



Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

## **Progress on Building a File Observatory for Secure Parser Development**

*PDFDays Europe 2022*

September 13, 2022

The research was carried out at the NASA (National Aeronautics and Space Administration) Jet Propulsion Laboratory, California Institute of Technology under a contract with the Defense Advanced Research Projects Agency (DARPA) SafeDocs program through an agreement with the National Aeronautics and Space Administration (80NM0018D0004). © 2022 California Institute of Technology. Government sponsorship acknowledged.



**Jet Propulsion Laboratory**  
California Institute of Technology

# The Team



Chris Mattmann  
PI; Division Mgr, AI  
AID Org



Tim Allison  
Files and Search



Wayne Burke  
Cognizant Engineer



Michael Fedell  
Data Scientist



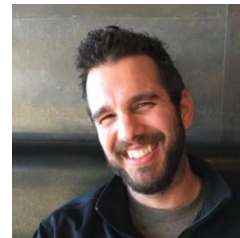
Dustin Graf  
Project Manager



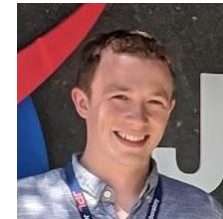
Anastasia Menshikova  
Data Scientist



Michael Milano  
Data Scientist  
UX Researcher



Phil Southam  
Trouble (Fun?)  
Maker



Ryan Stonebraker  
Data Scientist  
Alaskan

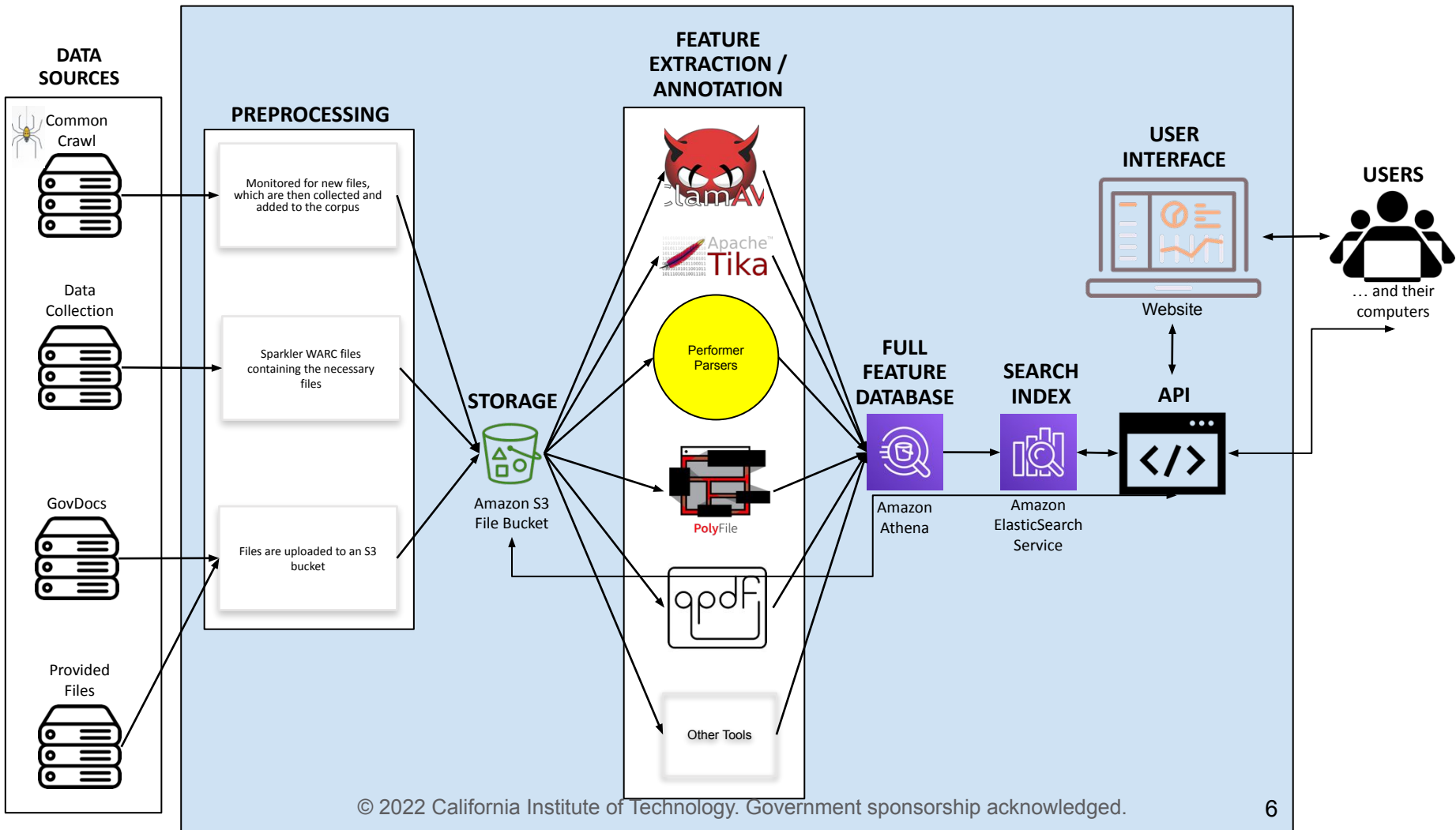
# Outline

- File Observatory
  - Backend updates – various scales, Tika updates, feature extraction
  - User Interface updates
- CC-MAIN-2021-31
- Next Steps

# File Observatory – Why?

- Demonstrate state of the possible
- Transfer techniques if not code
- Some uses: parser and specification development/improvement, market analysis, security analysis, corpus characterization/digital preservation assessments, machine learning for creator tool attribution and ???

