



PDFs as an Aggregate

What can thousands of PDFs tell us?

OctoberPDFest
ONLINE

Patrick Gallot

Senior Solutions Architect

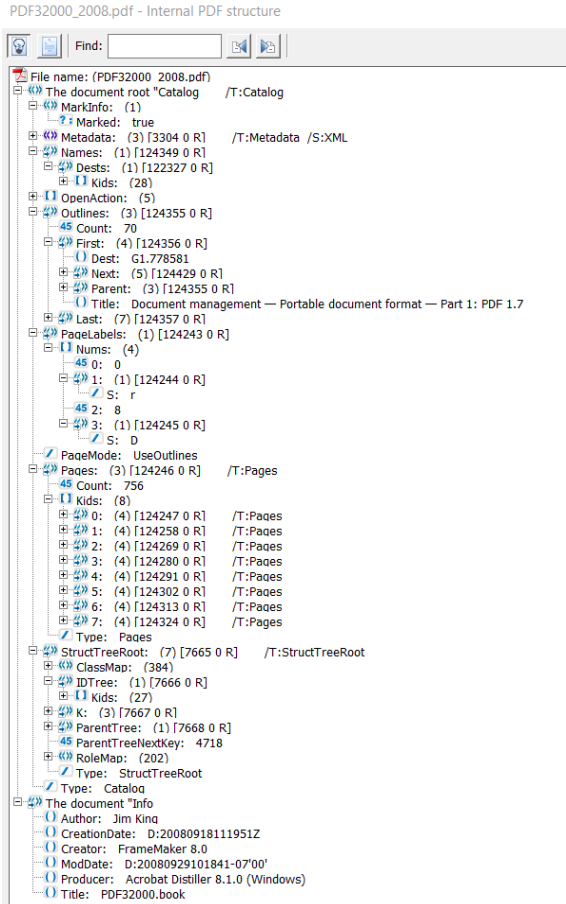


Datalogics

PDF is not 1 format

- **It is the format of untold thousands of different applications**
 - Each of which creates PDFs with their own quirks
- **Each PDF is constructed from a unique set of resources**
 - Similarly constructed PDFs may not be identical in appearance.
 - PDFs that are identical in appearance are not necessarily constructed identically.

How to think about the internal structure of PDFs?



