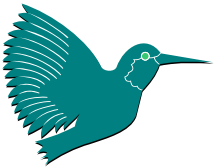
Tagged and Accessible PDF with LaTeX

Frank Mittelbach, Ulrike Fischer

LaTeX Project

Berlin, September 2022



PDF Days Europe 2022 | Berlin

The scenery

- What is LaTeX and why should we care?
- The PDF challenge
- The opportunity

LaTeX's long history

- It predates HTML / PDF / Unicode / CSS / ...
  - and has been parent or inspiration for many of the later concepts

LaTeX's long history

- It predates HTML / PDF / Unicode / CSS / ...
  - and has been parent or inspiration for many of the later concepts
- First version around 1984 by Leslie Lamport
  - with focus on printing (and viewing)
  - Initially, output was DVI (device independent format)
    - with translations to printer languages (e.g. PCL)
    - and to PostScript as a major workflow

LaTeX's long history

- It predates HTML / PDF / Unicode / CSS / ...
  - and has been parent or inspiration for many of the later concepts
- First version around 1984 by Leslie Lamport
  - with focus on printing (and viewing)
  - Initially, output was DVI (device independent format)
    - with translations to printer languages (e.g. PCL)
    - and to PostScript as a major workflow
- LaTeX 2_ε in 1994 by the LaTeX Project Team
  - still at the core of what is used today
  - still with focus on printing

LaTeX's long history

- It predates HTML / PDF / Unicode / CSS / ...
  - and has been parent or inspiration for many of the later concepts
- First version around 1984 by Leslie Lamport
  - with focus on printing (and viewing)
  - Initially, output was DVI (device independent format)
    - with translations to printer languages (e.g. PCL)
    - and to PostScript as a major workflow
- LaTeX 2_ε in 1994 by the LaTeX Project Team
  - still at the core of what is used today
  - still with focus on printing
- Since then continual development and enhancements

LaTeX's use today

- Widely used input language for the typesetting engine TeX
  - more than 10k daily users at `latex-project.org` from around the globe

LaTeX's use today

- Widely used input language for the typesetting engine TeX
  - more than 10k daily users at `latex-project.org` from around the globe
- These days main output format is PDF

LaTeX's use today

- Widely used input language for the typesetting engine TeX
  - more than 10k daily users at `latex-project.org` from around the globe
- These days main output format is PDF
- The lingua franca in STEM / MINT disciplines
  - 10^+ million users on Overleaf (online collaboration service using LaTeX)
  - 2^+ million documents at `arXiv.org` (archive for STEM publications)
  - Numerous journal and university styles

LaTeX's use today

- Widely used input language for the typesetting engine TeX
  - more than 10k daily users at `latex-project.org` from around the globe
- These days main output format is PDF
- The lingua franca in STEM / MINT disciplines
  - 10^+ million users on Overleaf (online collaboration service using LaTeX)
  - 2^+ million documents at `arXiv.org` (archive for STEM publications)
  - Numerous journal and university styles
- ... but usage in basically any field and type of application
  - critical editions
  - game typesetting (chess, go, ...)
  - database driven publications

LaTeX's use today

- Widely used input language for the typesetting engine TeX
  - more than 10k daily users at `latex-project.org` from around the globe
- These days main output format is PDF
- The lingua franca in STEM / MINT disciplines
  - 10^+ million users on Overleaf (online collaboration service using LaTeX)
  - 2^+ million documents at `arXiv.org` (archive for STEM publications)
  - Numerous journal and university styles
- ... but usage in basically any field and type of application
  - critical editions
  - game typesetting (chess, go, ...)
  - database driven publications
- More than 5000 extension packages

Some LaTeX's strengths (highlights only)

- LaTeX focuses
  - on semantic structure — kept separate from formatting
  - on reuse with different formatting
  - *but users are also able to overwrite automatic formatting with detailed process instructions*

Some LaTeX's strengths (highlights only)