# File Observatory







# Standard table for each tool (pdftinfo example)

## Query Editor

```
1 select path, exit_value, timeout, process_time_ms, stdout from pdfinfo
2 order by process_time_ms desc
3 limit 10;
4
```

also execute "EXPLAIN (FORMAT JSON) [QUERY]".

	 exit_value integer	 timeout boolean	 process_time_ms bigint	 stdout character varying (20000)
3876afe8f9ba0a9db	0	false	27624	Title: 60197265 Raceway 2.0 Phase I ESA FINAL
8b6aa3b696e6534cc	0	false	25726	Title: untitled
i02ebf2f06f6715cbd	0	false	25481	Title: One-Click-Pool-Light-Planner
d03e6bcd0b83a2	0	false	24435	Title:
8122b1d4c16a03b998	0	false	20880	Creator: Adobe InDesign 16.3 (Windows)
i50f334f89f1a5f32d	0	false	17391	Title: Apresentação do PowerPoint
dc6de84aa7018c2696	0	false	16904	Author: Graphic design: Gabinete Echeverría. gte@gt-echeverria.es
fcb1eba62142c447a3	0	false	15920	Title: 2017 WV Child Abuse and Neglect Judicial Benchbook
f0f7bccac8bee4c26	0	false	12331	Title: KOMET IL Programma
3fbec8b4a2d63ca64	0	false	10792	

Creator: Adobe InDesign CC 2015 (Macintosh)  
Producer: Adobe PDF Library 15.0  
CreationDate: Tue Dec 12 17:38:49 2017 UTC  
ModDate: Wed Dec 13 09:30:00 2017 UTC  
Tagged: no  
UserProperties: no  
Suspects: no  
Form: AcroForm  
JavaScript: no  
Pages: 97  
Encrypted: no

# From simple regexes, to more interesting items

## PDFInfo

---

Producer:    iText 2.1.7 by 1T3XT    →    pi\_producer: iText 2.1.7 by 1T3XT  
CreationDate: Thu Jul 29 05:33:30 2021 UTC    →    pi\_creation\_date: 2021-07-29T05:33:30.000Z  
ModDate:     Thu Jul 29 05:33:30 2021 UTC  
Tagged:       no  
UserProperties: no  
Suspects:     no  
Form:         none  
JavaScript:   no  
Pages:        1



# Structural bits...QPDF's JSON format

q\_keys=[/ProcSet, /Info, /Kids,...

q\_parent\_and\_keys=[/Kids->ARRAY, /ProcSet->ARRAY,..

q\_type\_keys=[/Pages->/Count, /Pages->/ProcSet,..

q\_key\_values=[/Producer->wkhtmltopdf,...

q\_filters=/ASCII85Decode, /FlateDecode->/CCITTFaxDecode

q\_max\_filter\_count=2

```
,
"14 0 R": {
  "/BitsPerComponent": 8,
  "/ColorSpace": "/DeviceRGB",
  "/Filter": "/FlateDecode",
  "/Height": 10,
  "/Length": "15 0 R",
  "/Mask": "12 0 R",
  "/Subtype": "/Image",
  "/Type": "/XObject",
  "/Width": 10
},
"140 0 R": [],
"141 0 R": {
  "/Ascent": 928.222656,
  "/CapHeight": 928.222656,
  "/Descent": -235.839843,
  "/Flags": 4,
  "/FontBBox": [
    -1020.50781,
    -415.039062,
    1680.66406,
    1166.50390
  ],
  "/FontFile2": "142 0 R",
  "/FontName": "/QLFAAA+DejaVuSans",
  "/ItalicAngle": 0,
  "/StemV": 43.9453125,
  "/Type": "/FontDescriptor"
},
```

# Frontend/User Interface

## SafeDocs File Observatory

This is an exploratory faceted search system for the DARPA SafeDocs file-observatory built by the NASA Jet Propulsion Laboratory for the DARPA SafeDocs program.

Choose a Tool/Source ▾ What would you like to search?



- 1) Initial UI was created focusing on the needs of Observatory earlier in its development
- 2) Use of a main search bar was intended for easy search entry

SafeDocs

What would you like to search?

Similarity Field: q\_keys ▾ Completion Field: q\_parent\_and\_keys ▾

Fetches 1010000 results in 1.358 seconds

Download	File	Collection	Format	Created
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a3/19a3afceb440e43cbb894d3b88b4fb52e1d9935d473317409e80e5232957bec	eval-three	application/pdf; version=1.4	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a2/19a2cb5f3ad0ac0be31dd72ddcdf6be01d74a0aba9cfd2d85cfa27955a03c4869	eval-three	application/pdf; version=1.4	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a3/19a36669cee6fb77887e54f5441326b8bcc6309095ac6d8f98789e673fb1a4f	eval-three	application/pdf; version=1.5	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a2/19a2e0e2126eb260416975711baac22f0955bf5cdfef0f31c966e94b65ecat	eval-three	application/pdf; version=1.5	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a3/19a34d3c9b71a7cf34329d46770a674b42cb7eae9df81a4f5006567317c406e	eval-three	application/pdf; version=1.4	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a3/19a30bc973360bc113b5af6be60822c04c35dd67f132d467a71daa9ecb77ada3	eval-three	application/pdf; version=1.4	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a3/19a30023d1c19c4eb4cbe33877178a60c2cf57ac570ce1176daf4be957359e	eval-three	application/pdf; version=1.3	
<input type="checkbox"/>	s3://safedocs-eval-three/eval-three/19/a3/19a37534fed74cd7c7b23a2d9eef342fe9cfcbb5d2c195122d79348aaaf916	eval-three	application/pdf; version=1.3	

