



**PDF Days  
Europe  
2025**

# Tagged PDF in the Wild

Evaluating quality of real world Tagged PDFs

Boris Doubrov | Director | Dual Lab



Dual Lab provides product development services in multiple domains:



### PDF technologies

Enabling the full power of PDF technology for your business.



### Process automation

Converting routine human tasks to algorithms.



### Scientific development

Converting modern Math to the working code.



### Digital transformation

Bringing your business processes in sync with modern technologies.

## HANCOM PDF Open Data Loader

SDK engine that **integrates with AI** to extract and structure data from PDFs

### Powerful Data Extraction

Accurately recognizes and extracts diverse data from PDF documents, including complex tables, text, and images.

### AI-Powered

Utilizes a combination of traditional rule-based methods and powerful AI models to overcome the limitations of existing models.

### Highly Flexible

Provides a flexible architecture that can be easily integrated with various AI models and libraries.

### Security

Process your documents with the complete security of local execution. Your data stays on your machine, always. Build powerful, private AI-driven document workflows with peace of mind.

# Objectives

---

- Main goal: Analyze tagged PDFs available on the web
  - Basic metadata
  - Structure tree stats
  - Correctness of parent-child relationship
  - Quality of (table) tagging
- Supplementary goals:
  - Setup reusable architecture for collecting stats on large collections of PDFs on demand
  - Evaluate AI models for checking appropriateness of tagging (on the example of tables)
- Data set:
  - **All** PDFs available in Common Crawl (~15M)

# Motivation

---

- Standards development:
  - better understanding of real world use cases when developing standards and best practice guides
- Build data collections for tests and AI training:
  - stress tests of existing tools (such as veraPDF)
  - new training data for AI-based layout recognition
- Returning to earlier work:
  - somewhat similar approach was used in SafeDocs project in analysing real world PDFs, but now with the focus on Tagged PDFs
- Technical challenges:
  - check how feasible is it to analyze ALL available PDFs on the Web

# All PDFs on the Web



Common Crawl maintains a free, open repository of web crawl data that can be used by anyone.

Latest crawl (CC-MAIN-2025-33) has

**14,138,664**

unique URLs of MIME type application/pdf

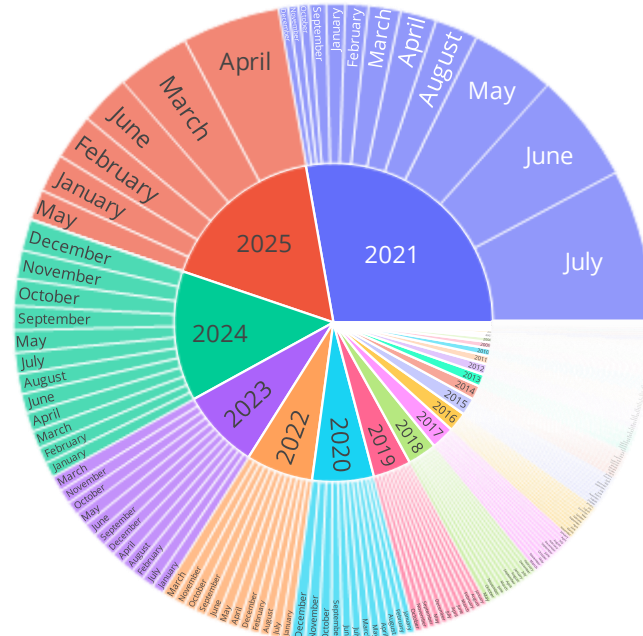
crawl	CC-MAIN-2025-26	CC-MAIN-2025-30	CC-MAIN-2025-33
mimetype	%	%	%
text/html	98.7237	98.7794	98.8693
application/pdf	0.6135	0.6394	0.5800
application/atom+xml	0.1283	0.1310	0.1185

