

**Approved for Public Release;
Distribution Unlimited.
Case Number 18-3138-8**

Text/Metadata Extraction

Tim Allison, Ph.D.

Principal Artificial Intelligence Engineer, MITRE

Library of Congress

July 2019

Agenda

- **Intro**
- **Content extraction: Overview**
- **tika-eval deep dive**
- **A note on open sourcing**
- **Discussion**

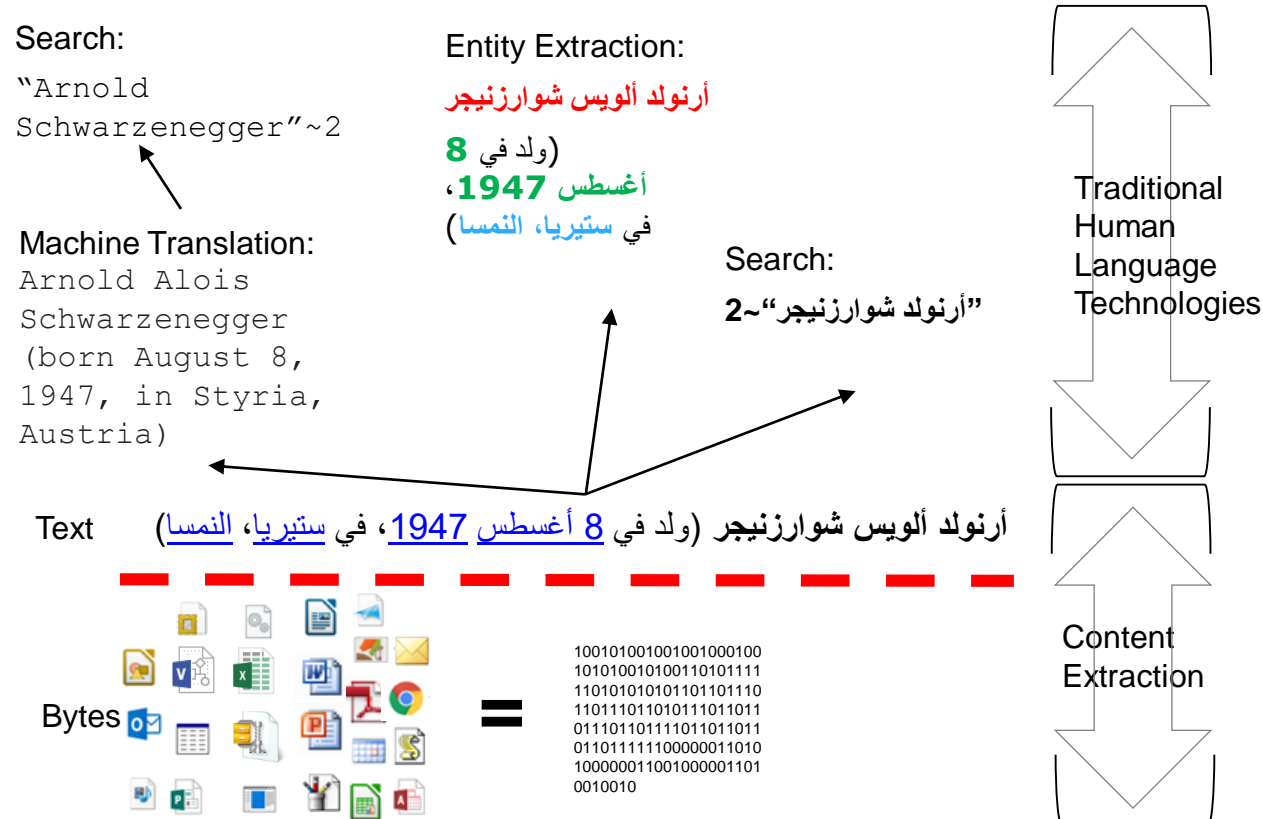
Disclaimer – Author Affiliations

- **Chair of Apache Tika**
- **Committer on Apache PDFBox, Apache POI and Apache Lucene/Solr**
- **Member Apache Software Foundation**
- **Member Apache Incubator Project Management Committee (PMC)**