PDF 32000-1:2008

Table 28 – Entries in the catalog dictionary

Key	Type	Value
-----	------	-------

Table 30 – Entries in a page object

Key	Type	Value
-----	------	-------

Table 33 – Entries in a resource dictionary

Key	Type	Value
-----	------	-------

Table 89 – Additional Entries Specific to an Image Dictionary

Key	Type	Value
-----	------	-------

Table 95 – Additional Entries Specific to a Type 1 Form Dictionary

Key	Type	Value
-----	------	-------



Two Collections of PDFs

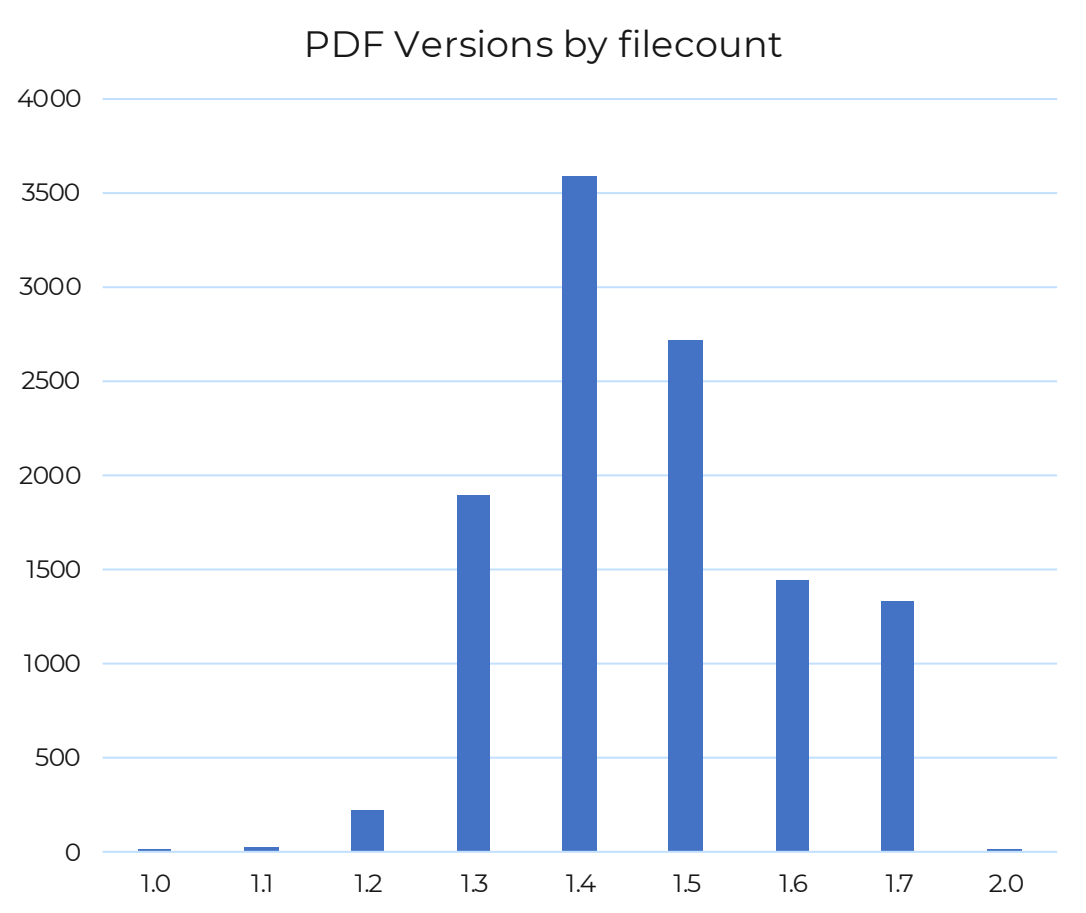
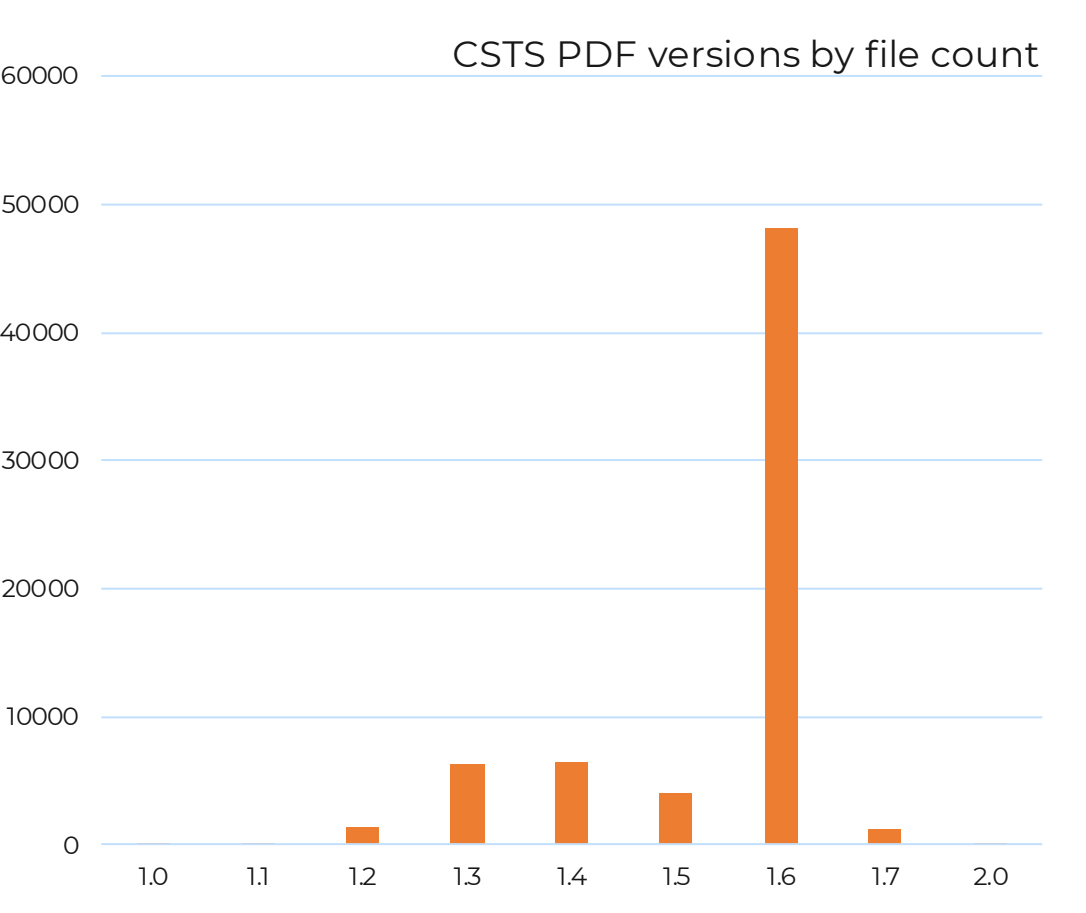
number of files	SELECT COUNT(*) as file_tally FROM FILEINFO
Total Filesize	SELECT SUM(FILESIZE) FROM FILEINFO
average filesize	SELECT AVG(FILESIZE) FROM FILEINFO
Distinct Producers	SELECT COUNT(*) as producer_tally from (select DISTINCT PRODUCER FROM DOCINFO)
date of most recent file	SELECT CREATIONDATE from DOCINFO ORDER BY CREATIONDate Desc limit 25
date of least recently modified	SELECT MODDate,CREATIONDATE From DOCINFO where MODDate !=" ORDER by MODDate DESC limit 25
oldest creation date	SELECT DISTINCT CREATIONDATE from DOCINFO ORDER BY CREATIONDate ASC limit 25
unique IDs	SELECT COUNT(Distinct INST_ID) FROM FILEINFO
null UIIDs	SELECT Count(*) FROM FILEINFO where INST_ID is " or INST_ID is null
unencrypted PDFs	SELECT COUNT(*) as unencrypted_tally from FILEINFO WHERE SECURED=0
Unencrypted, No Errors detected:	SELECT COUNT(*) as errorfree_tally from FILEINFO where SECURED=0 AND rowid not in (select distinct DOCID from ERRORS)
files with errors	SELECT COUNT(*) as badfile_count from (select DISTINCT DOCID FROM ERRORS)
Exceptions recorded:	SELECT COUNT(*) as errorcount FROM ERRORS
PDF/A files	SELECT COUNT(*) as pdfa_tally from XMPIDs WHERE pdfaid_part is not null
PDF/UA files	SELECT COUNT(*) as pdfua_tally from XMPIDs WHERE pdfuaid_part is not null
PDF/X files	SELECT COUNT(*) as pdfx_tally from XMPIDs WHERE pdfxid_GTS_PDFXVersion is not null
AcroForm PDFs	SELECT COUNT(*) from FILEINFO where rowid in (select distinct DOCID from ACROFORM WHERE Fields is not null GROUP by DOCID)
potentially signed	SELECT COUNT(*) as sigflags_tally from FILEINFO,ACROFORM where FILEINFO.rowid=docid and SigFlags>0
tagged PDFs	SELECT COUNT(*) as tagged_count from CATALOG where StructTreeRoot is not null and MarkInfo is not null
PDFs with NameTrees	SELECT COUNT(*) as nametree_tally from NAMETREE