# Reusing the SafeDocs corpus

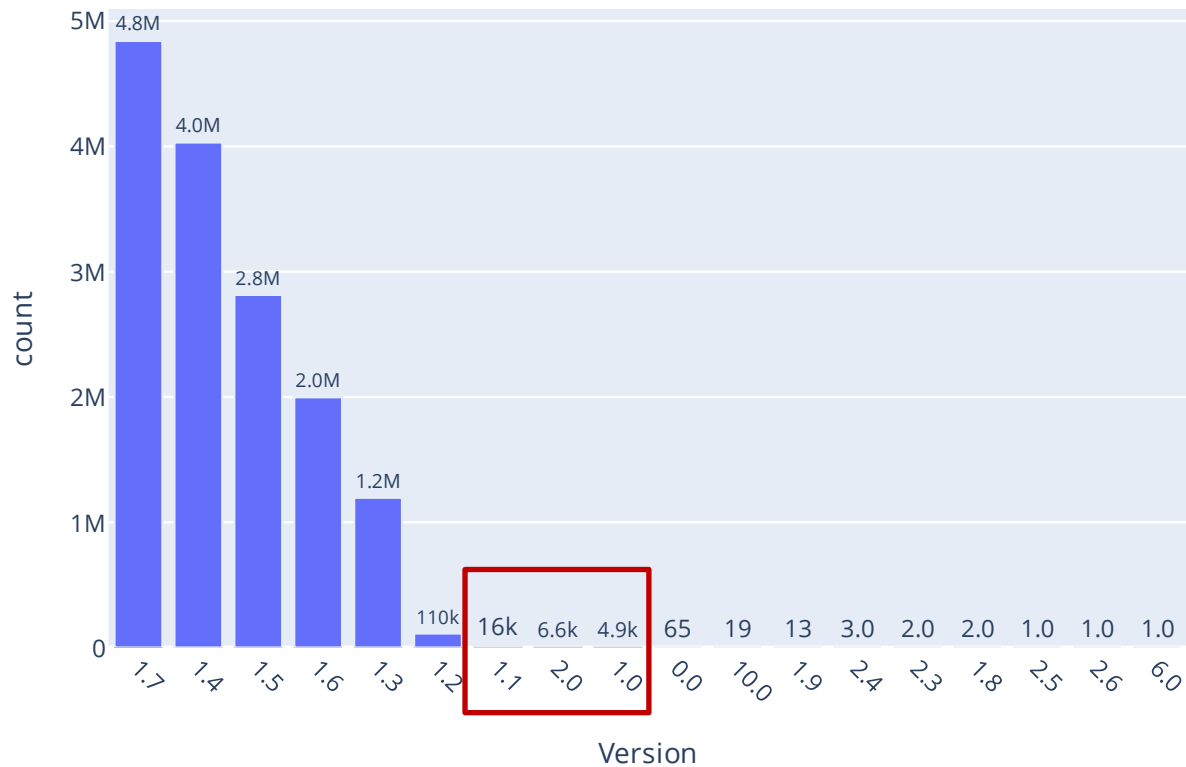
---

- The SafeDocs corpus:
  - CC-MAIN-2021-31-PDF-UNTRUNCATED:
  - contains ~8M PDFs totaling about 8 TB
  - Collected during July/August 2021 crawl
- Complement it with the data from the recent crawl
  - CC-MAIN-2025-21 (collected May 11-25, 2025)
  - To avoid duplication, only PDFs with ModificationDate >= September 2021
  - Total 7,211,883 additional PDFs totalling about 11.2Tb
  - 463,401 files (=3.5%) have no creation and modification date!!! (Skipped)
- In total selected for further analysis:
  - ~15.2M PDFs
  - Total size 19.2Tb

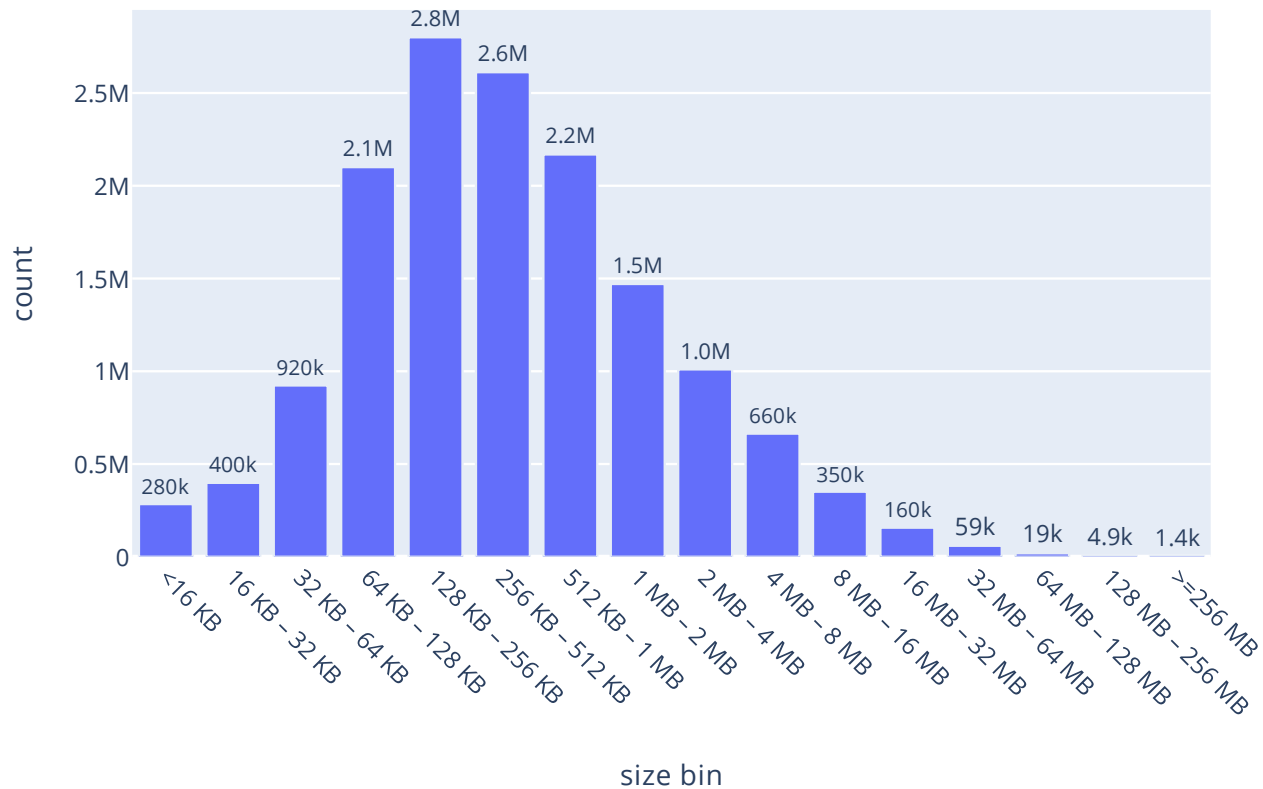
# PDF document modification dates by year/month



