactivity

- other targets than XHTML

- create same target using different systems

texmlbus: change of focus

- easy installation
  - $\Rightarrow$ use Docker images
- more interactivity
  - $\Rightarrow$ upload files via browser
  - $\Rightarrow$ import files directly from Overleaf
- other targets than XHTML
- create same target using different systems

texmlbus: change of focus

- easy installation
  - $\Rightarrow$ use Docker images
- more interactivity
  - $\Rightarrow$ upload files via browser
  - $\Rightarrow$ import files directly from Overleaf
  - $\Rightarrow$ schedule jobs via browser
- other targets than XHTML

- create same target using different systems

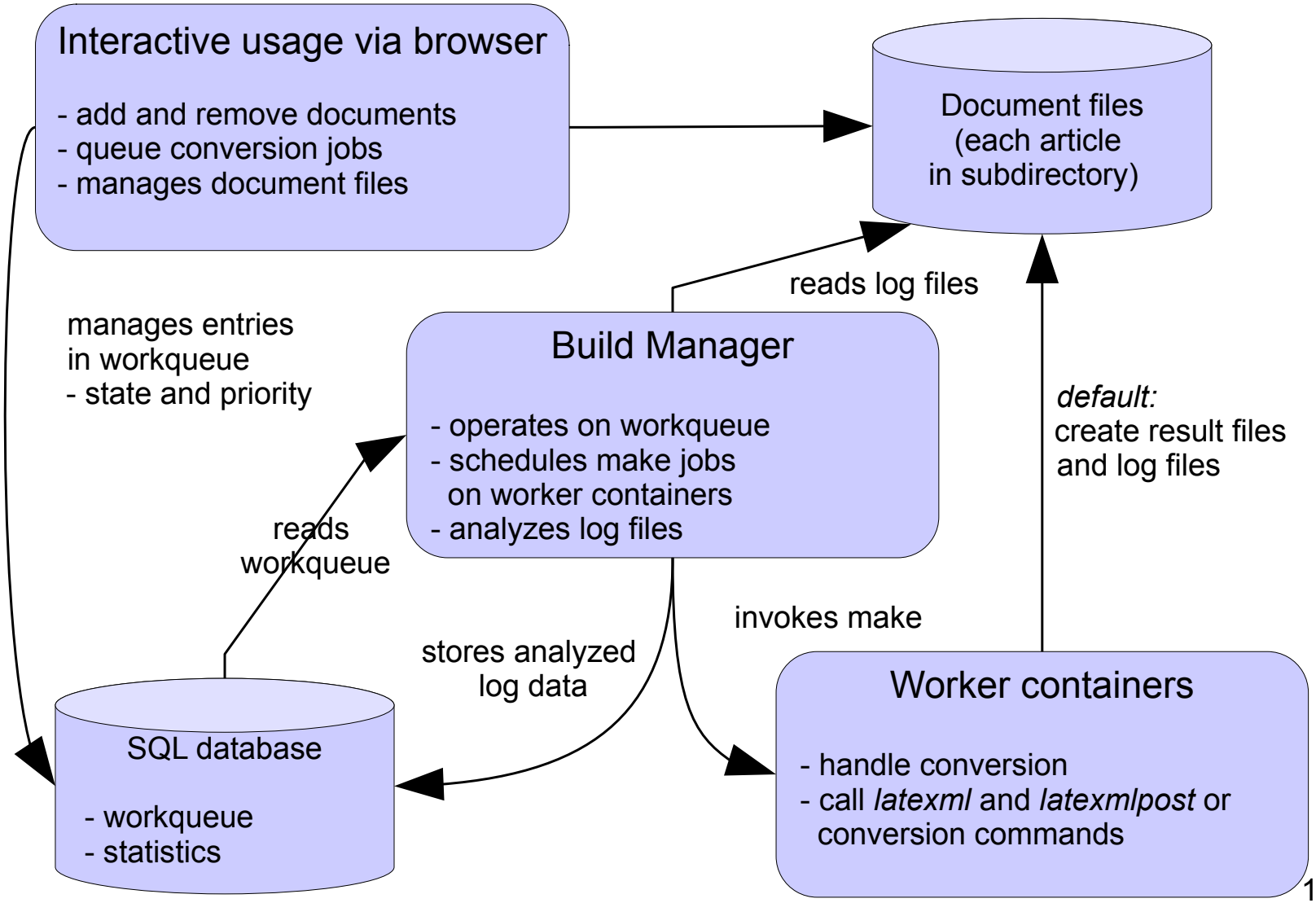
texmlbus: change of focus

- easy installation
  - $\Rightarrow$ use Docker images
- more interactivity
  - $\Rightarrow$ upload files via browser
  - $\Rightarrow$ import files directly from Overleaf
  - $\Rightarrow$ schedule jobs via browser
- other targets than XHTML
  - $\Rightarrow$ result table for each target
- create same target using different systems

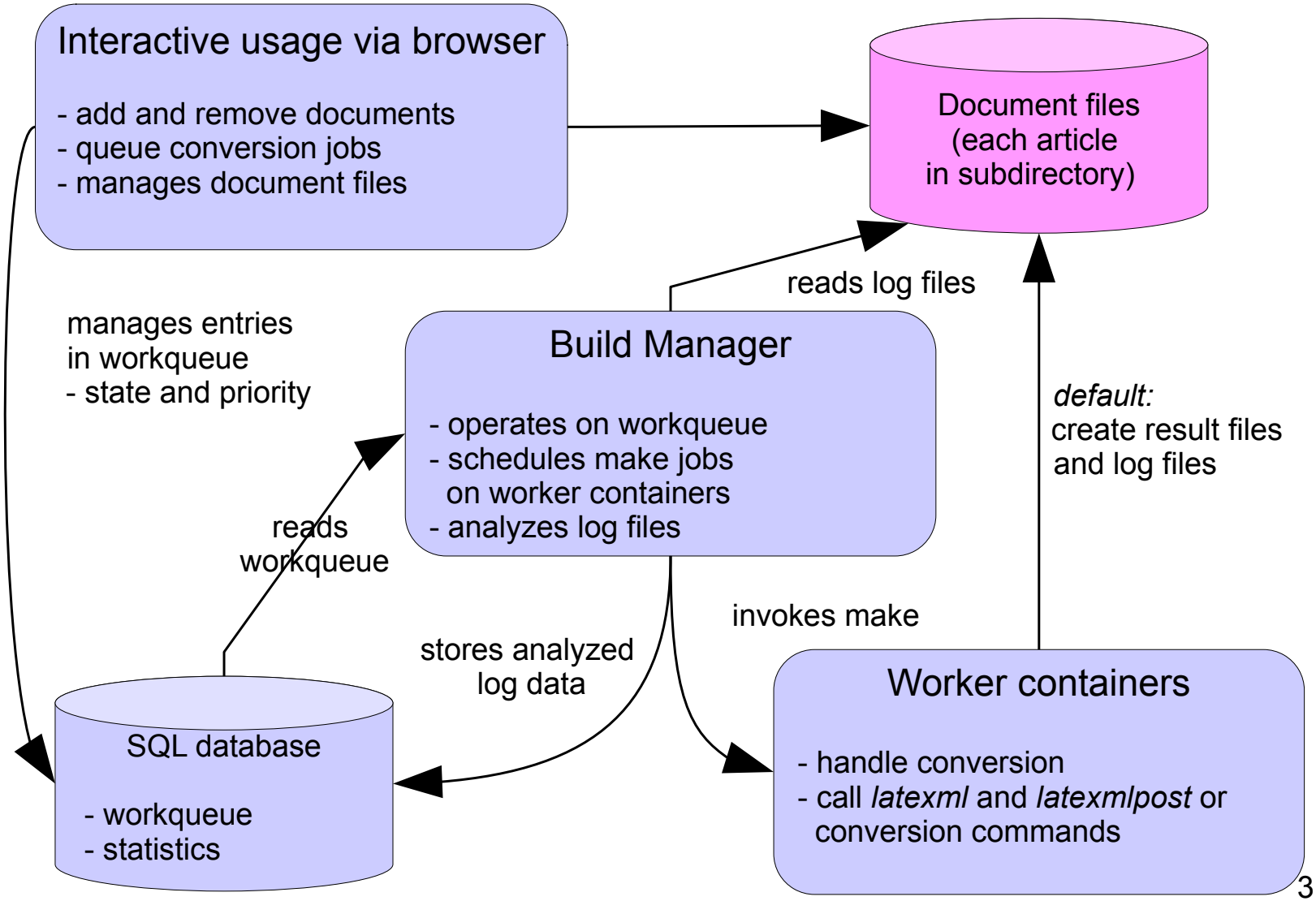
texmlbus: change of focus

- 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)

