Evaluating Text Extraction: Apache Tika's New tika-eval Module

Tim Allison
The MITRE Corporation

ApacheCon North America 2017
Miami, FL

May 18, 2017

Debts of Gratitude

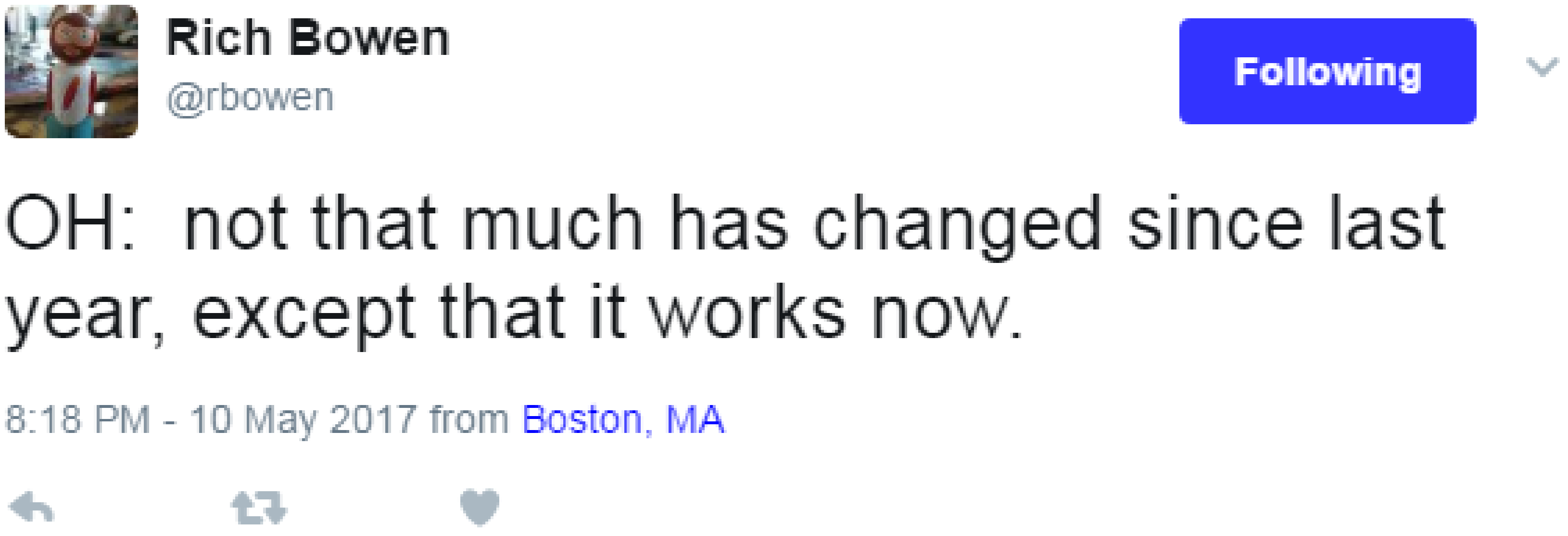
- **David Smiley**
- **Nick Burch**
- **Chris Mattmann**
- **Tilman Hausherr**
- **Dominik Stadler**
- **Fellow Apache Commons, Apache POI, Apache PDFBox, Apache Tika devs and users**
- **ASF Community!**

- **Common Crawl and govdocs1**
- **Rackspace**

Overview – tika-eval

- Content and metadata extraction in the ETL stack – overview
- Motivation for tika-eval: what can go wrong?
- tika-eval overview
- tika-eval workflow
- tika-eval on 1 TB public corpus
- Limitations

What’s new since 2015 talk on tika-eval?