Content Extraction

Overview

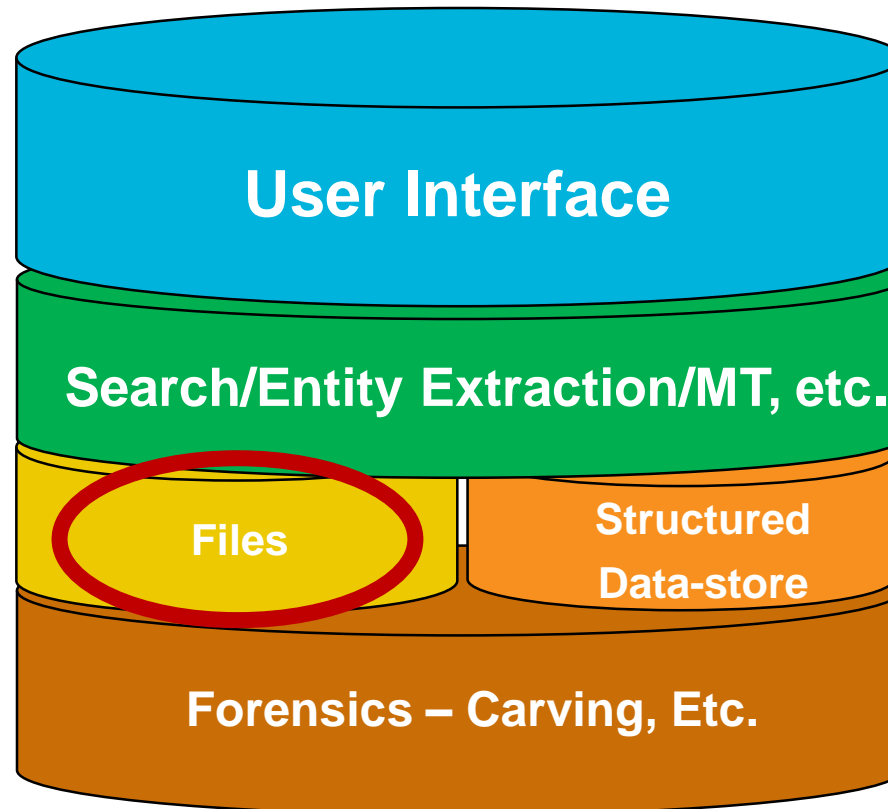
Content Extraction and HLT



High Level Components of a Media Processing Stack

| 6 |

**Text
Extraction
and
Metadata
Extraction**



Let's not forget Metadata!

- **Various formats store useful information**
- **Who:** author (first, last, commenters, editors), digital signature, company, from/to/cc/bcc (emails)
- **What:** hardware version/name, software version/name, globally unique file/heritage id (XMP), title, keywords, description
- **Where:** geo (latitude, longitude), file location (file paths embedded inside documents)
- **When:** created, last modified, last printed
- **Beyond the standard types...custom metadata**

Primary Considerations

- **Parser coverage** – which file types do you have? Which parsers do you need?
- **Attachments** – are they being extracted/processed?
- **Large files** – special preprocessing (zip, tar, pst, ost, mbox, etc.)
- **Duplicates** – exact, near – how to present to the user
- **Robustness and logging**
 - Normal exception handling
 - Permanent hangs/infinite loops, out of memory
- **Scaling**

Other Considerations – Advanced Processing

- **Zoning/form/image processing**
- **Optical Character Recognition**
- **Object Identification**
- **Captioning**
- **Automatic Speech Recognition/Audio Search**

When Things Go Wrong in the Foundation



http://www.flickr.com/photos/saffron_blaze

How well are you doing – Easy Metrics

- **File format (and format version) coverage**
- **Exception rates per file type**
- **Catastrophic errors per file type**

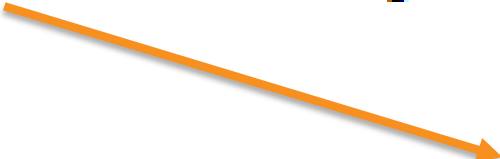
More Insidious Problems

- **Garbled text**
 - From slightly to...fully
- **Missing text/metadata**
 - From missing some text to ... no text at all
- **Missing attachments**

Corrupt Text (Upgrade from PDFBox 1.8.6 ->1.8.7)

| 13 |

Taking a close look at the forest or open meadows reveals that there are often subtle differences in plant species across a wide landscape. Unique micro-climates, exposure to the sun, soil types, moisture availability, and a variety of other factors influence the types of plant species present in any given location. Changes in any of these factors will cause changes to



BGQOTM G IRUYK RUUQ GZ ZNK LUXKYZ UX UVKT SKGJU]Y
XK\KGRY ZNGZ ZNKXX GXK ULZKT Y[HZRK JOLLKXKTIKY OT VRGTZ
YVKIOKY GIXUYY G]OJK RGTJYIGVK% CTOW[K SOIXU-
IROSGZKY\$ K^VUY[XK ZU ZNK Y[T\$ YUOR Z_VKY\$ SUOYZ[XK
G\GORGHOROZ_\$ GTJ G \GXOKZ_ UL UZKNX LGIZUXY OTLR[KTIK ZNK
Z_VKY UL VRGTZ YVKIOKY VXKYKTZ OT GT_ MO\KT RUIGZOUT%
4NGTMKY OT GT_ UL ZNKYK LGIZUXY]ORR IG[YK INGTMKY ZU

Content Comparison Example – Junk -> Better

	Tika 1.14	Tika 1.15-SNAPSHOT
Unique Tokens	786	156
Total Tokens	1603	272
LangId	zh-ch	de
Common Words	0	116
Alphabetic Tokens	1603	250
Top N Tokens	拙故: 18 獐档: 14 略獐: 14 m: 11 柿溪: 11 瑤 拙: 11 畚柿: 11 档溪: 10 捌敦: 9 敲涿: 9	die: 11 und: 8 von: 8 deutschen: 7 deutsche: 6 1: 5 das: 5 der: 5 finanzministerium: 5 oder: 5
Common Words/Alphabetic Tokens	0/1603 = 0%	116/250 = 46%

