Accessibility in the LaTeX kernel — experiments in Tagged PDF

Chris Rowley, Ulrike Fischer, Frank Mittelbach

Abstract

This is a brief summary of a talk given by the first author at the TUG'19 conference, together with some references for further reading and viewing.

1 Introduction

Accessibility requirements for PDF documents are described in two standards: PDF/UA and the more recent PDF2.0. One of the major features they mandate is that the PDF must be properly tagged, so we are investigating how LaTeX can be adapted to easily produce tagged PDF.

The main purpose of this talk was to introduce the experimental package *tagpdf* by the second author. But we start with a quick "bullet points introduction" to the structure of a PDF file and what is meant by "Tagged PDF".

2 PDF in bullets

The first thing to understand about PDF is: In a PDF file (almost) everything is ... *PDF Objects*! Here are two examples of important object types that we always find in a PDF file:

- many objects are Dictionaries, which are simply key-value (property) lists
- other objects are Streams: Text Streams are an important part of each Page Object

Some important particular objects are:

- Resource Objects: containing, for example, font and encoding information
- Page Objects: containing information about a page
- Navigation Objects: these enable quick access to all the important objects within the PDF file.

For Tagged PDF, in addition to Page Objects and their Text Streams, the following are required:

- a Structure Tree Object, whose nodes are PDF Objects (surprise?), with:
  - a root node
  - structure element nodes: each of these is a Dictionary Object containing references to its parent, siblings and child nodes
  - leaf nodes: each containing additional references, each of which is to:
    - a page, plus
    - a "marked part" of that page's Text Stream

The slides for the talk contain examples of the type of code used in a PDF file for defining the objects related to structure and tagging.

3 Philosophy

We believe that the production of documents that exploit the large range of functionality that can nowadays be incorporated into PDF is very important and is fundamental to what LaTeX, as a document processing system, is all about!

We also believe that the production system must pay close attention to the actual, detailed contents of the input LaTeX file: these details must not be ignored as they contain everything that the system knows about the author's intentions.

We are therefore certain that we first need to adapt and enhance the LaTeX kernel to better support all these new ideas. Furthermore we must go on to help package developers and maintainers in exploiting the new possibilities.

We are currently working primarily on getting right the low-level basic coding so that we can build on top of this to add necessary features into the LaTeX system. We do not want to add lots of new stuff on top of the current LaTeX, or to produce a parallel system that will most likely conflict badly with standard LaTeX processing.

4 Current work

Development has started on the experimental package *tagpdf*. Its purposes are summarised here:

- allow experimenters to identify problems they find with tagging, and to discover the support needed for other accessibility requirements;
- develop a code basis for the support of tagging in the LaTeX kernel.

Please note that, being experimental, it needs experimenters to use it, of at least these types:

- authors and users of documents:
  – What is truly needed in an accessible document?
- package maintainers/developers:
  – What is needed for a package to produce accessible output?
  – How can the LaTeX Team help ease the conversion of all packages?

The current (preliminary) version of the *tagpdf* packages provides low-level mark-up commands to support tagging. For example, commands to:

- add structure element nodes to the structure tree
- add "marked content" tags to the content stream

- add to the structure tree nodes all the necessary pointers to the marked content associated with a given node

The package also supports other aspects of accessibility, such as setting up links appropriately, and the input of essential document meta-data. It is well documented with descriptions of how to use it and of how to provide us with feedback. Please do!

The documentation contains more background information about accessibility and tagging, with descriptions of how PDF works and what makes a PDF file accessible. It also lists some currently known problems and how we plan to solve them.

5 Coming soon, we hope!

We of course hope to get many ideas from all you experimenters, but meanwhile we are looking in detail at how the LaTeX kernel can better support tagging and other aspects of accessibility. We are also looking at whether the various TeX engines need any enhancements to better support the production of full-featured PDF.

In the TeX community, Ross Moore and others in TUG's Accessibility Working Group have done considerable work on many of these problems, including the complex subject of how to represent formulas, etc., in accessible PDF. We are therefore actively exchanging ideas with them and we are pleased to thank TUG and DANTE e.V. for their current support of this work.

We are of course also very interested in collaboration with other organisations, individuals and companies who have engineering expertise in this area (from both the TeX perspective and the PDF perspective) and we intend to actively pursue such contacts.

As part of their program to position PDF as a prime source of accessible information in "value-added documents", the expert engineers at the "home of PDF", Adobe, are showing a high level of interest in the use of LaTeX to produce accessible PDF. This gives a clear indication of the importance of LaTeX for the production of PDF documents and we are therefore planning to collaborate closely with them.

References

The *tagpdf* package This is available at: `https://github.com/u-fischer/tagpdf` and `https://ctan.org/pkg/tagpdf`.

PDF Standards PDF/UA (PDF/Universal Accessibility) is the informal name for ISO 14289. On July 28, 2017, ISO 32000-2:2017 (PDF 2.0) was published. See: `https://www.iso.org/standard/64599.html` and `https://www.iso.org/standard/63534.html`

More on PDF Lots of information is available at: `https://en.wikipedia.org/wiki/PDF`

Moore on PDF Ross Moore has published many talks and articles, see: `https://maths.mq.edu.au/~ross/TaggedPDF` Videos of both his talk and this one at TUG 2019 can be seen at: `http://science.mq.edu.au/~ross/TaggedPDF/TUG2019-movies`

Slides for this talk These are available at: `https://latex-project.org/publications/indexbyyear/2019/`

TUG's Accessibility Working Group More information and a fuller bibliography on Accessibility is available at: `https://tug.org/twg/accessibility/`

◇ Chris Rowley
LaTeX3 Team
`chris.rowley (at) latex-project dot org`
`https://www.latex-project.org`

◇ Ulrike Fischer
LaTeX3 Team
`fischer (at) troubleshooting-tex dot de`
`https://www.latex-project.org`

◇ Frank Mittelbach
LaTeX3 Team
`frank.mittelbach (at) latex-project dot org`
`https://www.latex-project.org`