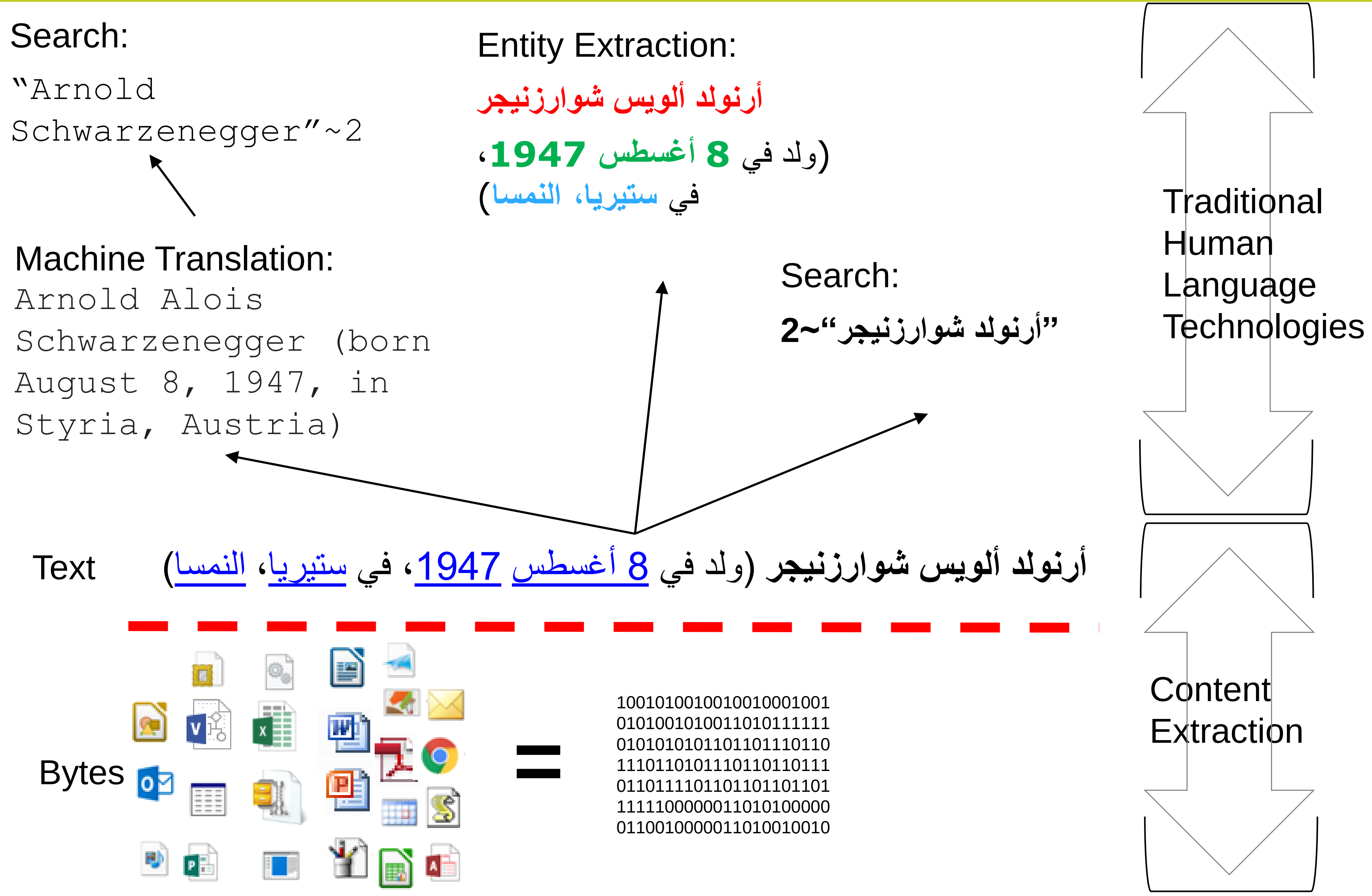
Rich Bowen
@rbowen

OH: not that much has changed since last year, except that it works now.

8:18 PM - 10 May 2017 from Boston, MA

Will be available in Tika 1.15; release to start soon!