## PDF files by version (header+Catalog)



## Size of PDF files



# Maximal document size

Number of pages: 858

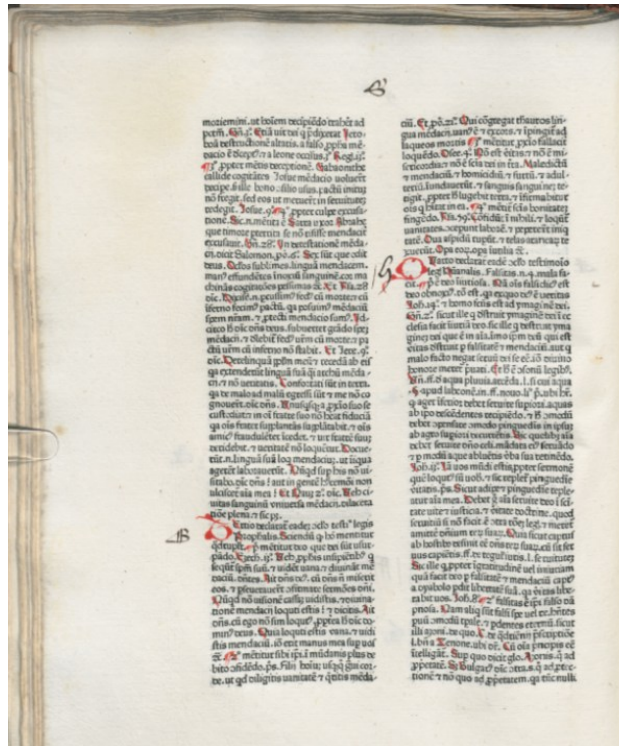
File size: 5.5Gb

Creator: N/A

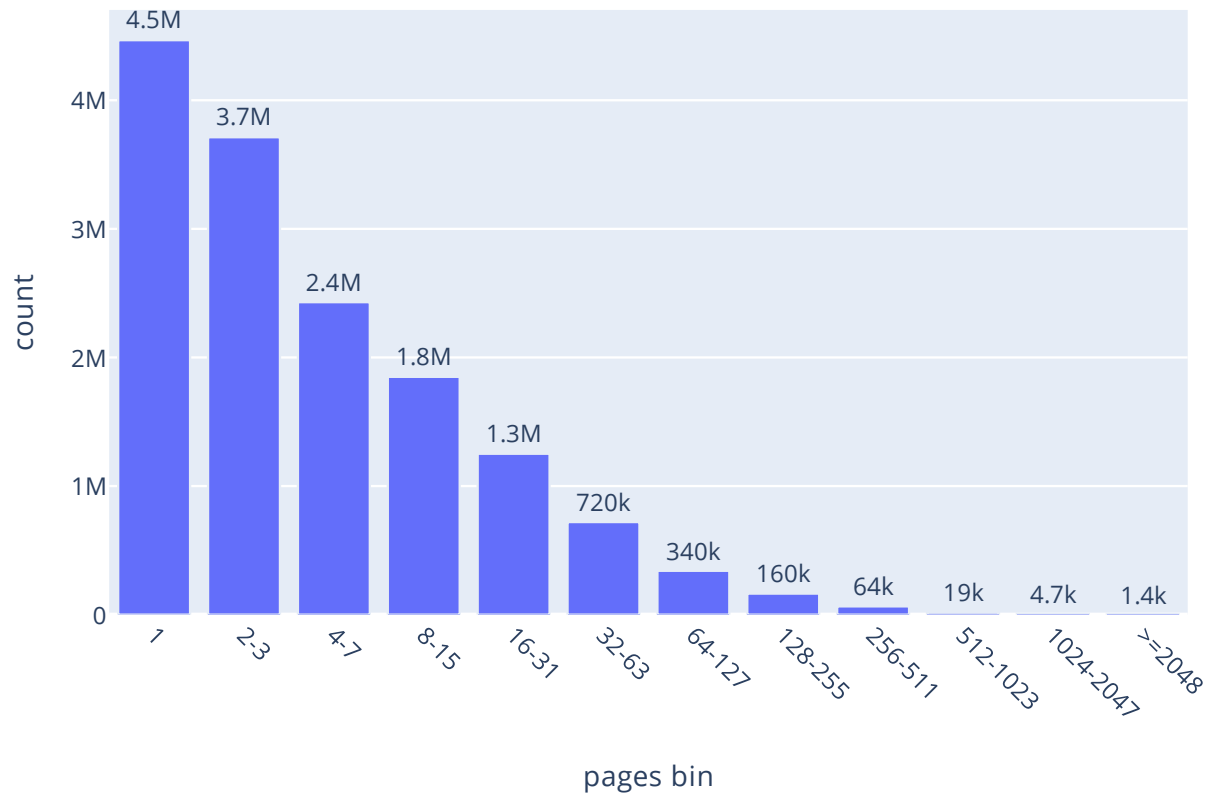
Producer: iText 5.5

Creation date: August 27, 2023

The size of 10 largest documents varies from 5.5Gb to 2.3Gb, all created by iText 5.5