number of files	67,472	
		59.85G
Total Filesize	64,265,618,607	B
average filesize	952,478.34	0.91MB
Distinct Producers	2,099	
date of most recent file	3/9/2020	
date of least recently modified	10/5/1994	
oldest creation date	2/18/1994	
unique IDs	24,511	36.33%
null UIIDs	1,397	2.07%
unencrypted PDFs	66,816	99.03%
Unencrypted, No Errors detected:	63,568	94.21%
files with errors	3,275	4.85%
Exceptions recorded:	117,509	
PDF/A files	533	0.79%
PDF/UA files	45	0.07%
PDF/X files	902	1.34%
AcroForm PDFs	5,232	7.75%
potentially signed	1,716	2.54%

number of files	11,246	
		20.88G
Total Filesize	22,414,557,390	B
average filesize	1,993,113.76	1.90MB
Distinct Producers	2,322	
date of most recent file	10/7/2019	
date of least recently modified	2/19/1998	
oldest creation date	2/15/1996	
unique IDs	7,850	69.80%
null UIIDs	2,433	21.63%
unencrypted PDFs	10,940	97.28%
Unencrypted, No Errors detected:	10,349	92.02%
files with errors	610	5.42%
Exceptions recorded:	79,029	
PDF/A files	366	3.25%
PDF/UA files	27	0.24%
PDF/X files	144	1.28%
AcroForm PDFs	1,243	11.05%

- SELECT
PDFVersion,count(*) FROM
FILEINFO GROUP BY
PDFVersion ORDER by
PDFVersion

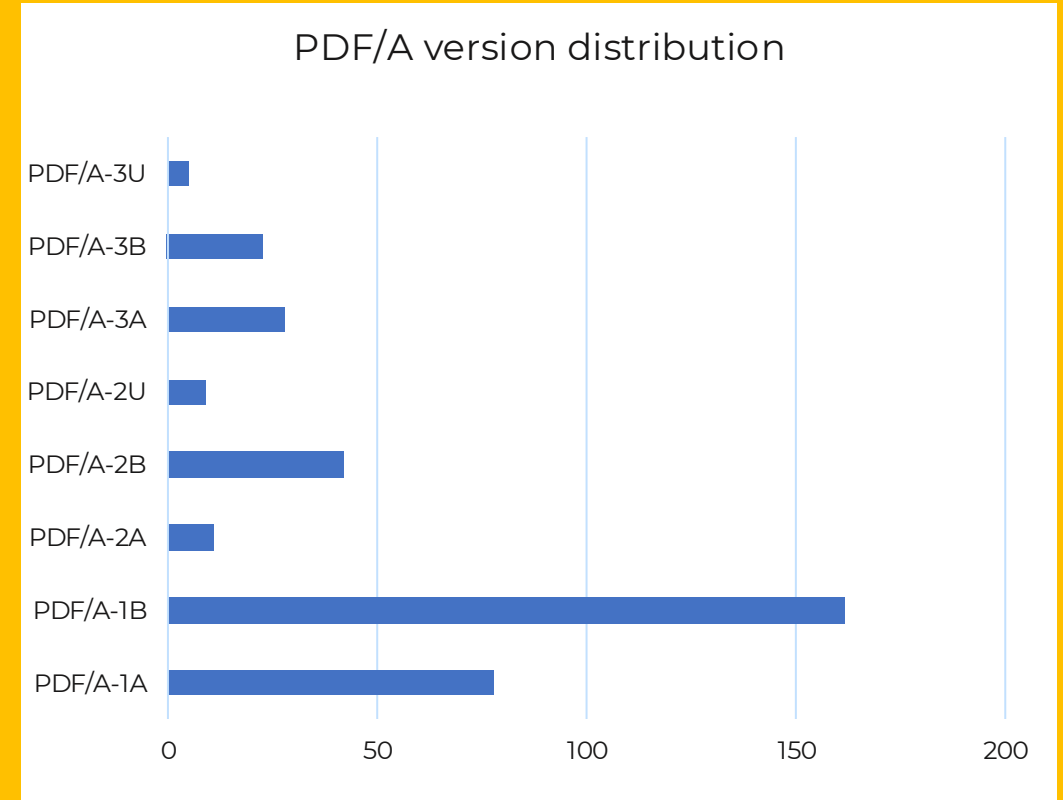
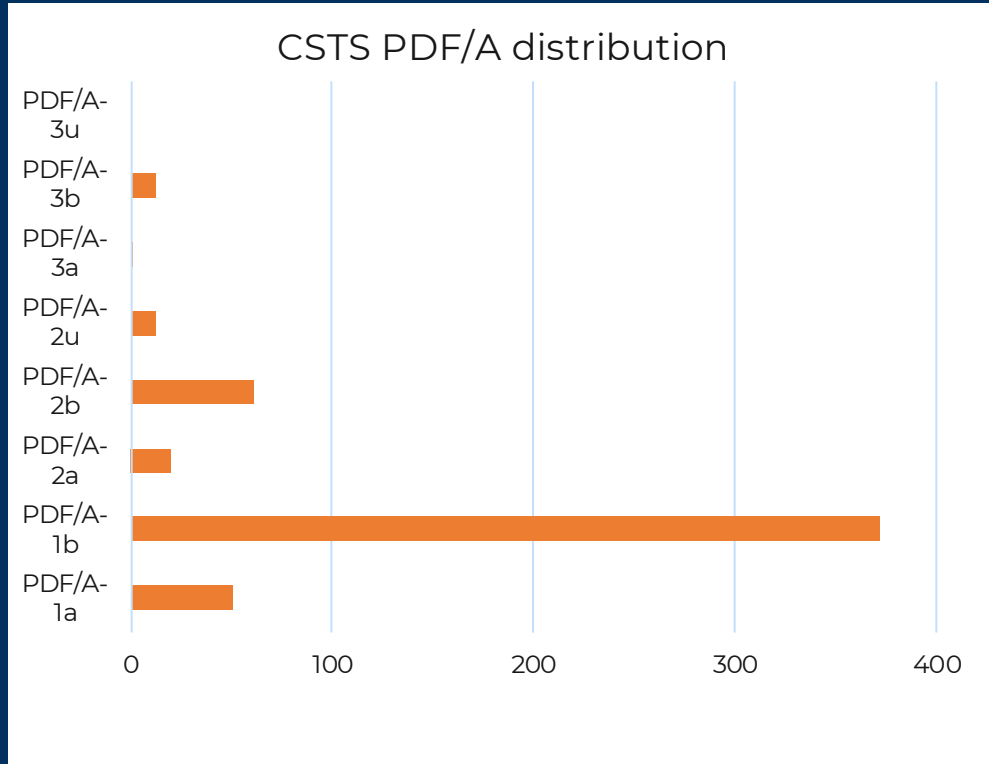
PDF Versions



