



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.

## **Making Sense of PDF Structures in the Wild at Scale**

*PDF Days 2021*

September 29, 2021

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. © 2021 California Institute of Technology. Government sponsorship acknowledged.



**Jet Propulsion Laboratory**  
California Institute of Technology

# The Team



Chris Mattmann  
PI; Chief Technology  
and Innovation Officer



Tim Allison  
Files and Search



Wayne Burke  
Cognizant Engineer



Michael Fedell  
Data Scientist



Dustin Graf  
Project Manager



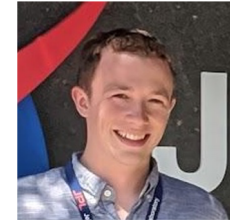
Anastasia Menshikova  
Data Scientist



Michael Milano  
UX/UI Researcher



Phil Southam  
Trouble (Fun?)  
Maker



Ryan Stonebraker  
Data Scientist  
Alaskan

# Debts of Gratitude

Sergey Bratus

Peter Wyatt and Duff Johnson, PDF Association

Kudu Dynamics, Trail of Bits, Galois, BAE and SRI

Common Crawl

# Goals

## Goals

- Demonstrate state of the possible
- Transfer techniques if not code
- Start a discussion for next steps

## Caveats

- Numbers/stats are on different sized samples and everything is preliminary...we **cannot** draw conclusions
- Some of the examples include synthetic data, labeled: **Synthetic data!**



# Motivation

- Inducing grammars
- Devtesting parsers during development
- Testing/profiling/tracing existing parsers
  - Literal files
  - Seeds for fuzzing
- Quickly identifying root causes, patterns

# Outline

1. Getting the PDFs -- URL descriptives and fetching
2. Observatory architecture
3. Feature Extraction
4. Basic Descriptives
5. Comparing Text Extraction Tools
6. Discovery with built-in Elasticsearch features
7. Next Steps

# Fetching Common Crawl CC-MAIN-2021-31

# Common Crawl



- Monthly open source crawls of large portions of the web: for July/August 2021 (CC-MAIN-2021-31), 3.2 billion pages (360 TB).
- Available via Amazon Web Services Public Datasets
- Searchable indexes available

<https://commoncrawl.org>

# Common Crawl -- known limitations

- Files truncated at 1MB
- Coverage/representativeness
  - Web contains web documents – likely missing subtypes designed for internal/proprietary/sensitive data (medical, 3D engineering, legal...)
  - Common Crawl is ~convenience sample of the web

Ref: [http://spw20.langsec.org/papers/corpus\\_LangSec2020.pdf](http://spw20.langsec.org/papers/corpus_LangSec2020.pdf)



# Overview of URLs

Random sample of ~3 million URLs with a '200'  
response in CC-MAIN-2021-31

# Detected File Types

Detected Mime	Percentage
<b>text/html</b>	<b>84.16%</b>
<b>application/xhtml+xml</b>	<b>13.20%</b>
<b>text/plain</b>	<b>1.85%</b>
<b>application/pdf</b>	<b>0.26%</b>
<b>image/jpeg</b>	<b>0.14%</b>
<b>application/rss+xml</b>	<b>0.09%</b>
<b>application/atom+xml</b>	<b>0.07%</b>
<b>application/xml</b>	<b>0.05%</b>
<b>image/png</b>	<b>0.03%</b>
<b>text/calendar</b>	<b>0.02%</b>

See also: <https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes>

# Top Level Domains (all file types)

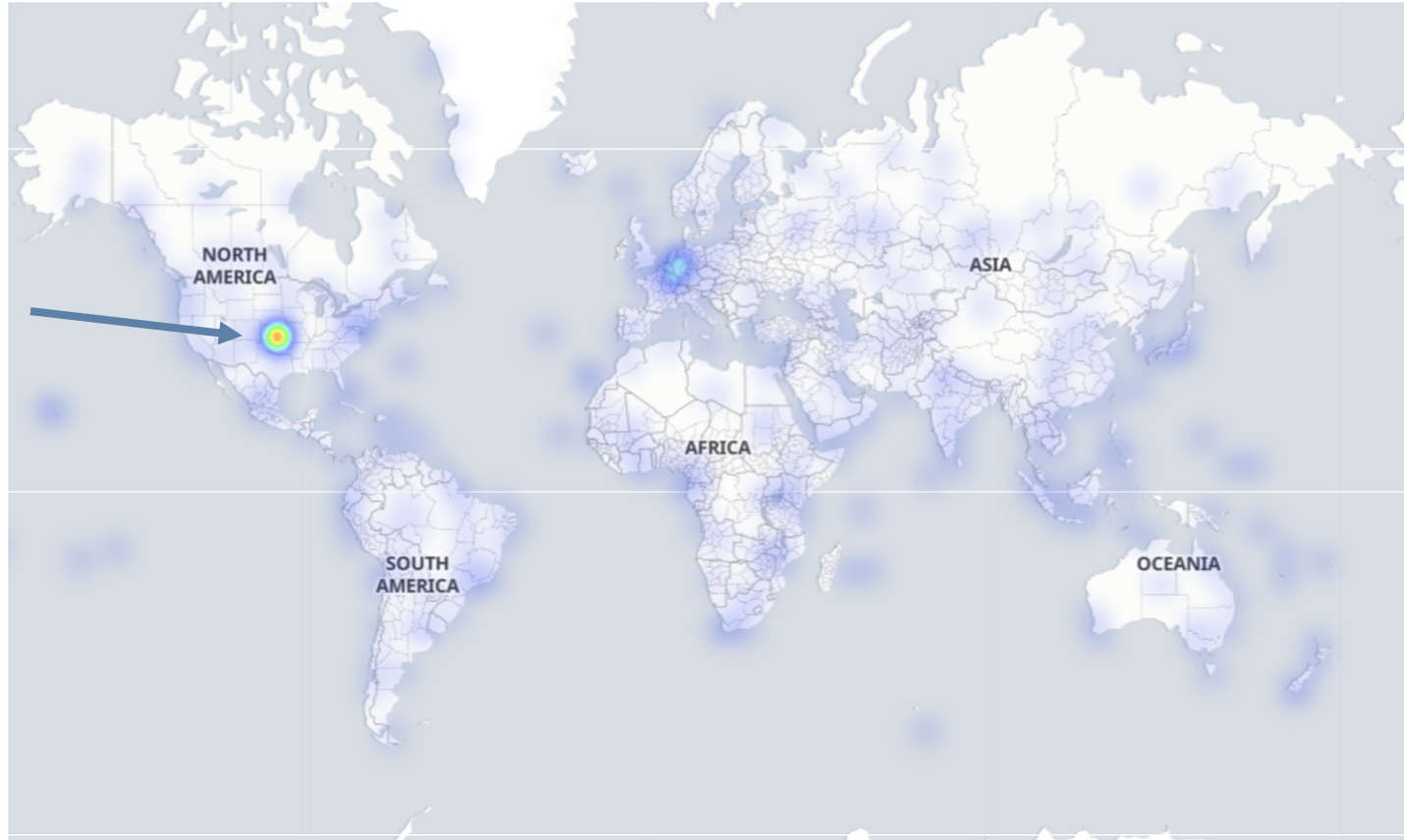
TLD	Percentage
com	45.1%
org	5.5%
ru	4.8%
de	4.3%
net	3.5%
uk	2.5%
jp	1.8%
fr	1.7%
it	1.7%
nl	1.6%