Overlap: 0%

Increase in Common Words: 116

Content Comparison Example – Small Regression

	Tika 1.14	Tika 1.15-SNAPSHOT
Unique Tokens	1916	1995
Total Tokens	14187	14302
LangId	en	en
Common Words	7498	7409
Alphabetic Tokens	13472	13587
Top 10 Unique Tokens	applicant's: 8 1.69: 1 arbitrary: 1 collecting: 1 constitution: 1 e112: 1 ei.b: 1 equating: 1 magnetically: 1 o: 1	ss: 106 applicantis: 8 ssss: 7 iactsi: 4 ithe: 4 imeansi: 3 iprocessi: 3 calculations.i: 2 iabstrack: 2 idata: 2
Common Words/Alphabetic Tokens	$7498/13472 = 56\%$	$7409/13587 = 55\%$

Overlap: 95.5%

Increase in Common Words: -89

Missing Text

| 16 |

Jane Coady

Statement	Seasoned professional with a skilled ability to connect co-workers and clients with the information, products and services they are seeking by utilizing professional experiences, organizational and client skills both as a team and an individual.	
Experience	OLS: Office Liquidations Solutions Sales and Project Administrator	May 2010 – May 2013
	Sales support and sales. Lead generation and follow up. Developed solutions for individual projects. Determine price schedules, budgets and profit margins. Created and streamlined forms and procedures. Located project specific furniture. Project Management. Plan and coordinate work schedules and duties for employees, freight companies and customers. Space planning/placement of systems furniture inventories into client's AutoCAD drawings with Giza. Coordinate project details and schedules with General Contractors, Building Engineers and Property Managers. Attend company meetings to exchange product information and coordinate work activities with other departments. Keep records and create reports regarding purchases, sales, bids and installation schedules. Coordinate marketing campaigns by compiling lists, marketing pieces to promote inventories. Inventory management. Resolve customer questions regarding sales, service and installations.	
	Bialek Healthcare Environments Design Associate, Client Services Coordinator	June 2001 – May 2010
	Furniture bid package review, quotation, response and presentation. Small office design, space planning, need assessment, presentation and quotation for commercial systems and freestanding furniture. Maintenance of client accounts including need assessment, quotation, order processing, purchasing, job costing, tracking and invoicing. Created streamlined procedures to reduce redundancies. Employee Training. Member of various committees including Process Streamlining, Marketing, and Fun.	
	Rhosymedre Design Group Office Manager	August 1998 – April 2001
	Processing and maintenance of accounts receivable, payable and payroll with Business Works Accounting System and QuickBooks Pro. Maintenance of client accounts including estimating, job costing, purchasing, tracking, and invoicing and project management. Establish and maintain vendor relations. Research new residential products.	
Education	University of Nebraska Bachelors of Science with a focus in Textiles, Clothing and Interior Design, with a minor in Business Honors: Gold Key Honorary Jan 1986, Sigma Phi Upsilon Honorary Officer – October 1985	August 1984 – May 1987

Jane Coady

Statement

OLS: Office Liquidations Solutions May 2010 – May 2013

Experience

Bialek Healthcare Environments June 2001 – May 2010

University of Nebraska August 1984 – May 1987

Education

Bachelors of Science with a focus in Textiles, Clothing and Interior Design, with a minor in Business

JC

2

Skills

Document available: <https://issues.apache.org/jira/browse/TIKA-1130>

Other Lessons Learned

- **Exceptions are not randomly distributed – a small problem for me could be a BIG problem for you**
- **Do not index and forget; ideally, schedule/plan for reindexing when updated versions of extraction software become available – be able to determine if there are improvements for your corpus**
- **Evaluate, evaluate, evaluate**

Take-away

- If you don't evaluate content extraction...