Filter Field ▾ collection ▾

collection: eval-three docs: 1010000 100.00%

Top 10 Collections

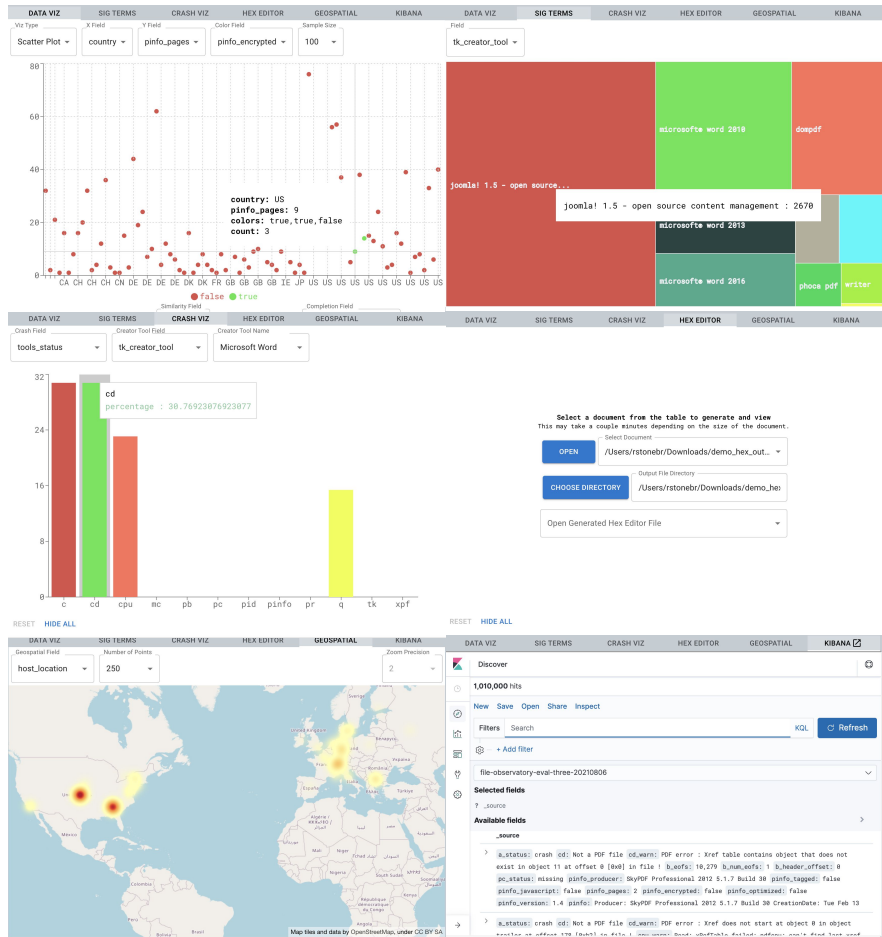
Collection	Count
eval-three	101000

< 1 2 3 4 5 >

- 1) As the Observatory started to mature and more features were added, it became less user friendly.
- 2) After a heuristic evaluation it became clear that we needed rethink the information hierarchy

# Complex Viz Support

- Data Viz Types
  - Donut, Bar Chart, Treemap, Scatterplot
  - Random sampling for scatterplot
- Integrated significant terms querying
- Crash Visualization Analytics for Creator Tools
- Dynamic geospatial mapping
- Trail of Bits Polyfile Hex Editor
- Kibana Integration



**CC-MAIN-2021-31**

# Data

- Common Crawl
  - ~8 million recent (August 2021) CC-MAIN-2021-31
- Bug tracker crawlers
  - ~30k PDFs
  - ~1 million total/40 projects (jpeg, icc, mp4 and others)
  - [https://corpora.tika.apache.org/base/docs/bug\\_trackers/](https://corpora.tika.apache.org/base/docs/bug_trackers/)
  - <https://www.pdfa.org/a-new-stressful-pdf-corpus/>

# Recognized Limitations of Common Crawl Data

- Not the whole web; ~convenience sample
- PDFs on the web likely have profoundly different features and distributions than “in-house” collections

# CC-MAIN-2021-31

- 3.2 billion pages (360 TB)
- 8.3 million PDFs
- 6.4 million retrieved from Common Crawl (1.6 TB)
- 1.9 million refetched (9.8 TB)
  - ~100k refetch failures
- 7.9 million unique hashes (10TB stored)