## Countries (all file types) -- GeoIP with free version of MaxMind

Country	Percentage
US	42.1%
DE	8.0%
UNKNOWN	5.3%
RU	4.7%
FR	4.5%
JP	4.3%
CA	3.9%
NL	2.7%
GB	2.6%
ES	1.4%

<https://www.maxmind.com/en/home>

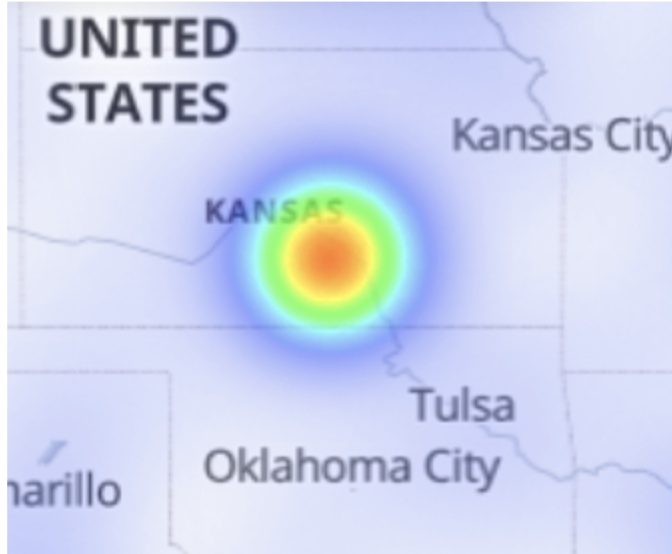
# PDFs around the world – all 8.3 million in CC-MAIN-2021-31



See next  
slide on geo  
precision!



# Note on GeoIP precision!



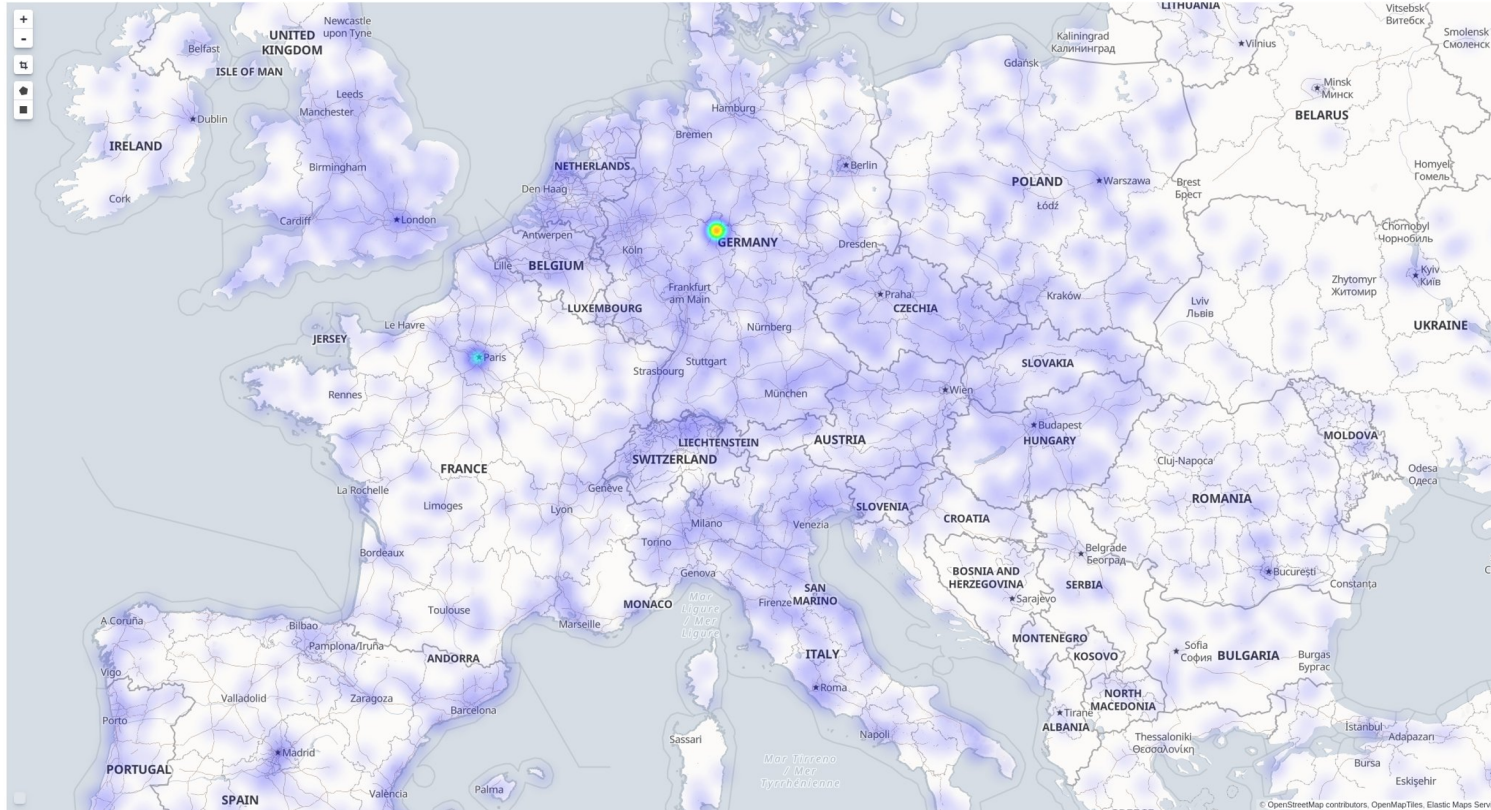
Cheney Reservoir does **NOT** host 1.6 million PDFs!

This is where MaxMind puts IP addresses that it can only identify as “somewhere in the U.S.”

[https://en.m.wikipedia.org/wiki/Cheney\\_Reservoir](https://en.m.wikipedia.org/wiki/Cheney_Reservoir)

<https://www.denverpost.com/2016/08/10/lawsuit-kansas-home-600-million-ip-addresses/>

# PDFs with Borders?



## The Math...See Extras section at the end of this deck

- Takeaway, yes Germany does have a high percentage of PDFs, but Switzerland and India have **more** as a percentage of URLs.
- Russia and China have fewer PDFs as a percentage of URLs.