**You don't know
what you can't find**

Deep Dive on tika-eval module

High-level overview

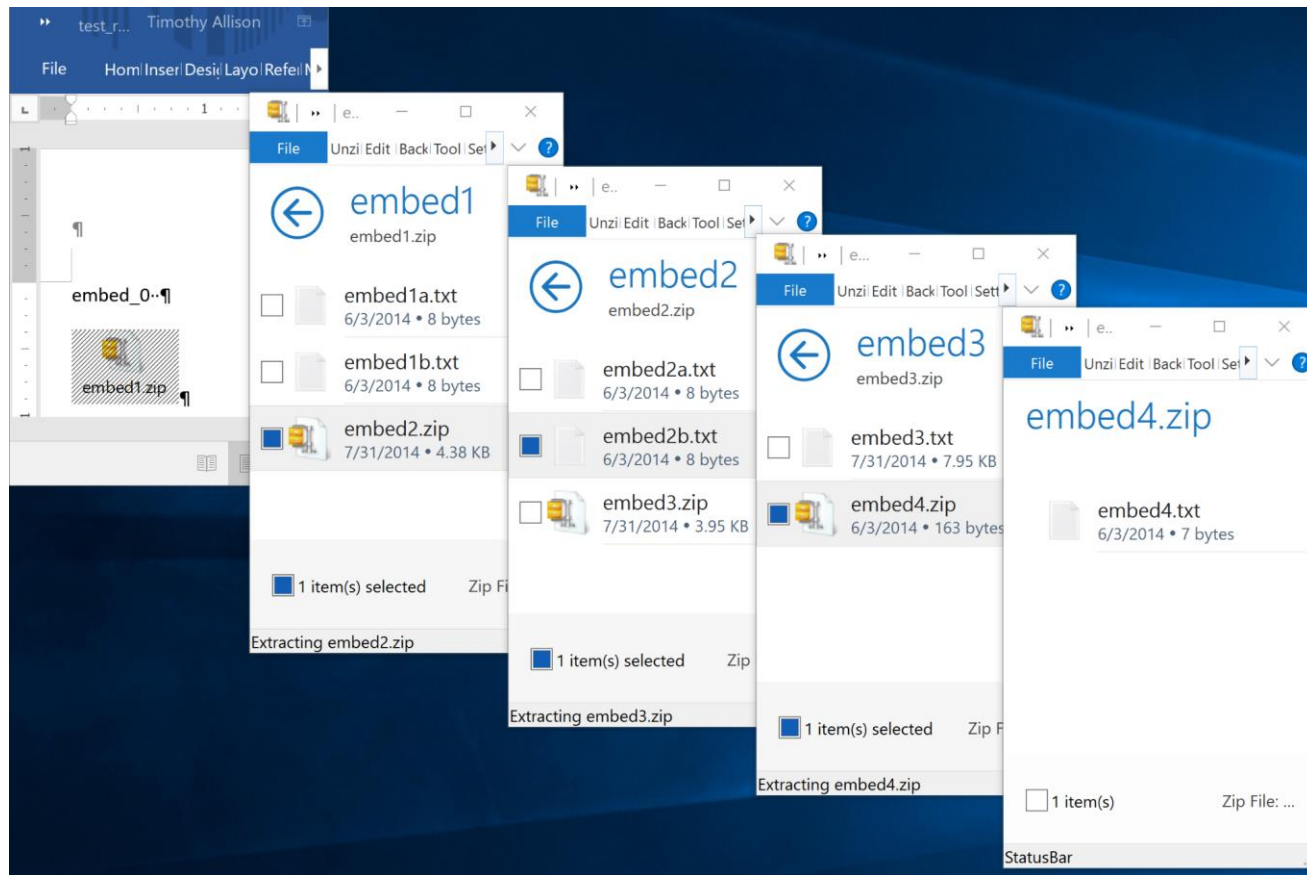
■ **tika-eval's scope**

- Single vm, file share to file share (with embedded H2 db), ~few million files is a reasonable size
- Not currently cloud-scale
 - Random sampling – should be good enough
 - Our Jira is open and committers are standing by!

■ **tika-eval's two modes**

- Profile single extraction run
- Compare two extraction runs
 - Ground truth vs. particular tool
 - Tool A vs. tool B
 - Tool A with settings X vs. Tool A with settings Y

Why the RecursiveParserWrapper?



Classic XHTML

```
<?xml version="1.0" encoding="UTF-8"?>
<meta name="Content-Type" .../>
...
<p>embed_0 </p>
<p><div class="embedded" id="rld7"/>
<p>embed1.zip</p>
<div class="embedded" id="embed1/embed1a.txt"/>
<div class="package-entry">
    <p>embed_1a</p>
</div>
...
```

- **Metadata from embedded docs is lost**
- **Exceptions from embedded docs are swallowed**
- **Metadata from the container document may be incomplete**

RecursiveParserWrapper

```
[
  {
    "Content-Type": "application/....wordprocessingml.document",
    ...
    "X-TIKA:content": "\n\n\nembed_0  \n\n\n\n\n\n\n\n\n\n"
    ...
  },
  {
    "Content-Type": "application/zip",
    "X-TIKA:content": "embed1/embed1a.txt embed1/embed1b.txt embed1/embed2.zip",
    "X-TIKA:embedded_resource_path": "/embed1.zip",
  },
  {
    "Content-Type": "text/plain; charset=ISO-8859-1",
    "X-TIKA:embedded_resource_path": "/embed1.zip/embed1a.txt",
    "X-TIKA:content": "embed_1a\n",
  }
  ...
]
```

- **Embedded metadata (e.g. mime/author/lat-long, etc.) are retained**
- **Embedded exceptions are stored in a metadata key**
- **All metadata is extracted stored**

Workflow – Profile

1. Generate extracts with parallel directory structure to original documents, append “.txt” or “.json” into, say `my_extracts` directory

2. Run profiler to populate in-process H2 DB

```
java -jar tika-eval.jar Profile  
      -extracts my_extracts  
      -db my_db
```