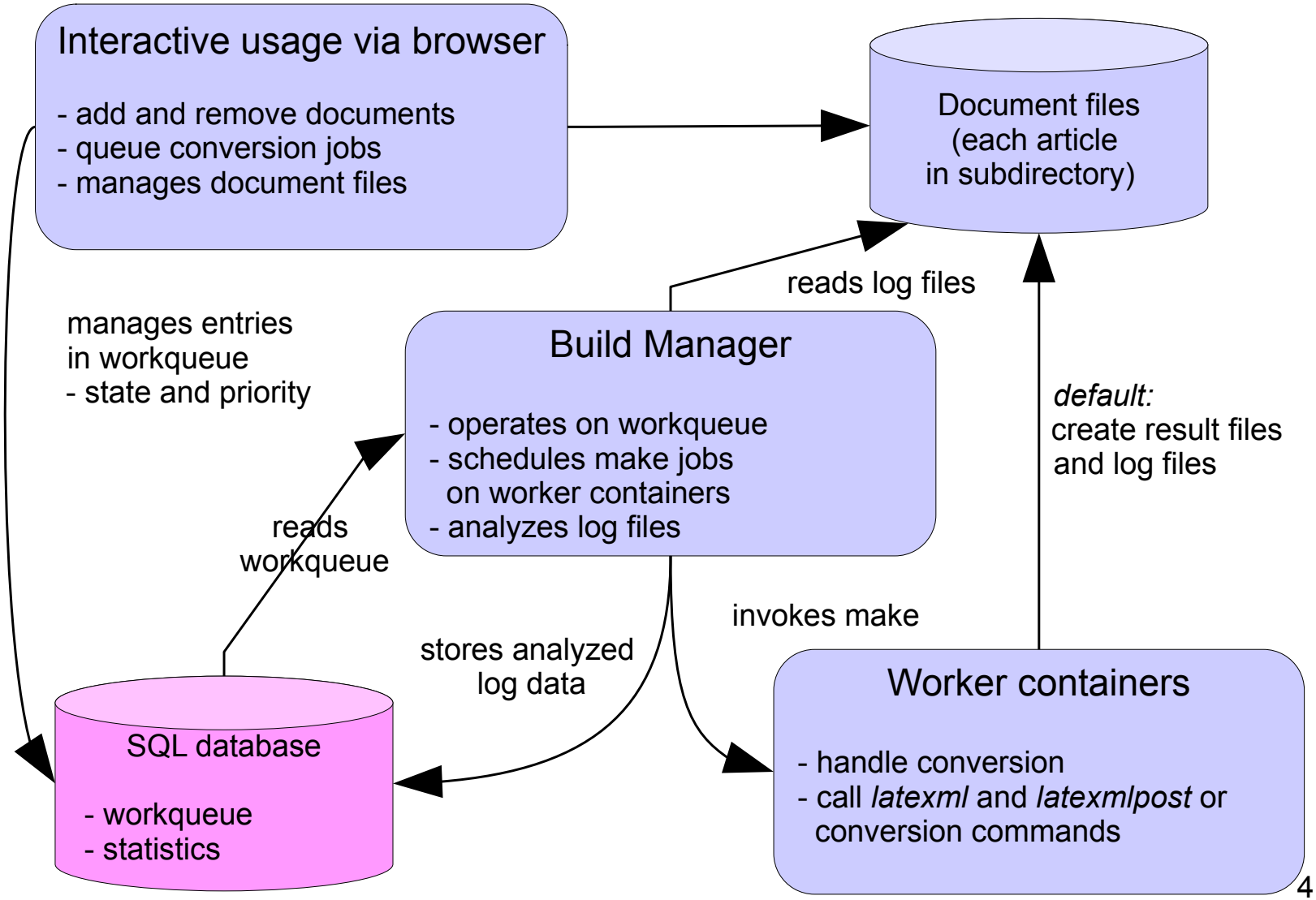
texmlbus build system



texmlbus build system

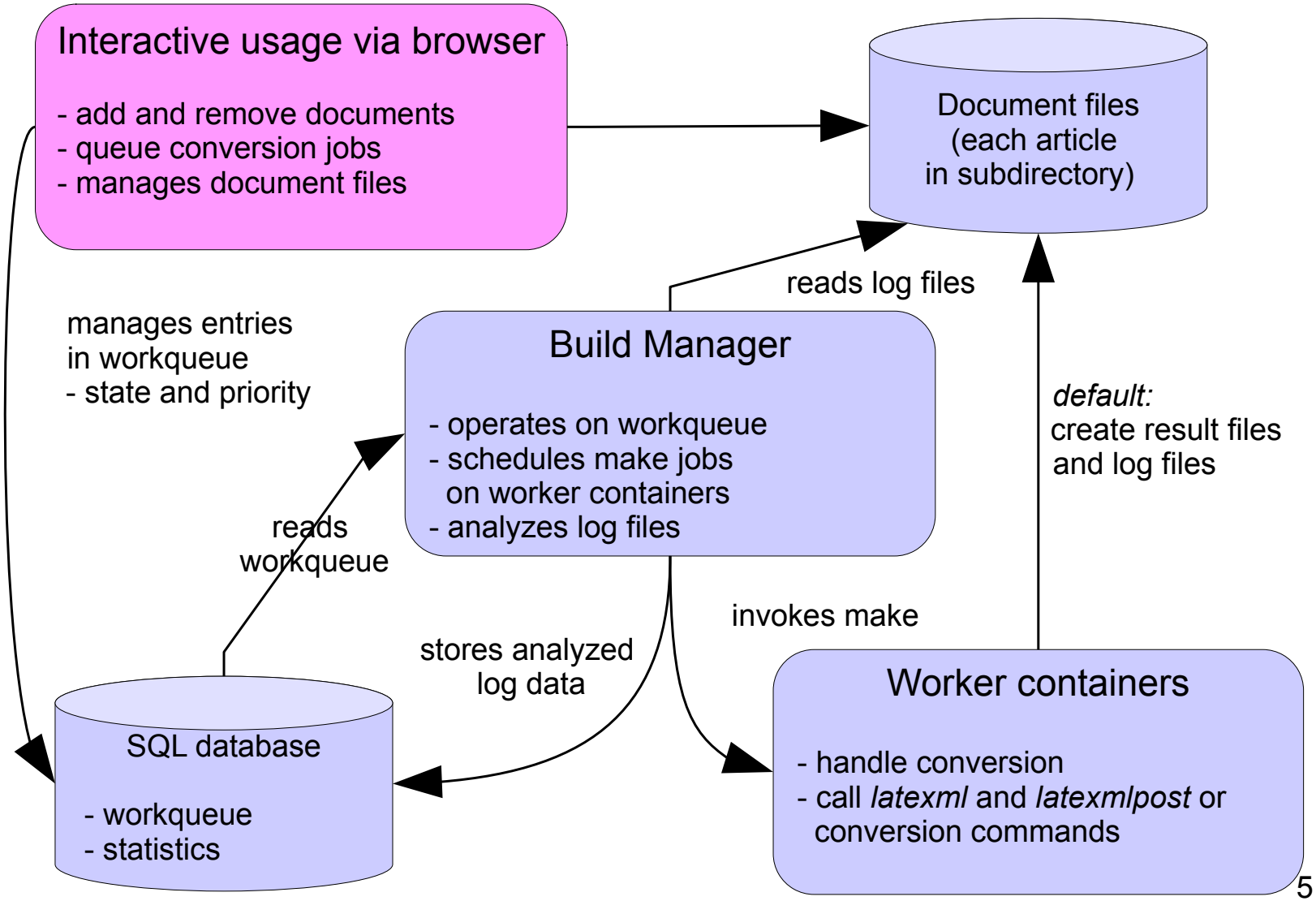


texmlbus build system



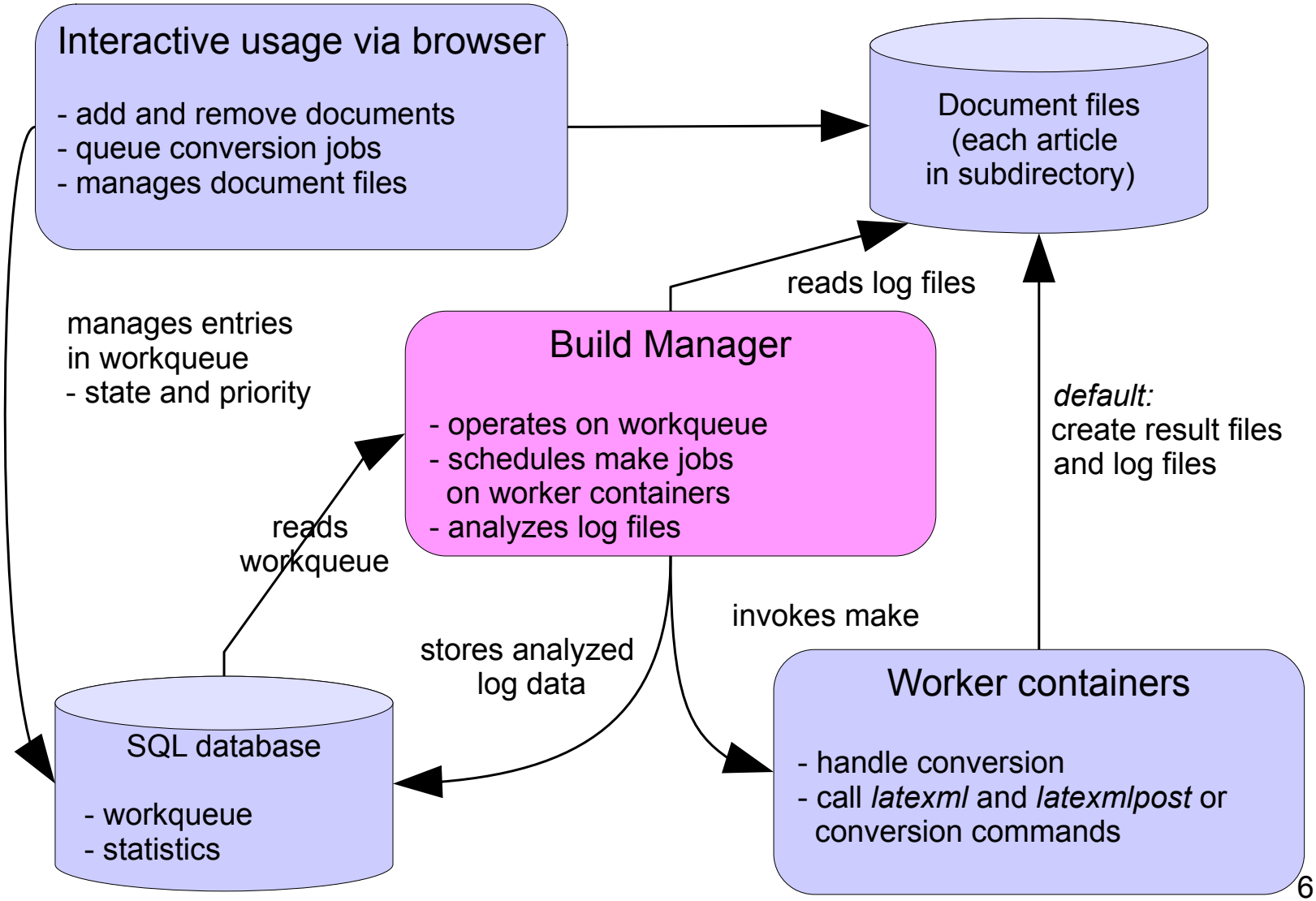