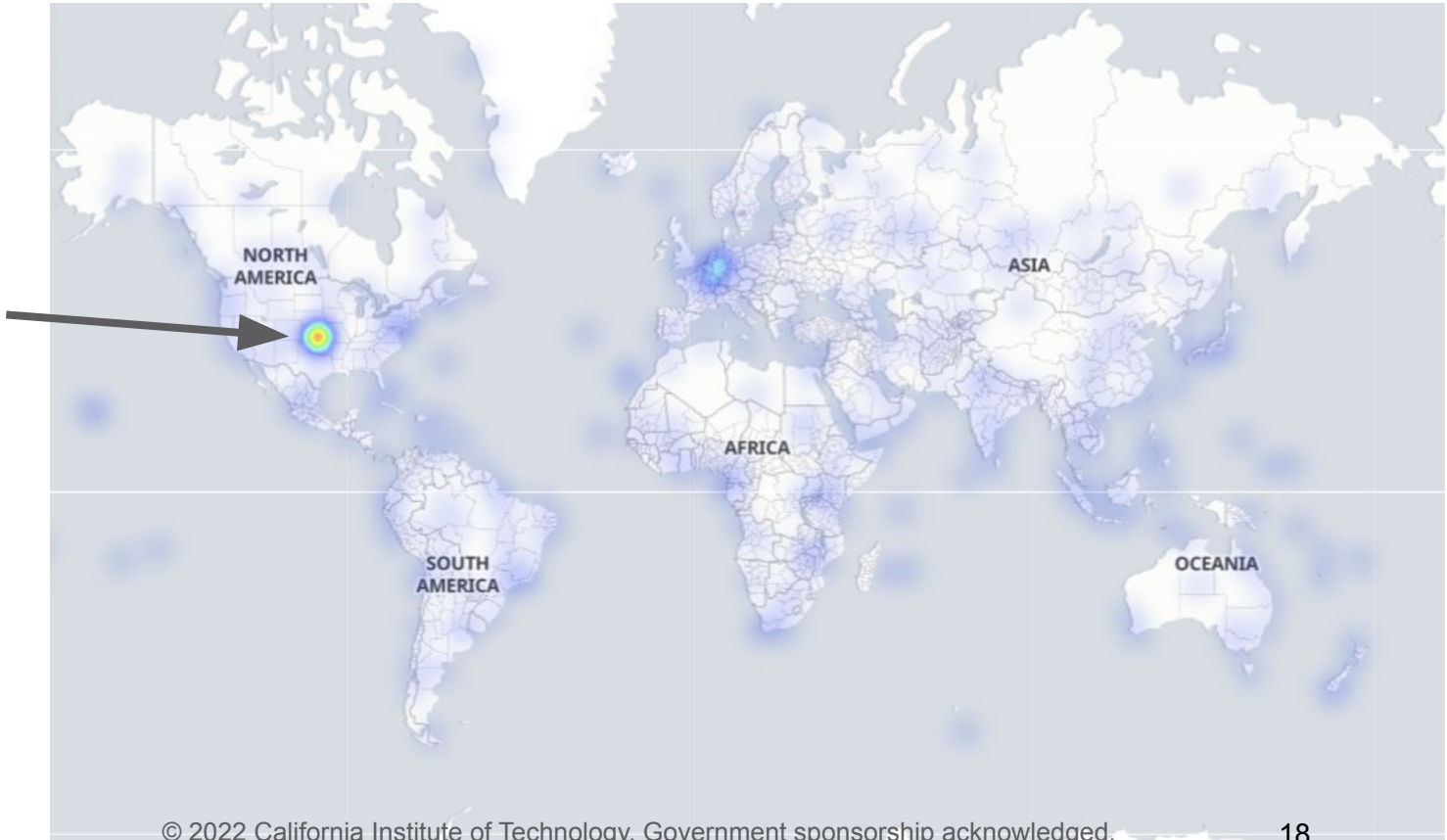


# CC-MAIN-2021-31 – ~7.9 unique million PDFs

Size	Counts
<1kb	7,275
<10kb	109,092
<100kb	1,671,726
<1mb	4,604,023
<10mb	1,692,893
<100mb	210,735
<=1gb	3,686
>1gb	2

# CC-MAIN-2021-31

**NOTE:**  
Country-level  
precision  
points are  
placed in  
country  
centroid



# Simple things are hard...

Different tools extract different types of information.  
Sometimes different tools or versions will disagree  
about the same feature.

Lesson:

Ask not: “What is the title?”

Ask: “What does tool X, version Y say the title is?”

# PDFInfo Boolean Features

Linearized: 2.2M (28%)

Tagged: 2.6M (34%)

Encrypted: 168K (2%)

Javascript: 31K (0.4%)

Custom metadata: 920K (12%)

Metadata stream: 4.7M (59%)

User Properties: 701 (0.009%)

# PDFInfo Top 10 Producers

Producer	Count
Adobe PDF Library 15.0	598,598
NULL	390,416
Microsoft® Word 2016	328,678
Microsoft® Word 2010	248,089
Microsoft® Word 2013	190,300
Microsoft® Word for Microsoft 365	173,702
Microsoft: Print To PDF	153,083
Microsoft® Office Word 2007	147,223
Adobe PDF Library 10.0.1	109,294
Microsoft® Word 2019	103,864

# PDFInfo Top 10 Creator Tools

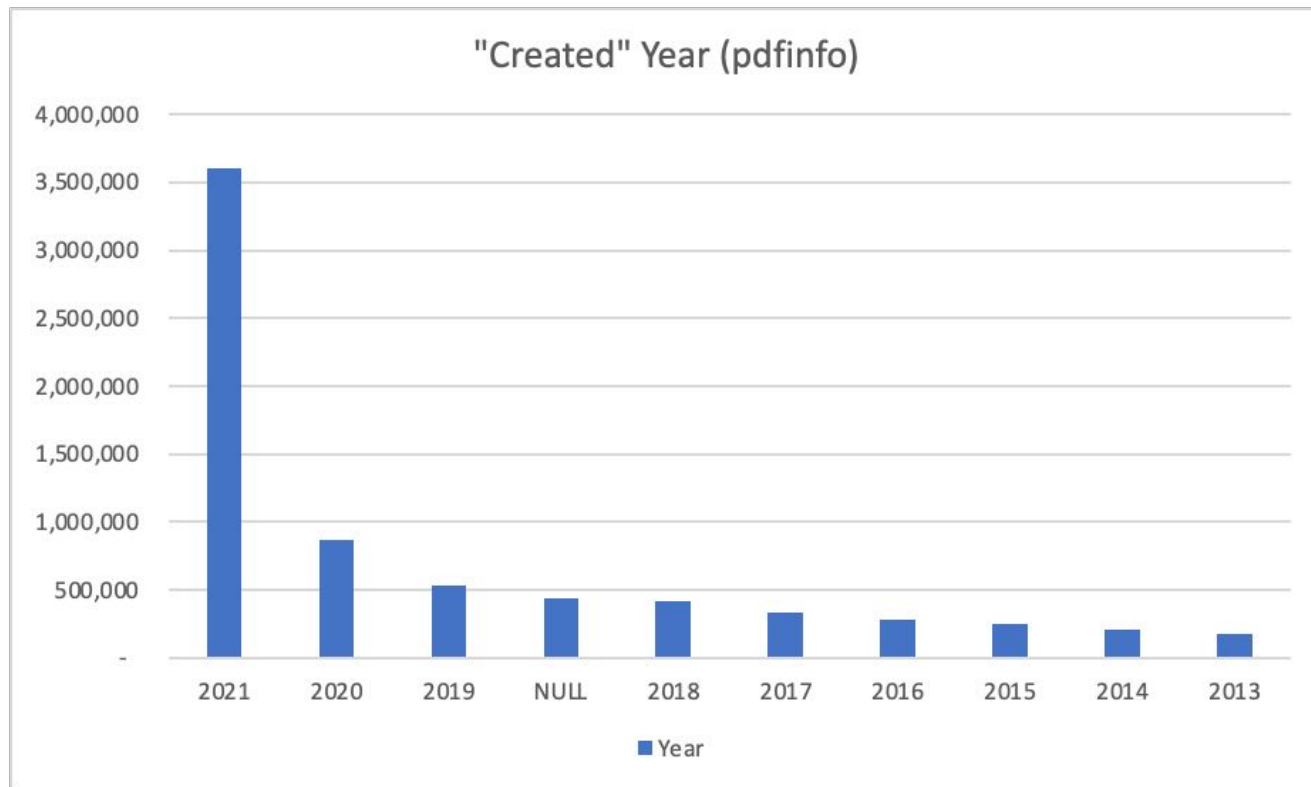
Creator Tool	Count
NULL	1,209,863
Microsoft® Word 2016	353,301
PScript5.dll Version 5.2.2	310,369
Microsoft® Word 2010	256,366
Microsoft® Word 2013	197,867
Microsoft® Word for Microsoft 365	177,634
Word	168,756
Microsoft® Office Word 2007	150,906
Microsoft® Word 2019	105,893
Writer	97,370

