Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
InIT Institut für angewandte
Informationstechnologie

Automatic Paragraph Detection for Accessible PDF Documents

Alireza Darvishy and Severin Münger

ICT-Accessibility Lab ZHAW

July 13th, 2016

Agenda

- About **PAVE**
- Live Demo
- Q&A

PAVE

Zürcher Hochschule
für Angewandte Wissenschaften
zh aw School of Engineering
InIT Institut für angewandte Informationstechnologie

PAVE is a web-based application to make PDF documents accessible (www.pave-pdf.org/index.en).

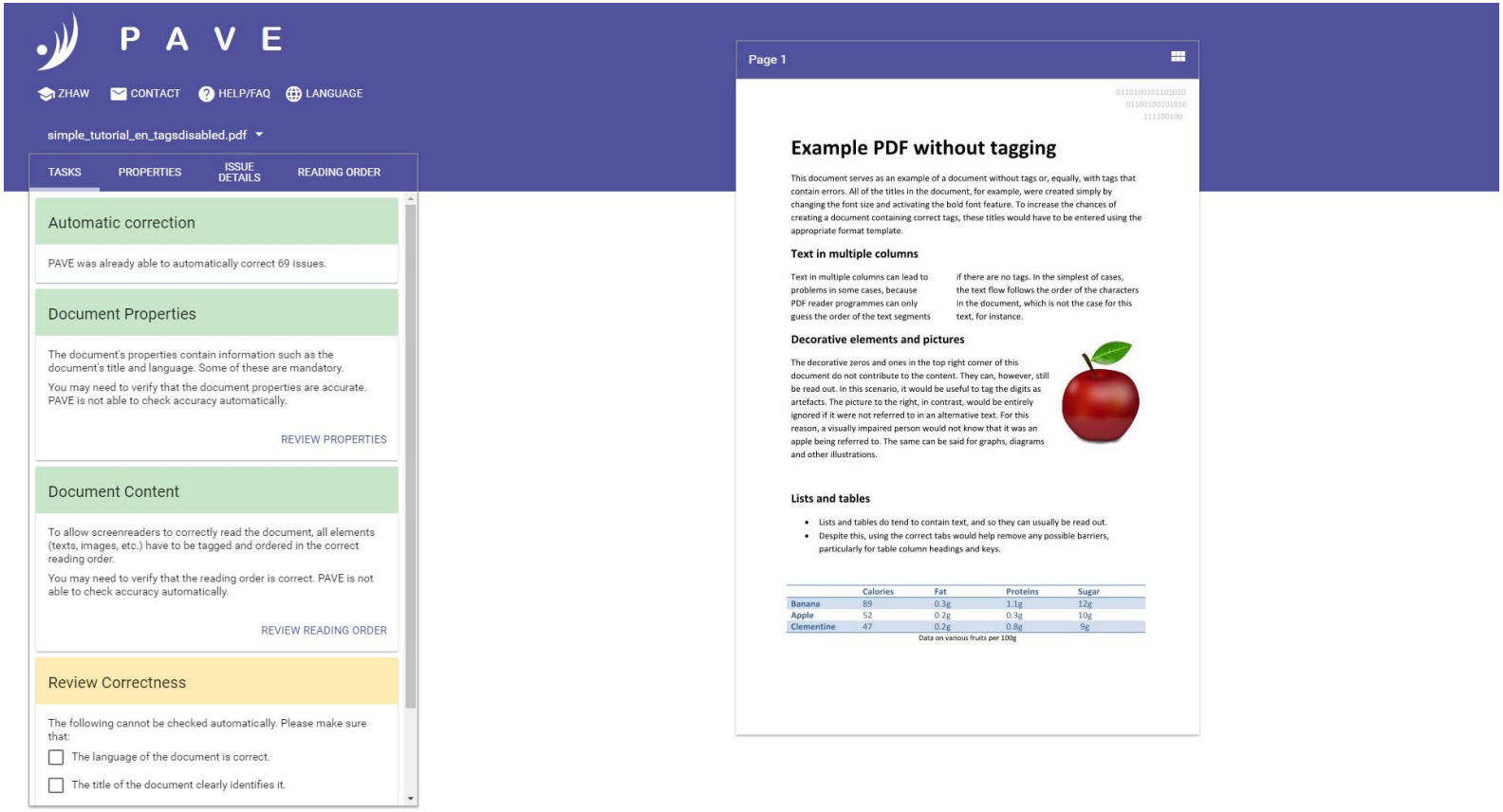
1. Upload your PDF document to **PAVE**
2. Automatic corrections (paragraphs, headings…)
3. Manual corrections
4. Download the accessible PDF