Content Extraction and HLT



Search:

```
"Arnold
Schwarzenegger"~2
```

Entity Extraction:

أرنولد ألويس شوارزنيجر

(ولد في **8 أغسطس 1947**،
في ستيريا، النمسا)

Machine Translation:

```
Arnold Alois
Schwarzenegger (born
August 8, 1947, in
Styria, Austria)
```

Search:

“أرنولد شوارزنيجر”~2

Traditional Human Language Technologies

Text أرنولد ألويس شوارزنيجر (ولد في 8 أغسطس 1947، في ستيريا، النمسا)

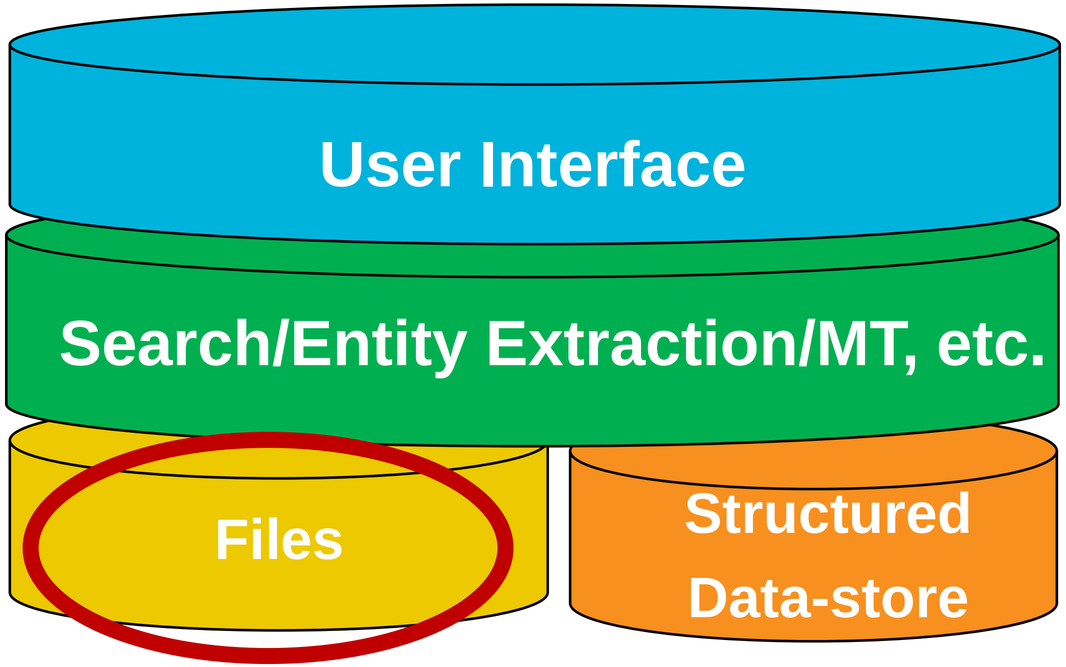
Bytes =

```
1001010010010010001001
0101001010011010111111
0101010101101101110110
1110110101110110110111
0110111101101101101101
1111100000011010100000
0110010000011010010010
```

Content Extraction

High Level Components of a Media Processing Stack

Text Extraction and Metadata Extraction



Let’s not forget Metadata!