# PDFInfo Top 10 Titles

Title	Count
NULL	3,698,634
	553,176
PowerPoint Presentation	34,137
untitled	26,921
Layout 1	12,786
PowerPoint プレゼンテーション	7,803
Slide 1	7,575
Untitled	7,574
Scanned Document	7,089
PowerPoint-Präsentation	6,633

# PDFInfo “created date”, Top 10

Year	Count
2021	3,603,206
2020	868,050
2019	538,157
NULL	437,135
2018	415,537
2017	335,621
2016	281,727
2015	252,303
2014	203,522
2013	179,403

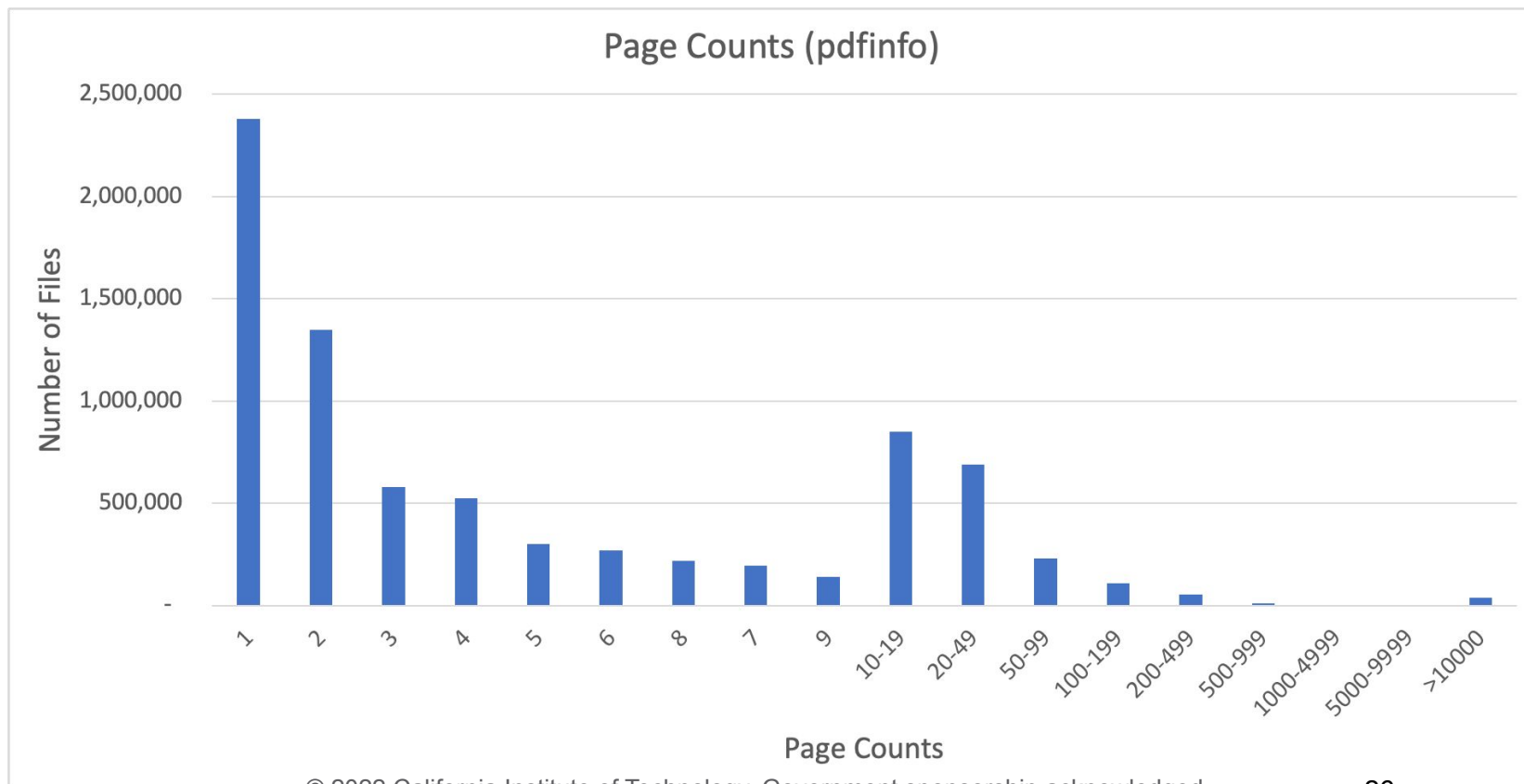




# PDFInfo “PDF version”

PDF Version	Count	Percentage
1.4	2,168,019	27.3%
1.7	2,111,933	26.6%
1.5	1,803,491	22.7%
1.6	958,473	12.1%
1.3	767,140	9.7%
1.2	70,074	0.9%
<b>NULL</b>	38,998	0.5%
1.1	9,865	0.1%
2.0	2,134	0.0%
1.0	2,125	0.0%
0.0	568	0.0%
1.9	11	0.0%
1.8	5	0.0%

# PDFInfo Page Counts



# PDFInfo Page Counts

<b>Pages</b>	<b>Number of Files</b>	<b>Pages</b>	<b>Number of Files</b>
<b>1</b>	2,380,818	<b>10-19</b>	850,670
<b>2</b>	1,346,855	<b>20-49</b>	689,192
<b>3</b>	580,029	<b>50-99</b>	228,998
<b>4</b>	522,442	<b>100-199</b>	109,851
<b>5</b>	298,861	<b>200-499</b>	54,281
<b>6</b>	267,783	<b>500-999</b>	9,262
<b>8</b>	219,429	<b>1000-4999</b>	2,983
<b>7</b>	193,737	<b>5000-9999</b>	62
<b>9</b>	138,533	<b>&gt;10000</b>	39,050

# iText – Boolean features

Encrypted: 167K (2.1%)

Fixed XREF: 278 (0.004%)

Rebuilt XREF: 242K (3.1%)

Hybrid XREF: 1.8M (23%)

XREF Stream: 3.5M (44%)

# Apache Tika – some features

Encrypted: 170K (2%)

Has Collection: 670 (.01%)