**And, of course, remember the caveats about generalizability of web data and Common Crawl!**

Bytes, bytes and more bytes

# Some stats for CC-MAIN-2021-31

- 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)



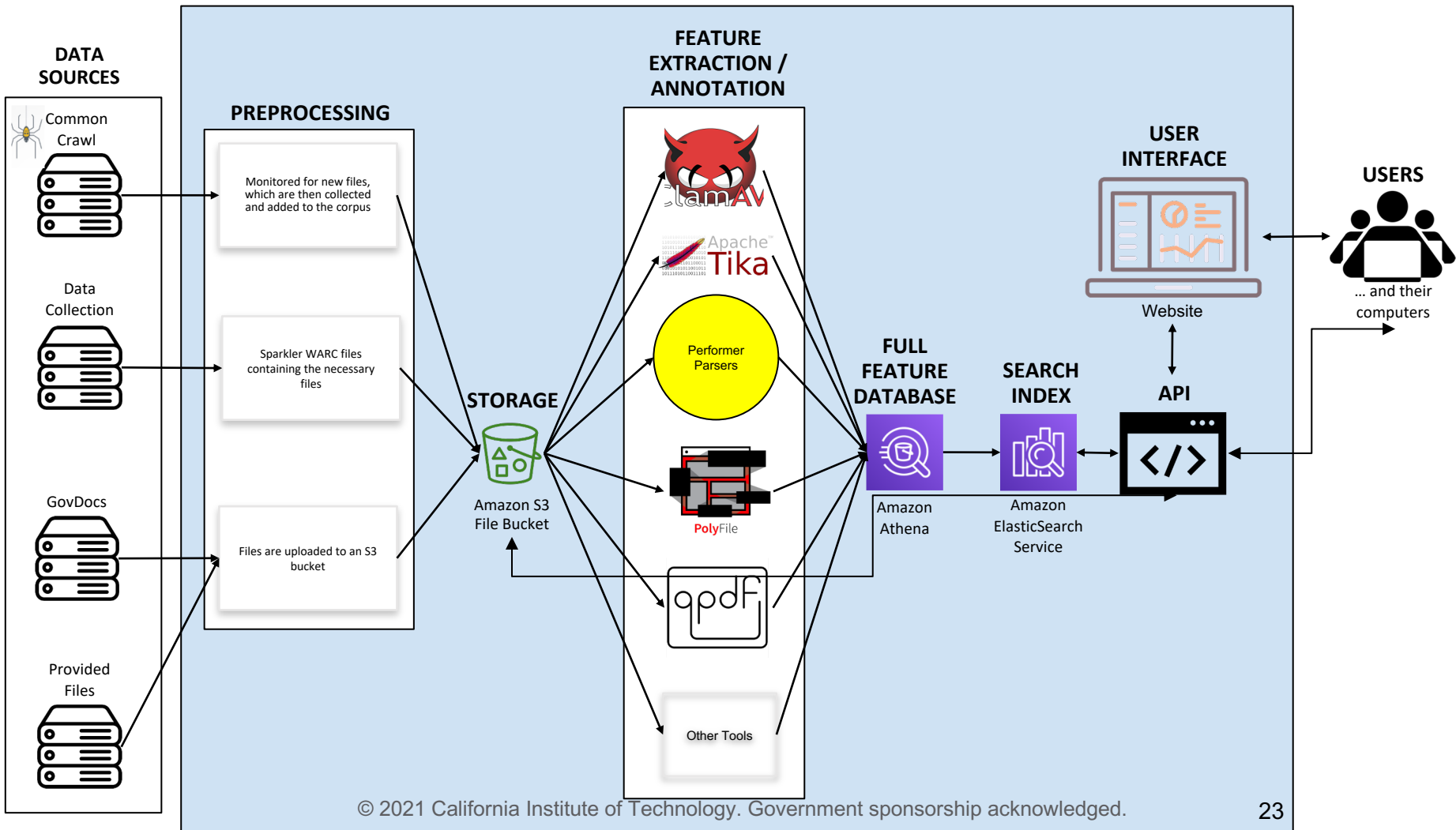
# PDF Sizes

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

# Observatory at various scales

# Various Scales

- Full Cloud -- AWS components (Athena, Kinesis, AWS Batch...), S3 for storage -> AWS Hosted Elasticsearch
- Desktop -- Postgresql, local files -> local Elasticsearch
- Hybrid -- hosted/local Postgresql, EC2 instance for processing, S3 for storage -> AWS Hosted Elasticsearch



# File Observatory in a Box

## Public github repository

- file-observatory ~/IntelliJ/file-observatory
  - .idea
  - batchlite
  - commoncrawl-fetcher
  - ingest
  - tika-containers
  - tool-runners
    - arlington
    - caradoc
    - clamav
    - fileprofiler
    - mutoolclean
    - mutooltext
    - pdfchecker
    - pdfcpu
    - pdfid
    - pdfimages
    - pdfinfo
    - pdfminerdump
    - pdfminertext
    - pdftoppm
    - pdftops
    - pdftotext
    - polyfile
    - qpdf
    - tika

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

## Standard Metadata Output Per Tool

- Tables (19)
  - arlington
  - caradoc
  - cc\_detected\_mimes
  - cc\_languages
  - cc\_mimes
  - cc\_truncated
  - cc\_urls
  - cc\_warc\_file\_name
  - clamav
  - mutoolclean
  - mutooltext
  - pdfchecker
  - pdfcpu
  - pdfid
  - pdfinfo
  - polyfile
  - profiles
  - qpdf
  - tika







# Standard table for each tool (pdfinfo 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 Echeverria. 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
3fbebcb8b4a2d63ca64	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

# Feature Extraction

# 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,...

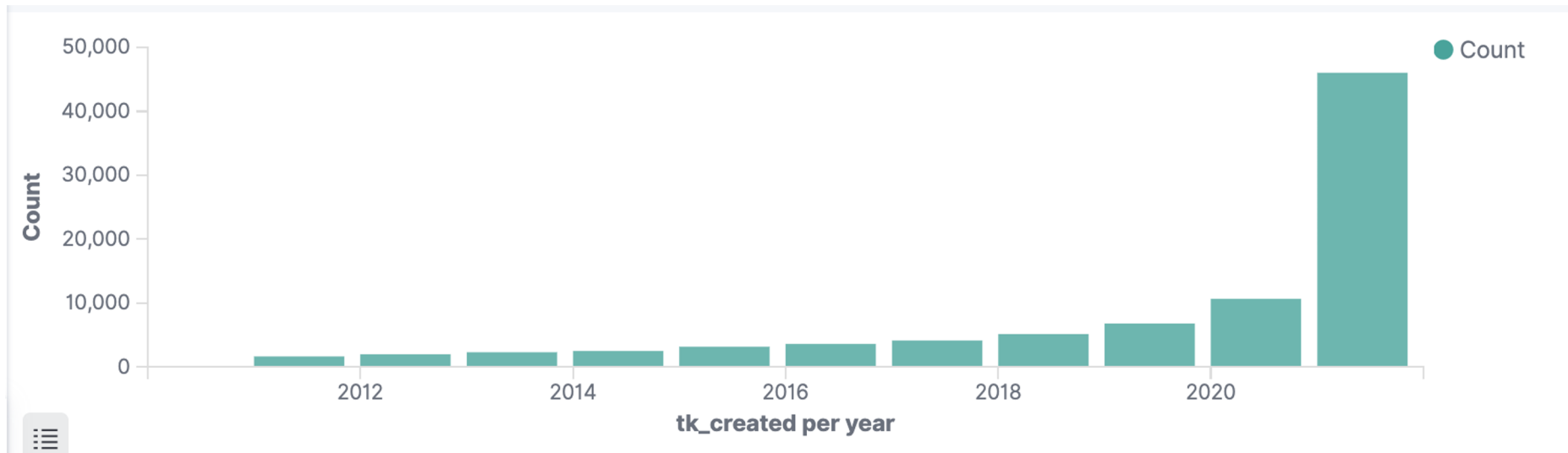
q\_filters=/ASCII85Decode, /FlateDecode->/CCITTFaxDecode