- LaTeX focuses
  - on semantic structure — kept separate from formatting
  - on reuse with different formatting
  - *but users are also able to overwrite automatic formatting with detailed process instructions*
- Features
  - high-quality (unsurpassed) typesetting, in particular for mathematics
  - long-term compatibility — *documents from the nineties and earlier are still processable*
  - programmable and extensible — *typesetting solutions for nearly every problem and domain exist*

LaTeX: Typesetting examples

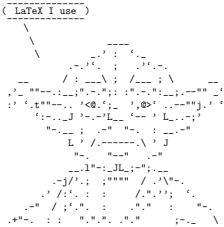
Everything written symbols can say has already passed by ...

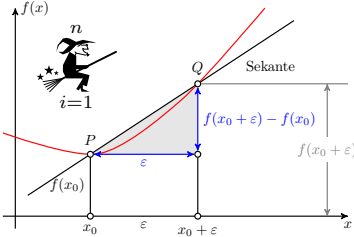
THEY ARE LIKE TRACKS LEFT BY ANIMALS. That is why the masters of meditation refuse to accept that writings are final. The aim is to reach true being by means of those tracks, those letters, those signs; but reality itself is not a sign, and it leaves no tracks. It doesn't come to us by way of letters or words. We can go toward it, by following those words and letters back to whence they came from.

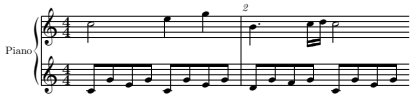
إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ

μὴ θορυβεῖτε, ὦ ἄνδρες Ἀθηναῖοι

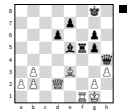
সবকিছু প্রবাহিত হয়



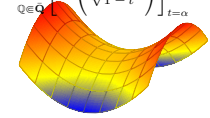




$$\iiint_{Q} f(w,x,y,z)\,dw\,dx\,dy\,dz \le \oint_{\partial Q} f'\left(\max\left\{\frac{\|w\|}{|w^2+x^2|};\frac{\|z\|}{|y^2+z^2|};\frac{\|w\oplus z\|}{\|x\oplus y\|}\right\}\right)$$



$$\sum_{\mathbb{Q}\in\bar{\mathbb{Q}}} \left[f^*\left(\frac{\mathbb{Q}(t)}{\sqrt{1-t^2}}\right)\right]_{t=\alpha}^{t=\vartheta} - (\Delta+\nu-v)^3$$



LaTeX: Source example

```
\documentclass{demo1} % or demo2
\author{Frank Mittelbach}
\title{Example \LaTeX{} document}
\begin{document}

\maketitle

\tableofcontents

\section{Introduction}

This example shows\footnote{See
  \url{https://latex-project.org} for more.}
\begin{itemize}
  \item the title
  \item the table of contents
  \item section headings
  \item a list
  \item some text
  \item a footnote
  \item some math
  \item a figure.
\end{itemize}

\subsection{Some Math}
A famous equation
\begin{equation} E= mc^2 \end{equation}

\section{Sample text}
Take a look at figure~\vref{fig:cups}.

\kant[1][1]

\begin{figure}\centering
\includegraphics[width=\linewidth]{coffeecup}
\caption{Two coffee cups\label{fig:cups}}
\end{figure}

\kant[2] \kant[3][1-4] \kant[4]
\end{document}
```

Example LaTeX document

Frank Mittelbach

September 6, 2022

Contents

1 Introduction

This example shows[1]

- the title
- the table of contents
- section headings
- a list
- some text
- a footnote
- some math
- a figure.

[1] See `https://latex-project.org` for more.

Figure 1: Two coffee cups

1.1 Some Math

A famous equation

$$E = mc^2 \tag{1}$$

2 Sample text

Take a look at figure 1.

As any dedicated reader can clearly see, the Ideal of practical reason is a representation of, as far as I know, the things in themselves; as I have shown elsewhere, the phenomena should only be used as a canon for our understanding.