3. Dump reports

```
java -jar tika-eval.jar Report -db my_db
```

Excel reports will be dumped to the `reports` directory

Workflow – Compare

1. Generate extracts with parallel directory structure to original documents, append “.txt” or “.json” into, say my_extractsA and my_extractsB directories

2. Run profiler to populate in-process H2 DB

```
java -jar tika-eval.jar Compare  
-extractsA my_extractsA  
-extractsB my_extractsB  
-db my_db
```

3. Dump reports

```
java -jar tika-eval.jar Report -db my_db
```

Excel reports will be dumped to the reports directory

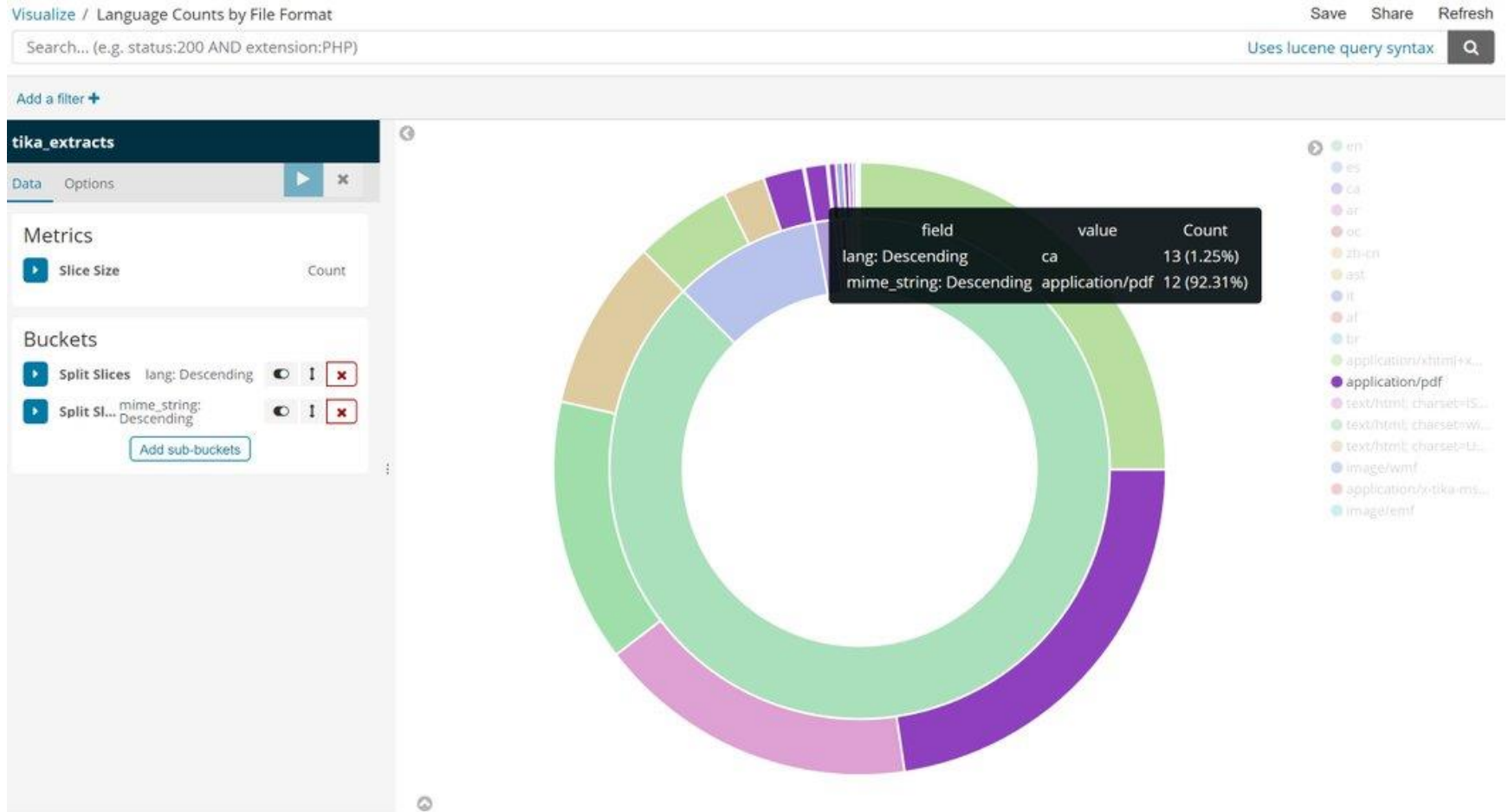
Taking tika-eval public

- Rackspace kindly hosts a vm for ongoing evals (TIKA-1302)
- 1 TB (~3 million files) from Common Crawl and govdocs1
- Collaborating with Apache PDFBox and Apache POI to run evals as part of the release process
- Critical to identifying regressions and building new parsers
- Stacktraces created by public documents are critical for the `hey-I'm-getting-this parse-exception-but-can't-share-the-document-with-you` problem
- See Dominik Stadler's Common Crawl download tool:
<https://github.com/centic9/CommonCrawlDocumentDownload>