Award for **PAVE**

C4C Award at ICCHP 2014 in Paris

PAVE

Web-based tool to make PDF documents accessible



Example of a non-accessible PDF

Irrelevant or confusing text is read out loud

0110100101101010
01100100101010
111100100

Headings are not recognized

Example PDF without tagging

This document serves as an example of a document without tags or, equally, with tags that contain errors. All of the titles in the document, for example, were created simply by changing the font size and activating the bold font feature. To increase the chances of creating a document containing correct tags, these titles would have to be entered using the appropriate format template.

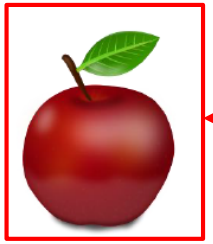
Text in multiple columns

Text is read out loud across the columns

Text in multiple columns can lead to	if there are no tags. In the simplest of cases,
problems in some cases, because	the text flow follows the order of the characters
PDF reader programmes can only	in the document, which is not the case for this
guess the order of the text segments	text, for instance.

Decorative elements and pictures

The decorative zeros and ones in the top right corner of this document do not contribute to the content. They can, however, still be read out. In this scenario, it would be useful to tag the digits as artefacts. The picture to the right, in contrast, would be entirely ignored if it were not referred to in an alternative text. For this reason, a visually impaired person would not know that it was an apple being referred to. The same can be said for graphs, diagrams and other illustrations.



Pictures are ignored completely and are ‘invisible’ for screen readers!

Lists and tables

- Lists and tables do tend to contain text, and so they can usually be read out.
- Despite this, using the correct tabs would help remove any possible barriers, particularly for table column headings and keys.

	Calories	Fat	Proteins	Sugar
Banana	89	0.3g	1.1g	12g
Apple	52	0.2g	0.3g	10g
Clementine	47	0.2g	0.8g	9g

Data on various fruits per 100g

Example of an accessible **PAVE** PDF

Irrelevant text is recognized and ignored

0110100101101010
01100100101010
111100100

Headings are recognized

Example PDF without tagging

This document serves as an example of a document without tags or, equally, with tags that contain errors. All of the titles in the document, for example, were created simply by changing the font size and activating the bold font feature. To increase the chances of creating a document containing correct tags, these titles would have to be entered using the appropriate format template.

Text in multiple columns

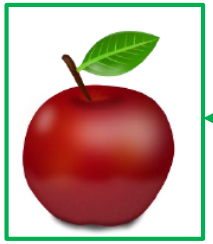
Text is read out loud columnwise

Text in multiple columns can lead to problems in some cases, because PDF reader programmes can only guess the order of the text segments

if there are no tags. In the simplest of cases, the text flow follows the order of the characters in the document, which is not the case for this text, for instance.

Decorative elements and pictures

The decorative zeros and ones in the top right corner of this document do not contribute to the content. They can, however, still be read out. In this scenario, it would be useful to tag the digits as artefacts. The picture to the right, in contrast, would be entirely ignored if it were not referred to in an alternative text. For this reason, a visually impaired person would not know that it was an apple being referred to. The same can be said for graphs, diagrams and other illustrations.



An alternative text is read out loud for figures

Lists and tables

- Lists and tables do tend to contain text, and so they can usually be read out.
- Despite this, using the correct tabs would help remove any possible barriers, particularly for table column headings and keys.

	Calories	Fat	Proteins	Sugar
Banana	89	0.3g	1.1g	12g
Apple	52	0.2g	0.3g	10g
Clementine	47	0.2g	0.8g	9g

Data on various fruits per 100g

Automatic Paragraph Detection

1. Filter out non-textual elements (based on element type or element height)
2. Clustering of single elements to text blocks
3. Determining reading order of the blocks using “Optimized XY-cut” [1]
4. Inside each block, classify each line to detect headings and paragraphs (based on left and right justification and element height)

[1] J. L. Meunier, "Optimized XY-cut for determining a page reading order," Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005, pp. 347-351 Vol. 1.

Results

- Basically very satisfying results
- Reduces the number of manual fixes significantly
- There is still need for human inspection
- Works best for structured papers such as technical papers or newspapers
- Improvement: also detect list and table structures

Are you interested in **PAVE**?

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

- Contact us during the break
- Directly write an E-Mail to the ICT-Accessibility Lab: alireza.darvishy@zhaw.ch