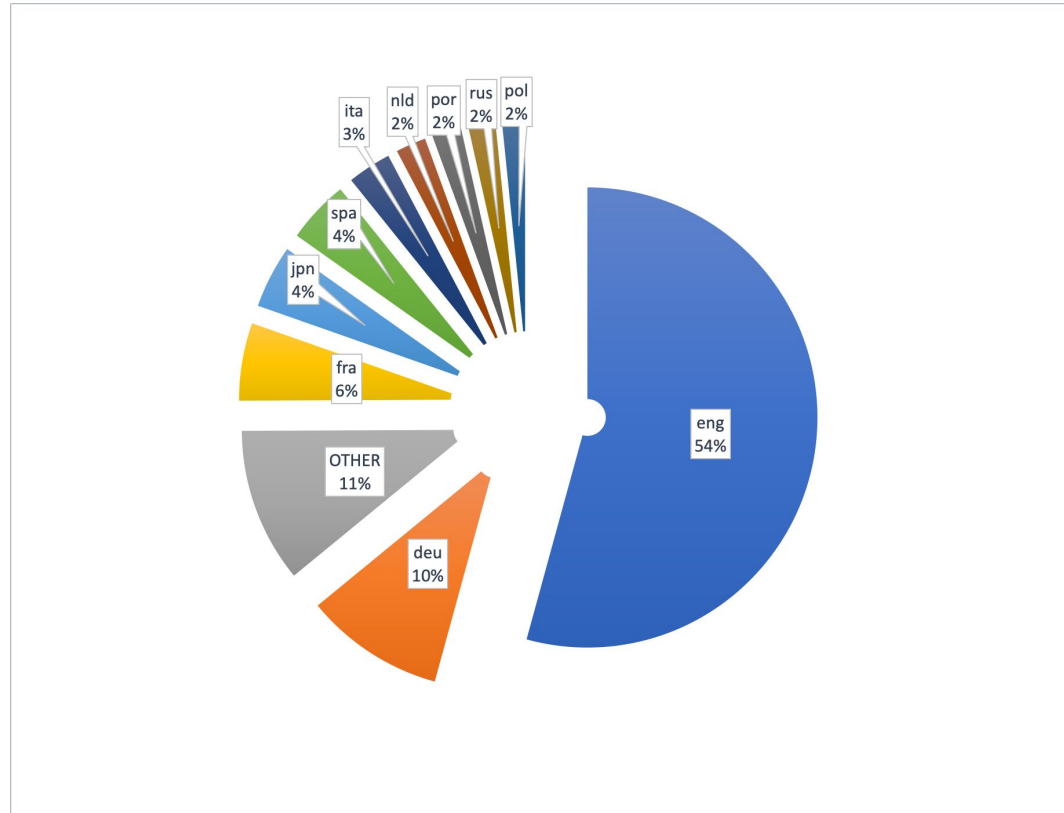
Has Marked Content: 2.8M (36%)

Has XFA: 4.3K (0.1%)

Has XMP: 4.6M (59%)

# Automatic Language Detection on text from Apache Tika

Language	Percentage
eng	54%
deu	10%
OTHER	11%
fra	5%
jpn	4%
spa	4%
ita	3%
nld	2%
por	2%
rus	2%
pol	2%



6.4M files had >  
100 words

# Apache Tika - PDF Subsets

PDF/A (any conformance level): 143K (1.8%)

PDF/UA: 43K (0.54%)

PDF/X: 88K (1.1%)

PDF/VT: 131 (0.002%)

# Apache Tika – Attached Files, Embedded Depth = 1

Mime	Count
text/plain; charset=ISO-8859-1	49,288
application/pdf	12,090
text/plain; charset=windows-1252	5,045
audio/mpeg	4,840
application/x-shockwave-flash	4,740
application/xml	4,727
text/html; charset=UTF-8	3,390
image/png	2,099
image/gif	1,656
image/svg+xml	1,476



# Apache Tika – Attached Files, Embedded Depth > 0

Mime	Count
text/plain; charset=ISO-8859-1	49,397
image/wmf	14,419
application/pdf	12,564
application/vnd.ms-equation	12,387
image/png	7,126
text/plain; charset=windows-1252	5,127
application/xml	4,959
audio/mpeg	4,886
application/x-shockwave-flash	4,753
text/html; charset=UTF-8	3,391

# Apache Tika – Maximum Number of Embedded Files

Embedded File Counts	PDF Count
1	42,054
2	1,416
3	884
4	563
6	423
5	303
8	193
7	176
9	171
16	145

One file has 3,852  
embedded files!

# Apache Tika – Embedded File Depths

Embedded Depth	Count
0	7,931,327
1	98,268
2	37,547
3	3,137
4	177

# Apache Tika (PDFBox)'s “Exit Values”

exit_value	Count
0	7,915,273
1 (caught exceptions)	15,523
3 (timeout)	129
2 (Out Of Memory Errors)	11

Caught exceptions:

- Malformed PDFs
- Runtime exceptions that should probably be fixed

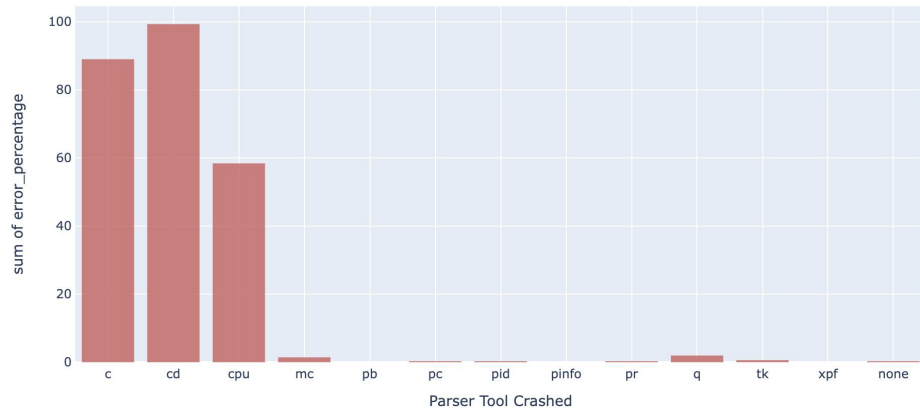
# Crash Analytics Task

## Example

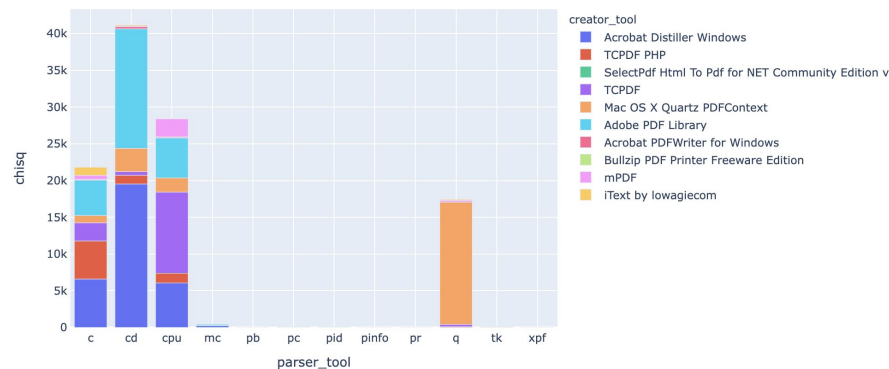
Performing analytics on the crash data from Eval 3

	id	tool	crashes
0	s3://safedocs-eval-three/eval-three/ae/74/ae74...	Acrobat Distiller Windows	[c, cd, cpu, mc, q, tk]
1	s3://safedocs-eval-three/eval-three/31/67/3167...	TCPDF PHP httpwwwtcpdforg	[c]
2	s3://safedocs-eval-three/eval-three/5e/86/5e86...	SelectPdf Html To Pdf for NET Community Edition v	[c, cd]
3	s3://safedocs-eval-three/eval-three/f4/94/f494...	TCPDF PHP httpwwwtcpdforg	[c]
4	s3://safedocs-eval-three/eval-three/0b/d3/0bd3...	Acrobat Distiller Windows	[cd, cpu]
...	...	...	...
81726	s3://safedocs-eval-three/eval-three/2b/f3/2bf3...	Acrobat Distiller Windows	[c, cd, cpu]
81727	s3://safedocs-eval-three/eval-three/3f/2f/3f2f...	Acrobat Distiller Windows	[c, cd]
81728	s3://safedocs-eval-three/eval-three/df/de/dfde...	TCPDF httpwwwtcpdforg	[c, cpu]
81729	s3://safedocs-eval-three/eval-three/29/8b/298b...	Mac OS X Quartz PDFContext	[c, cd, cpu]
81730	s3://safedocs-eval-three/eval-three/0a/f5/0af5...	TCPDF PHP httpwwwtcpdforg	[cd, cpu]