- SELECT pdfaid_part,
pdfaid_conformance,
count(*) as tally from
XMPIDs where pdfaid_part
is not null group by
pdfaid_part,
pdfaid_conformance

PDF/A versions

PDF/A Versions



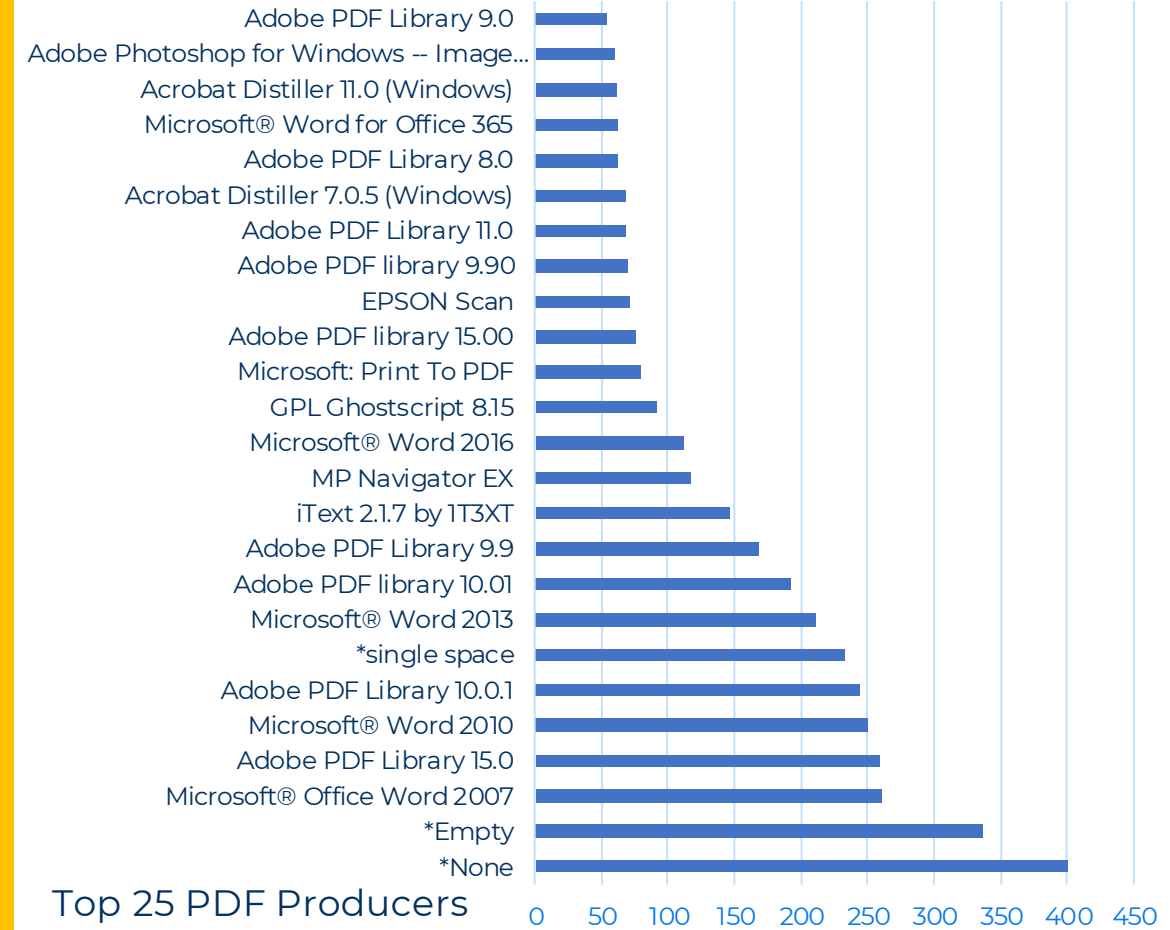
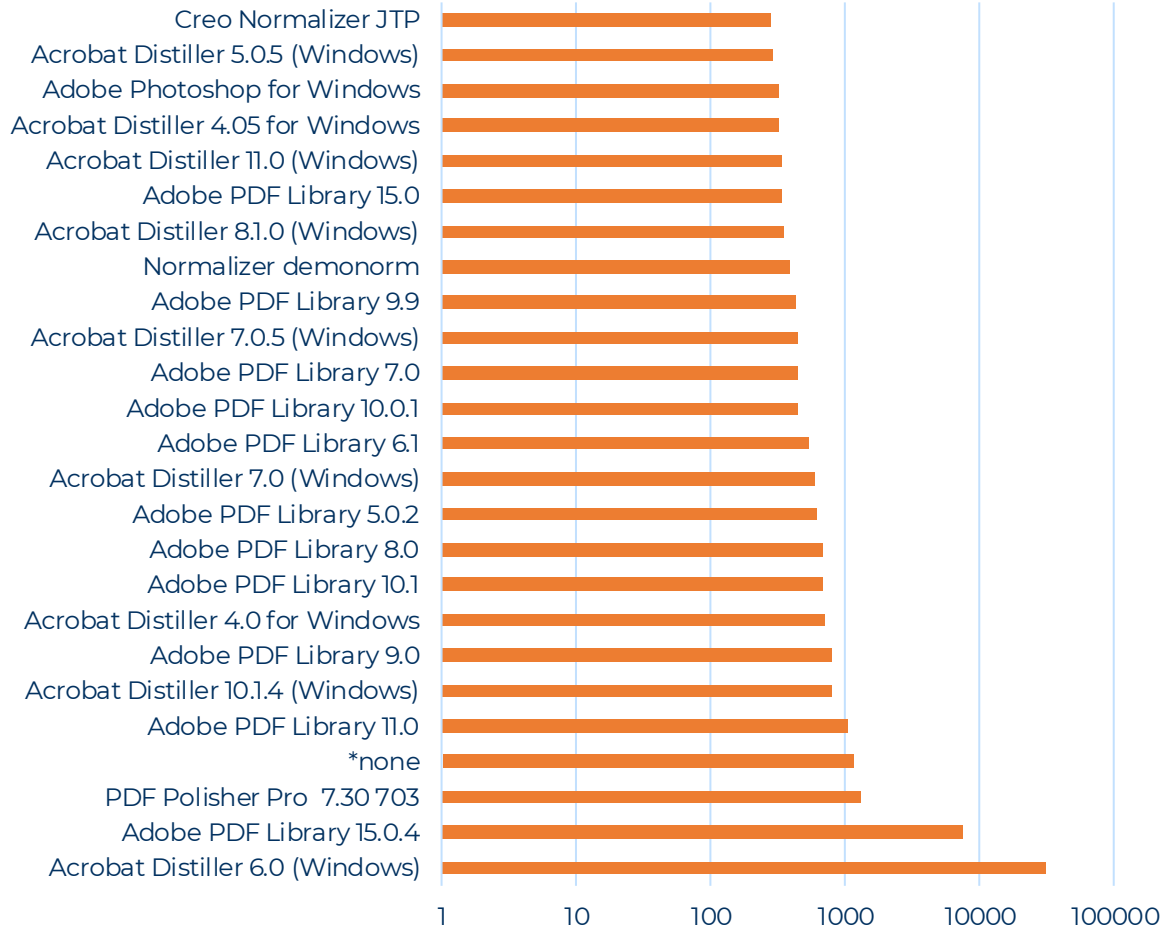