Limits of Automated Metrics without Ground Truth

- **More exceptions – We have a problem! Wait...**
 - New parser, we were entirely skipping those file types before
 - Parser was yielding junk before on this file, now it is letting us know there's a problem
- **Fewer exceptions – Great! Wait...**
 - Mime detection not working – skipping files that we used to parse (theoretical)
 - Now we're getting junk
- **More Common Words – Great! Wait...**
 - Serious bug that duplicates worksheets in some xlsx files (TIKA-2356...my fault...ugh!)
 - More non-html markup/xml tags incorrectly getting through
- **Fewer Common Words – Problem! Wait...**
- **More attachments, fewer attachments (Your turn!)**

Kibana – File Format by Lang Id



Kibana OOV% in English PDFs

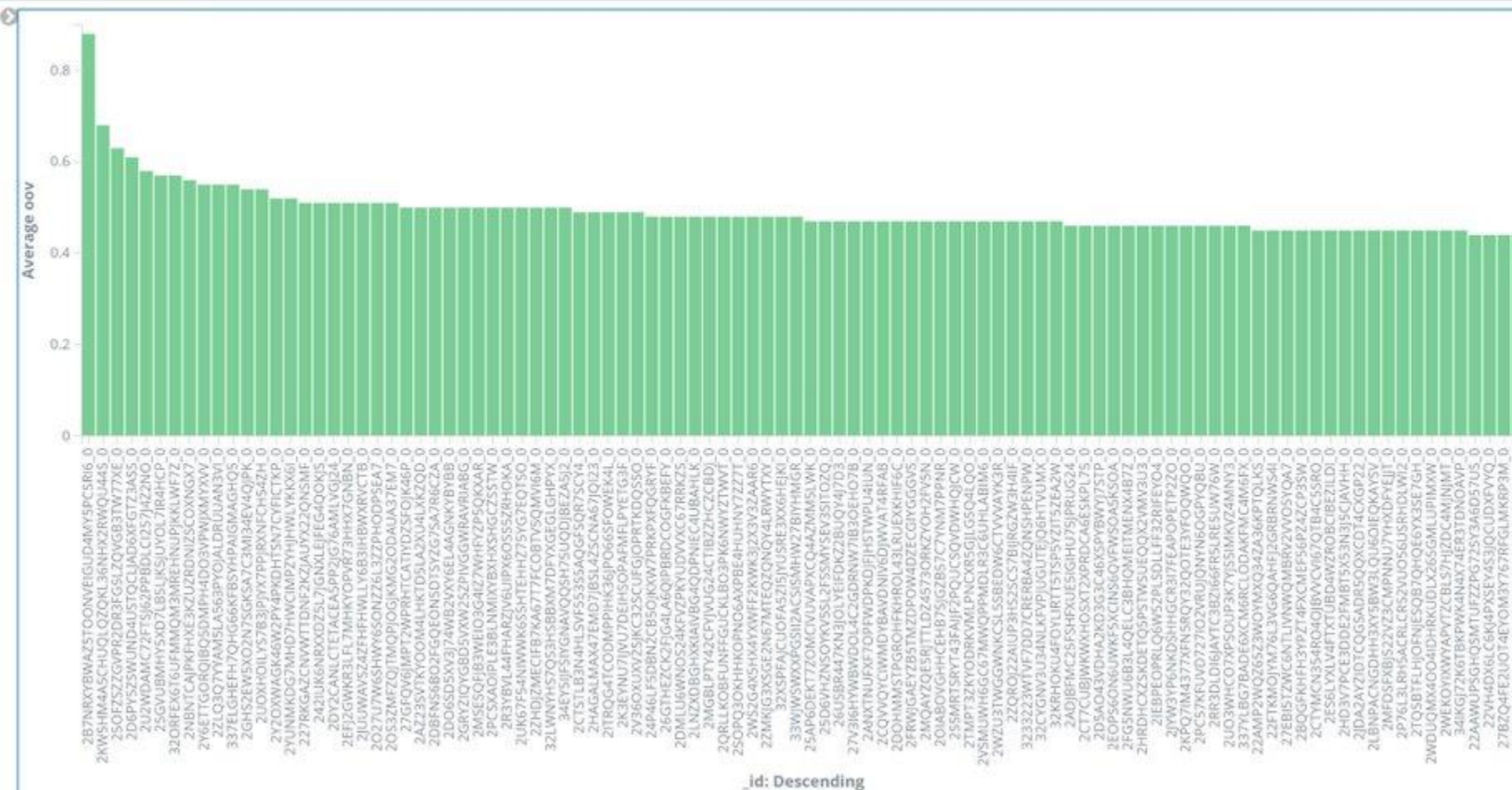
Visualize / OOV % in English PDFs

+mime:"application/pdf" +lang:en

Uses 1

gt 100 tokens

Add a filter +



_doc#2B7NRXYBWAZ5TOONVEIGUD4MY5PCSR16_0

JSON

[illegible]