## Number of pages in PDFs



# Maximal number of pages

---

Number of pages:	38,988
File size:	24Mb
Creator:	Microsoft Excel 2019
Creation date:	August 9, 2024
Tagged:	YES
Complies to ISO 32005:	YES

# Scanned PDFs

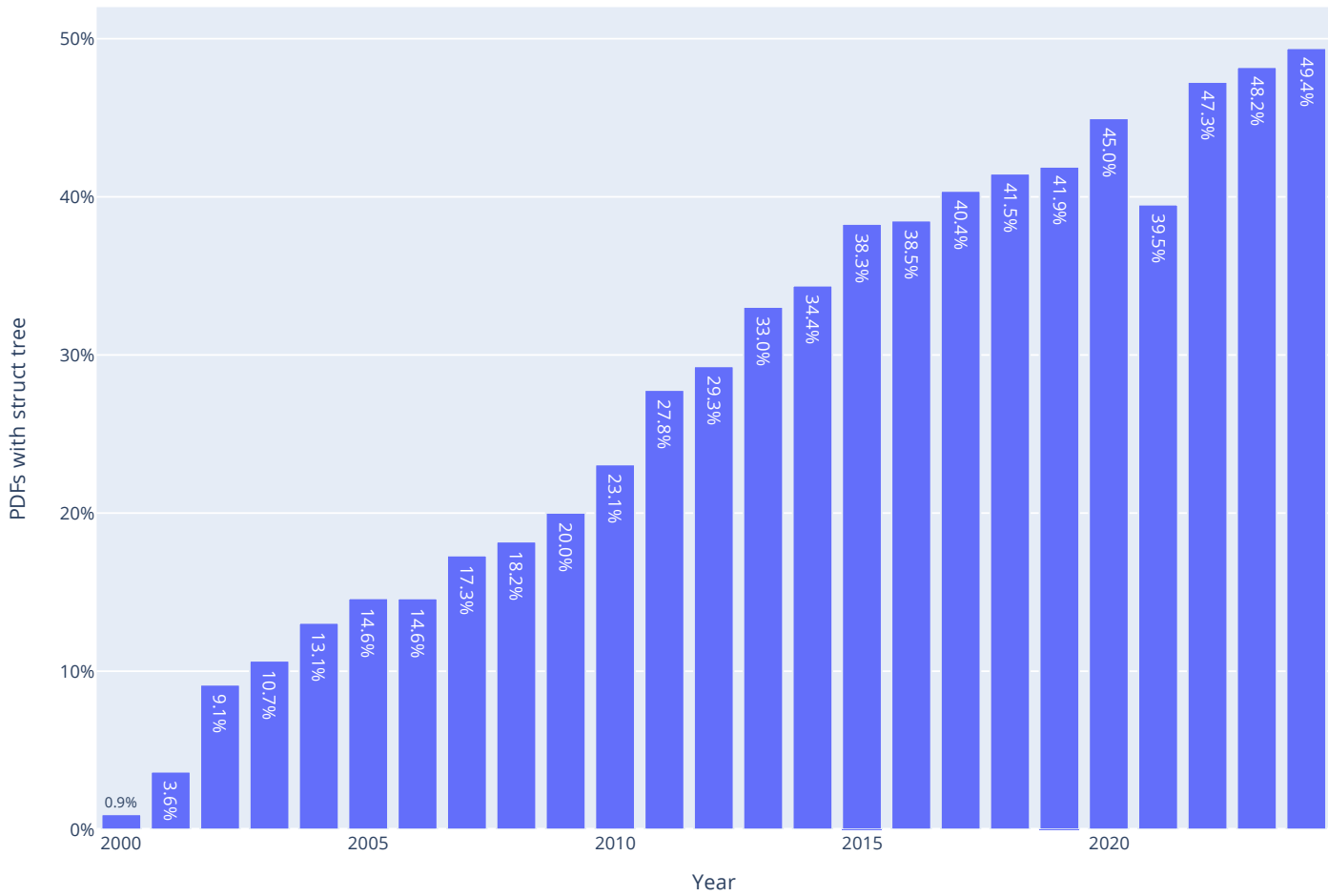
---

Document is *classified as scanned*, if it has *only images in page content*

Number of scanned PDFs:

1.6M, or ~12%

13. You will be notified by the corporation, in writing, after the facilities mentioned above are made available and are ready for commissioning the dealership. Immediately on receipt of the above notice from the corporation, you shall obtain each and every license necessary for operating your dealership as may be required under any central / state govt. / municipal or local authorities for the time being in force.
14. If we find that the progress made by you towards the above is not to our satisfaction, this offer is liable to be withdrawn.
15. Please note that you are required to fulfill the conditions with regard to inducting **Spouse as Co-owner** in the dealership **before issuance of Letter of Appointment**.
16. This letter of intent will stand automatically withdrawn and cancelled on the happening of any of the following events:-
  - a) If it is found that you have suppressed and / or misrepresented any material facts in your application.
  - b) In case you are found to be convicted for any criminal / economic offence involving moral turpitude.
  - c) In the event of death if you are an individual/partner.

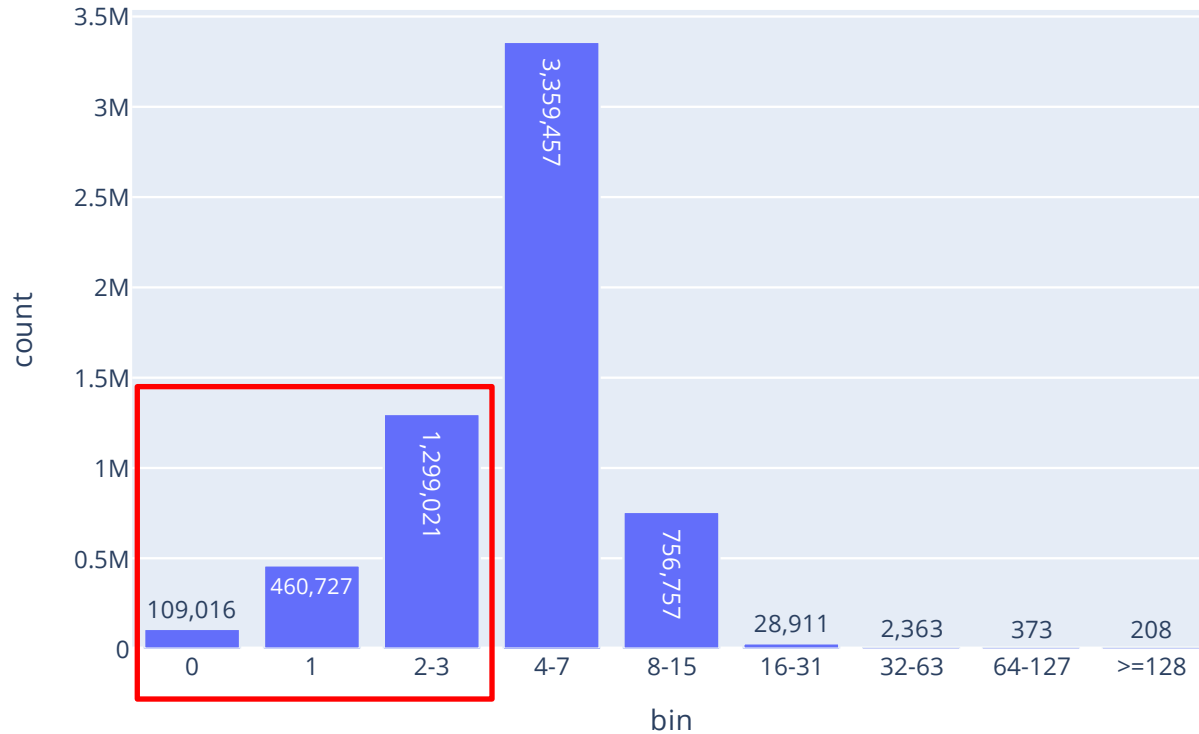


# Document with the largest number of structure elements

---

Number of structure elements:	11,724,216
Number of pages:	495
Producer:	ABBYY PDF Transformer 3.0
File size:	19Mb