PDF Producers



**SELECT
PRODUCER,
count(*) as tally
from DOCINFO
GROUP by
PRODUCER
ORDER by tally
DESC LIMIT 25**

CSTS Top 25 PDF Producers, logarithmic scale



Top 25 PDF Producers

More Methodology: Detecting errors in PDFs

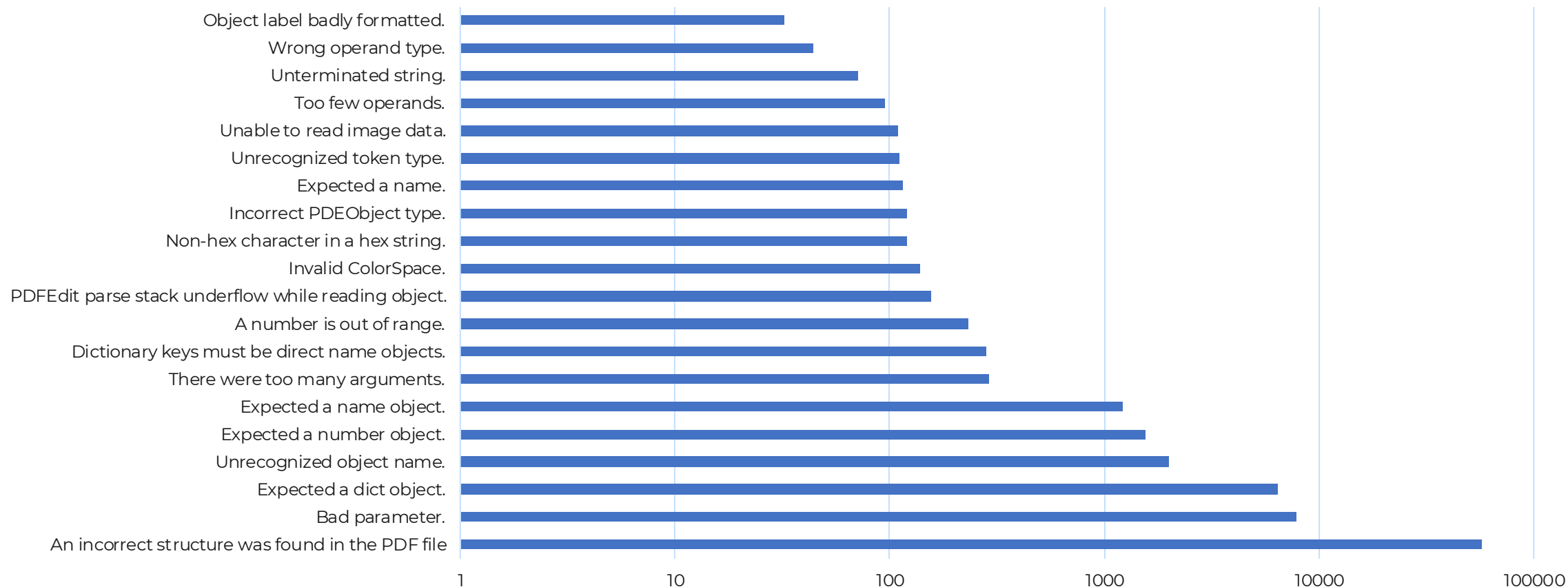
- **Why detecting errors is important and useful?**
 - Addressing errors helps other PDF consumers down the line.
 - Reduces frustration and time loss trying to fix PDF documents.
- **APDFL has several APIs for reading information from a PDF...**
 - Call them all, log the exceptions in the database, along with the function that triggered the exception, and the objectID of the function parameter in the database.