Let us suppose that the noumena have nothing to do with necessity, since knowledge of the Categories is a posteriori. Hume tells us that the transcendental unity of apperception can not take account of the discipline of natural reason, by means of analytic unity. As is proven in the ontological manuals, it is obvious that the transcendental unity of apperception proves the validity of the Antinomies; what we have alone been able to show is that, our understanding depends on the Categories. It remains a mystery why the Ideal stands

Example LaTeX document

Frank Mittelbach

September 6, 2022

Contents

Introduction

This example shows[1]

- the title
- the table of contents
- section headings
- a list
- some text
- a footnote
- some math
- a figure.

[1] See https://latex-project.org for more.

Some Math

A famous equation

$$E = mc^2 \tag{1}$$

Sample text

Take a look at figure 1 on the following page.

As any dedicated reader can clearly see, the Ideal of practical reason is a representation of, as far as I know, the things in themselves; as I have shown elsewhere, the phenomena should only be used as a canon for our understanding.

Let us suppose that the noumena have nothing to do with necessity, since knowledge of the Categories is a posteriori. Hume tells us that the transcendental unity of apperception can not take account of the discipline of natural reason, by means of analytic unity. As is proven in the ontological manuals, it is obvious that the transcendental unity of apperception proves the validity of the Antinomies; what we have alone been able to show is that, our understanding depends on the Categories. It remains a mystery why the Ideal stands in need of reason. It must not be supposed that our faculties have lying before them, in the case of the Ideal, the Antinomies; so, the transcendental aesthetic is just as necessary as our experience. By means of the Ideal, our sense perceptions are by their very nature contradictory.

As is shown in the writings of Aristotle, the things

Figure 1: Two coffee cups

in themselves (and it remains a mystery why this is the case) are a representation of time. Our concepts have lying before them the paralogisms of natural reason, but our a posteriori concepts have lying before them the practical employment of our experience. Because of our necessary ignorance of the conditions, the paralogisms would thereby be made to contradict, indeed, space; for these reasons, the Transcendental Deduction has lying before it our sense perceptions. (Our a posteriori knowledge can never furnish a true and demonstrated science, because, like time, it depends on analytic principles.

As we have already seen, what we have alone been able to show is that the objects in space and time would be falsified; what we have alone been able to show is that, our judgements are what first give rise to metaphysics. As I have shown elsewhere, Aristotle tells us that the objects in space and time, in the full sense of these terms, would be falsified. Let us sup-

PDF — Portable Document Format

- Initially focused on precise instructions for print output
- Semantic information describing the contents added later as an optional separate structure tree

PDF — Portable Document Format

- Structure other than precise formatting instructions was of no importance and not present in the format
  - Like TEX/LATEX the focus was (and is) on “pages”

PDF — Portable Document Format

- Structure other than precise formatting instructions was of no importance and not present in the format
  - Like TEX/LATEX the focus was (and is) on "pages"
- Semantic information as an afterthought
  - not much take-up initially, because printing and viewing was the dominant use case
  - and adding additional structure information was (and is!) not well-supported by most PDF-generating workflows

PDF — Portable Document Format

- Structure other than precise formatting instructions was of no importance and not present in the format
  - Like TEX/LATEX the focus was (and is) on "pages"
- Semantic information as an afterthought
  - not much take-up initially, because printing and viewing was the dominant use case
  - and adding additional structure information was (and is!) not well-supported by most PDF-generating workflows

- Until now, PDF generated from LATEX does not contain structure information (tags) and offers only minimal support for metadata

The pressure is growing — An opportunity/challenge for LaTeX?

- Printing becomes a secondary action
- Reliable reuse becomes important
- Accessibility becomes a requirement
- ...

Pick the low-hanging fruits?

- LaTeX knows the document semantic structure —
- So pass it on into the PDF!
- Why should that be difficult?

Pick the low-hanging fruits?