## Struct Tree depth



# Document with maximal depth of structure tree

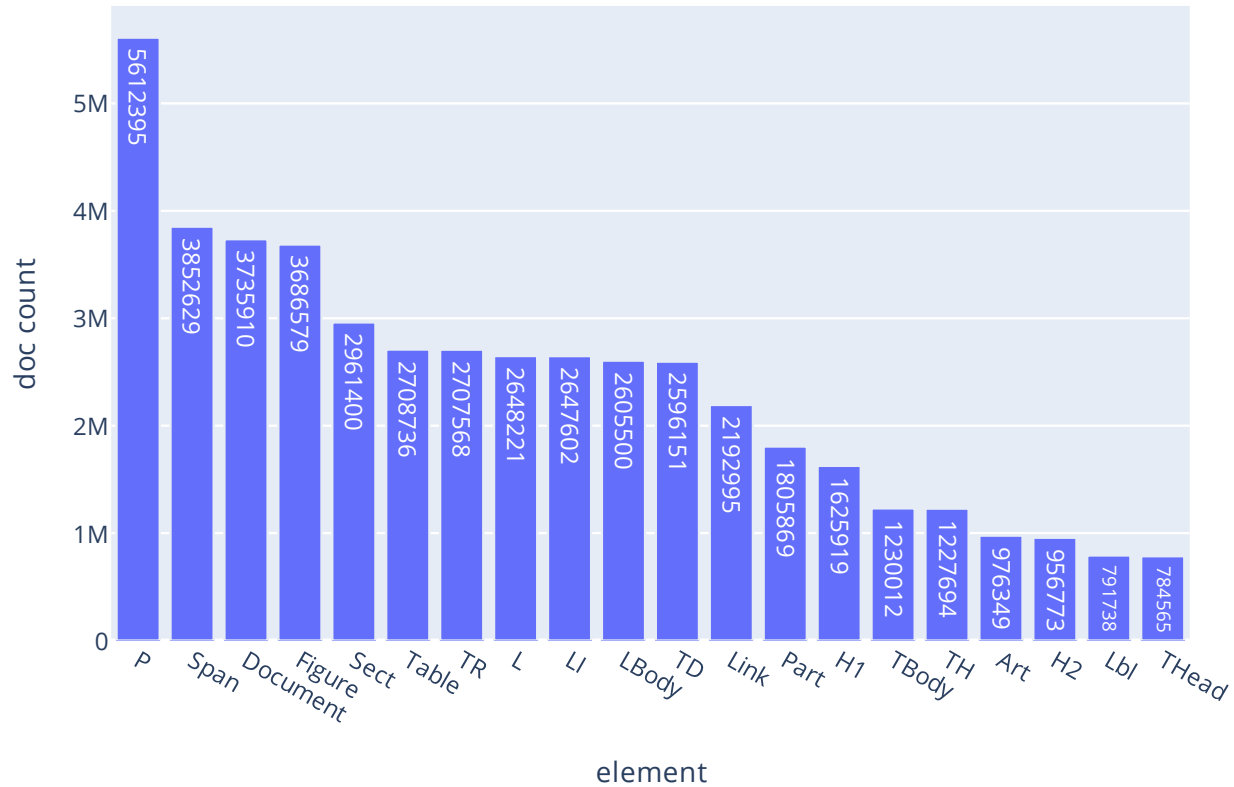
---

Depth of the structure tree:	490
File size:	284Kb
Producer:	Quartz Context
Number of pages:	2