q\_max\_filter\_count=2

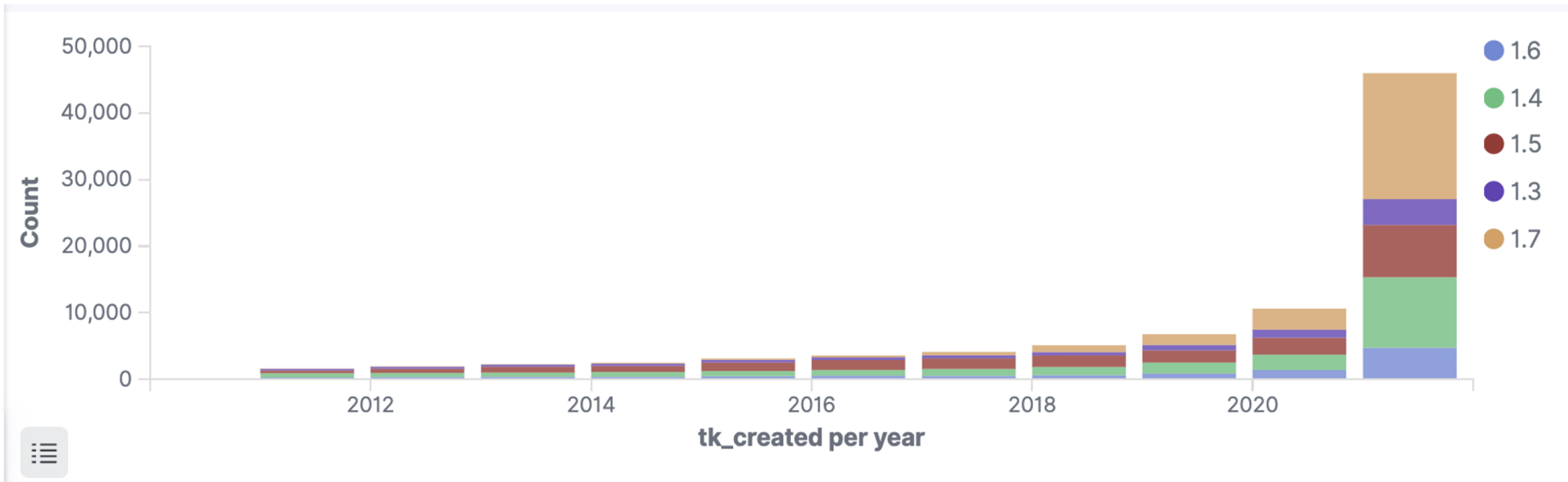
```
},
"14 0 R": {
  "/BitsPerComponent": 8,
  "/ColorSpace": "/DeviceRGB",
  "/Filter": "/FlateDecode",
  "/Height": 10,
  "/Length": "15 0 R",
  "/Mask": "12 0 R",
  "/Subtype": "/Image",
  "/Type": "/XObject",
  "/Width": 10
},
"140 0 R": [],
"141 0 R": {
  "/Ascent": 928.222656,
  "/CapHeight": 928.222656,
  "/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"
},
}
```

# Basic descriptives

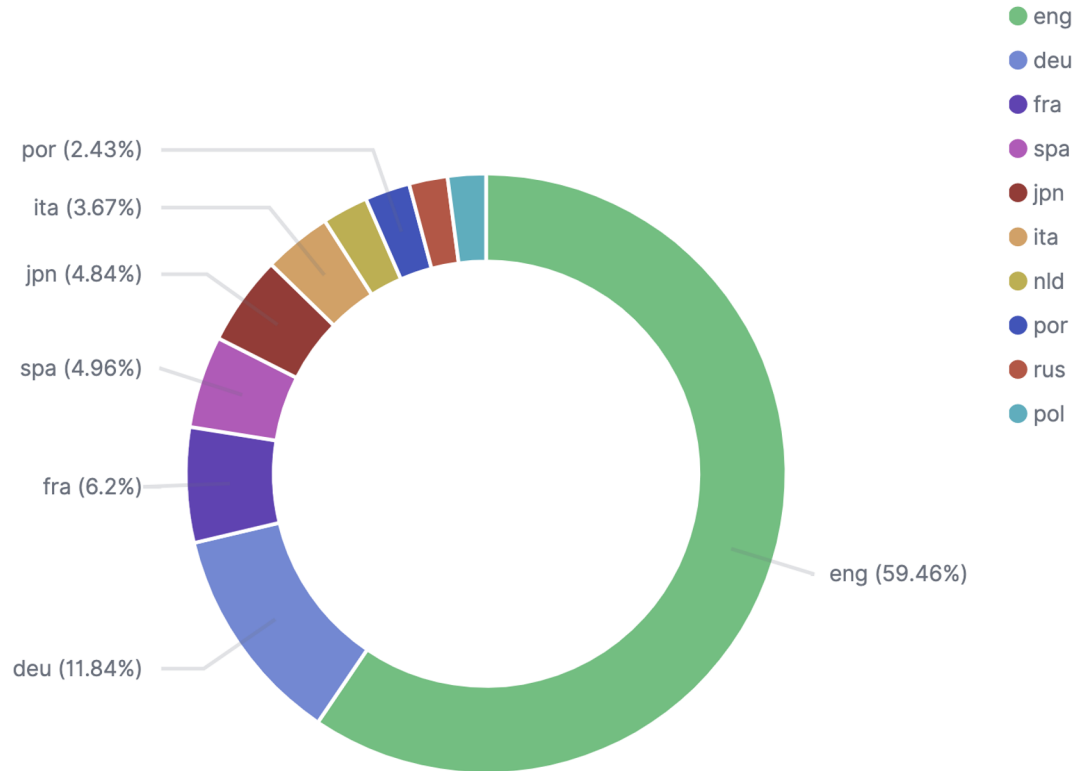
# PDFs by “created date”, where it exists



# PDF versions by year



# Automatic Language Detection on text from Apache Tika





# Comparing Text Extraction Tools

# Measuring Text

- Estimating quality without ground truth
  - Unmapped Unicode characters
  - Out-of-vocabulary (OOV %)
- Comparing output from two or more tools -- overlap, dice coefficient

See also: Popat, Ashok. “A panlingual anomalous text detector.” DocEng ‘09: Proceedings of the 9<sup>th</sup> ACM symposium on Document engineering. 2009.