Open Source

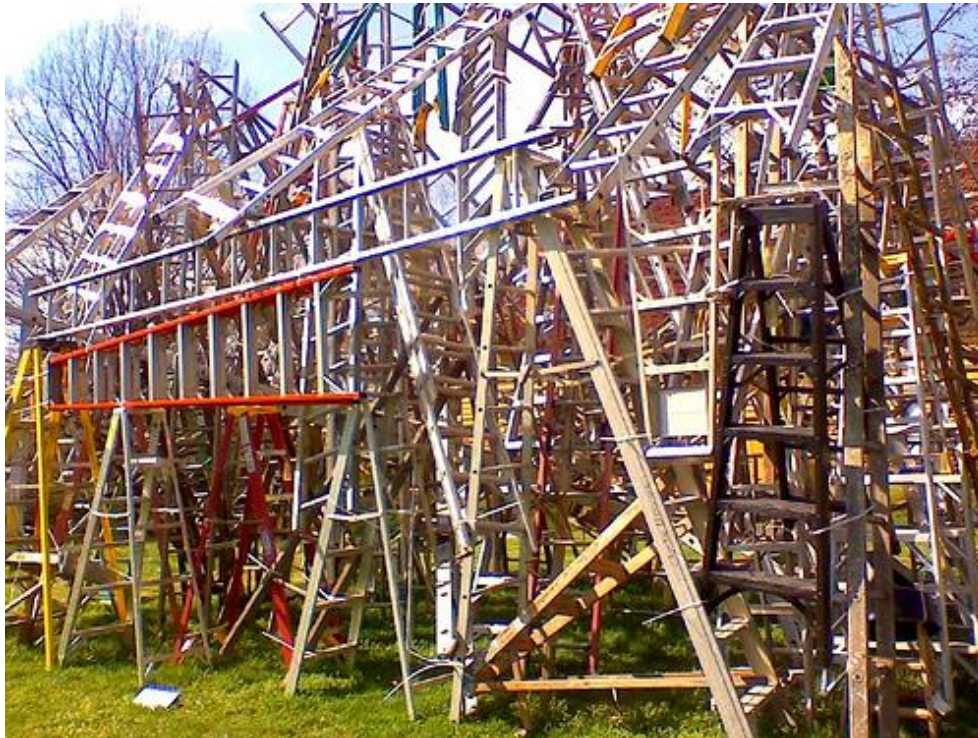
Help, please!