4

texmlbus build system

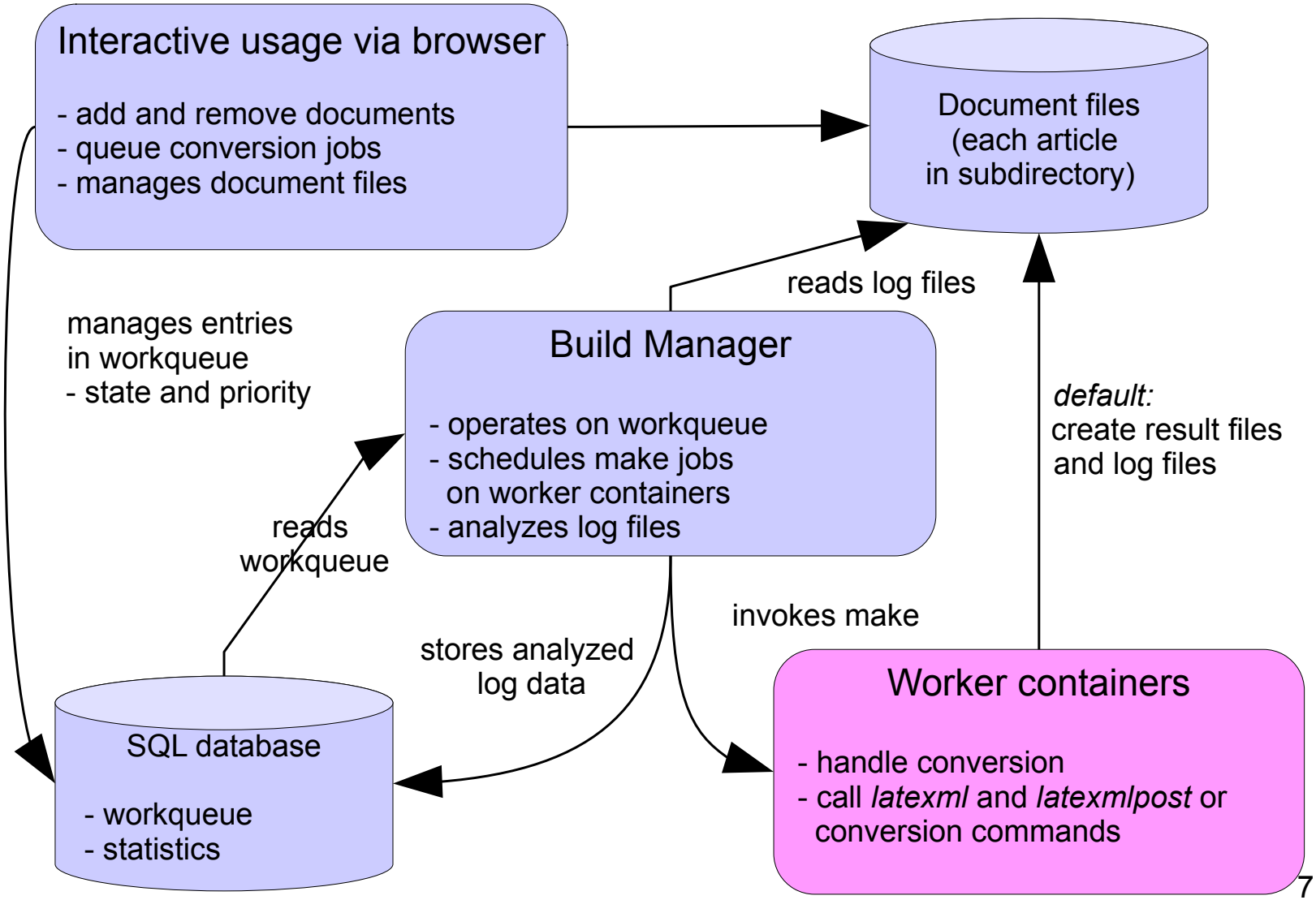


5

texmlbus build system

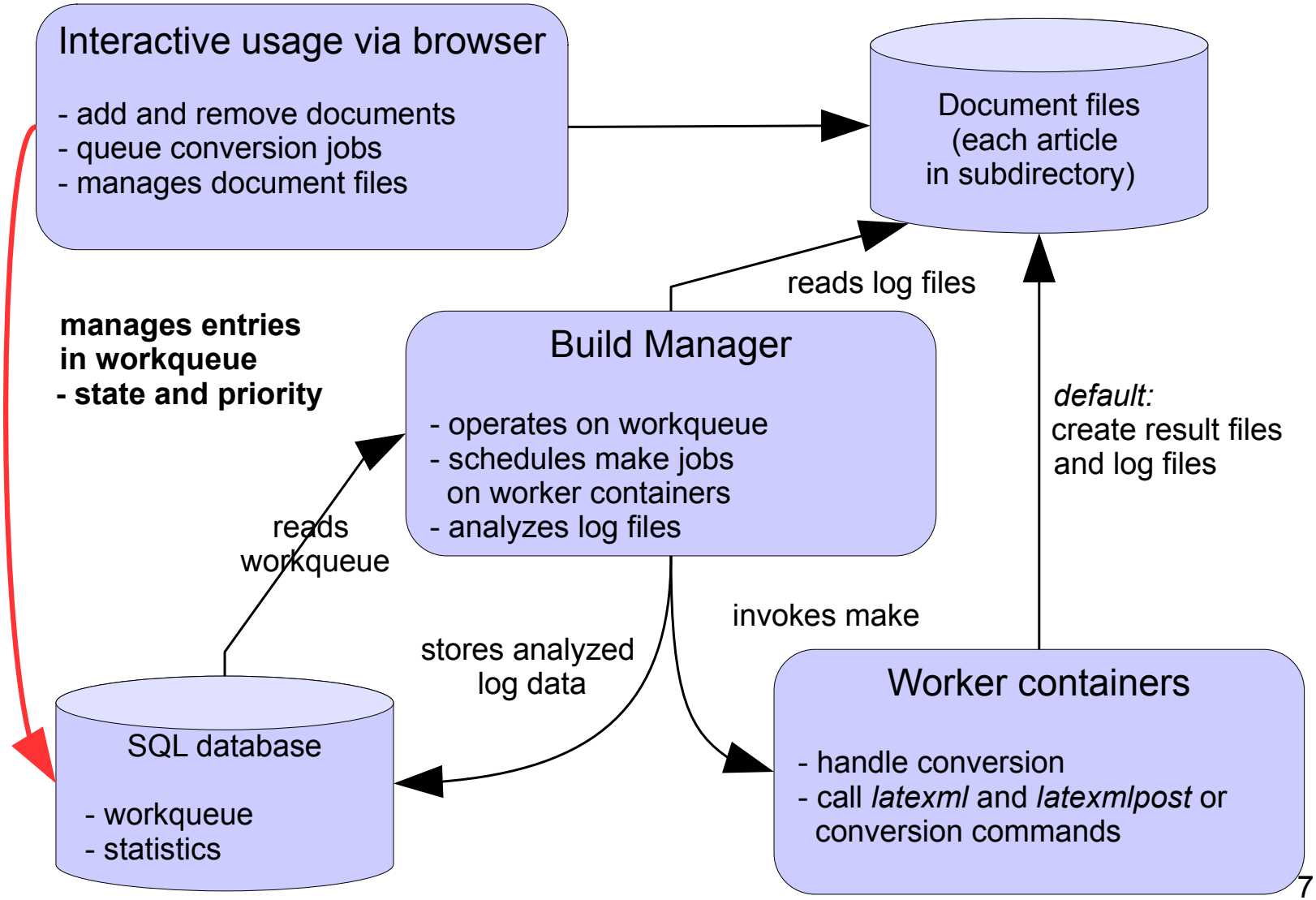


texmlbus build system

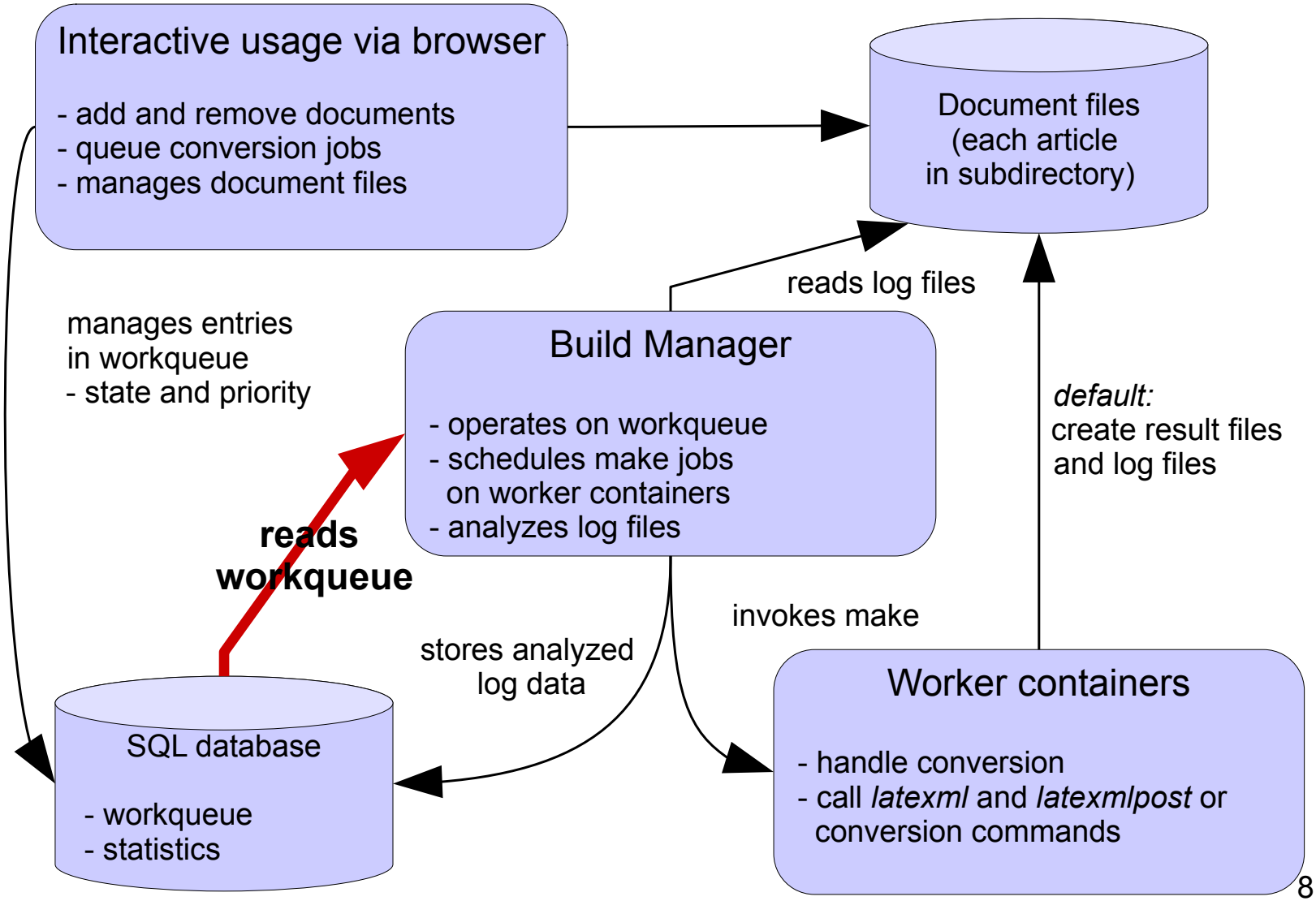