**Ashok C Popat**

Google, Inc., Mountain View, CA, USA



# Unmapped Unicode characters

File: 0000bd3e1cd25eaaacc56e30bdfd800a60018495a12e2d7d423f73f8da0d7416

Annullare il 25% del debito pubblico si può, ma ...

Creato: Domenica, 07 Febbraio 2021 00:00 Scritto da Rocco Artifoni

Da tempo e da più parti negli ultimi mesi ha preso slancio la proposta di cancellare la parte di debito (circa il 25%) che alcuni Paesi europei hanno nei confronti della Banca Centrale Europea (BCE). Questa tesi è stata sostenuta anche dal presidente del Parlamento europeo David Sassoli. Il recente appello sottoscritto da oltre un centinaio di economisti di vari Paesi europei è un ulteriore passo in questa direzione. Si tratta di una proposta tutto sommato ragionevole, che presenta alcune criticità, ma con vantaggi probabilmente superiori agli aspetti negativi.

```
"pdf:charsPerPage": [
  "1318",
  "2200"
],
"pdf:unmappedUnicodeCharsPerPage": [
  "1318",
  "2200"
],
```

**File: 0000bd3e1cd25eaaacc56e30bdfd800a60018495a12e2d7d423f73f8da0d7416**

	<b>Tika</b>	<b>Mutool</b>	<b>pdftotext</b>
<b>Detected Language</b>	<b>Romanized Urdu</b>	<b>Romanized Urdu</b>	<b>Romanized Bengali</b>
<b>OOV %</b>	<b>99.9%</b>	<b>99.9%</b>	<b>99.9%</b>

100% unmapped Unicode characters  
mutool

pdftotext

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

# Tika vs Mutool extract sort by overlap ascending

detected_lang_tika character varying (12)	num_tokens_tika integer	num_common_tokens_tika integer	detected_lang_mutool character varying (12)	num_tokens_mutool integer	num_common_tokens_mutool integer	tika_vs_mutool text
hin-rom	160571	5087	eng	197	138	.002
fas	645	552	pus	645	1	.004
fas	536	340	ckb	536	0	.016
eng	70881	46312	ceb	79	16	.018
heb	4068	2659	heb	4070	134	.025
urd-rom	1776	1	eng	311	214	.028
fas	501	372	pus	501	17	.038
heb	3240	2310	heb	3219	153	.041
fas	1460	920	pes	1457	110	.046
jpn	585	396	jpn	646	616	.052
pes	390	268	pes	378	31	.074
deu	9068	6539	urd-rom	43536	10	.092

**File: 3e652c7dfdd96db91ff85e67931b9ea35ff74bc5610e7f5458509f69f9bad471**

Profile

	Tika	Mutool extract	pdftotext
Detected Language	German	Romanized Urdu	German
Alphabetic Tokens	9068	43536	8952
Common Tokens	6539	10	6511
Out-of-Vocabulary (OOV)	28%	99.9%	27%

Compare

Tool A	Tool B	Overlap
Tika	Mutool extract	9%
Tika	pdftotext	95%
Mutool extract	pdftotext	9%

## 1 Der Hauptsatz der elementaren Zahlentheorie. Beweis des ersten Teils

Wir wollen jetzt die in der Einführung ausgesprochenen Behauptungen zu einem einzigen Satz, dem sogenannten Hauptsatz der elementaren Zahlentheorie, zusammenfassen.

Jede von Null verschiedene ganze Zahl kann als Produkt von Primzahlen dargestellt werden, wobei die Darstellung bis auf die Reihenfolge und die Vorzeichen der Faktoren eindeutig ist.

mutool

```
1 Der Hauptsatz
der elementaren
Zahlentheorie. Beweis
des ersten
Teils

Wir wollen jetzt die in der Einführung
ausgesprochenen Behauptungen zu
einem
einzigem Satz,
dem
sogenannten
Hauptsatz der elementaren
Zahlentheorie, zusammenfassen.
```

Tika

```
<p>1 Der Hauptsatz der elementaren Zahlentheorie. Beweis
des ersten Teils
</p>
<p>Wir wollen jetzt die in der Einführung ausgesprochenen
Behauptungen zu einem einzigen Satz,
dem sogenannten Hauptsatz der elementaren Zahlentheorie,
zusammenfassen.
</p>
<p>Jede von Null verschiedene ganze Zahl kann als Produkt von
Primzahlen dargestellt werden,
wobei die Darstellung bis auf die Reihenfolge und die
Vorzeichen der Faktoren eindeutig ist.
</p>
```



# Discovery with built-in Elasticsearch features

# FORCEDENTRY

Citizen Lab's post includes this image

```
. /Filter [/FlateDecode /FlateDecode /JBIG2Decode]
```

Three filters?! One repeat?

```
q_max_filter_count:[3 TO *]
```

Found in the wild: `/ASCII85Decode->/FlateDecode->/CCITTFaxDecode`

<https://citizenlab.ca/2021/09/forcedentry-nso-group-imessage-zero-click-exploit-captured-in-the-wild/>

# Elasticsearch terms aggregation (facets/group by)

country:DE

```
"buckets" : [  
  {  
    "key" : "microsoft® word 2016",  
    "doc_count" : 667  
  },  
  {  
    "key" : "microsoft® word 2010",  
    "doc_count" : 515  
  },  
  {  
    "key" : "pscript5.dll version 5.2.2",  
    "doc_count" : 484  
  },  
  {  
    "key" : "pdf24 creator",  
    "doc_count" : 338  
  },  
  {  
    "key" : "microsoft® word 2013",  
    "doc_count" : 307  
  },  
  {  
    "key" : "writer",  
    "doc_count" : 240  
  },  
  {  
    "key" : "microsoft® word fur microsoft 365",  
    "doc_count" : 238  
  },  
]
```

country:JP

```
"buckets" : [  
  {  
    "key" : "microsoft® word 2016",  
    "doc_count" : 290  
  },  
  {  
    "key" : "pscript5.dll version 5.2.2",  
    "doc_count" : 243  
  },  
  {  
    "key" : "microsoft® word 2013",  
    "doc_count" : 172  
  },  
  {  
    "key" : "microsoft® excel® 2016",  
    "doc_count" : 155  
  },  
  {  
    "key" : "microsoft® word 2019",  
    "doc_count" : 155  
  },  
  {  
    "key" : "microsoft® word 2010",  
    "doc_count" : 111  
  },  
  {  
    "key" : "microsoft® excel® 2019",  
    "doc_count" : 103  
  },  
]
```