Most Common Errors

- `SELECT ERRORS.ERRORCODE, MESSAGE, count(*) as tally from ERRORS, ERRORCODES where ERRORS.ERRORCODE=ERRORCODES.ERRORCODE GROUP BY ERRORS.errorcode ORDER BY tally, errors.errorcode`

Most common errors

Most Common errors by tally

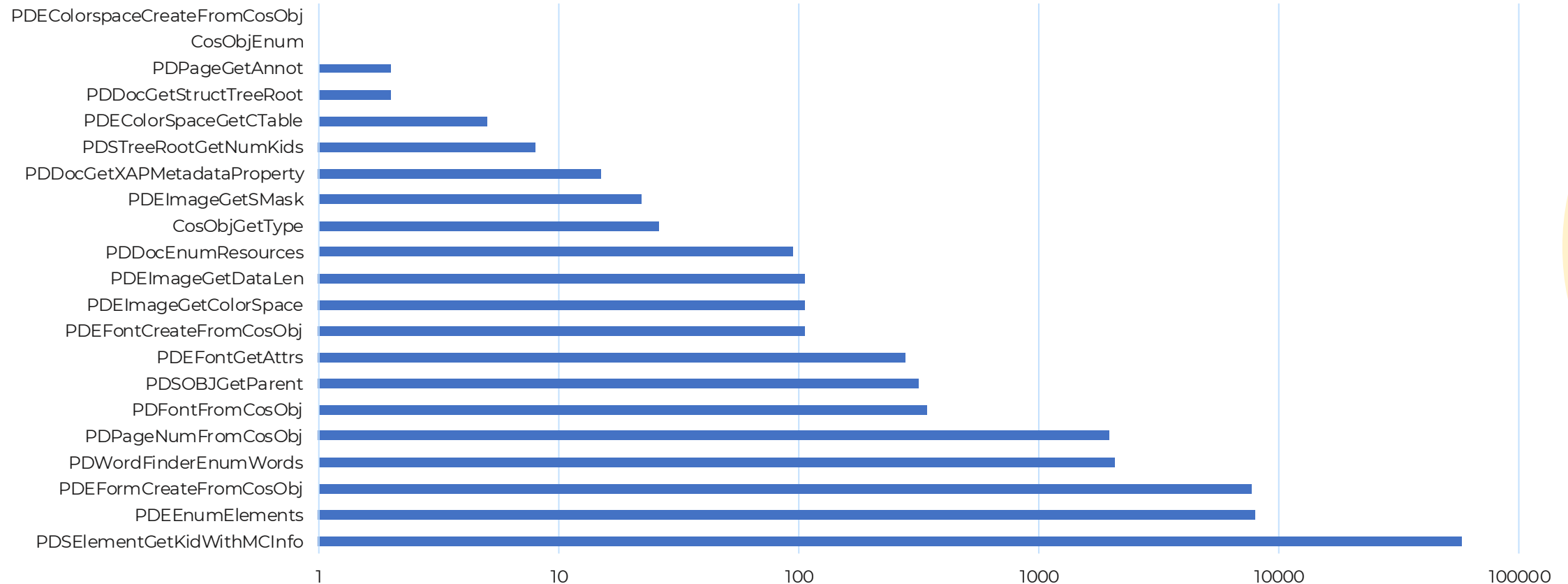


Errors by Function

- `SELECT FUNCNAME,COUNT(*) from
ERRORS GROUP BY FUNCNAME
order by funcname`

Errors by function

COUNT of Errors by Function



Looking for Objects that trigger more than one exception.

- `SELECT ERRORS.ERRORCODE, count() as tally, message, docid, objectid, funcname from ERRORS,ERRORCODES where ERRORS.ERRORCODE=ERRORCODES.ERRORCODE GROUP BY docid, objectid, ERRORS.ERRORCODE, funcname having tally > 1 order by tally, docid desc`

Malformed objects can trigger different errors in different API calls.