7

texmlbus build system

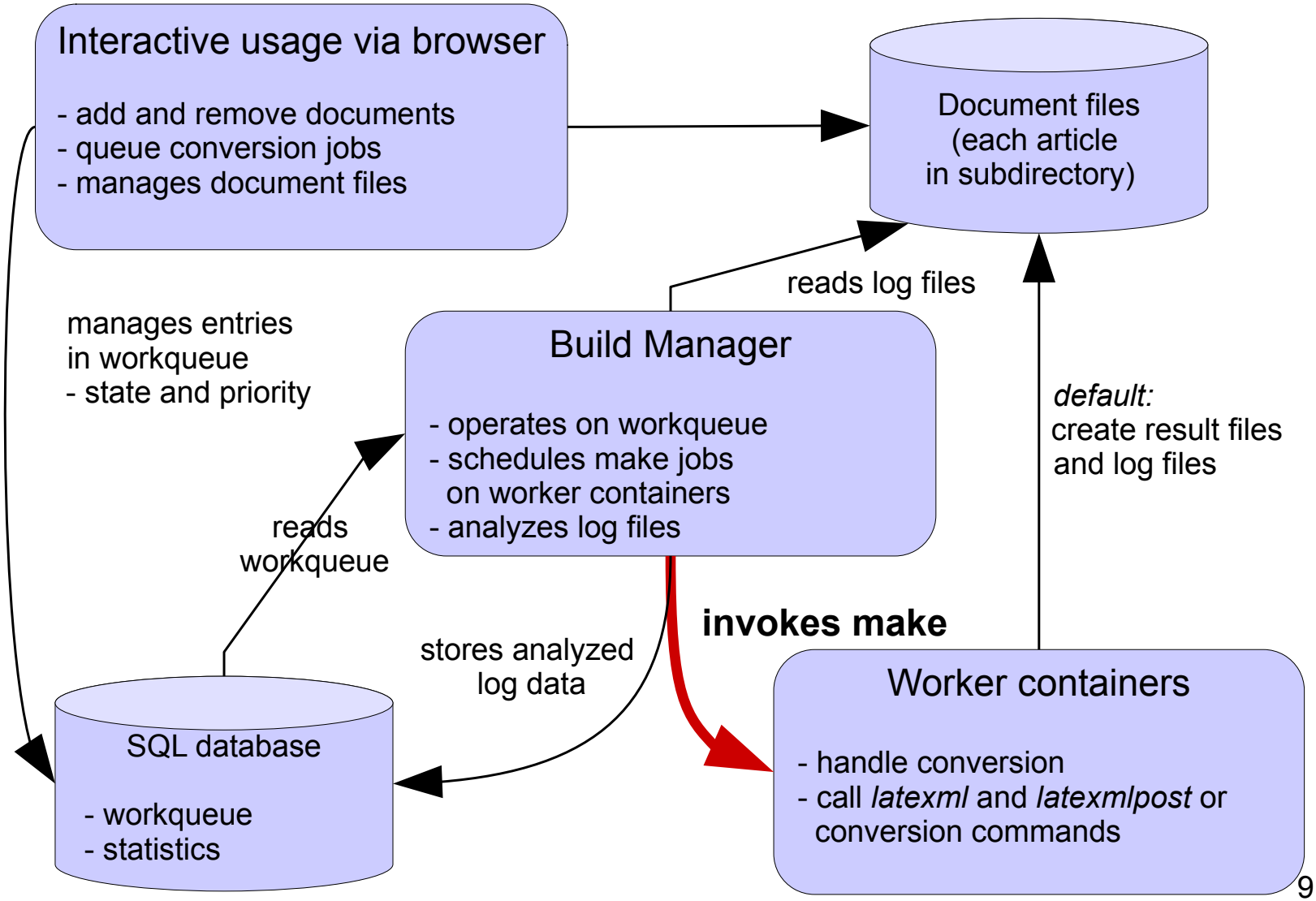


texmlbus build system



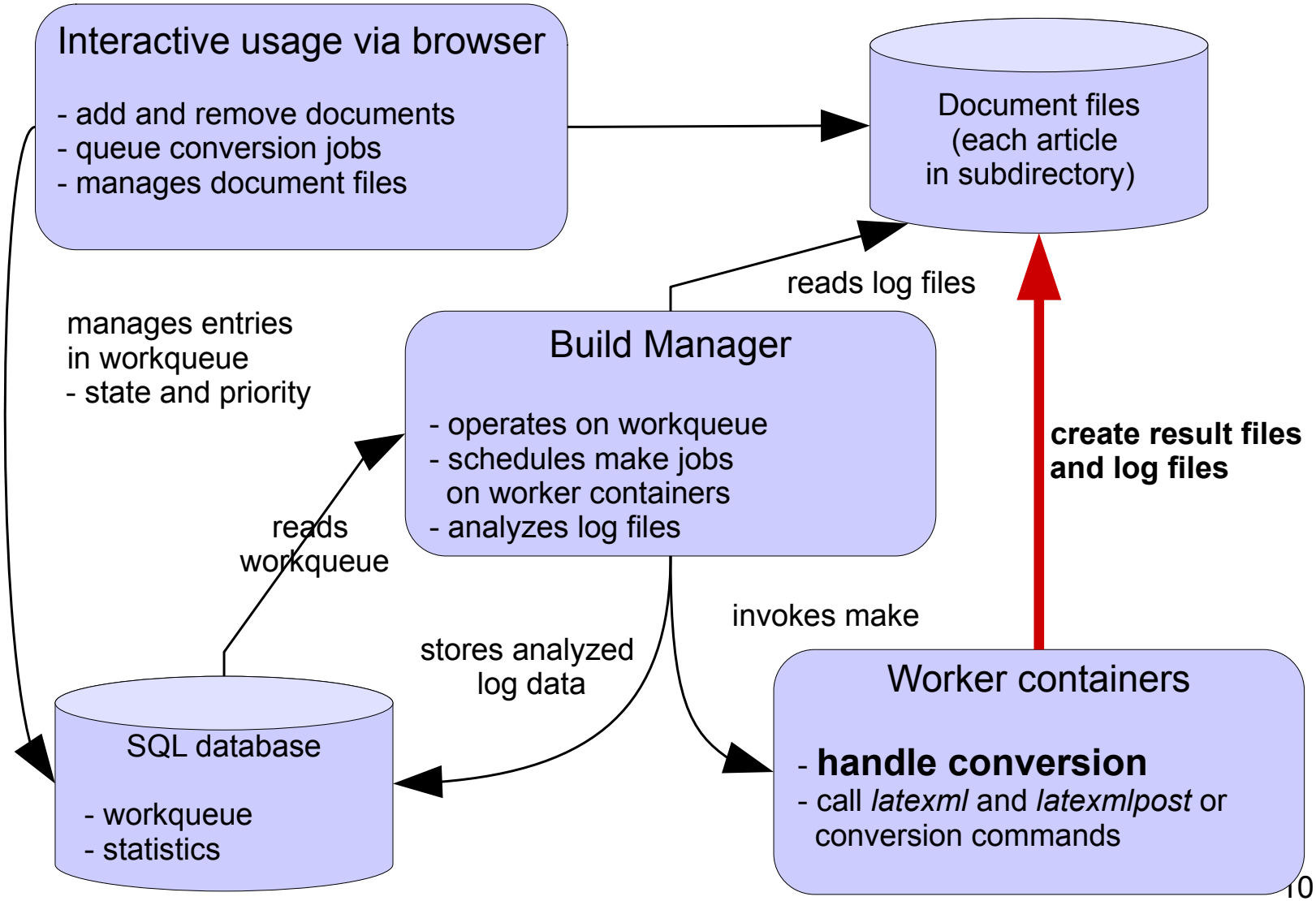
8

texmlbus build system

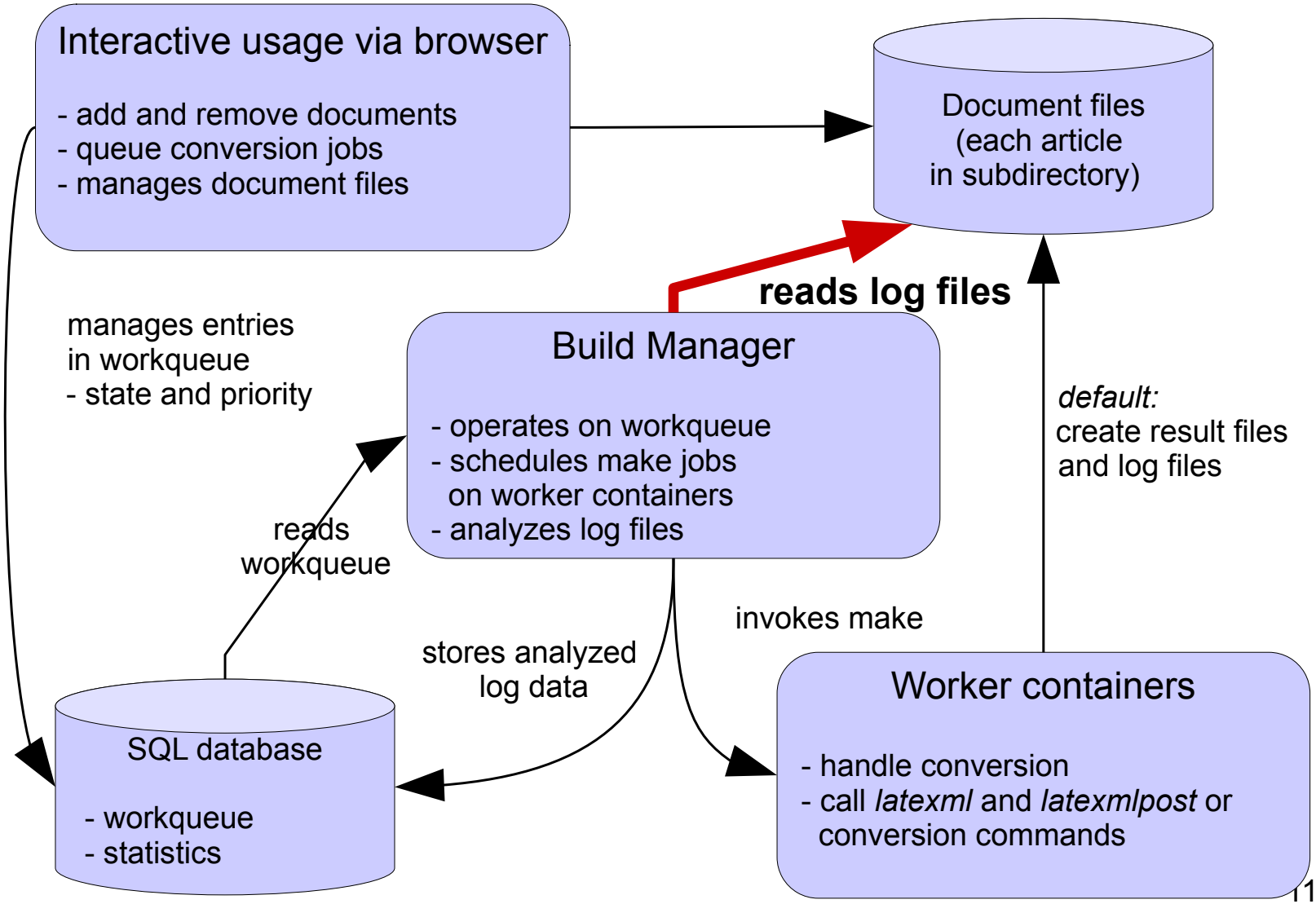


