



# Building a File Observatory for Secure Parser Development

*LangSec 2021*

May 27, 2021

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.



**Jet Propulsion Laboratory**  
California Institute of Technology

© 2021 California Institute of Technology. Government sponsorship acknowledged.

# The Team



Chris Mattmann  
PI; Division Mgr,  
AI AID Org



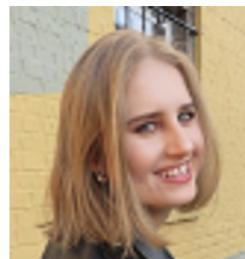
Tim Allison  
Files and Search



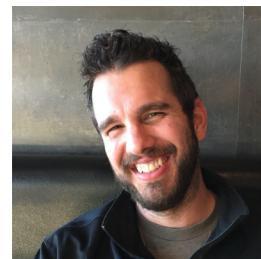
Wayne Burke  
Cognizant Engineer



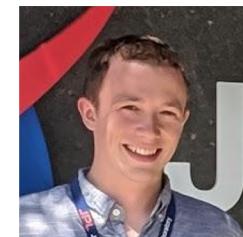
Michael Fedell  
Data Scientist



Anastasia Menshikova  
Data Scientist



Phil Southam  
Trouble (Fun?)  
Maker



Ryan Stonebraker  
Data Scientist  
Alaskan

# **Debts of Gratitude**

Sergey Bratus

Peter Wyatt and Duff Johnson, PDF Association

Dan Becker, John Kansky and team at Kudu Dynamics

Trail of Bits, Galois, BAE and SRI

# Outline

1. Motivation and Corpora
2. Fuzzy matching/advanced search
3. API/UI
4. Classification of Creator Tool
5. Next Steps

# Motivation and Corpora

© 2021 California Institute of Technology. Government sponsorship acknowledged.

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

# Motivation

- Inducing grammars
- Devtesting parsers during development
- Testing/profiling/tracing existing parsers
  - Literal files
  - Seeds for fuzzing

# Corpora

- Common Crawl
- Bug tracker corpora

# Common Crawl



- Monthly open source crawls of large portions of the web: for May 2021, 2.6 billion pages (280 TB).
- Available via Amazon Web Services Public Datasets
- Searchable indexes available

<https://commoncrawl.org>

## Common Crawl -- known limitations

- Files truncated at 1MB
- Coverage: ~convenience sample

# 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

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

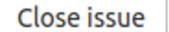
[https://corpora.tika.apache.org/base/packaged/pdfs/pdfs\\_202011/](https://corpora.tika.apache.org/base/packaged/pdfs/pdfs_202011/)

# Bugtrackers -- March 2021 Crawl (jpeg)

- 13 issue trackers
- 7k files (3.3 GB)



The screenshot shows a bug tracker interface for the GNOME project, specifically for the gdk-pixbuf module. The URL is [GNOME > gdk-pixbuf > Issues > #176](#). The issue is titled "OOMs and infinite loops on known problematic jpeg". It was created 2 months ago by Tim Allison. The description text discusses problems found while crawling jpeg parser issue trackers, mentioning OOMs and infinite loops on known problematic files. It also notes that gdk-pixbuf was tying up processors. A command line example is provided: `/usr/bin/gdk-pixbuf-thumbnailer -s 256 input output`. The footer of the screenshot includes a link to the issue's edit page.

Open  Created 2 months ago by  Tim Allison  

## OOMs and infinite loops on known problematic jpeg

 I recently crawled a bunch of jpeg parser issue trackers to grab problematic files for fuzzing and found that some known problematic files and variants of them are problematic for gdk-pixbuf. I did not set out to break gdk-pixbuff, but while watching other jpeg parsers have problems, I noticed that gdk-pixbuff was also tying up processors. 😅

The command I used: `/usr/bin/gdk-pixbuf-thumbnailer -s 256 input output`

I used this as above without `--g-fatal-warnings` because that's what was being called by Ubuntu in the background.

Here are some of the problematic files.