# Elasticsearch significant terms aggregation

country:DE

country:JP

```
"doc_count" : 10985,
"keywords" : {
  "doc_count" : 10985,
  "bg_count" : 100100,
  "buckets" : [
    {
      "key" : "PDF24 Creator",
      "doc_count" : 338,
      "score" : 743.5330885372576,
      "bg_count" : 587
    },
    {
      "key" : "Microsoft® Word für Microsoft 365",
      "doc_count" : 238,
      "score" : 524.112952240219,
      "bg_count" : 412
    },
    {
      "key" : "Acrobat PDFMaker 8.1 für Word",
      "doc_count" : 88,
      "score" : 302.65131076959705,
      "bg_count" : 94
    },
    {
      "key" : "Acrobat PDFMaker 21 für Word",
      "doc_count" : 58,
      "score" : 158.31469555800433,
      "bg_count" : 81
    },
    {
      "key" : "Acrobat PDFMaker 17 für Word",
      "doc_count" : 50
```

```
"doc_count" : 4396,
"keywords" : {
  "doc_count" : 4396,
  "bg_count" : 100100,
  "buckets" : [
    {
      "key" : "CubePDF",
      "doc_count" : 87,
      "score" : 777.1216043230356,
      "bg_count" : 111
    },
    {
      "key" : "Microsoft® Excel® 2016",
      "doc_count" : 155,
      "score" : 700.3757254960391,
      "bg_count" : 445
    },
    {
      "key" : "Microsoft® Excel® 2019",
      "doc_count" : 103,
      "score" : 633.8433236016839,
      "bg_count" : 213
    },
    {
      "key" : "Word 用 Acrobat PDFMaker 21",
      "doc_count" : 41,
      "score" : 372.1730314529601,
      "bg_count" : 51
    },
    {
```

# Using Elasticsearch's Prefix Completion

```
GET file-observatory-cc-dev/_search
{
  "_source": false,
  "suggest": {
    "fontWeight": {
      "prefix": "/FontWeight->",
      "completion": {
        "field": "q_keys_and_values.completion",
        "size": 2000,
        "skip_duplicates": true
      }
    }
  }
}
```

FontWeight	number	(Optional; PDF 1.5; should be used for Type 3 fonts in Tagged PDF documents) The weight (thickness) component of the fully-qualified font name or font specifier. The possible values shall be 100, 200, 300, 400, 500, 600, 700, 800, or 900, where each number indicates a weight that is at least as dark as its predecessor. A value of 400 shall indicate a normal weight; 700 shall indicate bold.
------------	--------	--

```
"options" : [
  {
    "text" : "/FontWeight->/Bold",
    "_index" : "file-observatory-cc-dev",
    "_type" : "_doc",
    "_id" : "190759",
    "_score" : 1.0
  },
  {
    "text" : "/FontWeight->0",
    "_index" : "file-observatory-cc-dev",
    "_type" : "_doc",
    "_id" : "8252444",
    "_score" : 1.0
  },
  {
    "text" : "/FontWeight->100",
    "_index" : "file-observatory-cc-dev",
    "_type" : "_doc",
    "_id" : "336682",
    "_score" : 1.0
  },
  {
    "text" : "/FontWeight->1000",
    "_index" : "file-observatory-cc-dev",
    "_type" : "_doc",
    "_id" : "5189420",
    "_score" : 1.0
  }
]
```

# Using Elasticsearch's Autosuggest

GET file-observatory-202102/\_search

```
{
  "_source": false,
  "suggest": {
    "my-suggest1": {
      "text": "/Subtype",
      "term": {
        "field": "q_keys",
        "suggest_mode": "always",
        "sort": "frequency",
        "size": 100,
        "max_edits": 2,
        "min_word_length": 2,
        "max_term_freq": 2000000
      }
    }
  }
}
```

```
"options" : [
  {
    "text" : "/SubType",
    "score" : 0.875,
    "freq" : 147
  },
  {
    "text" : "/Subtype2",
    "score" : 0.875,
    "freq" : 9
  },
  {
    "text" : "/Subtyp",
    "score" : 0.85714287,
    "freq" : 3
  },
  {
    "text" : "/Pubtype",
    "score" : 0.875,
    "freq" : 2
  }
]
```

## Synthetic data!

```
{
  "text" : "/Subtypd",
  "score" : 0.875,
  "freq" : 2
},
{
  "text" : "/Subtypg",
  "score" : 0.875,
  "freq" : 2
},
{
  "text" : "/subtype",
  "score" : 0.875,
  "freq" : 1
},
{
  "text" : "/SUBtyPe",
  "score" : 0.75,
  "freq" : 1
}
```

# Spelling Variants

Scripting autosuggest

/Subtype 8798		
/Subtype		8798
/SubType		41
/Subtype2		4
/subtype		1
/CapHeight 6489		
/CapHeight		6489
/CapHieght		8
/CVHeight		2
/ColorSpace 5748		
/ColorSpace		5748
/Colorspace		1
/FirstChar 5650		
/FirstChar		5650
/FirtsChar		1
/Widths 5646		
/Widths		5646
/Width		4625
/WXdths		1

**Synthetic  
data!**

# Syntax, erm, flexibility: /Root->/Dests with direct object

```
},
"136 0 R": {
  "/Dests": {
    "/__WKANCHOR_10": "94 0 R",
    "/__WKANCHOR_2": "21 0 R",
    "/__WKANCHOR_4": "22 0 R",
  },
  "/Pages": "2 0 R",
  "/Type": "/Catalog"
},
```

```
GET file-observatory-eval-three-20210806/_search
{
  "query": {
    "query_string": {
      "query": "q_parent_and_keys:/.Dests.*/"
    }
  },
  "size": 0,
  "aggregations": {
    "significant_queries": {
      "significant_terms": {
        "field": "pinfo_producer.keyword",
        "size": 50,
        "chi_square": {
          "background_is_superset": false
        }
      }
    }
  }
}
```

<b>Dests</b>	dictionary	(Optional; PDF 1.1; shall be an indirect reference) A dictionary of names and corresponding <i>destinations</i> (see 12.3.2.3, "Named Destinations").
--------------	------------	---



# Syntax, erm, flexibility: /Root->/Dests with direct object

```
"significant_queries" : {  
  "doc_count" : 9989,  
  "bg_count" : 1010000,  
  "buckets" : [  
    {  
      "key" : "wkhtmltopdf",  
      "doc_count" : 8405,  
      "score" : 414080.0519638461,  
      "bg_count" : 8780  
    },  
    {  
      "key" : "dvips + gpl ghostscript git prerelease 9.22",  
      "doc_count" : 241,  
      "score" : 11949.740948720284,  
      "bg_count" : 241  
    },  
    {  
      "key" : "s&p global inc. using abcpdf",  
      "doc_count" : 96,  
      "score" : 4733.0721670074,  
      "bg_count" : 97  
    },  
    {  
      "key" : "vnext technologies via abcpdf",  
      "doc_count" : 62,  
      "score" : 3073.1280857436413,  
      "bg_count" : 62  
    }  
  ]  
}
```



wkhtmltopdf / wkhtmltopdf

Public

v0.11.0 rc2:

\* #635: Make /dests an indirect object

v0.12.0 released 2014-02-06

<https://www.nuget.org/packages/ABCpdf.ABCWebKit> :

**ABCpdf.ABCWebKit 12.1.0.7 - NuGet Gallery**

The **ABCpdf** .NET ABCWebKit runtime for HTML to PDF conversion using the **WebKit** rendering engine. The ABCWebKit engine uses a signed version of WkHtmlToPdf.