Open Source: Land of the Free



<http://www.writeonnewjersey.com/2011/03/if-its-free-its-for-me/>

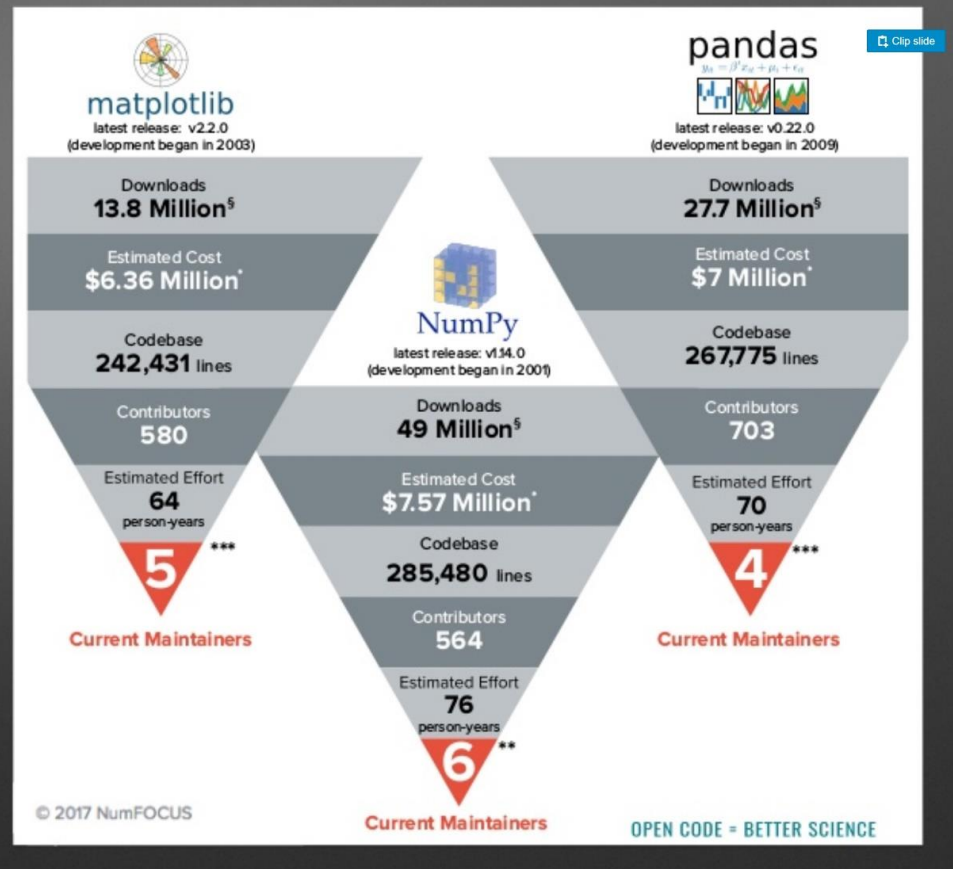
Home of the Brave



<http://www.contractortalk.com/f59/ladder-safety-has-come-along-way-71332/>

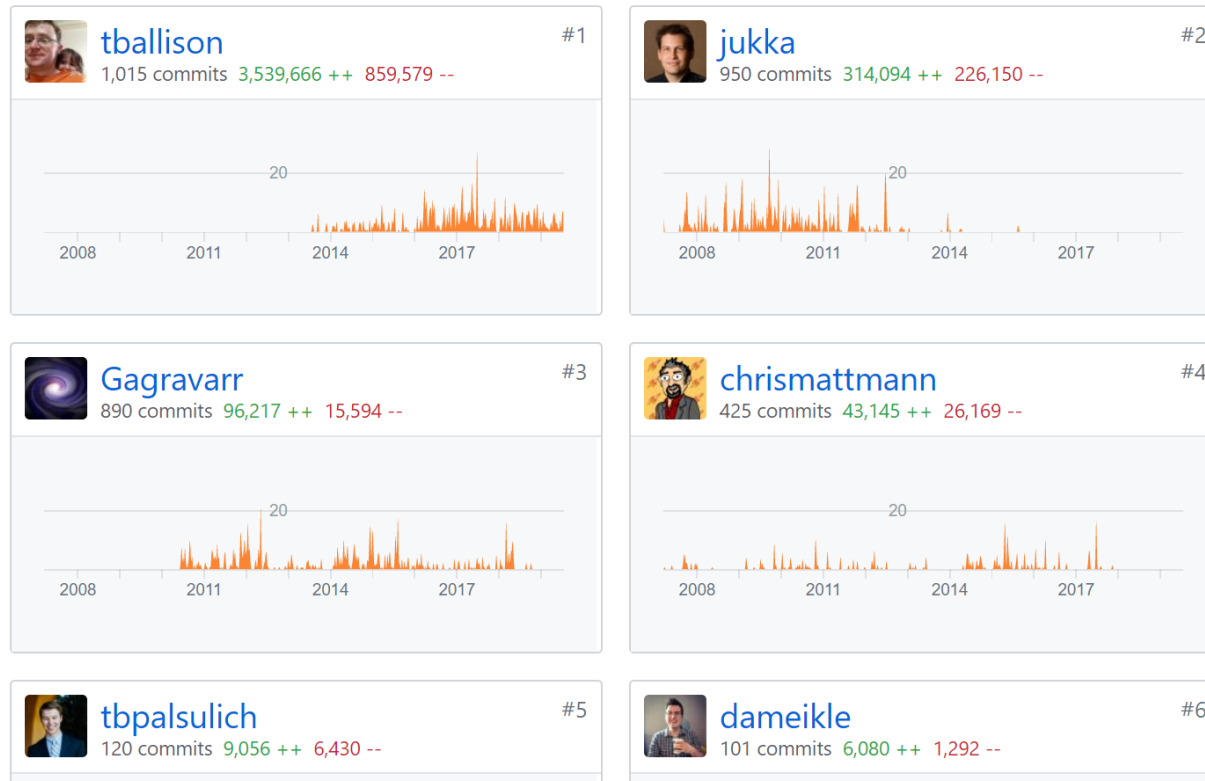
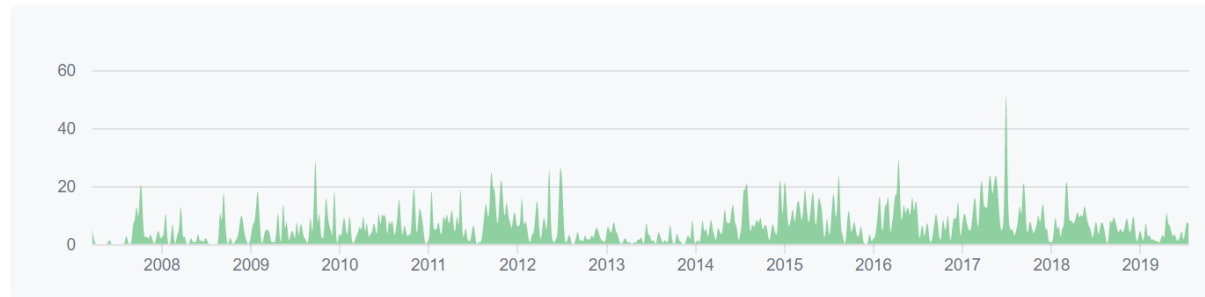
Grim Numeric Realities

**Vital code
is
maintained
by a small
number of
people**



Kelle Cruz. "Collaborations in the Extreme: The rise of open code development in the scientific community." <https://www.slideshare.net/KelleCruz/collaborations-in-the-extreme-the-rise-of-open-code-development-in-the-scientific-community>

Apache Tika – Commits



Contact Info

- tallison@mitre.org
- tallison@apache.org
- **@_tallison**

Discussion