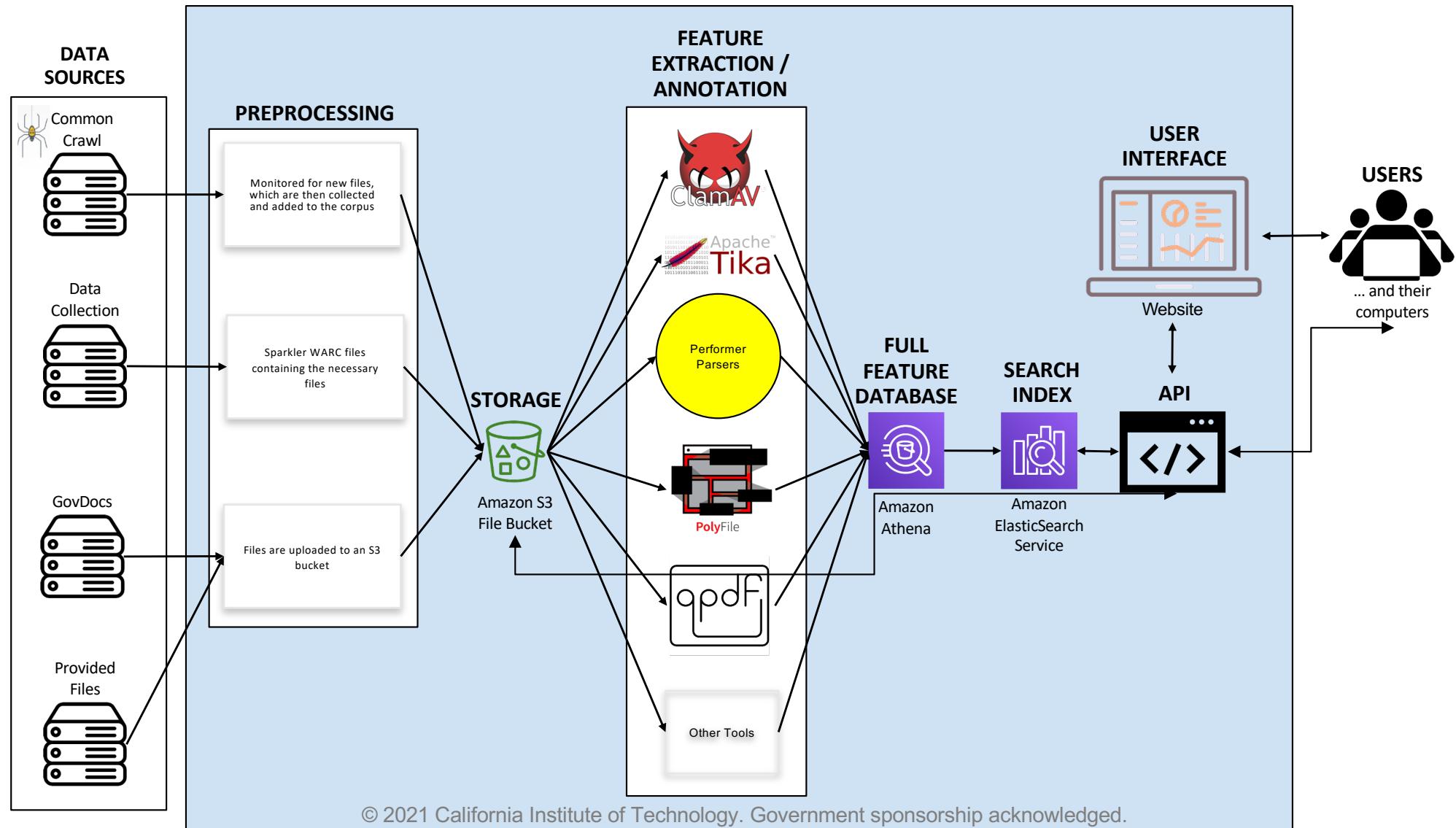
# Observatory at 2 scales

© 2021 California Institute of Technology. Government sponsorship acknowledged.

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

# Various Scales

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

© 2021 California Institute of Technology. Government sponsorship acknowledged.

## 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

jpl.nasa.gov

# Standard table for each tool

Explain    Query Editor

```
1 select path, exit_value, process_time_ms, stdout, stderr from caradoc
```

See "EXPLAIN (FORMAT JSON) [QUERY]."