9

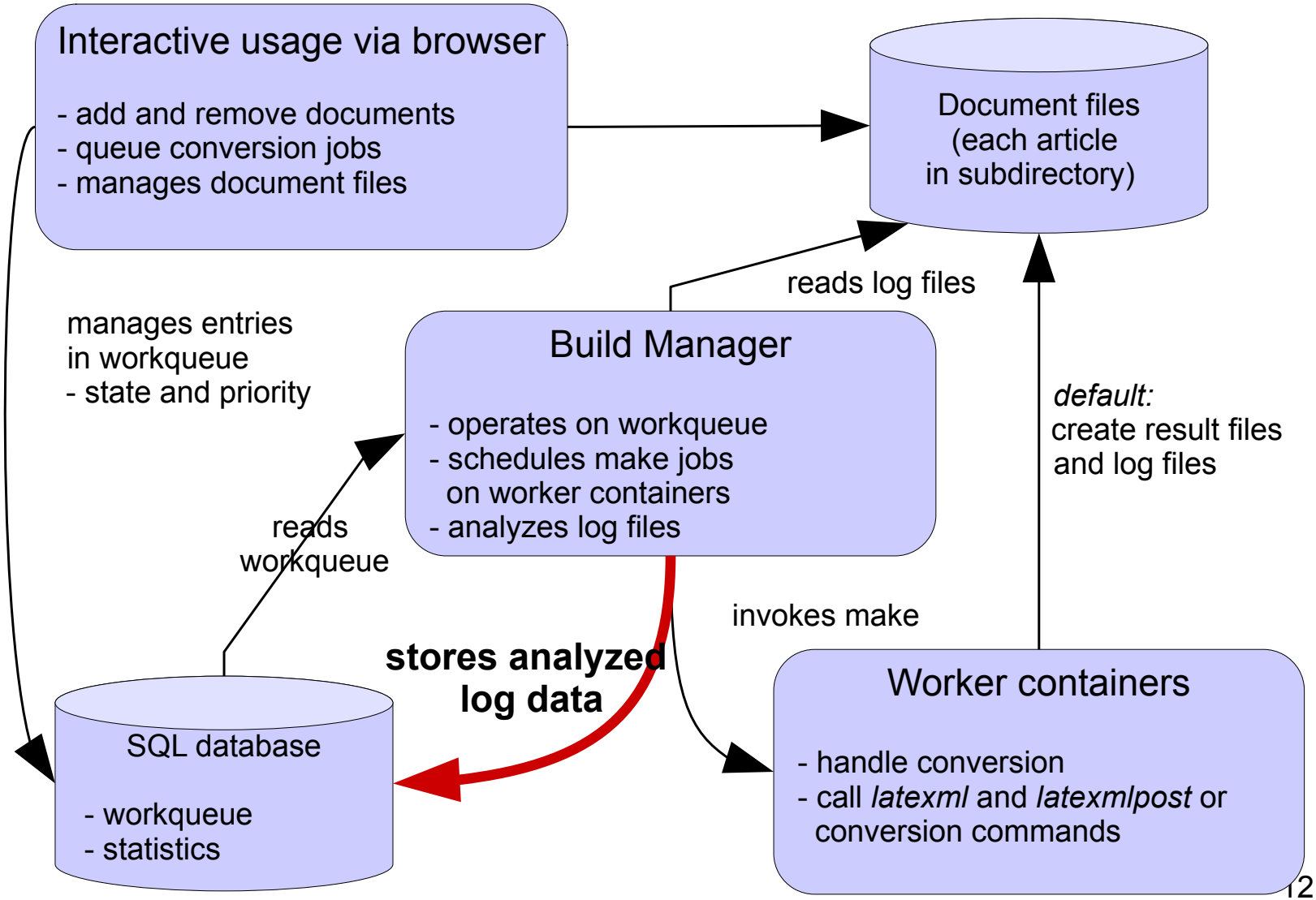
texmlbus build system



texmlbus build system

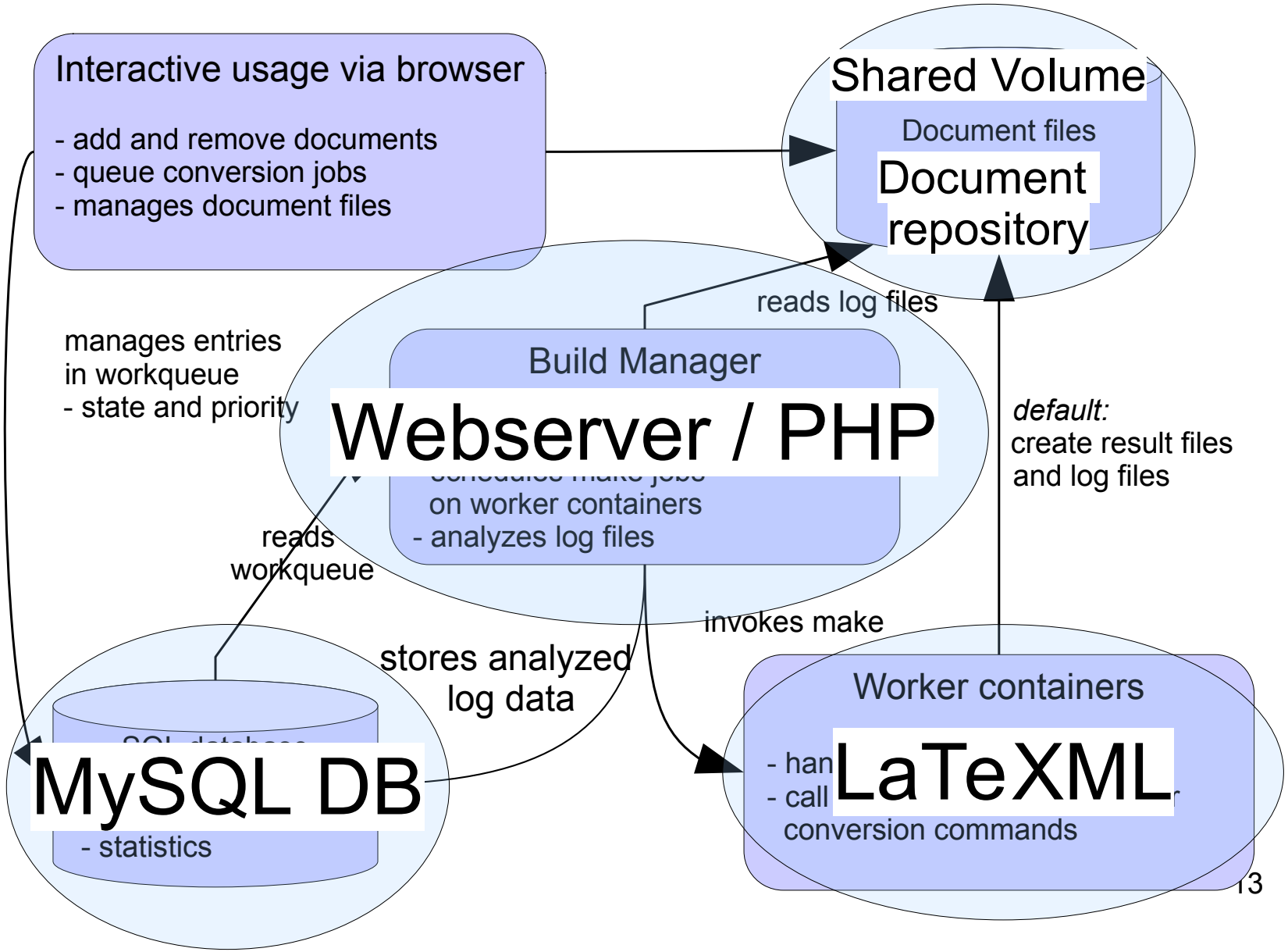


texmlbus build system