# Syntax, erm, flexibility: /Root->/Dests with direct object

## Why we need the Arlington DOM!

```
Warning: possibly wrong value (predicates NOT supported): Size ("FileTrailer")
should be: integer [fn:Eval(@Size>0)] and is integer==174 (direct-obj)
Trailer->Root
Error: not an indirect reference as required: Dests ("Catalog")
Trailer->Info
Trailer->Root->Pages
Warning: possibly wrong value (predicates NOT supported): Count
("PageTreeNodeRoot") should be: integer [fn:Eval(@Count>=0)] and is integer==25
(direct-obj)
Warning: unknown key 'ProcSet' is not defined in Arlington for PageTreeNodeRoot
Trailer->Root->Dests
Trailer->Root->Outlines
Trailer->Root->Pages->Kids
Trailer->Root->Dests->__WKANCHOR_2 (as DestXYZ)
Warning: possibly wrong value (predicates NOT supported): 0 ("DestXYZ") should
be: dictionary;number [];[fn:Eval(@0>=0)] and is integer==0 (direct-obj)
Trailer->Root->Outlines->First
```



wkhtmltopdf

```
Trailer->Root->Pages->Kids
Trailer->Root->Names->Dests->[cite.3MOT (as DestDict)]
Trailer->Root->Names->Dests->[cite.Bergeman (as DestDict)]
Trailer->Root->Names->Dests->[cite.BigelowMolecules (as DestDict)]
Trailer->Root->Names->Dests->[cite.Bijlsma (as DestDict)]
Trailer->Root->Names->Dests->[cite.Bongs (as DestDict)]
Trailer->Root->Names->Dests->[cite.BongsLattice (as DestDict)]
Trailer->Root->Names->Dests->[cite.BongsMol (as DestDict)]
Trailer->Root->Names->Dests->[cite.Burke (as DestDict)]
```



dvips + GPL  
Ghostscript

# Next Steps

# Next Steps

Refactor tool wrappers/integration

Identify final features

Releasing “observatory in a box”

# References

Tika Eval: <https://cwiki.apache.org/confluence/display/TIKA/TikaEval>

Evaluating Content Extraction:

<https://www.pdfa.org/presentation/evaluating-text-extraction-at-scale/>

Contact info:

[timothy.b.allison@jpl.nasa.gov](mailto:timothy.b.allison@jpl.nasa.gov), [tallison@apache.org](mailto:tallison@apache.org) @\_tallison



**Jet Propulsion Laboratory**  
California Institute of Technology

---

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

# Extras

# PDFs by Country -- which countries stand out?

Country	Total URLs
Switzerland	17,335
India	12,334
Russia	150,865
Germany	258,775
China	42,928
NULL	170,749
Canada	124,448
Colombia	909
Italy	46,187
Luxembourg	1814



# PDFs by Country

Country	Total URLs	PDFs Observed
Switzerland	17,335	159
India	12,334	100
Russia	150,865	167
Germany	258,775	934
China	42,928	17
NULL	170,749	272
Canada	124,448	191
Colombia	909	13
Italy	46,187	191
Luxembourg	1814	18

# PDFs by Country

Country	Total URLs	PDFs Observed	PDFs Expected
Switzerland	17,335	159	45
India	12,334	100	32
Russia	150,865	167	392
Germany	258,775	934	673
China	42,928	17	112
NULL	170,749	272	444
Canada	124,448	191	324
Colombia	909	13	2
Italy	46,187	191	120
Luxembourg	1814	18	5

**Expected =  
0.26% \* Total URLs**

# PDFs by Country

Country	Total URLs	PDFs Observed	PDFs Expected	chi value	Higher/Lower
Switzerland	17,335	159	45	288	Higher
India	12,334	100	32	144	Higher
Russia	150,865	167	392	129	Lower
Germany	258,775	934	673	101	Higher
China	42,928	17	112	80	Lower
NULL	170,749	272	444	67	Lower
Canada	124,448	191	324	54	Lower
Colombia	909	13	2	48	Higher
Italy	46,187	191	120	42	Higher
Luxembourg	1814	18	5	37	Higher

# PDFs by Top Level Domain (TLD)

TLD	Total URLs	PDFs Observed	PDFs Expected	chi value	Higher/Lower
gov	10,722	239	28	1599	Higher
org	176,671	923	459	468	Higher
com	1,452,337	2517	3776	420	Lower
ru	154,378	125	401	190	Lower
de	138,341	587	360	144	Higher
ch	19,239	129	50	125	Higher
in	14,586	106	38	122	Higher
us	6,486	56	17	91	Higher
jp	59,521	271	155	87	Higher
edu	35,249	170	92	67	Higher

# Generalizations per language -- sum of common tokens

Language	Tika	mutool	pdftotext
deu	469,218	427,576	463,817
eng	4,591,875	4,499,390	4,533,080
fra	292,285	286,659	288,634
ita	124,081	193,515	190,452
jpn	268,508	264,195	254,348
kor	128,923	128,954	128,656
nld	156,353	144,415	155,143
por	232,705	232,823	232,965
slv	103,489	102,626	95,068
spa	377,255	371,049	378,145

# Select Right-to-Left Languages, Sum of Common Tokens

**Example only. NOT SUFFICIENT DATA TO GENERALIZE!**

	Tika	mutool	pdftotext
ara	551	118	1112
fas	3222	11	1449
heb	4969	287	5076

## Side note: Bugtrackers -- November 2020 Crawl

- 35 issue trackers
- 32 tools (3 tools have 2 issue trackers -- legacy and current)
- 1.2 million files (311 GB)
- PDF-centric subset ~33k files

Data

[https://corpora.tika.apache.org/base/docs/bug\\_trackers/](https://corpora.tika.apache.org/base/docs/bug_trackers/)  
[https://corpora.tika.apache.org/base/packaged/pdfs/pdfs\\_202011/](https://corpora.tika.apache.org/base/packaged/pdfs/pdfs_202011/)

And Peter Wyatt's posts:

<https://www.pdfa.org/a-new-stressful-pdf-corpus/>  
<https://www.pdfa.org/stressful-pdf-corpus-grows/>

# Stored vs. Refetched

Total: 10 TB (distinct files)

Stored in CC: 1.6 TB (not distinct)

Refetched: 9.8 TB (not distinct)



# Process

- Identify PDF files via Common Crawl index files (240 GB compressed)
- Fetch the non-truncated files from Common Crawl
- Refetch the truncated files from URLs in Common Crawl
- Store files by their hashes in AWS S3, e.g. `cc-MAIN-2021-31/00/00/00000836a41510292765a86e57a80f1e182360ca5c6ff0841482b31c8f25ab84`