Data Output

	path [PK] character varying (500)	exit_value integer	process_time_ms bigint	stdout character varying (20000)	stderr character varying (20000)
1	EvalTwo/dangerous/01b6856a7e6...	0	58	Version : 1.4	Warning : Flate/Zlib stream with appended newline in o...
2	EvalTwo/dangerous/023e918b4e1...	2	53	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
3	EvalTwo/dangerous/0291076dcac...	2	56	Not a PDF file	PDF error : Xref does not start at object 0 in object traile...
4	EvalTwo/dangerous/019734dc8e7...	2	24	Version : 1.5	Warning : Flate/Zlib stream with appended newline in o...
5	EvalTwo/dangerous/025aecde10b...	2	30	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
6	EvalTwo/dangerous/026929a0433...	2	296	Not a PDF file	PDF error : Lexing error : unexpected word at offset 493 ...
7	EvalTwo/dangerous/00af2c403b20...	2	56	Version : 1.7	Type error : Invalid type : expected s"metadata_stream" i...
8	EvalTwo/dangerous/00ee0b8dba2...	2	113	Not a PDF file	PDF error : Lexing error : unexpected word at offset 53 [...
9	EvalTwo/dangerous/007e86e378e...	2	14	Version : 1.4	Type error : Unexpected entry /ITXT in instance of class ...
10	EvalTwo/dangerous/02935e7a7e2...	2	20	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
11	EvalTwo/dangerous/02fc5ed61c7e...	0	39	Version : 1.3	Warning : Flate/Zlib stream with appended newline in o...
12	EvalTwo/dangerous/02ed8c4bc27...	2	20	Version : 1.4	PDF error : Invalid stream filter : AHx in object 4 at entry ...
13	EvalTwo/dangerous/00eb20ea546...	2	29	Not a PDF file	PDF error : Lexing error : unexpected character : 0x25 at...
14	EvalTwo/dangerous/02aedcb47ea...	2	85	Version : 1.7	Type error : Invalid variant type : expected (c"pagenode" ...
15	EvalTwo/dangerous/02aedcb47ea...	0	67	Version : 1.7	...

# Discovery with built-in Elasticsearch features

# Search – Information Leakage

New Save Open Share Inspect

Filters q\_keys\_and\_values:\*\Users\/\* Lucene

+ Add filter

file-observatory-202103

**Selected fields**

? \_source

**Available fields**

\_source

> q\_keys\_and\_values: /AIS->FALSE, /AU->REF, /AU->  
file:///C|/Users/[REDACTED]Desktop/html\_for\_pdf/5\_images\_for\_pdf/160x600/160x600.html, /AU->  
file:///C|/Users/[REDACTED]Desktop/html\_for\_pdf/5\_images\_for\_pdf/728X90/728x90.html, /AcroForm->  
>/DeviceRGB, /Annots->REF, /Article->/Art, /Ascent->NUMBER, /BM->/Normal, /BaseFont->/Arial-Black,

# Using Elasticsearch's Prefix Completion

```
GET file-observatory-202102/_search🔧
{
  "_source": false,
  "suggest": {
    "YOUR_SUGGESTION": {
      "prefix": "/Rotate->",
      "completion": {
        "field": "q_keys_and_values
.completion",
        "size": 2000,
        "skip_duplicates": true
      }
    }
  }
}
```

```
"options" : [
  {
    "text" : "/Rotate->-10",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/PDFBOX/PDFBOX-3353-0.pdf",
    "_score" : 1.0
  },
  {
    "text" : "/Rotate->-13",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/PDFIUM/PDFIUM-1380-1.pdf",
    "_score" : 1.0
  },
  {
    "text" : "/Rotate->-180",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-689418-0.zip-0
.pdf",
    "_score" : 1.0
  },
  {
    "text" : "/Rotate->-90",
    "_index" : "file-observatory-202102",
    "_type" : "_doc",
    "_id" : "s3://safedocs-dev/bugtrackers/PDFBOX/PDFBOX-3785-0.pdf",
    "_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  
  },  
  {  
    "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  
  }]
```

# Significant Terms Query for /SubType

```
GET file-observatory-202103/_search
{
  "query" : {
    "match" : {
      "q_keys" : {
        "query": "\SubType"
      }
    }
  },
  "size" : 0,
  "aggregations" : {
    "significant_queries" : {
      "significant_text" : {
        "field" : "tk_creator_tool",
        "size": 10,
        "chi_square": {
          "background_is_superset" : false
        }
      }
    }
  }
}
```

```
22 ▾
23 ▾
24
25
26
27
28 ▾
29 ▾
30
31
32
33
34 ▾
35 ▾
36
37
38
39
40 ▾
41 ▾
42
43
44
45

BUCKETS . L
{
  "key" : "arcmap",
  "doc_count" : 23,
  "score" : 3740.5023984417203,
  "bg_count" : 36
},
{
  "key" : "esri",
  "doc_count" : 23,
  "score" : 3677.4677551236864,
  "bg_count" : 37
},
{
  "key" : "6.0",
  "doc_count" : 50,
  "score" : 2755.4940054732288,
  "bg_count" : 322
},
{
  "key" : "prinectsignastation",
  "doc_count" : 6,
  "score" : 1073.003716686631,
  "bg_count" : 8
}
```

