

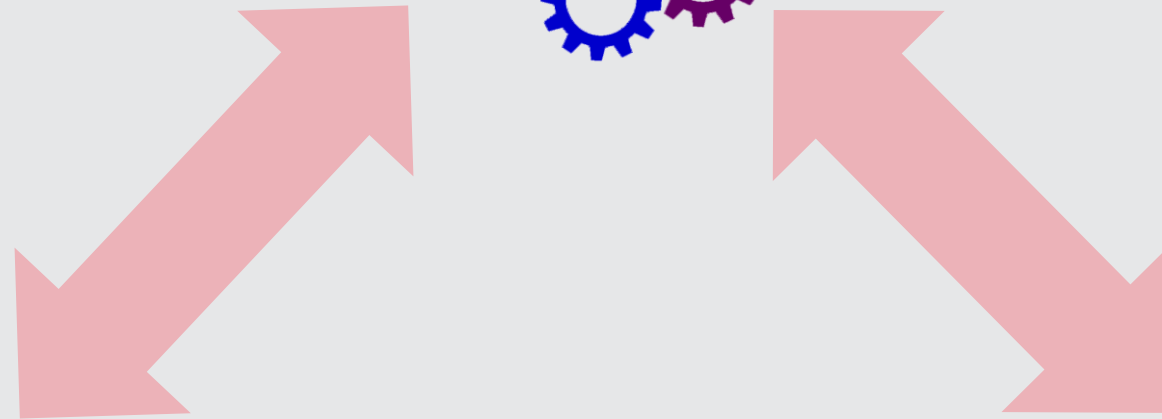
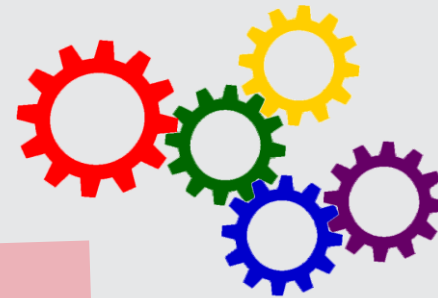


# “SafeDocs” Update to industry

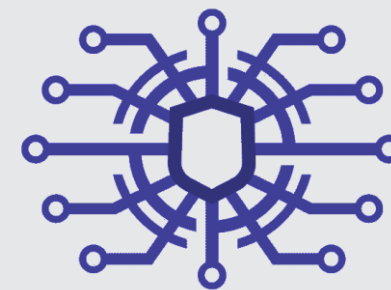
Where PDF, parsing and cyber-security intersect

**OctoberPDFest**  
**ONLINE**

## Parsers



**Extant Data**  
("in the wild")



**Cybersecurity**

<https://www.darpa.mil/program/safe-documents>

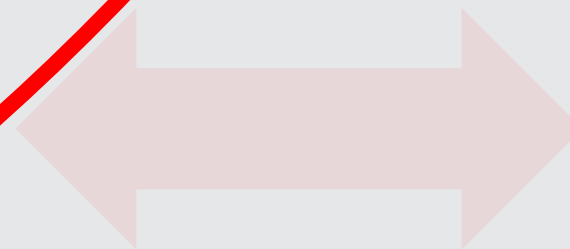
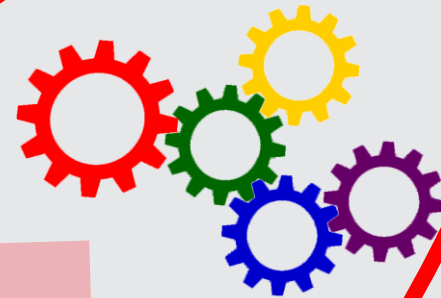
Copyright © 2020, PDF Association

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0079.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Approved for public release.

# Parsers & extant data

Parsers



Cybersecurity

Extant Data  
("in the wild")





- Limitations of internet search corpora
  - Common Crawl data is truncated
  - Internet search results are pre-filtered
  - Internet search results vary greatly by search engine
- New stressful “Issue Track” PDF Corpus  
<https://corpora.tika.apache.org/base/packaged/pdfs/>

Issue Tracker	No. of PDFs	Size
GhostScript	5,279	5.3 GB
Libre Office	5,183	1.3 GB
Mozilla pdf.js	2,400	4.9 GB
OCR-my-PDF	187	437 MB
Apache Open Office	3,109	688 MB
Open PDF	31	3.2 MB
Apache PDFBOX	3,577	2.6 GB
Chromium pdfium	379	212 MB
POI	10	908 KB
qpdf	68	33 MB
Sumatra PDF	213	487 MB
Apache TIKA	140	140 MB
<b>TOTAL</b>	<b>20,576</b>	<b>16 GB</b>