# Duplicates

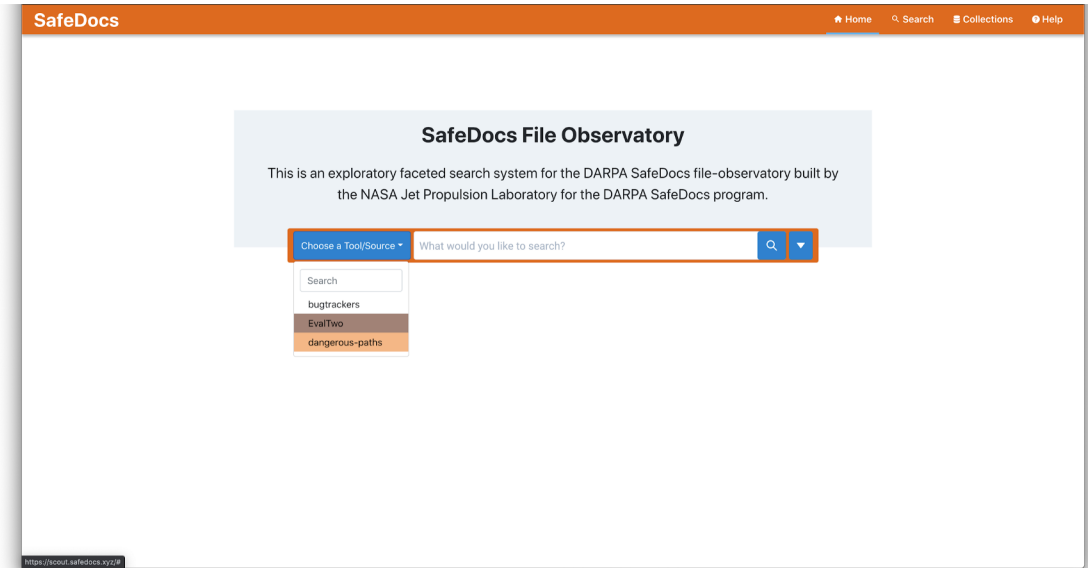
- 8.3 million PDFs
- 7.9 distinct hashes
- 370k exact duplicates
- Few handfuls of less than optimal url design, e.g. 744 copies of the same file

<b>fetchd_digest</b> character varying (64)	 <b>url</b> character varying (10000)
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=0gYyThp2t8SCTXhwyVol
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=0jsKniDgXHpc2o9Q5Pw
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=0KYNbOWNwvj3009cWh
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=0l4eZzuEsUqo60s8JoMh
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=0NwyAr3lO-zYmRnsvrg79
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=0Q8AMK1TljLC75F6Sd2l
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=11YWObbeYAvEBVOPMs
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=15ojJgB8lLu3krvd3jhhqV
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=1DUWn0bf1virvsf9ySmin
ac6a16507cd6f0727ee253fa39bbe481d1cfd8becc8fd5873e204d1a82747610	https://www.cellc.co.za/cellc/static-content/PDF/Acceptable_Use_Policy.pdf;jsessionid=1eYBSrWPKHkXS6ANqlh

# Search and Analytics at Scale: Observatory UI

# Search UI

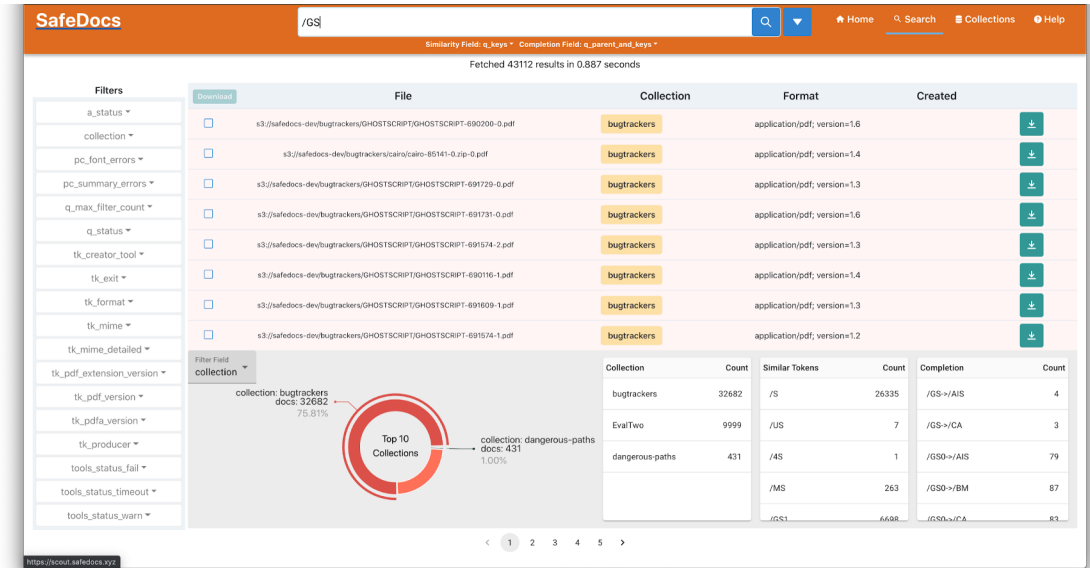
- Interactive way to search through the file-observatory, visualize content, and download files of interest
- Built with a large-volume of data in mind
- Currently live and working



# Search UI

## Search/Results Page

- Dynamic filtering based on 18 different fields and search text
- Allows for downloading of individual or multiple PDFs
- Provides table of similar tokens within a 2 character distance
- Provides table of suggested completions and the number of documents in which they occur
- Dynamic filter field breakdown visualization



# Search UI

## File View

- Provides in-depth view of all fields stored in elasticsearch for a document by clicking on name

... ..
s3://safedocs-corpa/a2/75/a2758a930605fc89a32a5b176851daaaddeed63f1ef363bec1180f1fe0e5aa8e
original_name
jpl_demo201912/govdocs/061/061164.pdf
collection
jpl_demo201912
size
1440276
shasum_256
a2758a930605fc89a32a5b176851daaaddeed63f1ef363bec1180f1fe0e5aa8e
tk_status
success
tk_mime
application/pdf
tk_mime_detailed
application/pdf
tk_format
application/pdf; version=1.4
tk_pdf_version

# Search UI

## Collections Page

- Provides a real-time view of all collections in the file-observatory
- Pulls the latest ElasticSearch file-observatory mapping schema with all available fields

Property	Type	Analyzer
a_status	keyword	
a_warn	text	text_basic
c_status	keyword	
cd	text	text_basic
cd_status	keyword	
cd_warn	text	text_basic
clamav	text	text_basic
collection	keyword	
cpu_status	keyword	
cpu_warn	text	text_basic
fname	keyword	
mc_status	keyword	
mc_warn	text	text_basic
mt_status	keyword	
mt_warn	text	text_basic
original_fname	keyword	
pc_font_errors	keyword	
pc_summary_errors	keyword	
pc_summary_info	text	text_basic
pinfo	text	text_basic
pinfo_creator	text	text_basic
pinfo_encrypted	boolean	
pinfo_javascript	boolean	
pinfo_optimized	boolean	