# Significant keys for creator tools “arcmap” or “esri”

```
GET file-observatory-202103/_search
{
  "query" : {
    "match" : {
      "tk_creator_tool" : {
        "query": "arcmap OR esri"
      }
    }
  },
  "size" : 0,
  "aggregations" : {
    "significant_queries" : {
      "significant_text" : {
        "field" : "q_keys",
        "size": 50,
        "chi_square": {
          "background_is_superset" : false
        }
      }
    }
  }
}
```

```
GET file-observatory-202103/_search
```



```
21
22   "by_count" : 0.0119,
23   "buckets" : [
24     {
25       "key" : "/0C0",
26       "doc_count" : 19,
27       "score" : 10762.751337739273,
28       "bg_count" : 25
29     },
29     {
30       "key" : "/WKT",
31       "doc_count" : 15,
32       "score" : 8431.315225385713,
33       "bg_count" : 20
34     },
34     {
35       "key" : "/0C4",
36       "doc_count" : 14,
37       "score" : 8033.2824452632785,
38       "bg_count" : 18
39     },
39     {
40       "key" : "/ESRINorth",
41       "doc_count" : 13,
42       "score" : 7917.152340761275,
43       "bg_count" : 15
44     },
44     ...
45   }
```



```
"key" : "/SubType",
"doc_count" : 23,
"score" : 4766.166123231515,
"bg_count" : 122
```

# 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

# Observatory API

© 2021 California Institute of Technology. Government sponsorship acknowledged.

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

# Observatory API

## SafeDocs Observatory API 1.0.0 OAS3

[/openapi.json](#)

Search and Retrieval API for the file observatory built by NASA JPL for the DARPA SafeDocs program.

### Search

`GET /v1/elasticsearch/{index}` Es Search

### Retrieval

`GET /v1/files` Get Files

- Consists of two endpoints;
  - Elasticsearch pass-through
  - File Retrieval

# Observatory API – Elasticsearch pass through

- Acts as a pass-through for any query through the API's query parameter
- Returns the total number of document matches
- Upcoming:
  - Will also be accepting json queries through the request body

Search

GET /v1/elasticsearch/{index} Es Search

Arguments  
index: the Elasticsearch index to be searched.  
query: the JSON format Elasticsearch query.

Returns: A dictionary that contains the result of a search for the initial set of results.

Parameters

Name Description

index \* required string (path)  
query \* required string (query)

Server response

Code Details

200 Response body

```
{ "took": 9, "timed_out": false, "shards": { "total": 2, "successful": 2, "skipped": 0, "failed": 0 }, "hits": { "total": { "value": 27941, "relation": "eq" }, "max_score": 1, "hits": [ { "index": "file-observatory", "type": "doc", "id": "265f887fbcbadeb3ae96137517a5df15787781eaef9c7248efbb95ec93d54cecd", "score": 1, "source": { "file": "265f887fbcbadeb3ae96137517a5df15787781eaef9c7248efbb95ec93d54cecd", "original_fname": "jpl_demo201912/govdocs/234/234540.pdf", "collection": "jpl_demo201912", "size": 244108, "shnum": 27941, "shnum_256": 265f887fbcbadeb3ae96137517a5df15787781eaef9c7248efbb95ec93d54cecd" } } ] }
```

Download

Response headers

```
content-length: 3177
content-type: application/json
date: Tue, 26 May 2020 23:03:36 GMT
server: unicorn
```

# Observatory API – s3 file retrieval

- Input: List of S3 urls returned by the “fname” parameter of ES.
- Returns a zip folder containing all the requested files

## Retrieval

The screenshot shows a configuration interface for the 'Retrieval' endpoint. At the top, it displays a 'GET' method and the URL '/v1/files Get Files'. Below this, there is a section for 'Arguments' with a note about 'paths' and 'Returns'. Under 'Parameters', there is a table with two entries: 'paths' (array[string]) containing 's3://safedocs-corpa/ff/ea/feab23ae1191d83a' and 's3://safedocs-corpa/00/00/000000cbd1d90c'. An 'Add item' button is available to add more entries. A 'Cancel' button is located in the top right corner of the parameters section.