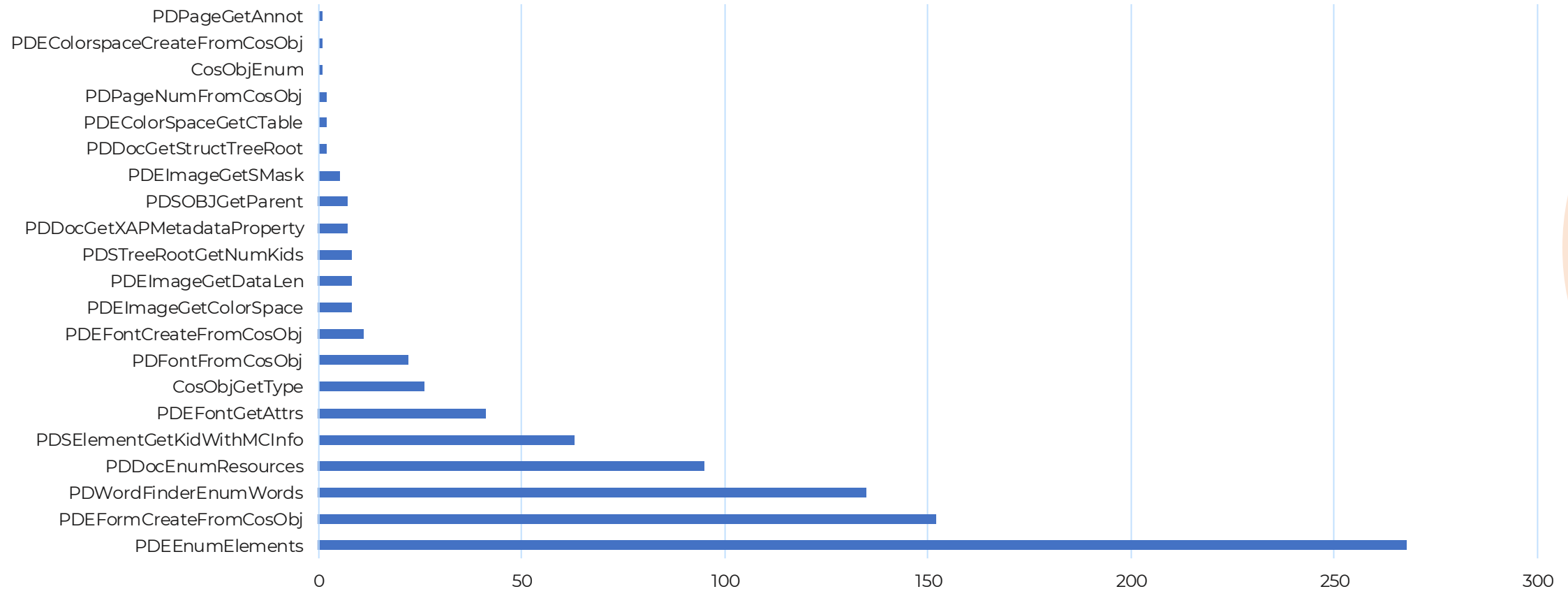
- PDEEnumElements: there were too many arguments.
- PDEEnumElements: Non-hex character in a hex string.
- PDEEnumElements: Unterminated string.
- PDEEnumElements: Expected a name.
- PDEFormCreateFromCosObj: Bad parameter.
- PDEFormCreateFromCosObj: Bad parameter.
- PDEFormCreateFromCosObj: Bad parameter.
- PDEFormCreateFromCosObj: Bad parameter.

Document count of Functions triggering Errors

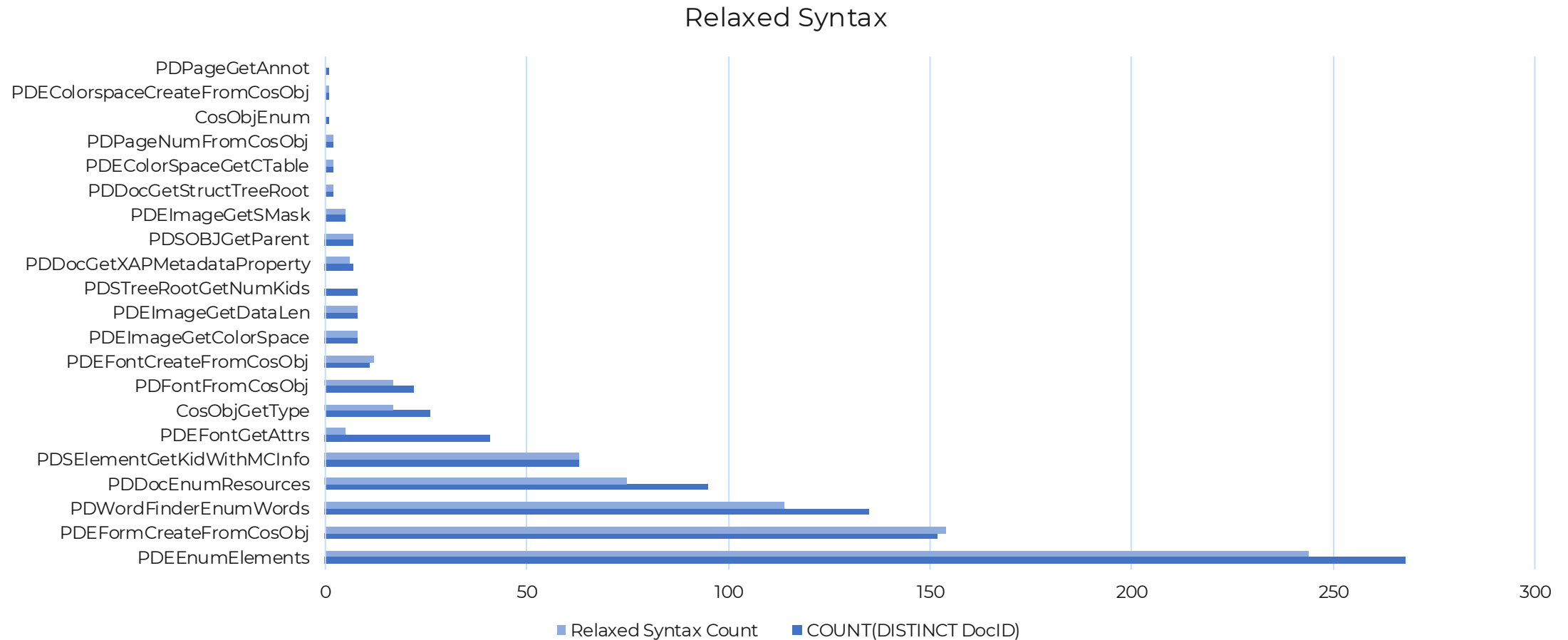
- `SELECT FUNCNAME,
COUNT(DISTINCT DocID) from
ERRORS GROUP BY FUNCNAME
order by funcname`