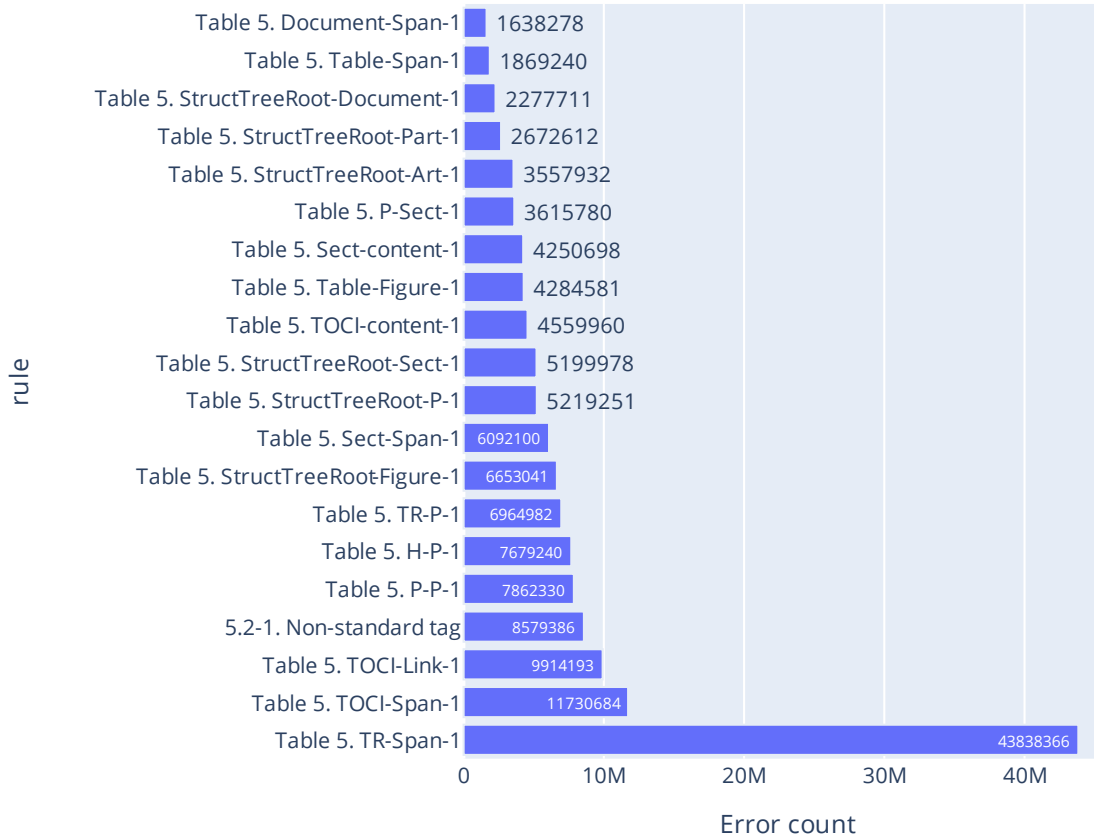
# Total count of structure elements

Element	Count	Element	Count
P	2,115,167,345	Code	48,980
TD	1,386,492,189	Index	34,670
Span	1,209,782,240	Ruby	33,256
TR	300,143,486	Private	18,868
LI	135,898,505	BibEntry	6,878
LBody	133,458,391	Warichu	132
Sect	119,029,266	Em, Strong, RP, WP	0
Figure	118,211,982	DocumentFragment, FENote, Title, Aside, Sub, Artifact (PDF 2.0)	0

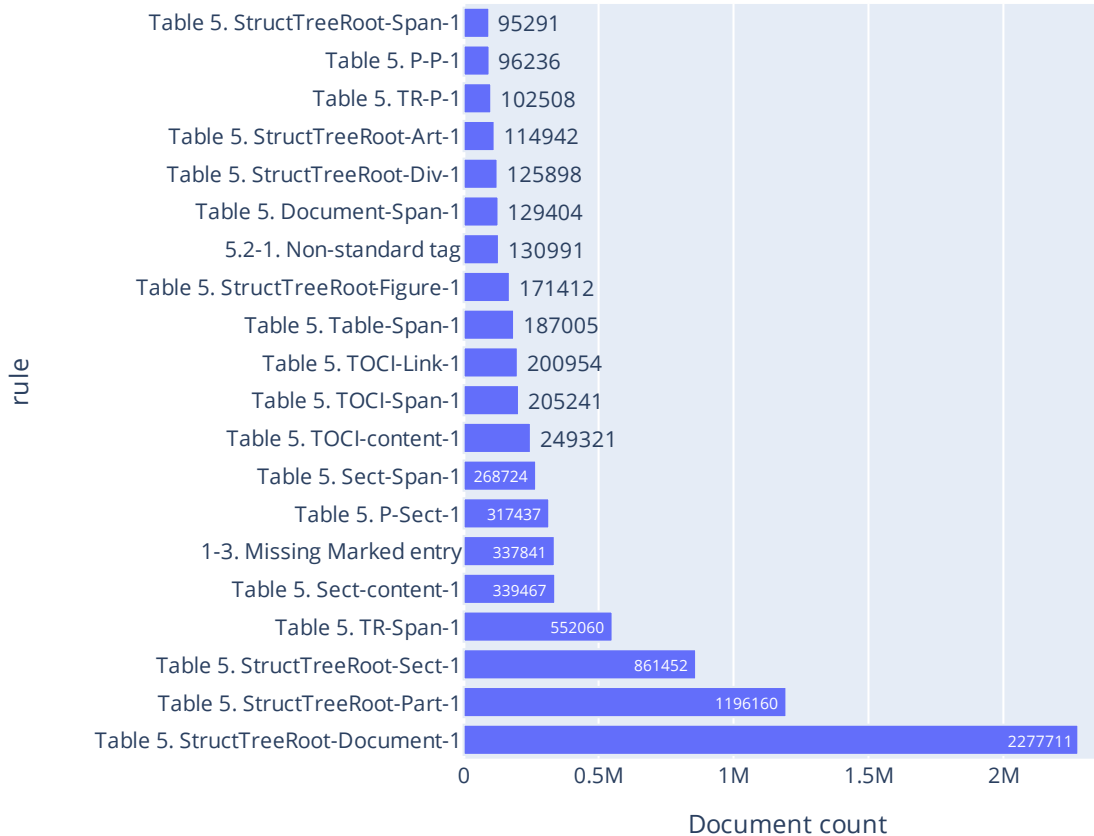
## Docs by structure element



## Top-20 failed ISO 32005 rules, validated by veraPDF



## Top-20 failed ISO 32005 rules by documents, validated by veraPDF



# Validation against ISO 32005

---

**44%** of all Tagged PDFs are fully compliant with ISO 32005

Most common errors:

- **Span** inside **TR** (by large!)
- **Span / Link / content** inside **TOCI** (shall we permit them?)
- **P** inside **P / H / TR**
- **P / Figure / Sect / Art / Part / Span** inside **Document** (**Part, Art, Sect** are OK in PDF 1.7)
- **StructTreeRoot** has no **Document** or has more than one **Document** (OK in PDF 1.7)
- Non-standard structure elements (not role mapped to standard ones)

What if we ignore errors specific to PDF 2.0 (as above):