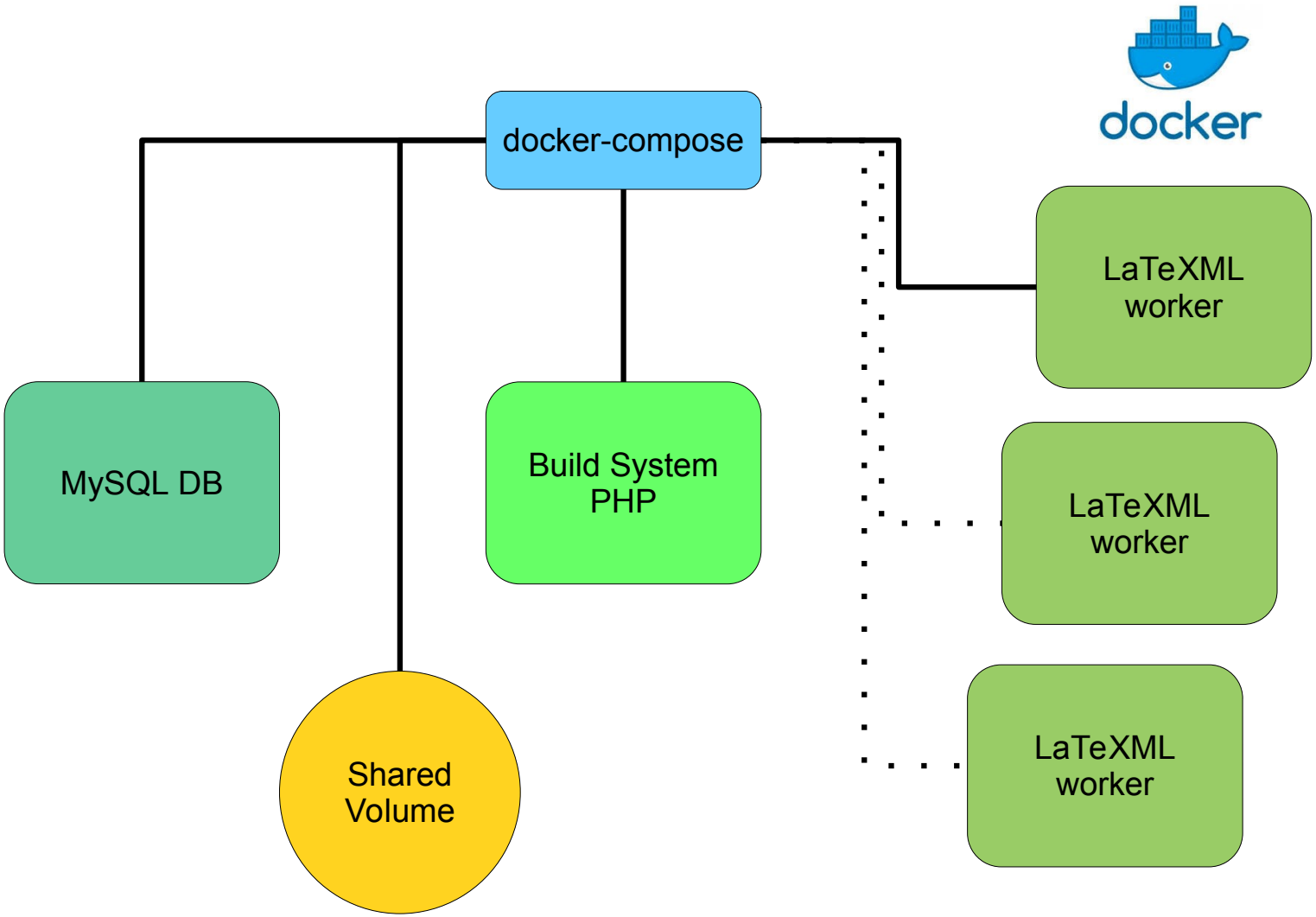


12

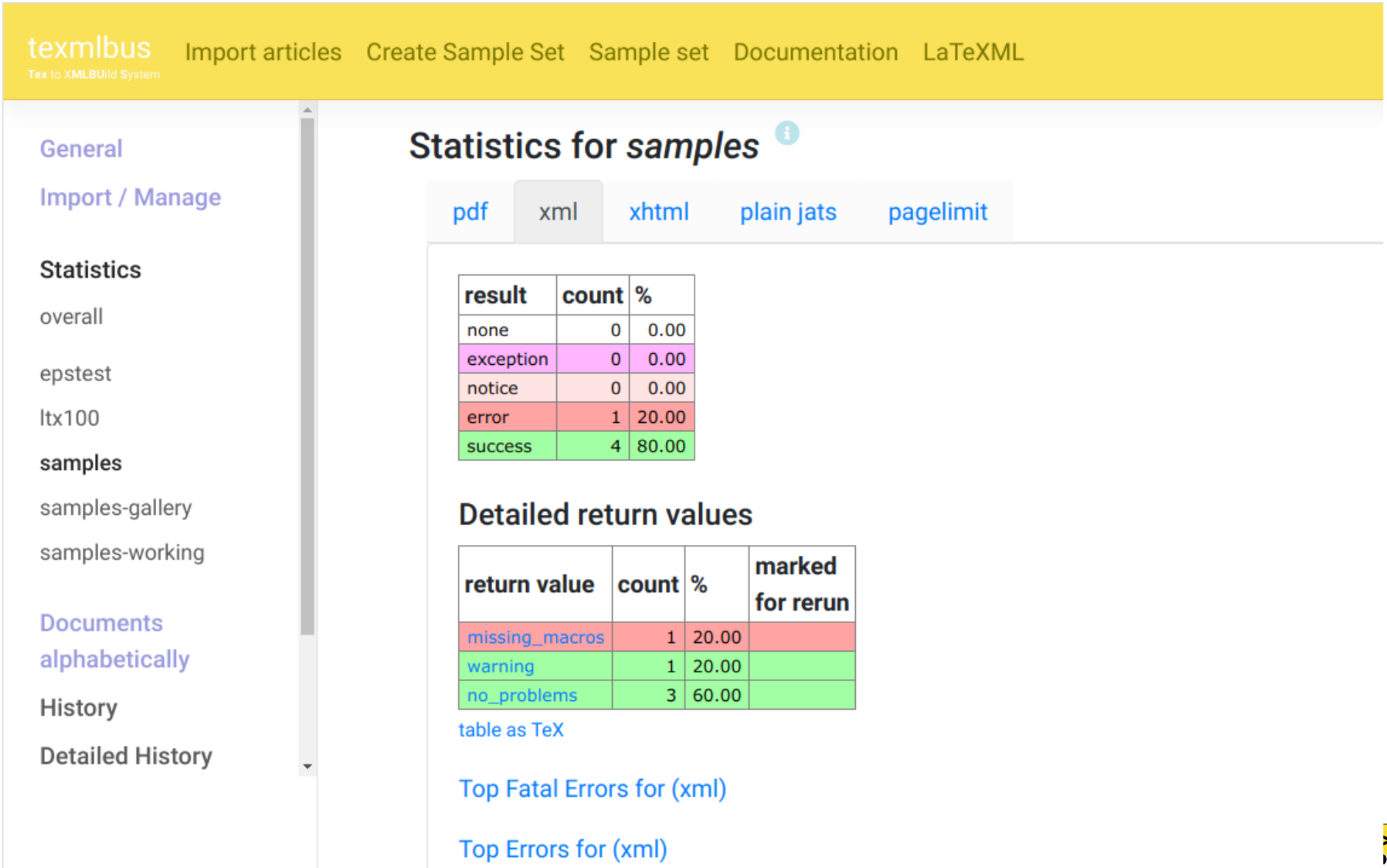
texmlbus build system



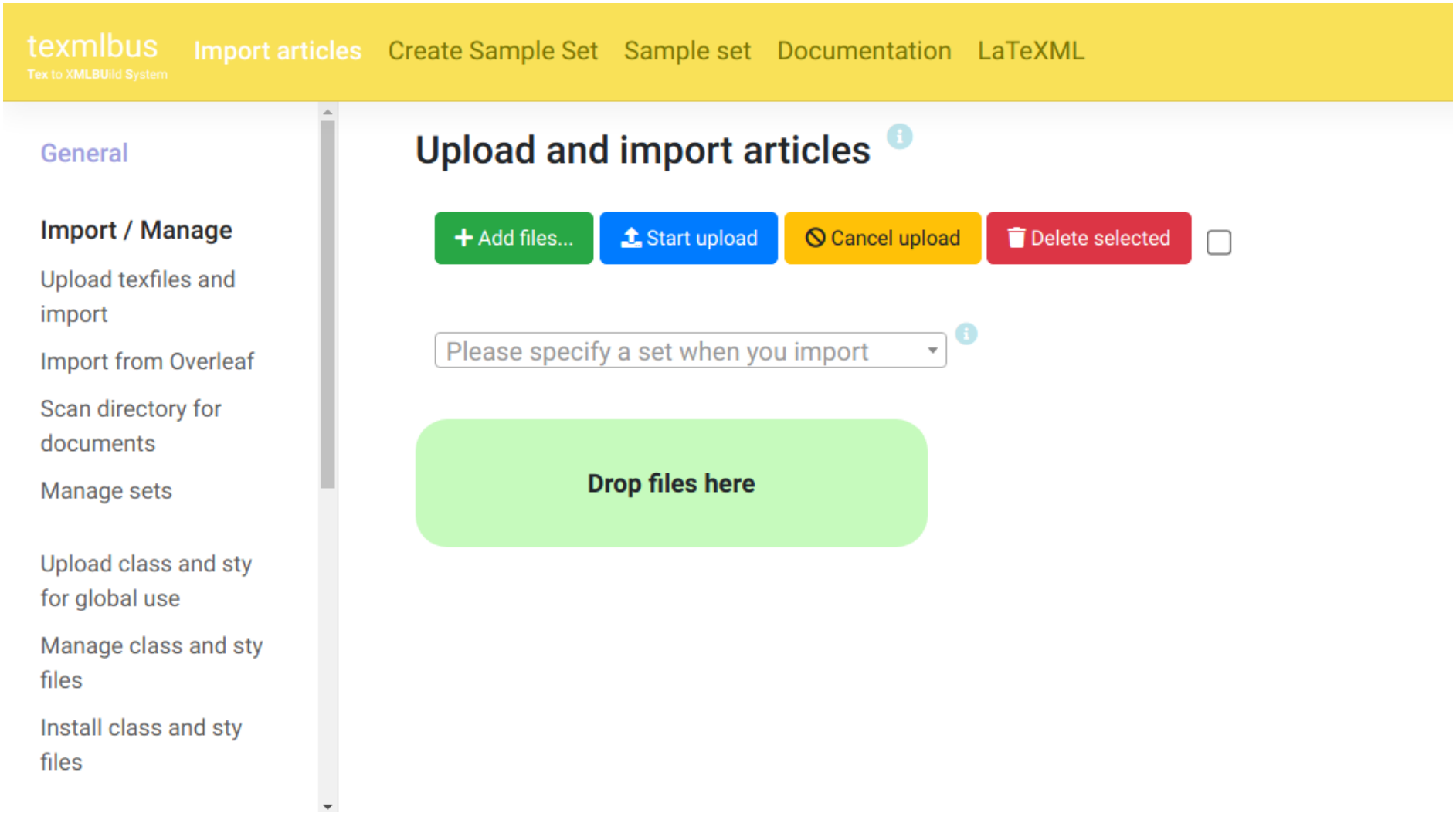
texmlbus build system



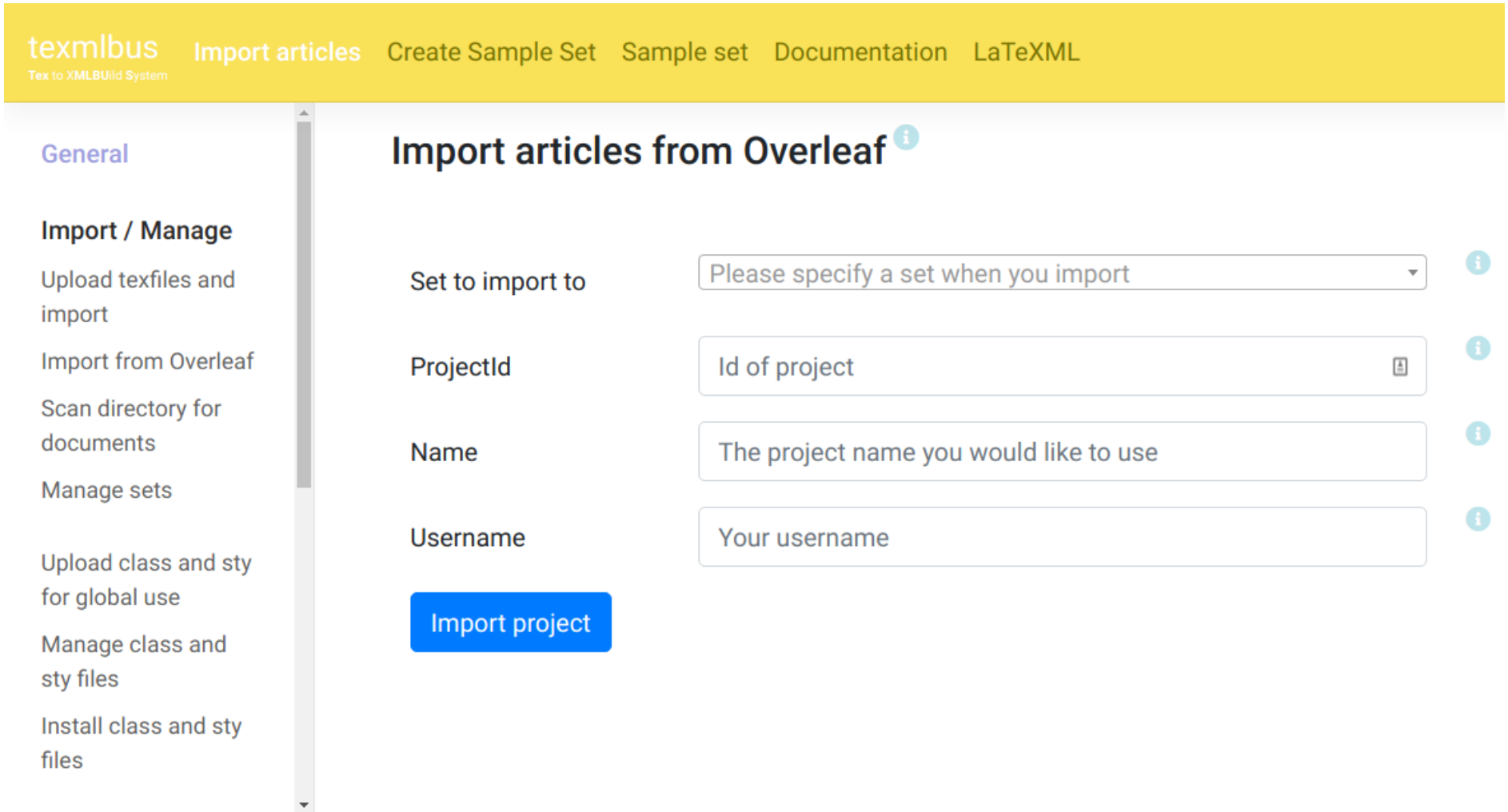