Name	Description
paths array[string] (query)	s3://safedocs-corpa/ff/ea/feab23ae1191d83a s3://safedocs-corpa/00/00/000000cbd1d90c

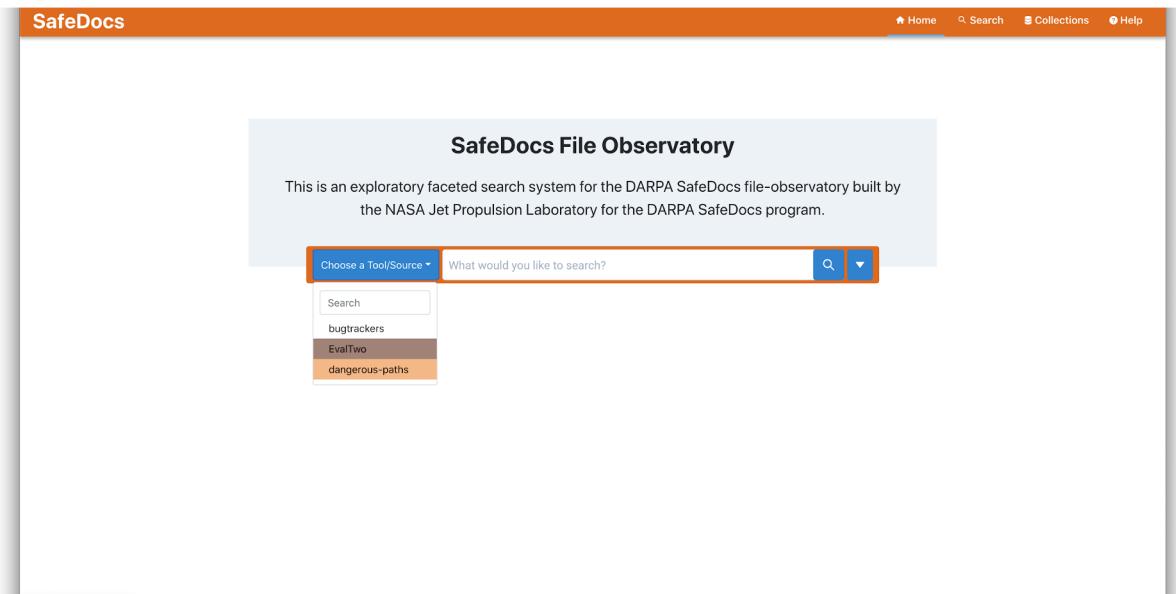
# Search and Analytics at Scale: Observatory UI

© 2021 California Institute of Technology. Government sponsorship acknowledged.

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

# Search UI

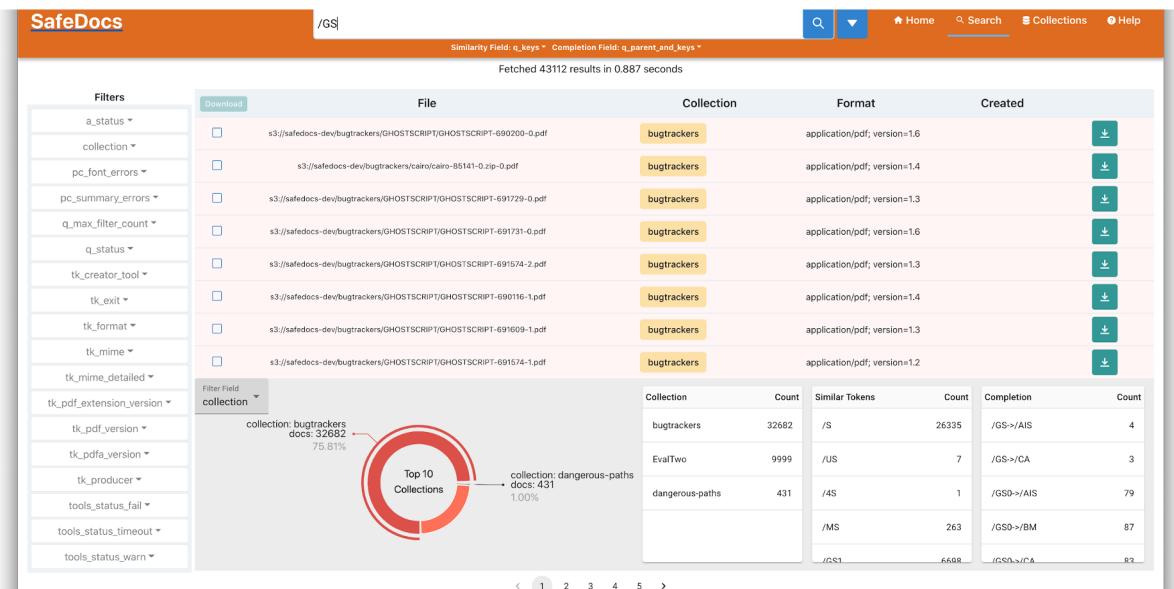
- Interactive way to search through the file-observatory, visualize content, and download files of interest
- Built with a large-volume of data in mind
- Currently live and working