**63%** of Tagged PDF 1.7 complying to ISO 32005 inclusion rules

# Dataset for quality of table tagging

---

- Tagged PDF
- Potential positive cases:
  - Has Table tags
  - These Tables comply to ISO 32005
  - Tables are regular
  - Bounding boxes of cells go top to bottom (by rows) and left to right (by columns)
- Potential negative cases:
  - Has no Table tags
  - But the AI model detects tables
- AI model used for table detection:
  - Hancom PDF Open Data Loader
- In total 440K documents selected for further “Human/AI” checks

# Emulating Human checks with AI

- Not feasible to manually check 440K PDF documents (within reasonable time, reasonable number of people and no riot)
- Use Docling (IBM Research) to emulate human behavior
- Strategy:
  - Detect tables with AI
  - True positives (tp): table is present in the structure tree and is detected by Docling
  - False positives (fp): table is present in the structure tree, but not detected by Docling
  - False negatives (fn): table is detected by Docling, but not present in the structure tree
- Result:

True positives	False positives	False negatives	Precision = $tp / (tp+fp)$	Recall = $tp / (tp+fn)$
835,586	632,074	397,907	56.9%	67.7%

# But can we trust AI?

---

- Finally, use human (manual) checks to verify if AI is correct
- Select random 500 tables and check if AI is correct
- According to a Human:

AI is correct	AI is incorrect	Undefined
77%	13%	10%

- Typical undefined cases:
  - (almost) empty tables with borders
  - Table of contents: is it a table?
  - Use of tables for a text frame
  - Tables vs lists disambiguation

Development Management  
Milton Keynes Council  
Civic Offices  
1 Saxon Gate East  
Milton Keynes  
Buckinghamshire  
MK9 3EJ

**Our ref:** AC/2020/129112/02-L01  
**Your ref:** 20/00133/OUTEIS  
**Date:** 2 September 2020

**FAO: David Buckley**

Dear Sir/Madam

**OUTLINE PLANNING APPLICATION (ALL MATTERS RESERVED EXCEPT ACCESS) FOR THE DEMOLITION OF THE EXISTING FARM BUILDINGS ON SITE AND THE DEVELOPMENT OF UP TO 930 DWELLINGS (INCLUDING AFFORDABLE DWELLINGS), PRIMARY SCHOOL, LOCAL CENTRE, OPEN SPACE, SPORTS PITCHES, PLAY AREAS, PAVILION/WELLBEING CENTRE AND OTHER ASSOCIATED WORKS.  
TICKFORD FIELDS, FARM NORTH CRAWLEY ROAD, NEWPORT PAGNELL,  
MK16 9HG**

Legend: blue box - Table structure element, red box - Table detected by AI

Mr. Ric Goss, Planning Director, stated that this was reviewed by all of the departments in the city and no objections were raised. He explained that there was no proposed use or current use for the property. He noted that all of the utility companies were contacted. He stated that staff recommended that the notice of intent to vacate resolution be approved.

**Commissioner Boehm moved, seconded by Commissioner Stowers, to approve Resolutions No. 2016-99, as read by title only.**

Call Vote:	Commissioner Stowers	Yes
	Commissioner Kent	Yes
	Commissioner Boehm	Yes
	Commissioner Partington	Yes
Carried.	Mayor Kelley	Yes

Item 9D – Pineland Preliminary Plat

City Clerk Scott McKee read by title only:

Legend: blue box - Table structure element, red box - Table detected by AI

Läs informationen "[Dygnet-runt-omsorg – underlag för behovsprövning](#)" innan du/ni fyller i svaren. Om utrymmet under frågorna inte räcker, gör en pil och fortsätt skriva på baksidan.

1. I vilken utsträckning har du/ni försökt påverka dina/era arbetstider?
2. Vilka egna lösningar har du/ni prövat?
3. Vilka speciella omständigheter i familjens situation vill du/ni åberopa? Det går bra att hänvisa till bilaga om du/ni inte vill ange de speciella omständigheterna direkt i blanketten.

Legend: blue box - Table structure element, red box - Table detected by AI

## Öffentlicher Teil

1	Eröffnung, Feststellung der ordnungsgemäßen Ladung und der Beschlussfähigkeit
2	Bestätigung der Tagesordnung
3	Bestätigung der Niederschrift vom 18.10.2017
4	Informationen zur Kultur in der Hansestadt Stralsund
5	Opernale - Rückblick auf die vergangenen sieben Jahre und Informationen zu einem neuen Projekt
6	Verständigung auf die weiteren Sitzungstermine in 2018
7	Anfragen
8	Mitteilungen

Legend: blue box - Table structure element, red box - Table detected by AI

# Summary

---

- Half of PDFs today are Tagged. And their share keeps increasing
- Many syntax issues are still here: more best practice guides, samples, validation tools needed?
- Tagged PDF 2.0 is not yet here
- AI (on the example of Docling model) is progressing fast and is already very useful as a Human assistant (but not yet a replacement!): table recognition is incorrect in ~10% of cases on real world PDFs (another ~10% is undecided)
- Future plans: repeat such evaluation of (Tagged) PDFs periodically (once a year, every new Common Crawl?)

# Interested in more data?

---

Submit your queries at:

