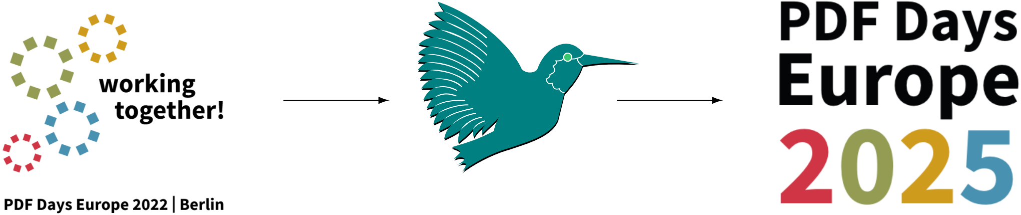
Tagged and Accessible PDF with LaTeX — revisited

Frank Mittelbach, Ulrike Fischer

LaTeX Project

Berlin, September 2025



Overview

- LaTeX's role in producing STEM documents
- STEM documents are a challenge for accessibility
- The accessibility of PDF/UA-1 documents with STEM content
- The accessibility of PDF/UA-2 documents with STEM content
- An end-to-end (E2E) workflow for accessible STEM documents

LaTeX's role in producing STEM documents

- LaTeX is dominant in academic and technical writing in many STEM areas, particularly in mathematics, physics, and computer science
- Overleaf (online service for LaTeX) has more than 20 million user accounts
- Many technical publishers require or offer LaTeX as a publishing workflow
- Many academic/technical archives use PDFs produced from LaTeX
- This includes, for example, 95% of the documents on Cornell Tech's `arXiv.org`, which currently holds roughly 2.2 million scholarly articles

Mathematics in STEM documents — A simple example

You hear: Some formulas:

$$\sum_{i=1}^{n} i = \frac{(n+1)n}{2} \tag{1}$$

You hear: n X (n + 1)n i = (1) 2 i=1

$$(a+b)(a-b) = a^2 - ab + ba - b^2 \tag{2}$$

$$= a^2 - b^2 \tag{3}$$

You hear: (a + b)(a - b) = a 2 - ab + ba - b2 (2) = a (3) 2 - b2

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 3 & 7 \end{pmatrix}$$

You hear: 1211 13 = 3401 37

Observations & conclusions

- Currently, most STEM documents are untagged PDF documents!
- Untagged PDF documents are not accessible if they contain math (they are't very accessible if they don't contain math either)

“PDF is evil”, as often voiced in the accessibility community, is warranted

But does it have to be?

Is tagged PDF any better? — The PDF/UA-1 approach

- PDF/UA-1 mandates the use of tagged PDF but it is based on PDF 1.7
- PDF 1.7 does not understand the structure of math formulas
- PDF/UA-1 therefore mandates the addition of descriptive “alternative text” to each and every formula

Can this work?

Is tagged PDF any better? — The PDF/UA-1 approach

Manually adding alternative text

- A typical page of a math article contains several dozen (usually between 50 and 100) inline and display formulas
- It it unrealistic to expect that authors annotate them (unless forced to)
- Verbal descriptions of formulas are likely to be inaccurate and incomplete
- If done in a post-processing step then any change in the document invalidates that work

Bottom line: **This means most documents remain not accessible (even if they formally comply with the PDF/UA-1 standard)**

A PHASE SPACE LOCALIZATION OPERATOR IN NEGATIVE BINOMIAL STATES 9

The PDF/UA-1 approach with automation

Automatic addition of alternative text with LaTeX

Instead of a verbal description the LaTeX source can be used as a representation of a formula!

Pro:

- This can be done automatically
- It does not require post-processing steps and allows for document changes
- LaTeX source is understood by many STEM users

Con:

- Accessible, but with reduced quality
- Fairly unusable with braille devices (too verbose)
- All users of AT have to understand LaTeX formula source syntax
- It requires some discipline by the document author
(formula source has to be understandable and self-contained)

The PDF/UA-2 approach

- PDF/UA-2 is based on PDF 2.0
- MathML allows for a granular representation of formulas and is, for example, successfully used to produce accessible formulas in Web browsers
- PDF 2.0 introduces the idea of using MathML for formula data
  - Either as MathML embedded in “Associated Files” or
  - through the use of PDF Structure Elements named after MathML elements

So are there any road blocks for accessible STEM PDFs?

The PDF/UA-2 approach — The road blocks (before 2025)

Underspecification of MathML inside PDF!

- PDF 2.0 introduced the idea of using MathML
- However, the details are left undefined
  - A precise mapping between PDF MathML Structure Elements and the MathML specification is missing
  - The PDF 2.0 spec also gives no indication on how formulas containing text, links and nested math, should be represented

Result: No producer and no consumer implementations

The PDF/UA-2 approach — The road blocks (before 2025)

No implementation of an end-to-end workflow!

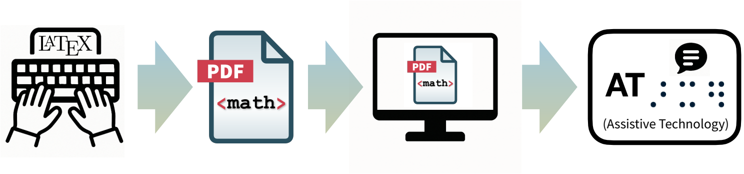
- No producer that generated PDF with MathML inside
- No reader that passed MathML to the AT tools
- No AT tool that consumed MathML in the PDF context

The Accessibility checker problem

- Most Accessibility checkers test for PDF/UA-1 rules only
- As a consequence they incorrectly fail PDF 2.0 and PDF/UA-2 documents

Any change in 2025?

New in 2025: An E2E workflow for STEM — Finally!



Ingredients

- A suitable creation software (such as LaTeX)
- A PDF format with the necessary features (PDF 2.0)
- PDF reader software with the ability to process MathML (Foxit, Acrobat, . . .)
- Assistive technology that can handle MathML in the PDF context (NVDA & MathCAT)

A simple example — revisited

$$\sum_{i=1}^{n} i = \frac{(n+1)n}{2} \tag{1}$$

Now you hear: 1 line with label 1 / the sum from i is equal to 1 to n of i is equal to the fraction with numerator open paren n plus 1 close paren times n and denominator 2

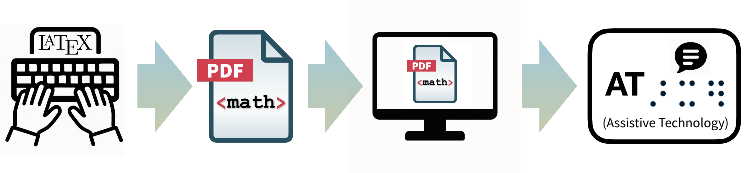
$$\begin{align} (a+b)(a-b) &= a^2 - ab + ba - b^2 \tag{2} \\ &= a^2 - b^2 \tag{3} \end{align}$$

Now you hear: 2 lines / line 1 with label 2 / open paren a plus b close paren times open paren . . . / line 2 with label 3 / is equal to a squared minus b squared

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 3 & 7 \end{pmatrix}$$

Now you hear: The 2 by 2 matrix row 1 1 2 / row 2 3 4 / times . . .

An E2E workflow for STEM — Caveats



There is still work to do . . .

- PDF 2.0 is still poorly supported by consumer applications
  - Foxit supports MathML both as AF or as SEs
  - Acrobat to date only supports MathML Structure Elements but not AF
  - Other readers: ???
- No AT support other than NVDA (Windows only) so far
- The Accessibility checker problem remains

But despite all the work that is still needed — The desert is finally starting to bloom . . .

Thank you for your attention