texmlbus build system



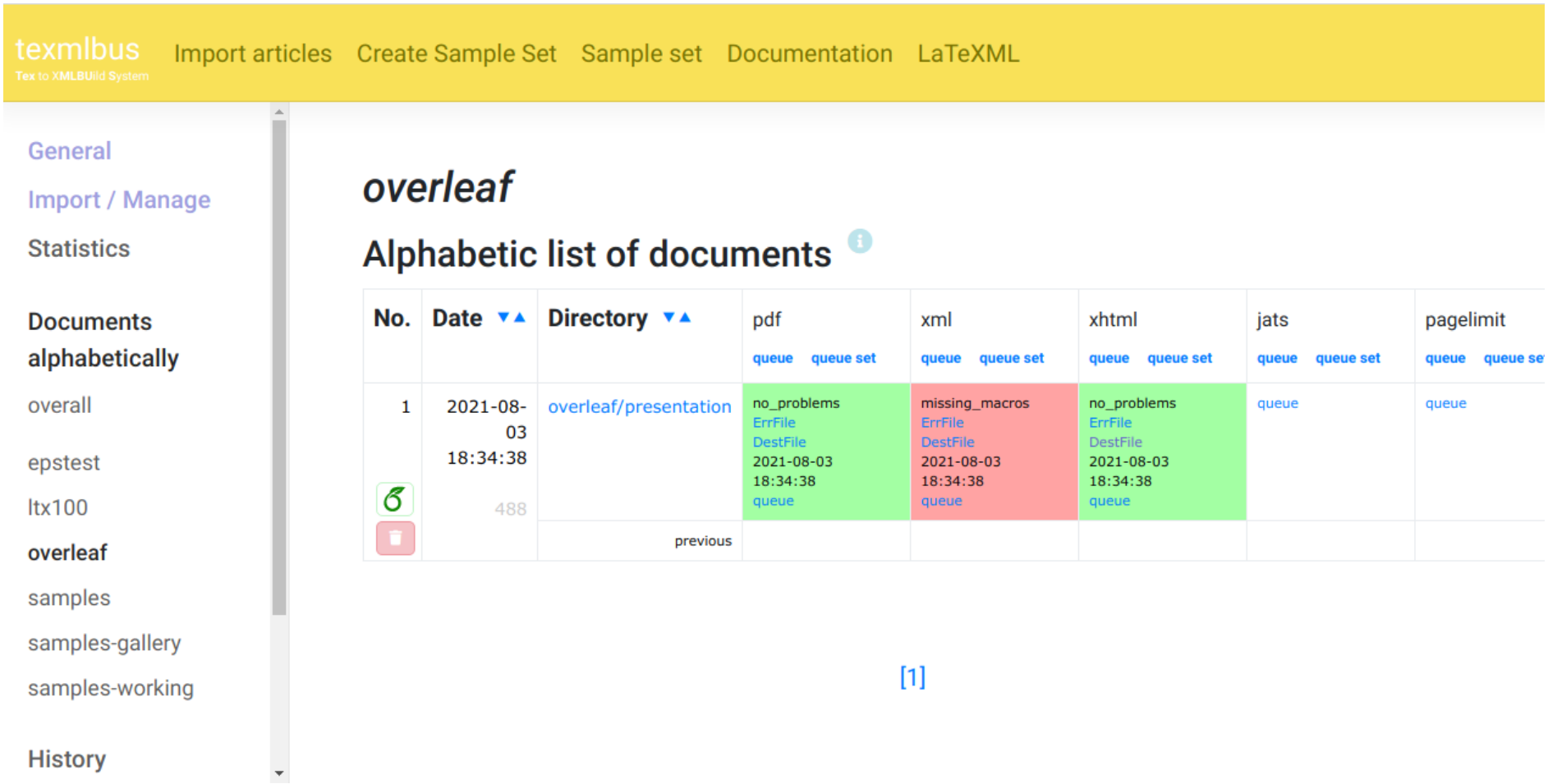
texmlbus build system



texmlbus build system

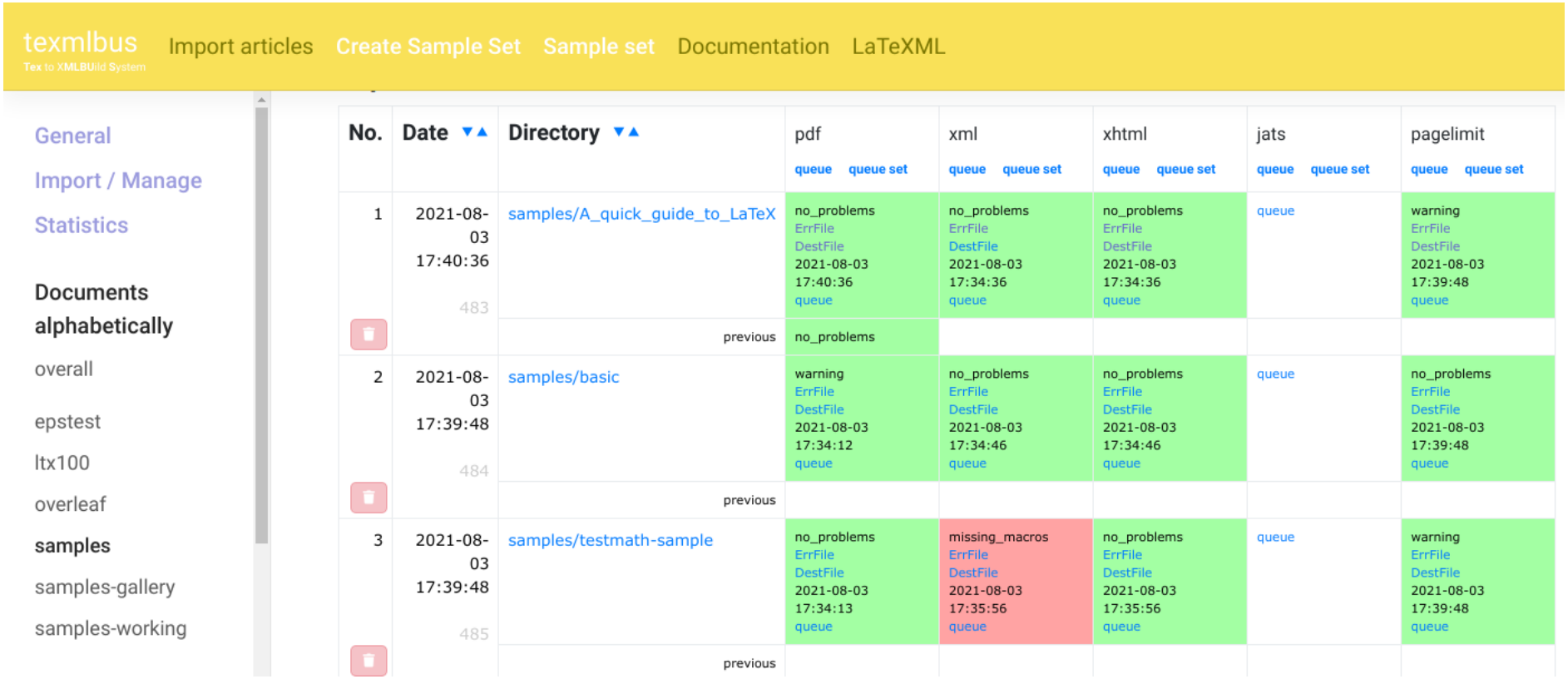


texmlbus build system



[1]