# 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

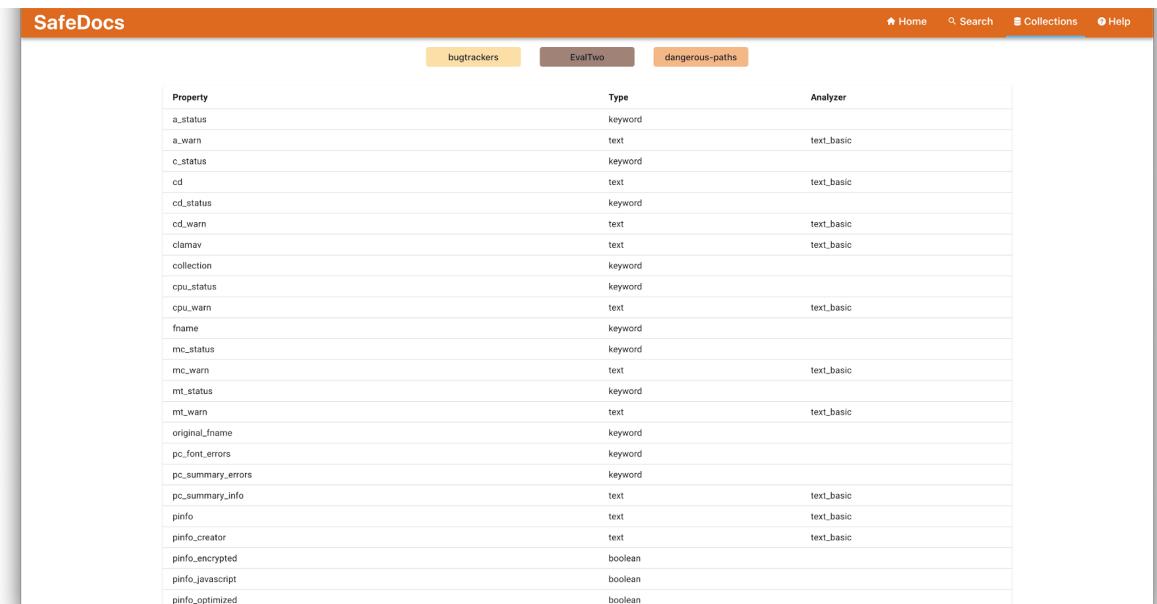
The screenshot shows a search results page for a PDF document. At the top right, there are navigation links: Home, Search, Collections, and Help. The main content area is titled "SafeDocs". It displays various metadata fields for a document:

- fname: [s3://safedocs-corpa/a2/75/a2758a930605fc89a32a5b176851daeaddeed63f1ef363bec1180fffe0e5aa8e](#)
- original\_fname: [jpl\\_demo201912/govdocs/061/061164.pdf](#)
- collection: [jpl\\_demo201912](#)
- size: 1440276
- shasum\_256: [a2758a930605fc89a32a5b176851daeaddeed63f1ef363bec1180fffe0e5aa8e](#)
- tk\_status: success
- tk\_mime: application/pdf
- tk\_mime\_detailed: application/pdf
- tk\_format: application/pdf; version=1.4
- tk\_pdf\_version: 1

# 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



The screenshot shows a table of field mappings for the 'bugtrackers' collection. The table has three columns: Property, Type, and Analyzer. The 'Analyzer' column contains 'text\_basic' for most fields, except for 'a\_status', 'a\_warn', 'c\_status', 'cd', 'cd\_status', 'cd\_warn', 'clamav', 'collection', 'cpu\_status', 'cpu\_warn', 'fname', 'mc\_status', 'mc\_warn', 'mt\_status', 'mt\_warn', 'original\_fname', 'pc\_font\_errors', 'pc\_summary\_errors', 'pc\_summary\_info', 'pinfo', 'pinfo\_creator', 'pinfo\_encrypted', 'pinfo\_javascript', and 'pinfo\_optimized', which have 'text' type.

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	

# **Document Classification Task**

© 2021 California Institute of Technology. Government sponsorship acknowledged.

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

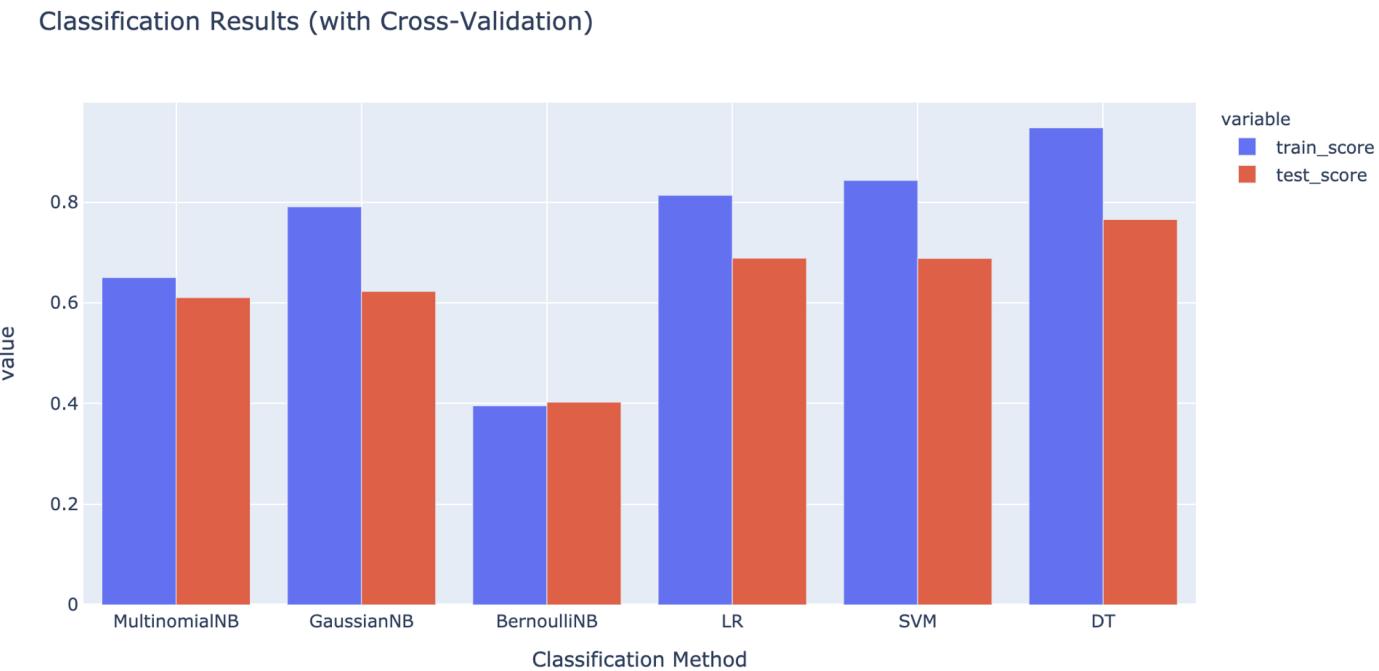
# Document Classification Task: Setup

- Classifying the *creator tool* based on *keys*
  - 2021 dataset: q keys, key-value pairs, parent-key pairs
- Top 10 most occurring creator tools
- Splitting the dataset 80:20
- Label = creator tool, features = keys
- Using CountVectorizer to vectorize the data
- Testing out different classifiers (with cross-validation):
  - MultinomialNB
  - GaussianNB
  - Bernoulli NB
  - Logistic Regression
  - SVM
  - Decision Trees