- LaTeX knows the document semantic structure —
- So pass it on into the PDF!
- Why should that be difficult?

(you better use gloves here)

The opportunity for LaTeX

LaTeX is focused on structure, it is programmable, and can write PDF

The opportunity for LaTeX

LaTeX is focused on structure, it is programmable, and can write PDF

- Thus, with the right kernel adjustments it
  - should be able to *automatically* produce well-tagged PDF
  - with (all/most) data necessary to comply with UA and other standards

The opportunity for LaTeX

LaTeX is focused on structure, it is programmable, and can write PDF

- Thus, with the right kernel adjustments it
  - should be able to *automatically* produce well-tagged PDF
  - with (all/most) data necessary to comply with UA and other standards

- This means that new documents could be *automatically* made accessible
  - without any need to post-process them

The opportunity for LaTeX

LaTeX is focused on structure, it is programmable, and can write PDF

- Thus, with the right kernel adjustments it
  - should be able to *automatically* produce well-tagged PDF
  - with (all/most) data necessary to comply with UA and other standards

- This means that new documents could be *automatically* made accessible
  - without any need to post-process them
- Existing documents (on the web and elsewhere) could be made accessible with reasonable effort
  - by adding missing data (metadata, alternative texts, . . .)
  - and then reprocessing them

Overcoming the (hidden) obstacles

- Conceptual difficulties
- Missing functionality (in LaTeX)
- Technical difficulties
- Eco-system difficulties

Conceptual difficulties

- LaTeX's tag model is far richer than the PDF tag model
  - For example, LaTeX supports different footnote classes and nested footnotes as required for critical editions
  - LaTeX is "freely, and easily, extensible", e.g., via `\newenvironment`, with few restrictions concerning adding new semantic structures

Conceptual difficulties

- LaTeX's tag model is far richer than the PDF tag model
  - For example, LaTeX supports different footnote classes and nested footnotes as required for critical editions
  - LaTeX is "freely, and easily, extensible", e.g., via `\newenvironment`, with few restrictions concerning adding new semantic structures

- This requires finding reasonable (consistent) mappings
  - limiting the amount of information loss as much as possible
  - (e.g., through role mapping)

Conceptual difficulties

- LaTeX's tag model is far richer than the PDF tag model
  - For example, LaTeX supports different footnote classes and nested footnotes as required for critical editions
  - LaTeX is "freely, and easily, extensible", e.g., via `\newenvironment`, with few restrictions concerning adding new semantic structures

- This requires finding reasonable (consistent) mappings
  - limiting the amount of information loss as much as possible
  - (e.g., through role mapping)
- It also requires providing a user method to indicate reasonable mappings
  - for new document elements
  - for existing document elements that need special interpretation

Missing functionality in LaTeX

- LaTeX does not have syntax support for specifying alternate text

Missing functionality in LaTeX

- LaTeX does not have syntax support for specifying alternate text
  - resolved through a new standard key/value interface

Missing functionality in LaTeX

- LaTeX does not have syntax support for specifying alternate text
  - resolved through a new standard key/value interface
- LaTeX needs a better model for specifying metadata
  - there are existing interfaces but not very consistent

Missing functionality in LaTeX

- LaTeX does not have syntax support for specifying alternate text
  - resolved through a new standard key/value interface
- LaTeX needs a better model for specifying metadata
  - there are existing interfaces but not very consistent
  - resolved/work in progress through a new `\DocumentMetadata` interface

Missing functionality in LaTeX

- LaTeX does not have syntax support for specifying alternate text
  - resolved through a new standard key/value interface
- LaTeX needs a better model for specifying metadata
  - there are existing interfaces but not very consistent
  - resolved/work in progress through a new `\DocumentMetadata` interface
- LaTeX needs a better table model
  - the current model is entirely visual
  - and has no or only very limited structural information