Acrobat Distiller Windows Crash Percentages



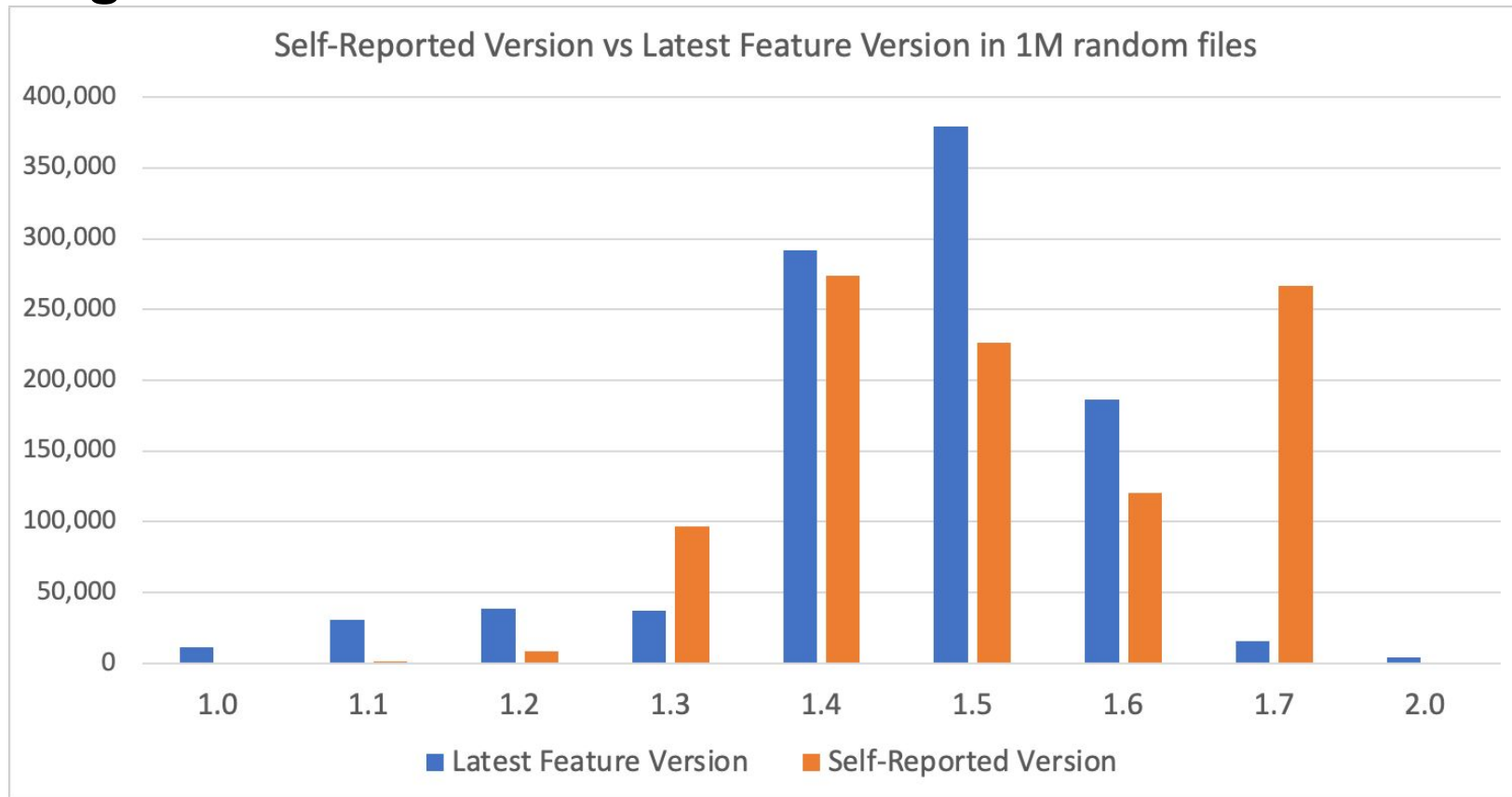
Chi Square Values for 10 Creator Tools



# PDF Versions

Version	PDF Info	Arlington Model TestGrammar's "Self-Reported Version"
1.0	0.03%	0.03%
1.1	0.12%	0.12%
1.2	0.89%	0.87%
1.3	9.72%	9.75%
1.4	27.47%	27.48%
1.5	22.85%	22.79%
1.6	12.14%	12.12%
1.7	26.76%	26.82%
1.8	0.00%	0.00%
1.9	0.00%	0.00%
2.0	0.03%	0.03%

# Arlington Model – Grammar Checker



# Versions Confusion Matrix

Self-Identified Version



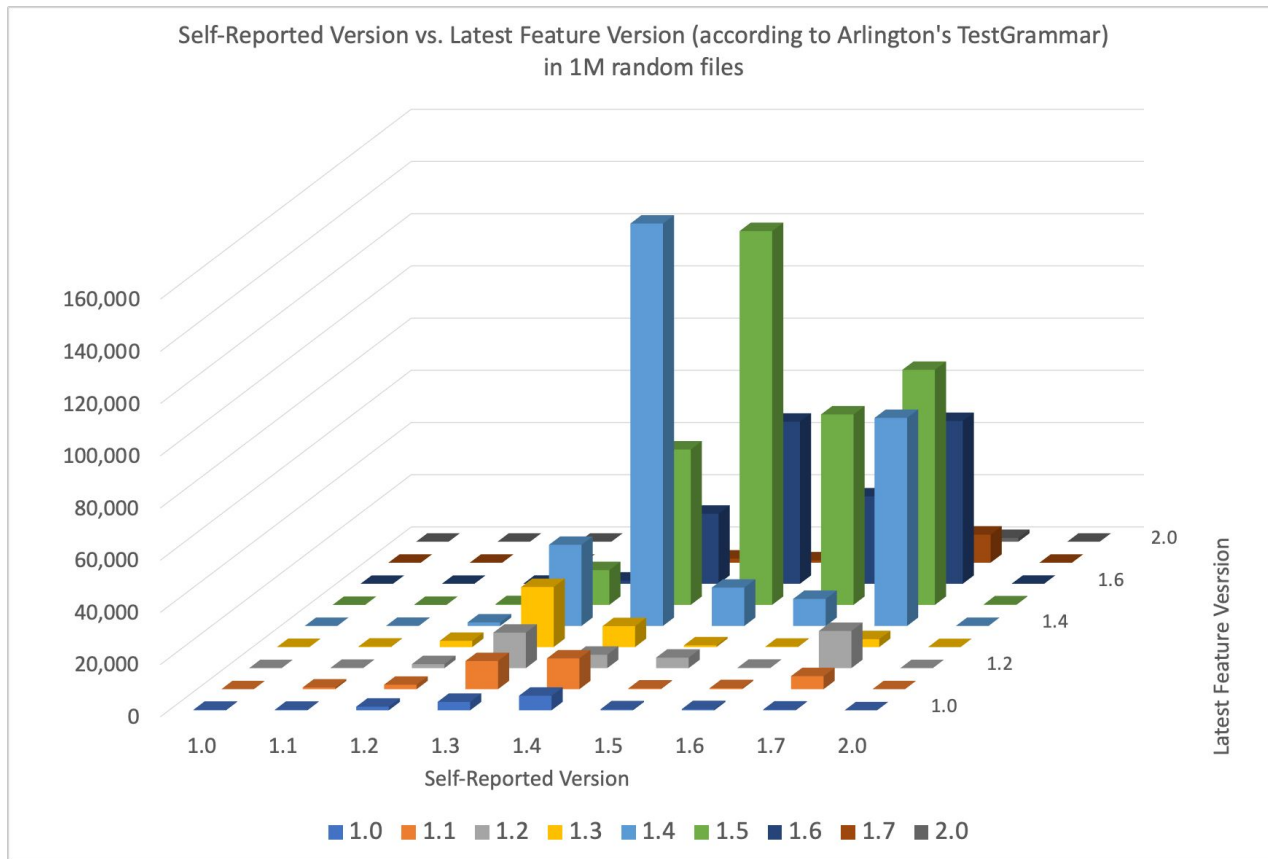
Latest Feature Version



	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.0
1.0	208	218	1,389	3,232	5,578	2 37	452	279	-
1.1	2	656	1,712	10,787	11,797	8 27	397	5,007	-
1.2	2	103	1,547	13,599	5,131	4,008	149	14,235	-
1.3	7 1	169	2,379	22,998	8,022	5 64	148	3,034	-
1.4	8 1	2 7	1,351	31,091	154,162	14,691	10,374	79,704	3 2
1.5	-	1	276	13,339	59,530	143,205	72,959	90,049	202
1.6	-	-	32	1,244	26,847	62,120	33,504	62,396	2 2
1.7	-	-	11	394	1,610	7 90	1,695	10,785	-
2.0	2	10	13	398	829	599	822	1,440	12



# Versions – Confusion Matrix



# Arlington Model – Grammar Checker

At least one warning: 16%

At least one error: 25%

Buckle up!

Start using it and contributing to it!

<https://github.com/pdf-association/arlington-pdf-model>

# Next Steps

# Next Steps

- Continue to work with potential end users on back end and front end
- Find a home for the raw data and features
- Documentation, documentation and documentation

# Some Resources

<https://github.com/tballison/file-observatory>

<https://github.com/pdf-association/arlington-pdf-model>

<https://tika.apache.org/>

Contact Info: [timothy.b.allison@jpl.nasa.gov](mailto:timothy.b.allison@jpl.nasa.gov), @\_tallison



**Jet Propulsion Laboratory**  
California Institute of Technology

---

[jpl.nasa.gov](https://jpl.nasa.gov)