# Document Classification Task: Data

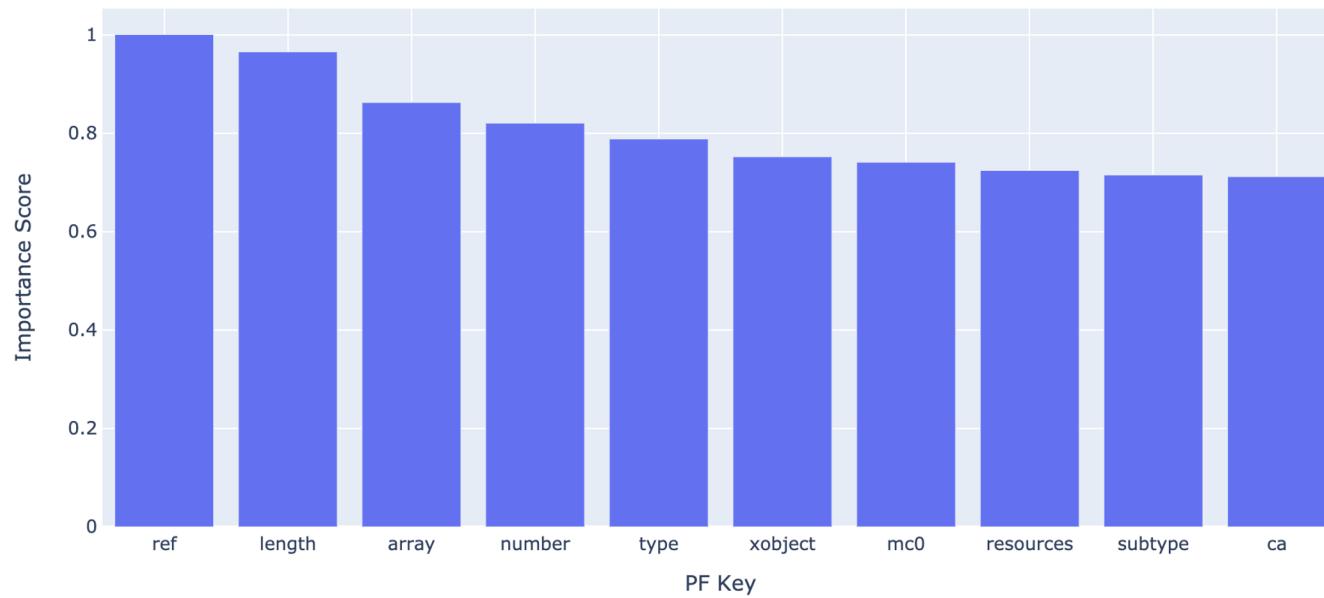
		id	tool	keys
0	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-688979-0.pdf		PScriptdll	[], A', 'Annots', 'Ascent', 'Author', 'BG...
2	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-696715-0.pdf		PScriptdll	[], Ascent', 'Author', 'BaseFont', 'BitsPe...
5	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-690548-2.r...		PScriptdll	[], Ascent', 'Author', 'BaseFont', 'BitsPe...
9	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-691045-0.b...		Adobe InDesign CS	[], Ascent', 'BaseFont', 'BitsPerComponent'...
14	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-692988-0.pdf		Adobe Illustrator CS	[], Ascent', 'BaseFont', 'BitsPerComponent'...
...	...	...	...	...
1224	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-699947-0.t...		TeX	[], Ascent', 'BaseFont', 'BitsPerComponent'...
1240	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-693588-1.pdf	Microsoft Office Word		[], Ascent', 'Author', 'BaseFont', 'CapHei...
1253	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-695334-0.pdf		PScriptdll	[], Ascent', 'Author', 'BaseFont', 'CapHei...
1263	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-697350-0.pdf		Adobe InDesign CS	[], Author', 'BaseFont', 'BitsPerComponent'...
1266	bugtrackers/GHOSTSCRIPT/GHOSTSCRIPT-695390-0.pdf		TeX	[], ColorSpace', 'Contents', 'Count', 'Cre...

# Document Classification Task: Results



# Document Classification Task: Feature Importance (an example, for MultinomialNB)

Adobe Illustrator CS Top 10 Features (PF Keys)



# Next Steps

© 2021 California Institute of Technology. Government sponsorship acknowledged.

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

# Next Steps

Settling on features for extraction

Simplifying code to enable faster integration of new tools

Releasing “observatory in a box”



**Jet Propulsion Laboratory**  
California Institute of Technology

---

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

# Extras

# Map of PDFs in the July 2020 Common Crawl Data -- GeoLocation of URL/IPs via MaxMind's GeoIP City Database



© 2021 California Institute of Technology. Government sponsorship acknowledged.

jpl.nasa.gov

# Top 10 “Top Level Domains” for PDFs in the July 2020 Common Crawl Data