<https://www.pdfa.org/a-new-stressful-pdf-corpus/>  
"Building a Wide Reach Corpus for Secure Parser Development" by Allison et. al. <http://spw20.langsec.org/papers.html>



<https://www.eff.org/observatory>  
<https://digitalcorpora.org/corpora/files>

# /Type vs /type

86 out of ~550,000 PDFs

Case sensitive key  
name search

Instant summary of  
PDF Creator metadata  
(as determined by Apache Tika)

Filenames identify sub-  
corpus (e.g. GovDocs1)

The screenshot shows the 'Discover' interface with the following components:

- Discover** header with a search icon and a close button.
- 86 hits** displayed below the header.
- Actions:** New, Save, Open, Share, Inspect.
- Filters:** +pf\_keys.case:\type (Lucene search engine, Refresh button).
- Selected fields:**
  - original\_fname**: A list of PDF file paths starting with 's3://safedocs-prod-raw-file-input/govdocs1/...'.
  - tk\_creator\_tool**: A list of PDF creator tools, including 'I.R.I.S.', 'Adobe InDesign CS2 (4.0.5)', 'PScript5.dll Version 5.2', 'Smart touch 1.3', and '-'. A callout points to this field, stating 'Instant summary of PDF Creator metadata (as determined by Apache Tika)'.
- Top 5 values in 66 / 86 records:**
  - I.R.I.S. (36.4%)
  - Appligent pdfHarmony 2.0 (31.8%)
  - Readiris Build 5965 (6.1%)
  - PScript5.dll Version 5.2.2 (6.1%)
  - Smart touch 1.3 (4.5%)
- Available fields:**
  - Popular: fname, pf\_filters, pf\_keys, pi\_content.

# /Type vs /type

```
2485 /F10 24 0 R
2486 /F20 29 0 R
2487 /F30 34 0 R
2488 /F40 39 0 R
2489 /F50 44 0 R
2490 >>
2491 endobj
2492 46 0 obj
2493 <<
2494 /Rect [225 135 318 126]
2495 /Subtype /Link
2496 /Type /Annot
2497 /Border [0 0 0]
2498 /A <<
2499 /S /URI
2500 /type /Action
2501 /URI (mailto:e-miltangp@tc.gc.ca)
2502 >>
2503 >>
2504 endobj
2505 47 0 obj
2506 [46 0 R]
2507 endobj
2508 48 0 obj
2509 <</Application <5363616E2053746174696F66
```

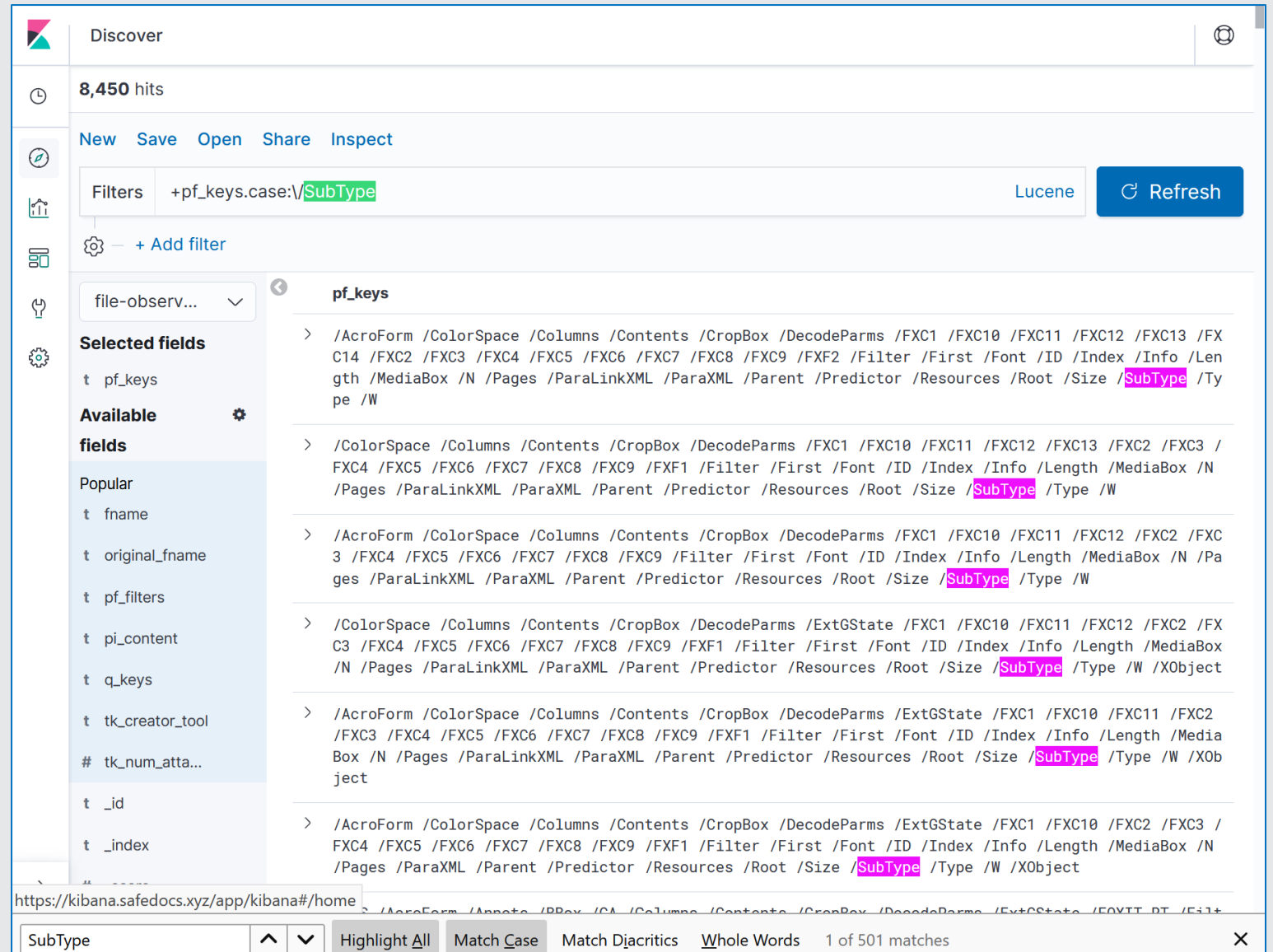
```
6420 endobj
6421 44 0 obj
6422 [45 0 R]
6423 endobj
6424 45 0 obj
6425 <<
6426 /Rect [172 117 232 110]
6427 /Subtype /Link
6428 /Type /Annot
6429 /Border [0 0 0]
6430 /A <<
6431 /S /URI
6432 /type /Action
6433 /URI (mailto:building.bridges@)
6434 >>
6435 >>
6436 endobj
6437 46 0 obj
6438 <<
6439 /Type /Page
6440 /Parent 1 0 R
6441 /MediaBox [0 0 620 814]
```

```
847 32 0 obj
848 <<
849 /Rect [308 679 397 667]
850 /Subtype /Link
851 /Type /Annot
852 /Border [0 0 0]
853 /A <<
854 /S /URI
855 /type /Action
856 /URI (mailto:wildlife@peak.org)
857 >>
858 >>
859 endobj
860 3
861 <
```

Clearly a recurring issue in Link Annotation inline URI Action dictionaries!

# /Subtype vs /SubType

- “/SubType” = 8,450 hits

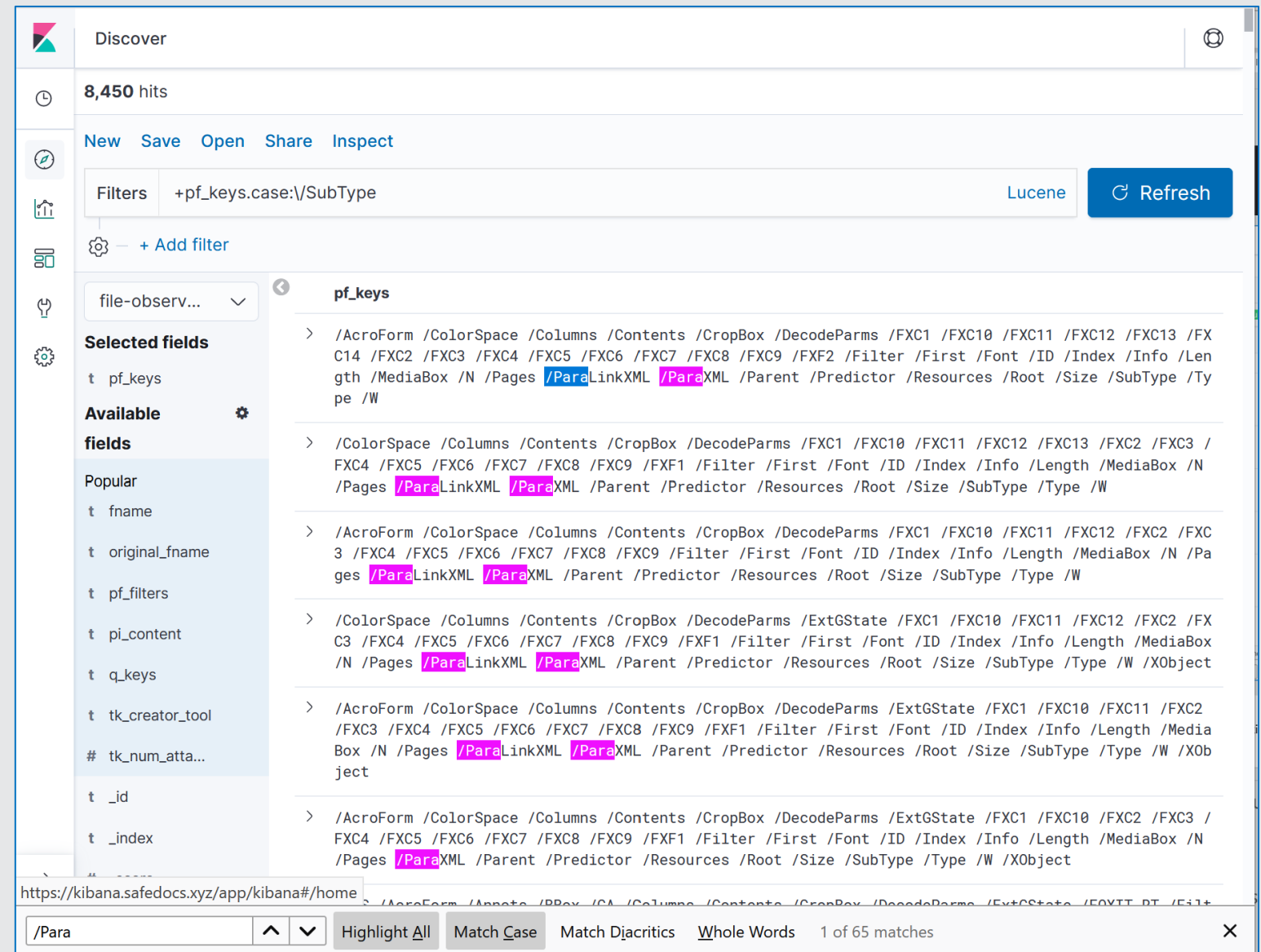


The screenshot shows the Kibana Discover interface. The top bar indicates 8,450 hits. The filter bar shows a filter: `+pf_keys.case:\SubType`. The left sidebar shows the 'Selected fields' list with `pf_keys` selected. The main panel displays a list of search results, each showing a path structure where `/SubType` is highlighted in pink. The bottom status bar shows the search term `SubType` and the number of matches: 1 of 501 matches.



# /Subtype vs /SubType

- “/SubType” = 8,450 hits
- Highlight “/Para...”
  - Proprietary extensions



Discover

8,450 hits

New Save Open Share Inspect

Filters +pf\_keys.case:/SubType Lucene Refresh

+ Add filter

file-observ... ▾

**Selected fields**

- pf\_keys

**Available fields**

Popular

- fname
- original\_fname
- pf\_filters
- pi\_content
- q\_keys
- tk\_creator\_tool
- tk\_num\_atta...
- \_id
- \_index

**pf\_keys**

- > /AcroForm /ColorSpace /Columns /Contents /CropBox /DecodeParms /FXC1 /FXC10 /FXC11 /FXC12 /FXC13 /FXC14 /FXC2 /FXC3 /FXC4 /FXC5 /FXC6 /FXC7 /FXC8 /FXC9 /FXF2 /Filter /First /Font /ID /Index /Info /Length /MediaBox /N /Pages /ParaLinkXML /ParaXML /Parent /Predictor /Resources /Root /Size /SubType /Type /W
- > /ColorSpace /Columns /Contents /CropBox /DecodeParms /FXC1 /FXC10 /FXC11 /FXC12 /FXC13 /FXC2 /FXC3 /FXC4 /FXC5 /FXC6 /FXC7 /FXC8 /FXC9 /FXF1 /Filter /First /Font /ID /Index /Info /Length /MediaBox /N /Pages /ParaLinkXML /ParaXML /Parent /Predictor /Resources /Root /Size /SubType /Type /W
- > /AcroForm /ColorSpace /Columns /Contents /CropBox /DecodeParms /FXC1 /FXC10 /FXC11 /FXC12 /FXC2 /FXC3 /FXC4 /FXC5 /FXC6 /FXC7 /FXC8 /FXC9 /Filter /First /Font /ID /Index /Info /Length /MediaBox /N /Pages /ParaLinkXML /ParaXML /Parent /Predictor /Resources /Root /Size /SubType /Type /W
- > /ColorSpace /Columns /Contents /CropBox /DecodeParms /ExtGState /FXC1 /FXC10 /FXC11 /FXC12 /FXC2 /FXC3 /FXC4 /FXC5 /FXC6 /FXC7 /FXC8 /FXC9 /FXF1 /Filter /First /Font /ID /Index /Info /Length /MediaBox /N /Pages /ParaLinkXML /ParaXML /Parent /Predictor /Resources /Root /Size /SubType /Type /W /XObject
- > /AcroForm /ColorSpace /Columns /Contents /CropBox /DecodeParms /ExtGState /FXC1 /FXC10 /FXC11 /FXC2 /FXC3 /FXC4 /FXC5 /FXC6 /FXC7 /FXC8 /FXC9 /FXF1 /Filter /First /Font /ID /Index /Info /Length /MediaBox /N /Pages /ParaLinkXML /ParaXML /Parent /Predictor /Resources /Root /Size /SubType /Type /W /XObject
- > /AcroForm /ColorSpace /Columns /Contents /CropBox /DecodeParms /ExtGState /FXC1 /FXC10 /FXC11 /FXC2 /FXC3 /FXC4 /FXC5 /FXC6 /FXC7 /FXC8 /FXC9 /FXF1 /Filter /First /Font /ID /Index /Info /Length /MediaBox /N /Pages /ParaXML /Parent /Predictor /Resources /Root /Size /SubType /Type /W /XObject

https://kibana.safedocs.xyz/app/kibana#/home

/Para ^ ▾ Highlight All Match Case Match Diacritics Whole Words 1 of 65 matches X

# /Subtype vs /SubType

- “/SubType” = 8,450 hits
- Highlight “/Para...”
  - Proprietary extensions
- Exclude “/Para\*” = **8,365 hits**
  - No single Creator/Producer

The screenshot shows the 'Discover' interface with the following details:

- Discover** header with a search icon.
- 8,365 hits** displayed below the header.
- Buttons: **New**, **Save**, **Open**, **Share**, **Inspect**.
- Filters** section: `+pf_keys.case:\SubType -pf_keys.case:\Para*` with a **Lucene** dropdown and a **Refresh** button.
- + Add filter** button.
- file-observer...** dropdown menu.
- Selected fields** section: `? ppin_info`.
- Available fields** section with a settings icon:
  - Popular** fields:
    - `t fname`
    - `t original_fname`
    - `t pf_filters`
    - `t pf_keys`
    - `t pi_content`
    - `t q_keys`
    - `t tk_creator_t...`
    - `# tk_num_atta...`
  - Other fields** (partially visible):
    - `t _id`
    - `t _index`
    - `# _score`
- ppin\_info** results table:

>	Creator:	Esri ArcCatalog 10.3.1.4959
	Producer:	Adobe PDF Library 9.0
	CreationDate:	Tue May 17 02:20:33 2016 UTC
	ModDate:	Tue May 17 02:20:33 2016 UTC
	Tagged:	no
	UserProperties:	no
	Suspects:	no
>	Creator:	Esri ArcCatalog 10.3.1.4959
	Producer:	Adobe PDF Library 9.0
	CreationDate:	Tue May 17 01:06:54 2016 UTC
	ModDate:	Tue May 17 01:06:54 2016 UTC
	Tagged:	no
	UserProperties:	no
	Suspects:	no
>	Creator:	Esri ArcCatalog 10.3.1.4959
	Producer:	Adobe PDF Library 9.0
	CreationDate:	Tue May 17 01:06:07 2016 UTC
	ModDate:	Tue May 17 01:06:07 2016 UTC
	Tagged:	no
	UserProperties:	no
	Suspects:	no
>	Title:	MarathiMitra's Jump Start Guide
	Author:	MarathiMitra
	Creator:	Acrobat PDFMaker 6.0 for Word
	Producer:	Acrobat Distiller 6.0 (Windows)
	CreationDate:	Wed Jun 29 02:36:49 2005 UTC

# /Subtype vs /SubType



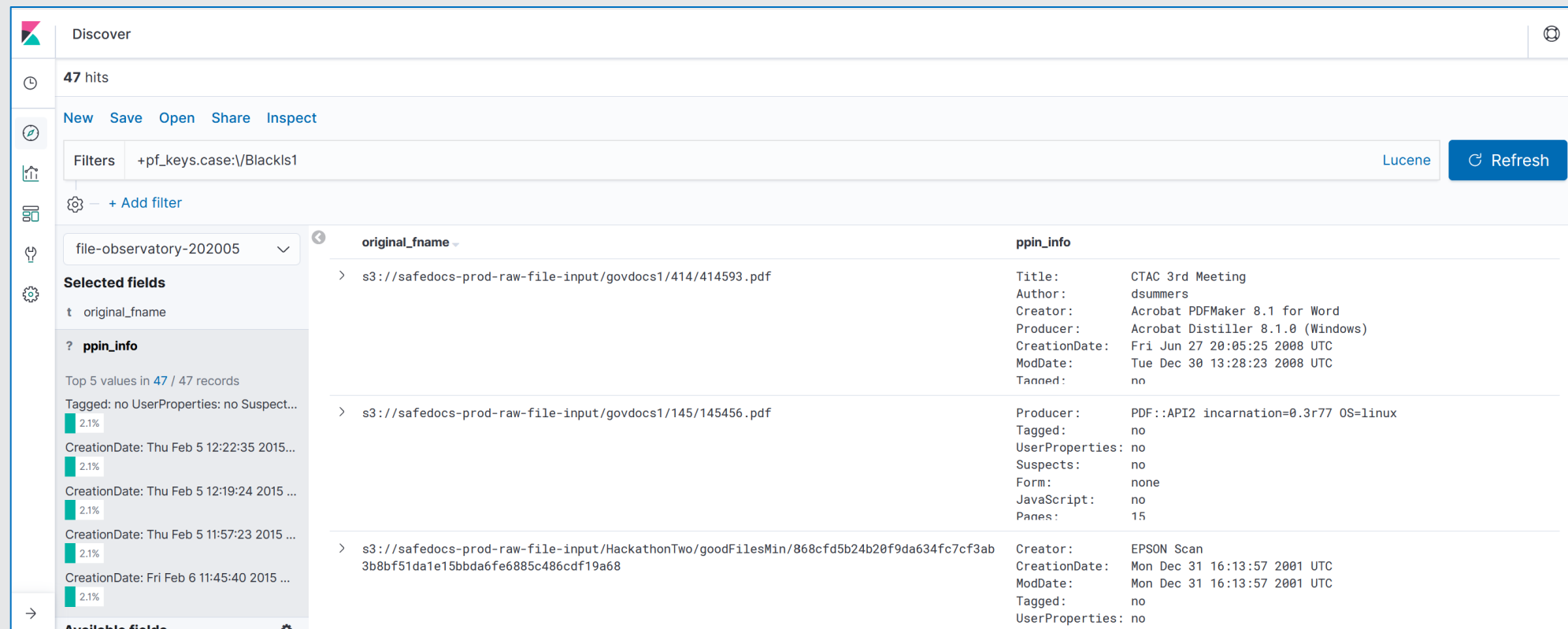
OctoberPDFest ONLINE

```
9991 >>
9992 endobj
9993 24 0 obj
9994 <</Type/OCG/Name (ypARCTIC) /Usage<</CreatorInfo<</Creator (ESRI ArcMap 9.2.2.1350)/SubType (Layer)>>>>
9995 endobj
9996 25 0 obj
9997 <</Type/OCG/Name (ypARCTIC_ANNO) /Usage<</CreatorInfo<</Creator (ESRI ArcMap 9.2.2.1350)/SubType (Layer)>>>>
9998 endobj
9999 26 0 obj
1000 <</Type/OCG/Name (yp<Default>) /Usage<</CreatorInfo<</Creator (ESRI ArcMap 9.2.2.1350)/SubType (Layer)>>>>
1001 endobj
1002 27 0 obj
1003 <</Type/OCG/Name (ypOther) /Usage<</CreatorInfo<</Creator (ESRI ArcMap 9.2.2.1350)/SubType (Layer)>>>>
1004 endobj
1005 28 0 obj
```

```
46 endobj
47 1981 0 obj
48 <</MarkInfo<</LetterspaceFlags 0/Marked true>>/Outlines 363 0 R/Metadata 449 0 R/PieceInfo<</MarkedPDF<</LastModified(D:201209191711
eLayout/SinglePage/OCProperties<</D<</RBGroups[]/Order[1982 0 R]>>/OCGs[1982 0 R]>>/StructTreeRoot 451 0 R/Type/Catalog/LastModified
/PageLabels 442 0 R>>
49 endobj
50 1982 0 obj
51 <</Usage<</CreatorInfo<</Creator(Acrobat PDFMaker 8.1 for PowerPoint)>>/PageElement<</SubType/BG>>>/Name (Background) /Type/OCG>>
52 endobj
53 1983 0 obj
54 <</CropBox[36.0 36.0 576.0 756.0]/Parent 445 0 R/StructParents 0/Contents 1985 0 R/Rotate 90/MediaBox[0 0 612 792]/Trans 1991 0 R/Re
```

# CCITTFaxDecode “Black is 1”

- /Black1s1 (“lowercase L” + “lowercase S”) – instead of /BlackIs1 (“uppercase i” + “lowercase S”)
- 47 hits vs 4379 hits for correct spelling
- No identifiable creator or producer



original_fname	ppin_info
s3://safedocs-prod-raw-file-input/govdocs1/414/414593.pdf	Title: CTAC 3rd Meeting Author: dsummers Creator: Acrobat PDFMaker 8.1 for Word Producer: Acrobat Distiller 8.1.0 (Windows) CreationDate: Fri Jun 27 20:05:25 2008 UTC ModDate: Tue Dec 30 13:28:23 2008 UTC Tanned: no
s3://safedocs-prod-raw-file-input/govdocs1/145/145456.pdf	Producer: PDF::API2 incarnation=0.3r77 OS=linux Tagged: no UserProperties: no Suspects: no Form: none JavaScript: no Pages: 15
s3://safedocs-prod-raw-file-input/HackathonTwo/goodFilesMin/868cfd5b24b20f9da634fc7cf3ab3b8bf51da1e15bbda6fe6885c486cdf19a68	Creator: EPSON Scan CreationDate: Mon Dec 31 16:13:57 2001 UTC ModDate: Mon Dec 31 16:13:57 2001 UTC Tagged: no UserProperties: no



- /PageDirection: incorrectly named key instead of /Direction used by Acrobat versions 4.0 – 4.05<sup>Δ</sup> – **190 hits**
- /DP instead of DecodeParams. Appendix C Implementation Note #7 in legacy Adobe PDF specifications: “*Acrobat viewers accept the name DP as an abbreviation for the DecodeParms key in any stream dictionary*” – **42 hits**
- /XUID: only defined in Adobe PDF 1.0 to Adobe PDF 1.2 specifications – **1 hit** (*Kibana-whack!*) GovDocs1/125347.pdf:

```
1510 endobj
1511 45 0 obj
1512 << /Type /XObject /Subtype /Form /FormType 1 /BBox [ 0 792 612 0 ] /Matrix [ 1 0 0 1 0 0 ]
1513 /Name /Watermark1 /XUID [ 1000000 886 2 27669840 1 2147483647 14791616 ]
1514 /Resources << /Font << /F2 228 0 R >> >> /Length 44 0 R /Filter /FlateDecode >>
1515 stream
1516 xœ...Ž1 Â0...÷@bÃ u0^RÓl¥V  'Lšfô
```

<sup>Δ</sup> <http://web.archive.org/web/20071130095404/http://support.adobe.com/devsup/devsup.nsf/docs/50757.htm>

FINAL  
DRAFT

Licensed to: Wyatt Peter Mr.  
Downloaded: 2020-09-17T07:06:29.332  
Single user licence only, copying and networking prohibited  
INTERNATIONAL  
STANDARD

ISO/FDIS  
32000-2

ISO/TC 171/SC 2

Secretariat: ANSI

Voting begins on:  
2020-09-30

Voting terminates on:  
2020-11-25

**Document management — Portable  
document format —**

**Part 2:  
PDF 2.0**

*Gestion de documents — Format de document portable —  
Partie 2: PDF 2.0*

RECIPIENTS OF THIS DRAFT ARE INVITED TO  
SUBMIT, WITH THEIR COMMENTS, NOTIFICATION  
OF ANY RELEVANT PATENT RIGHTS OF WHICH  
THEY ARE AWARE AND TO PROVIDE SUPPORTING  
DOCUMENTATION.

