

# Automated tagging of $\LaTeX$ documents — what is possible today?

Ulrike Fischer, Bonn  
 $\LaTeX$  Project Team

15.7.2023  
TUG 2023

# What is tagging

- Tagging adds structure to a PDF
- Tagging improves accessibility and reuse of data
- More info
  - Documentation tagpdf package
  - Various tugboat articles

# Motivation: A tagged bank statement

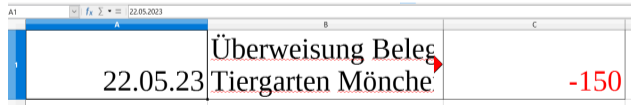
The screenshot shows a window titled "Tags für die Barrierefreiheit" (Tags for accessibility). The interface displays a tree view of HTML tags used in a document. The root tag is `<TR>`, which is expanded to show its children: another `<TR>` tag, a `<TD>` tag, and a `<P>` tag. The `<TD>` tag is further expanded to show its content: the date "22.05.2023", another `<TD>` tag, and a `<P>` tag. The `<TD>` tag is expanded to show its content: "Überweisung Beleglos". The `<P>` tag is expanded to show its content: "Tergarten Mönchengladbach Ver". The `<TD>` tag is expanded to show its content: "-150,00". The `<TR>` tag is expanded to show its children: another `<TR>` tag, a `<TD>` tag, and a `<P>` tag.

a tagged bank statement

- basic, simple tagging
- a table with three columns: date, description, amount

# Motivation: A tagged bank statement – benefits

- reading
  - untagged: order is messy
  - tagged: order is correct and knows the row and column numbers
- copy & paste into a spreadsheet
  - untagged: dumps everything into one cell
  - tagged: splits the data into three columns



The image shows a screenshot of a spreadsheet application. The active cell is A1, containing the date '22.05.23'. The adjacent cell B1 contains the text 'Überweisung Beleg' on the top line and 'Tiergarten Mönche' on the bottom line. The cell C1 contains the value '-150'. The spreadsheet interface includes a formula bar at the top showing '22.05.2023' and column headers 'A', 'B', and 'C' above the respective columns.

A	B	C
22.05.23	Überweisung Beleg Tiergarten Mönche	-150

# Goals of the Tagged PDF project

- Make tagging with  $\text{\LaTeX}$  *possible*  $\Rightarrow$  done!
- Make tagging with  $\text{\LaTeX}$  *automatic* and *easy* to use

## Problem: Tags are not visible

- free PDF viewer do not show tags  
(a few exceptions: PDF-XChange, PDFix Desktop Lite)
- what can't be seen is not requested and so not used
- testing is difficult for user

## Problem: Missing PDF 2.0 support

PDF 2.0 is important for good tagging

- better tag set (Title, Aside, FEnote ...)
  - structure destinations for links to structures
  - associated files to add data to structures
  - MathML name space
- } important for math tagging

$\text{\LaTeX}$  can create tagged PDF 2.0 but

**PDF viewers don't fully support PDF 2.0 features**

# Problem: Parent-Child-rules

- PDF declares various containment rules:

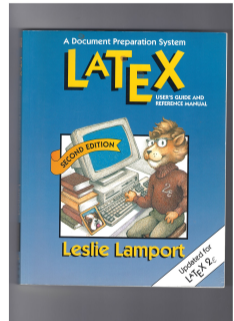
Structure Type	Children		Parents	
	Occ.	Structure Type	Occ.	Structure Type
P	0..n	NonStruct	0..n	Document
	0..n	Private	0..n	DocumentFragment
	0..n	Note	‡	Part
	0..n	Code	‡	Div
	0..n	Sub	0..n	Art
	0..n	Lbl	0..n	Sect
	0..n	Em	0..n	TOCI
	0..n	Strong	0..n	Aside

- the rules do not always fit the  $\text{\LaTeX}$  structures
- the rules must be checked and violations must be handled



# Next milestone: Tagging of “Leslie Lamport Documents”

- standard classes
- standard commands and environments
- restricted set of packages
- hyperref



# What is possible today?

- today = L<sup>A</sup>T<sub>E</sub>X 2023-06-01
- supported engines: pdfL<sup>A</sup>T<sub>E</sub>X or luaL<sup>A</sup>T<sub>E</sub>X
- new code in the latex-lab bundle
- Code is loaded with a testphase keys:

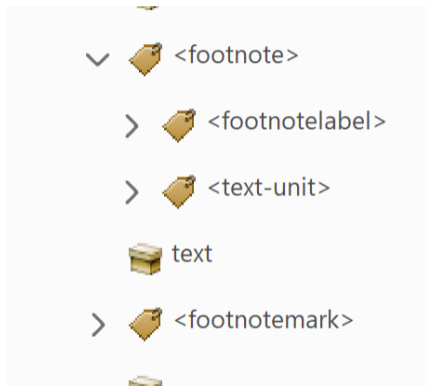
```
\DocumentMetadata{testphase=phase-III}  
\documentclass{article}
```

# Paragraphs and links



- implemented with the para hooks

# Footnotes

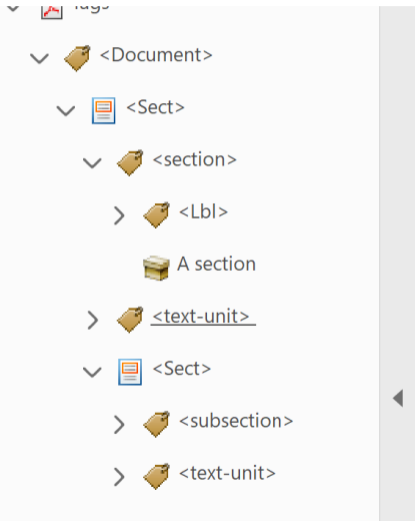


- new implementation of footnotes
- `footmisc` is already compatible



- minipage footnotes not yet fully handled

# Sectioning



- Sect structure surrounds heading and body

- changed:










`\@startsection`, `\@sect` etc



- currently incompatible:

e.g. memoir, KOMA-classes, titlesec

# Table of contents and similar lists









- >  <H1>
- ✓  <TOC> toc
  - ✓  <TOCI> Heading on le
    - >  <Reference>
  - ✓  <TOC>
    - >  <TOCI> Heading or
    - >  <TOC>
  - >  <TOCI> Lists
  - >  <TOC>

- changed:  
`\@starttoc`, `\addcontentsline`,  
`\@dottedtocline` and `\l@chapter` etc



- currently incompatible:  
e.g. memoir, KOMA-classes, titletoc

# Display environments and lists

- ✓  <text-unit>
- ✓  <text>
  -  centered text
- ✓  <text-unit>
- ✓  <enumerate>
- ✓  <LI>
- >  <Lbl>
- >  <LBody>
- >  <LI>

- LaTeX environments built on trivlist:

- center, flushleft, flushright
- quotation, quote, verse
- verbatim
- theorems
- enumerate, itemize, description
- list, trivlist

- new implementation based on xtemplate

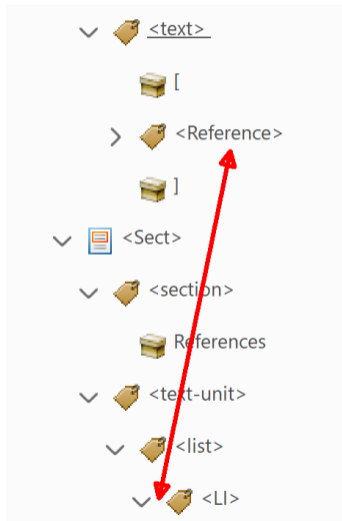
- block environments like center or verbatim are no longer lists



- currently incompatible:

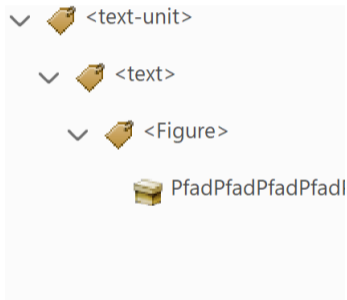
enumitem, enumerate, fancyvrb, listings, theorem packages ...

# Citations and bibliography



- `thebibliography` is supported
- `natbib` is supported
- `biblatex` with `hyperref` is supported
- still unsupported:  
 `biblatex` without `hyperref`





- keys for alternative text:

```
\includegraphics[alt={This shows a  
duck}]{duck}
```

```
\begin{picture}[alt={This shows a duck  
too}](100,100)
```

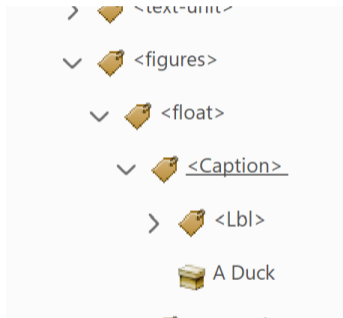
...

```
\end{picture}
```



- tikz is not yet supported

# Floats



- changed:  
`\@xfloat` and `\@makecaption`
- caption package more or less compatible

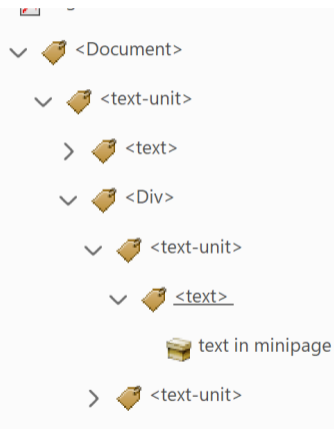


- float package currently incompatible



- `\marginpar` not yet supported

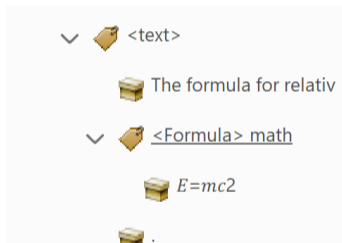
# minipage and \parbox



- tagging is enabled



- adjustments in special cases needed



- experimental prototype for math tagging

- loaded as additional module:

```
\DocumentMetadata  
  {testphase={phase-III,math}}
```

- math is grabbed and then processed



- affects also “non-math” text like superscripts or urls



- only rough tagging, best tagging still unclear (missing PDF 2.0 support)

## Additional support for packages

- `firstaid` key loads small fixes for classes and packages

```
\DocumentMetadata{testphase={phase-III,firstaid}}
```

# What is missing? Reusing boxes



- `\savebox` and `\usebox` not yet supported

# What is missing? Tables

- Manual tagging is possible

-  ● How to identify and markup the header cells for automatic tagging?

# Summary

- Automatic tagging of various standard documents is now enabled

```
\DocumentMetadata {testphase={phase-III,math,firstaid}}  
\documentclass{book}  
\usepackage[math,toc]{blindtext}  
\begin{document}  
\Blinddocument  
\end{document}
```

- Tester and feedback welcome!
- Report issues at  
<https://github.com/latex3/tagging-project>





Thank you for  
your attention!