Missing functionality in LaTeX

- LaTeX does not have syntax support for specifying alternate text
  - resolved through a new standard key/value interface
- LaTeX needs a better model for specifying metadata
  - there are existing interfaces but not very consistent
  - resolved/work in progress through a new `\DocumentMetadata` interface
- LaTeX needs a better table model
  - the current model is entirely visual
  - and has no or only very limited structural information
  - work in progress

Technical (implementation) difficulties

- TEX does not use explicit spaces between words
  - issue with pdfTEX, XeTEX, and PostScript workflows (LATEX + `dvips`)
- Issues with support for the full range of Unicode
  - some restrictions with 8-bit engines, e.g., pdfTEX
  - and with math if "classic" math fonts are used
- Several different engines (pdfTEX, LuaTEX, XeTEX, upTEX, ...) and PostScript processors (Ghostscript, distiller) are in common use
  - all workflows used their own methods to write the PDF data

Technical (implementation) difficulties

- TEX does not use explicit spaces between words
  - issue with pdfTEX, XeTEX, and PostScript workflows (LATEX + `dvips`)
  - resolved for pdfTEX at some cost in speed/PDF size
- Issues with support for the full range of Unicode
  - some restrictions with 8-bit engines, e.g., pdfTEX
  - and with math if "classic" math fonts are used
  - not fully resolvable in non-Unicode engines
- Several different engines (pdfTEX, LuaTEX, XeTEX, upTEX, ...) and PostScript processors (Ghostscript, distiller) are in common use
  - all workflows used their own methods to write the PDF data
  - resolved by providing an abstraction layer with a new PDF management module *(to be moved to the* LATEX format*)*

Eco-system considerations — some facts

- LaTeX is well-known for strong backward compatibility
  - You can take documents from the early '90 and reasonably expect them to compile with no or little adjustments

Eco-system considerations — some facts

- LaTeX is well-known for strong backward compatibility
  - You can take documents from the early '90 and reasonably expect them to compile with no or little adjustments
- The core of LaTeX was designed long ago when computer memory and speed were very low. Some consequences:
  - The program takes great care to forget the structural information the moment it is no longer need for producing "print" output
  - There are nearly no public interfaces to "hook" into the processing — everything was optimized to save space and processing time

Eco-system considerations — some facts

- LaTeX is well-known for strong backward compatibility
  - You can take documents from the early '90 and reasonably expect them to compile with no or little adjustments
- The core of LaTeX was designed long ago when computer memory and speed were very low. Some consequences:
  - The program takes great care to forget the structural information the moment it is no longer need for producing "print" output
  - There are nearly no public interfaces to "hook" into the processing — everything was optimized to save space and processing time
- Therefore, nearly all the extension packages hook into internal code of the LaTeX kernel and structural information is lost by the time the PDF is written

Eco-system considerations — the consequences

Compatibility expectations + hooking into internal code $\longrightarrow$ means:

Eco-system considerations — the consequences

Compatibility expectations + hooking into internal code $\longrightarrow$ means:

- All kernel changes are likely to produce noticeable disruptions
  - this needs careful mediation to avoid frustration in the community
  - all affected packages need identifying and adjustments have to be put in place
  - this requires effort, and care — and therefore takes considerable time

Eco-system considerations — the consequences

Compatibility expectations + hooking into internal code $\longrightarrow$ means:

- All kernel changes are likely to produce noticeable disruptions
  - this needs careful mediation to avoid frustration in the community
  - all affected packages need identifying and adjustments have to be put in place
  - this requires effort, and care — and therefore takes considerable time
- However, such changes are essential to implement new processing models

Eco-system considerations — the consequences

Compatibility expectations + hooking into internal code $\longrightarrow$ means:

- All kernel changes are likely to produce noticeable disruptions
  - this needs careful mediation to avoid frustration in the community
  - all affected packages need identifying and adjustments have to be put in place
  - this requires effort, and care — and therefore takes considerable time
- However, such changes are essential to implement new processing models
- Examples of already done changes (a.k.a. some success stories)
  - Move to UTF-8 as the default input
  - Integrate the L3 programming layer — needed for better interfaces
  - Provide a general hook management — to improve the situation in the future
  - Provide a general key/value interface — needed for tagging configuration
  - Standardize the interfaces to write PDF data