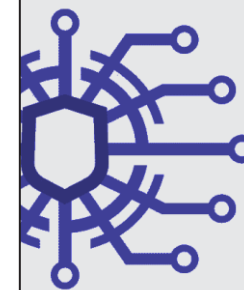
IN ADDITION TO THEIR EVALUATION AS  
BEING ACCEPTABLE FOR INDUSTRIAL, TECHNO-  
LOGICAL, COMMERCIAL AND USER PURPOSES,  
DRAFT INTERNATIONAL STANDARDS MAY ON  
OCCASION HAVE TO BE CONSIDERED IN THE  
LIGHT OF THEIR POTENTIAL TO BECOME STAND-  
ARDS TO WHICH REFERENCE MAY BE MADE IN  
NATIONAL REGULATIONS.



Reference number  
ISO/FDIS 32000-2:2020(E)

© ISO 2020

Extant Data  
("in the wild")



Cybersecurity

# How standards interconnect

## 2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

Undated reference

ISO 3166-1, *Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*.

ISO/IEC 8824-1, *Information technology – Abstract Syntax Notation One (ASN.1): Specification of basic notation*

ISO/IEC 10646, *Information technology – Universal Coded Character Set (UCS)*

“Family” reference

ISO/IEC 10918 (all parts), *Information Technology – Digital Compression and Coding of Continuous-Tone Still Images: Requirements and guidelines* (informally known as the JPEG standard, for the Joint Photographic Experts Group, the ISO group that developed the standard)

Dated reference

ISO/IEC 14492:2019, *Information technology – Lossy/lossless coding of bi-level images*


ISO/IEC 14496-22:2019, *Information technology – Coding of audio-visual objects — Part 22: Open Font Format*

ISO 14711, *Document management – 3D use of Product Representation (PC) format*

**From ISO/FDIS 32000-2:2020**



Selected references:





# Machine-readable definition

## Focus on the Body

- Each of the specialized dictionary objects are well defined
- Each of the values of the keys are well defined
- Could we use that information to build a tool for validating all of those dictionaries and pointing out when they point?

Yes!

© 2012 Adobe Systems Incorporated. All Rights Reserved. Adobe Confidential.

9

## Not only Validation



- We could do more

### EXAMPLES

- Click through documentation (like JavaDoc or doxygen)
- Easily create diffs of PDF versions  
(what exactly was the difference between 1.5 and 1.7?)



## Formal Representation!



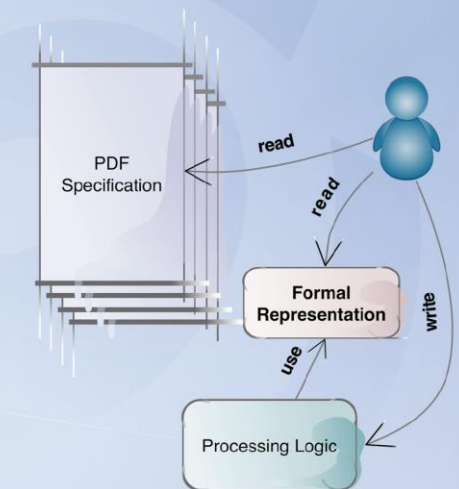
TOMORROW (or so)

Resources for references

Specification  
Representation

IS CHANGE  
BIG?

YES!



[https://youtu.be/-K\\_yBHw3C0U](https://youtu.be/-K_yBHw3C0U)

- Specification-derived for PDF 2.0 Document Object Model
- “Written rarely by very few / read often by very many”
- TSV format
  - Columnar structured data
  - Trivial to process
- Internal declarative grammar
- Verified against Adobe DVA and extant PDFs
- Corrections and clarifications made to ISO/FDIS 32000-2:2020

```
$ |
```



[peter.wyatt@pdfa.org](mailto:peter.wyatt@pdfa.org)