texmlbus build system

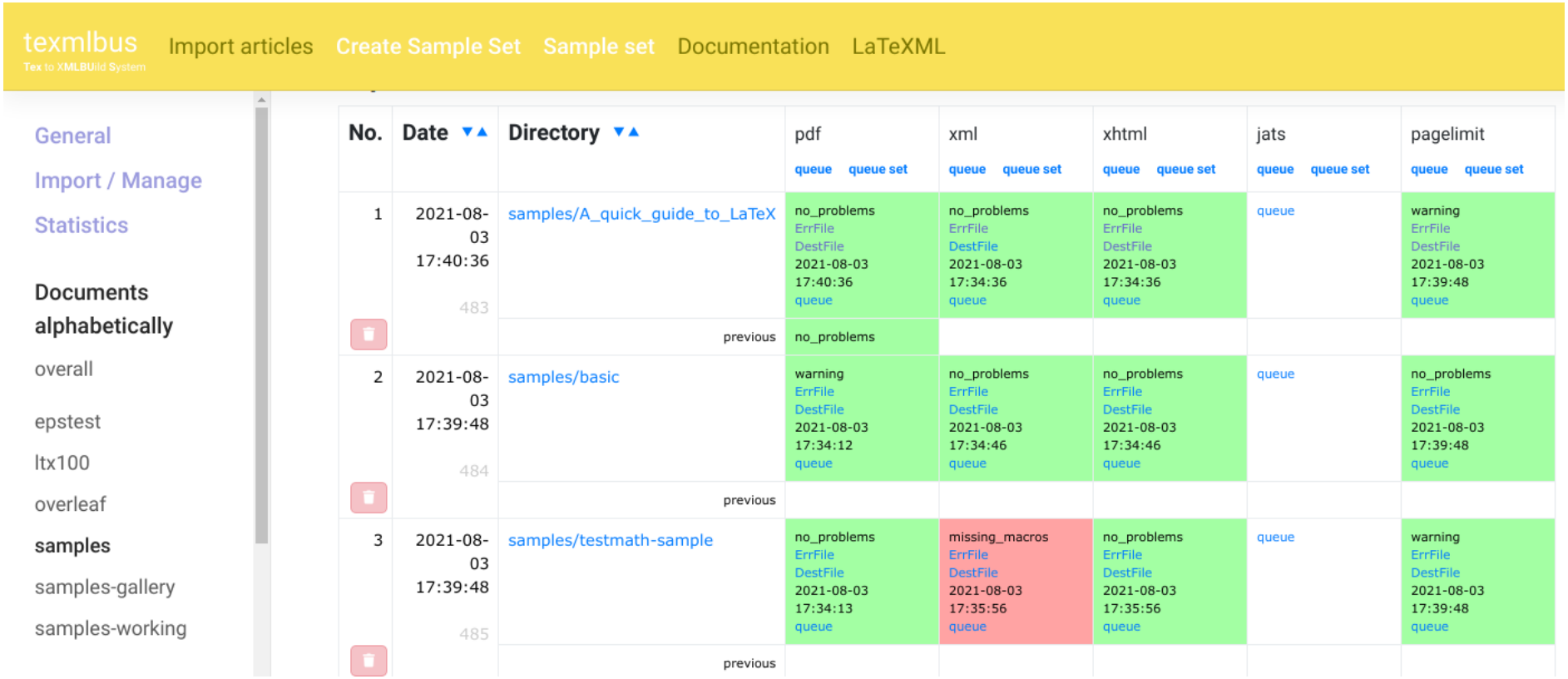


texmlbus build system

```
This is pdfTeX, Version 3.14159265-2.6-1.40.20 (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>
(/usr/share/texmf-dist/tex/latex/base/article.cls
Document Class: article 2019/12/20 v1.4l Standard LaTeX document class
(/usr/share/texmf-dist/tex/latex/base/size10.clo
File: size10.clo 2019/12/20 v1.4l Standard LaTeX file (size option)
)
\c@part=\count167
\c@section=\count168
\c@subsection=\count169
\c@subsubsection=\count170
\c@paragraph=\count171
\c@subparagraph=\count172
\c@figure=\count173
\c@table=\count174
\abovecaptionskip=\skip47
\belowcaptionskip=\skip48
\bibindent=\dimen134
)
(/usr/share/texmf-dist/tex/latex/amsfonts/amssymb.sty
Package: amssymb 2013/01/14 v3.01 AMS font symbols

(/usr/share/texmf-dist/tex/latex/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.
))
(/usr/share/texmf-dist/tex/latex/amsmath/amsmath.sty
Package: amsmath 2020/01/20 v2.17e AMS math features
```

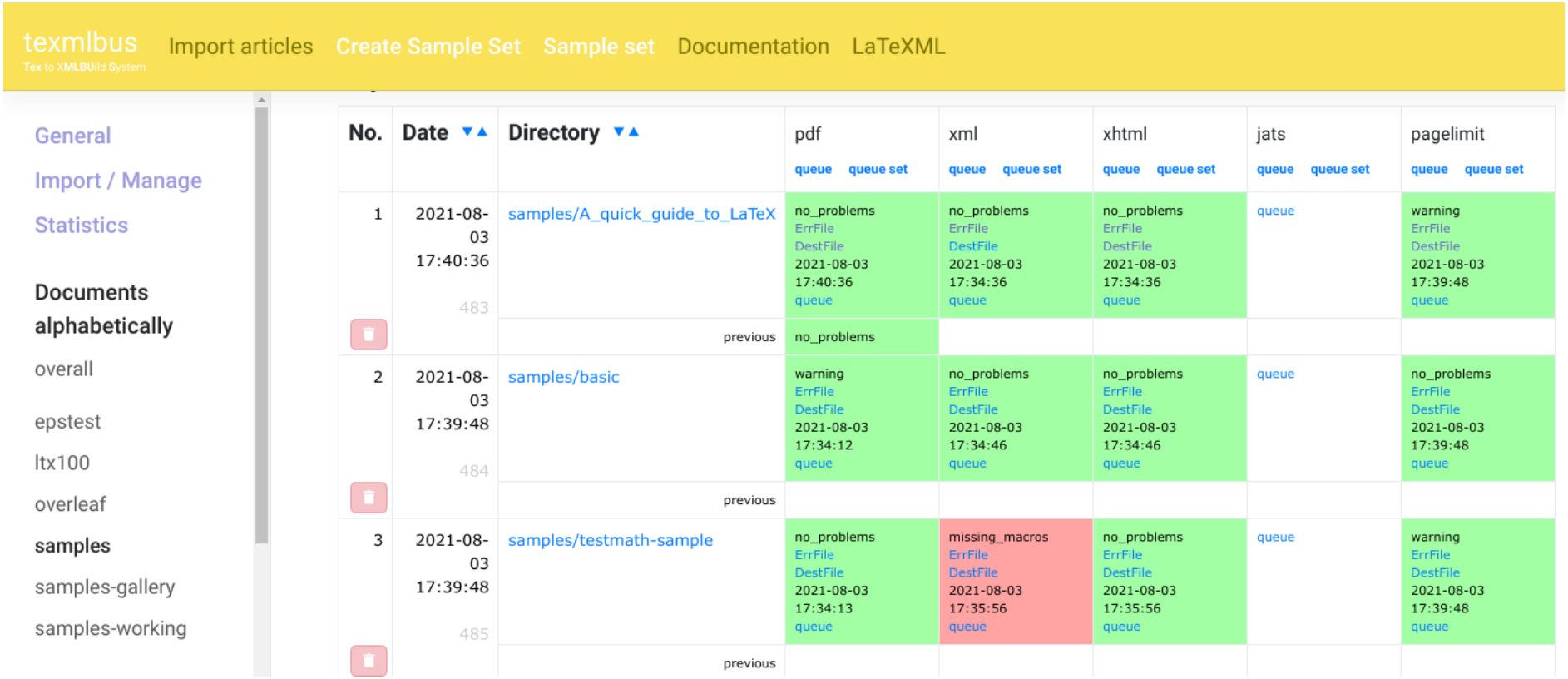
texmlbus build system



This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
<?latexml searchpaths="/srv/texmlbus/articles/samples/A_quick_guide_to_LaTeX/__texmlbus_worker,/srv/texmlbus/articles/samples/A_quick_guide_to_LaTeX/_
<!--   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%   -->
<!--   %writeLaTeX Example: A quick guide to LaTeX  -->
<!--   %Source: Dave Richeson (divisbyzero.com), Dickinson College  -->
<!--   %A one-size-fits-all LaTeX cheat sheet. Kept to two pages, so it  -->
<!--   %can be printed (double-sided) on one piece of paper  -->
<!--   %Feel free to distribute this example, but please keep the referral  -->
<!--   %to divisbyzero.com  -->
<!--   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%   -->
<!--   %How to use writeLaTeX:  -->
<!--   %You edit the source code here on the left, and the preview on the  -->
<!--   %right shows you the result within a few seconds.  -->
<!--   %Bookmark this page and share the URL with your co-authors. They can  -->
<!--   %edit at the same time!  -->
<!--   %You can upload figures, bibliographies, custom classes and  -->
<!--   %styles using the files menu.  -->
<!--   %If you're new to LaTeX, the wikibook is a great place to start:  -->
<!--   %**** main.tex Line 25 ****  -->
<!--   %http://en.wikibooks.org/wiki/LaTeX  -->
<!--   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%   -->
<?latexml class="article" options="10pt,landscape"?>
<?latexml package="amssymb,amsmath,amsthm,amsfonts"?>
<?latexml package="multicol,multirow"?>
<?latexml package="calc"?>
<?latexml package="ifthen"?>
<?latexml package="geometry" options="landscape"?>
<?latexml package="hyperref" options="colorlinks=true,citecolor=blue,linkcolor=blue"?>
<!--   %x  -->
<!--   %**** main.tex Line 50 ****  -->
<!--   %__  -->
<?latexml RelaxNGSchema="LaTeXML"?>
▼<document xmlns="http://dlmf.nist.gov/LaTeXML">
   <resource class="ltx_align_left" src="LaTeXML.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>
 ▼<para align="center" class="ltx_align_left" xml:id="p1">
```

texmlbus build system



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 mathe 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 use of LaTeX programmers created the current version, LaTeX 2ε.

Math vs. text vs. functions

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

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 math expression between dollar signs (`$`). For example, typing `$90^{\circ}$ is` yields 90° 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 showcase equations that cannot fit inline. To produce an inline expression, place the mathematical expression between the symbols `\[` and `\]`. Typing `\[x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}\]`

The web interface

Result statistics

result	count	%
none	0	0.00
exception	1	2.33
notice	0	0.00
error	8	18.60
success	34	79.07

Detailed return values

return value	count	%	marked for rerun
fatal_error	1	2.33	
error	8	18.60	
warning	24	55.81	
no_problems	10	23.26	

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

Outlook

Things to be done

- add converters more easily
- help to improve `LaTeXML`

https://github.com/stamer/texmlbus

Thanks to Overleaf for their support!