Document count of Functions triggering Errors

Error Functions COUNT(DISTINCT DocID)



Errors of allowing relaxed syntax on exceptions, by document



Trends & Final Thoughts

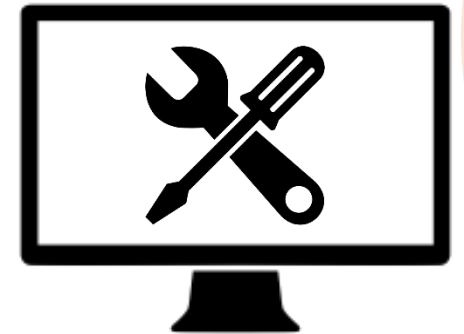
- **Document error rates might trend downward...**
 - But defect-free PDFs as an aggregate will probably always be out of reach.
 - But reducing errors in documents helps extends their useful life.
- Tagged PDFs are growing in importance.
 - But PDF/UA so far seems to have a much slower adoption rate.
- PDF/A is also growing in Importance.
- PDF 2.0? Not so much in this dataset.

How can you streamline and automate this?



Ensure that your input files are **checked** and **fixed** before consumption

Leverage **scriptable server tools** in your workflows that can automate processes



Scan your PDFs for Errors

1. Analyze your PDFs

- There can be many issues lurking within your PDF files
- Could impact the ability and reliability of tools and software to view or process PDF files
- Can cause a negative impact on end users interacting with your PDFs
- Use a tool that can be automated to analyze multiple PDF files.

Assess what can be fixed

2. Assess problems found from your analysis

- Gain insight about how your PDFs are structured
- Decide on what to do next with your PDF files
 - PDF can be further optimized
 - Contains metadata
 - Contains transparencies
 - Uses fonts not embedded

Scanning your PDFs with PDF Checker

Checks that can be completed:

- Optimization assessment
- Check for conformance to: PDF/A, PDF/E, PDF/UA, PDF/VT, PDF/X
- Metadata
- Font analysis
- JavaScript
- Color and image resolution



Resolve your PDF Errors with PDF Checker + PDF Optimizer

3. Fix PDF errors with PDF Optimizer

- **PDF Optimizer** can fix many issues flagged by **PDF Checker**
- **PDF Optimizer** is a powerful tool that provides fast processing speeds and offers user-specific optimization profiles
- Customizable PDF capability needs such as:
 - Optimal PDF size reduction
 - Optimize PDFs for viewing and printing
 - Archive PDFs for long term preservation

Resolve your PDF Errors with PDF Checker + PDF Optimizer

- **Our PDF Checker tool now comes with a FREE 14 day trial of PDF Optimizer**
- **PDF Checker now offers JSON report output capabilities streamlining the process for programmatically driving PDF Optimizer processes based on conditions reported by PDF Checker enabling you to have more *power* over your PDFs.**
- **Go to www.Datalogics.com to get your FREE PDF Checker download!**

Who is Datalogics?



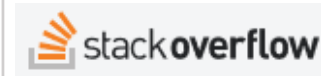
Established in
1967



Located in
Chicago



Providing Adobe
PDF toolkits to
developers since
1997



Active in the
Developer
Community



Member of the PDF
Association

Full Suite of Scriptable Server Tools & SDKs



PDF Checker

Identify common problems with your PDF files



FLIP2PDF

Easily convert PDF files to a variety of filetypes



PDF2PRINT

Automate your printing processes



PDF Forms Flattener

Flatten XFA & AcroForms to PDF



PDF Optimizer

Reduce file size, enable archival compliance



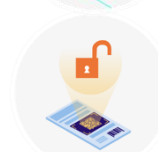
PDF2IMG

Convert PDF pages to a variety of image formats



PDF Alchemist

Easily convert PDFs to HTML, XML, and EPUB



Adobe Experience Reader Extensions

Reader extend PDF documents

Software development kits

Ideal for SaaS and OEM implementations



ADOBE PDF LIBRARY



PDF JAVA TOOLKIT

A graphic with a light gray background. In the center is a large dark blue circle containing the text "Q & A" in white. Four light blue diagonal lines cross the background, two on the left and two on the right. Four light blue circular shapes are partially visible at the corners: top-left, top-right, bottom-left, and bottom-right.

Q & A

Thank You!



Patrick Gallot

Senior Solutions Architect

datalogics.com | info@datalogics.com | +1.312.853.8200