- **Various formats store useful information**

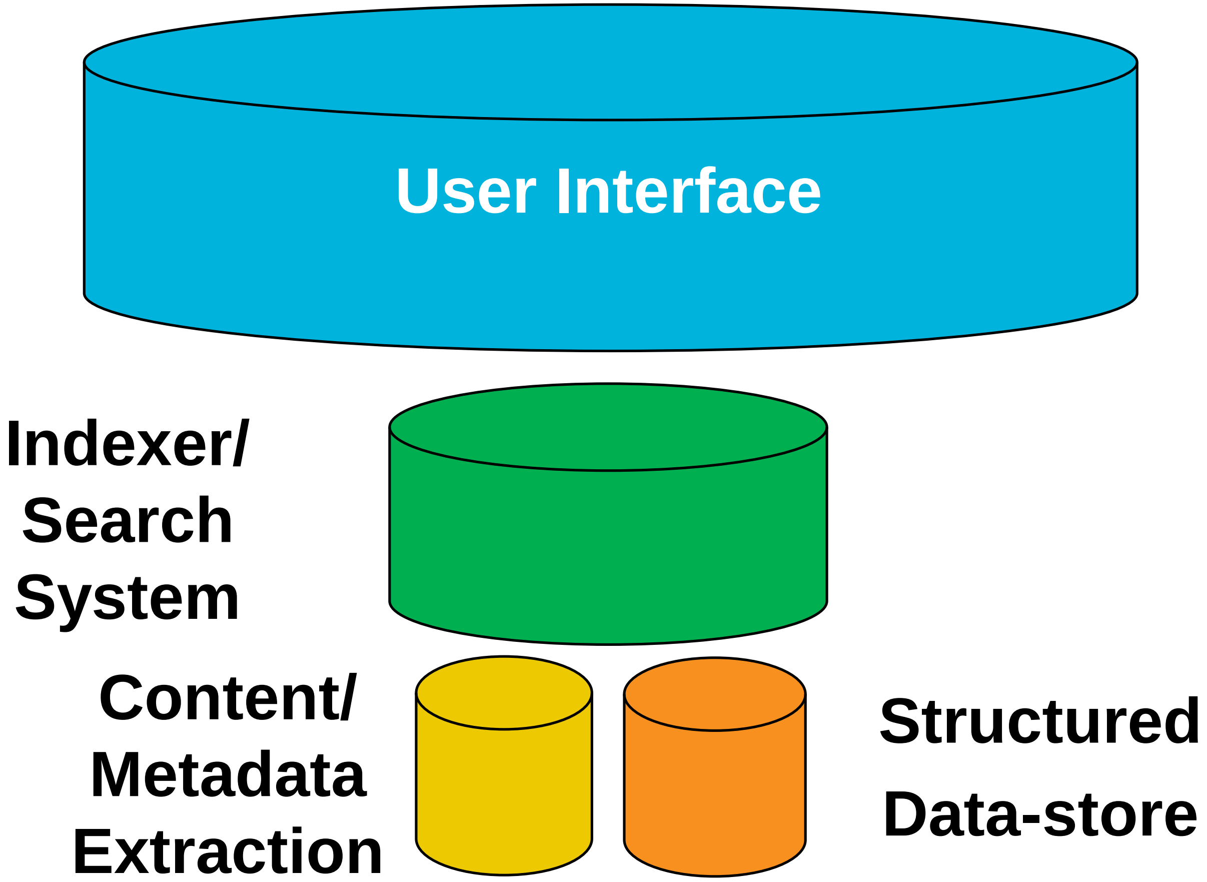
- **Who:** author (first, last, commenters, editors), digital signature, company, from/to/cc/bcc (emails)
- **What:** hardware version/name, software version/name, globally unique file/heritage id (XMP), title, keywords, description
- **Where:** geo (latitude, longitude), file location (file paths embedded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

- **Beyond the standard types…custom metadata**

When Things Go Wrong with Text Extraction

Example Application: Search

What the User Sees in a Search System



When Things Go Wrong with a Foundation

W. Lloyd MacKenzie, via Flickr
@http://www.flickr.com/photos/saffron_blaze/

What can go wrong? Basic problems

- **Completely expected Exceptions – no need to fix parsers**
  - Truncated files
  - Password/access protected files
  - Format version not handled (add new parser?)
  - Corrupt files – can’t be opened by primary application or parsed by other parsers
- **Somewhat expected Exceptions – might be able to fix parsers**
  - Parser has a problem with non-corrupt file (and admits it…thank you!!!)
  - Corrupt files – slight variant from spec/other parsers can handle it

Note: some text/metadata may or may not be extracted before the exception is thrown