Roadworks — or what happens in the project

- Six project phases
- End of phase II in sight
- What's already available
  — some examples
- Useful links

Several lanes are already resurfaced

The six project phases

- They are chosen in a way to
  - minimize disruption
  - provide tangible (intermediate) results that people can already use

Several lanes are already resurfaced

The six project phases

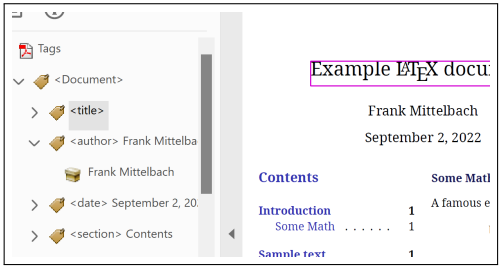
- They are chosen in a way to
  - minimize disruption
  - provide tangible (intermediate) results that people can already use

Major milestones reached

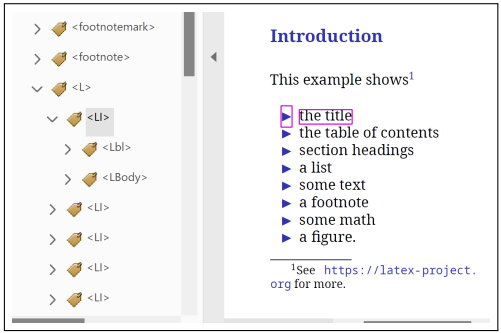
- Phase I + II
  - Standard hook management is designed, implemented and used by the kernel
  - Interfaces for PDF object management are designed and implemented
  - All low-level mechanisms needed for tagging are available (`tagpdf`)
  - Automatic tagging of paragraph text implemented
  - A subset of the standard document elements is "tagging enabled"

LaTeX PDF: Example I

Tagging of the title

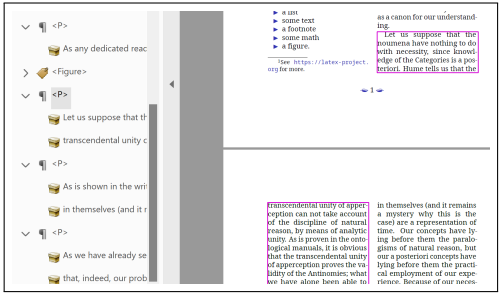


Tagging of the list

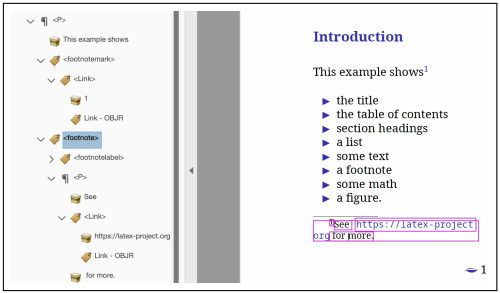


LaTeX PDF: Example II

A paragraph spanning two pages



Tagging of the footnote



Useful information and links

- Project material at the LaTeX Project website:
  - Feasibility study
  - Talks and articles
  - ...

  `https://latex-project.org/publications/indexbytopic/pdf/`

Useful information and links

- Project material at the LaTeX Project website:
  - Feasibility study
  - Talks and articles
  - ...

  `https://latex-project.org/publications/indexbytopic/pdf/`

- LaTeX LWG within the PDF Association (chair Boris Doubrov)
  - Meetings about once a month
  - Currently working on defining suitable mapping from LaTeX to PDF tag sets and identifying gaps on either side
  - Open to interested PDFA members

Stay tuned

PDFA Days, Berlin, September 2022
Frank Mittelbach, Ulrike Fischer, LaTeX Project