What can go wrong? Catastrophic problems

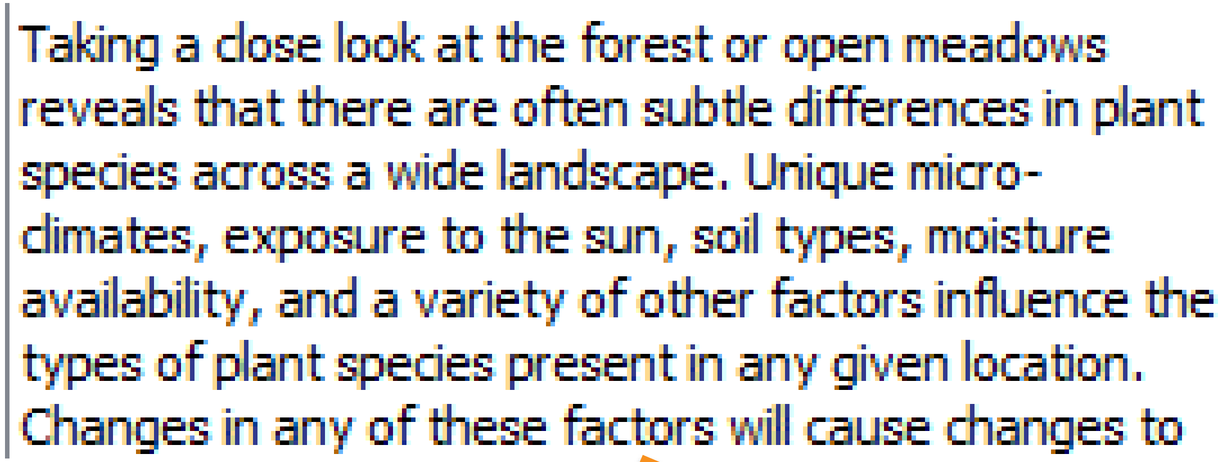
- **OutOfMemoryError – potentially corrupting the JVM**
  - Inefficient parsers DOM vs SAX on rare docx (TIKA-2170) and pptx (TIKA-2201). **NOTE:** with multithreaded garbage collection, a single thread running Tika can cause a quad-core system to grind to a snail’s pace before hitting OOM.
  - Four bytes of a compressed file (TIKA-2330)
- **Slowly building memory leak**
  - See above on quad-core, gc and snails (TIKA-2180?)
- **Permanent Hang**
  - TIKA-1132
- **Security Vulnerabilities**
  - XXE (CVE-2016-4334), arbitrary code execution (CVE-2016-6809)

These are extremely rare, and we try to fix them when we’re aware of them!

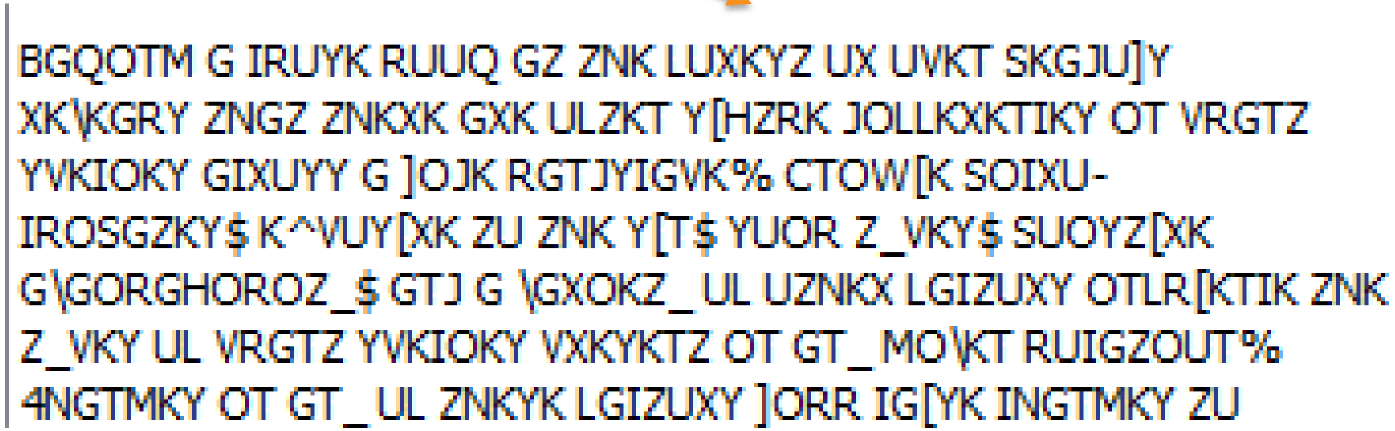
What can go wrong? Usually hidden problems

- **No Exception But…**
  - Garbled text
    - From slightly to…fully
  - Missing text/metadata
    - From missing some text to … no text at all
  - Missing attachments
  - Silently swallowed exceptions of embedded documents
    - **Classic Tika xhtml/text extraction silently swallows embedded exceptions!!!**

Corrupt Text (Upgrade from PDFBox 1.8.6->1.8.7)

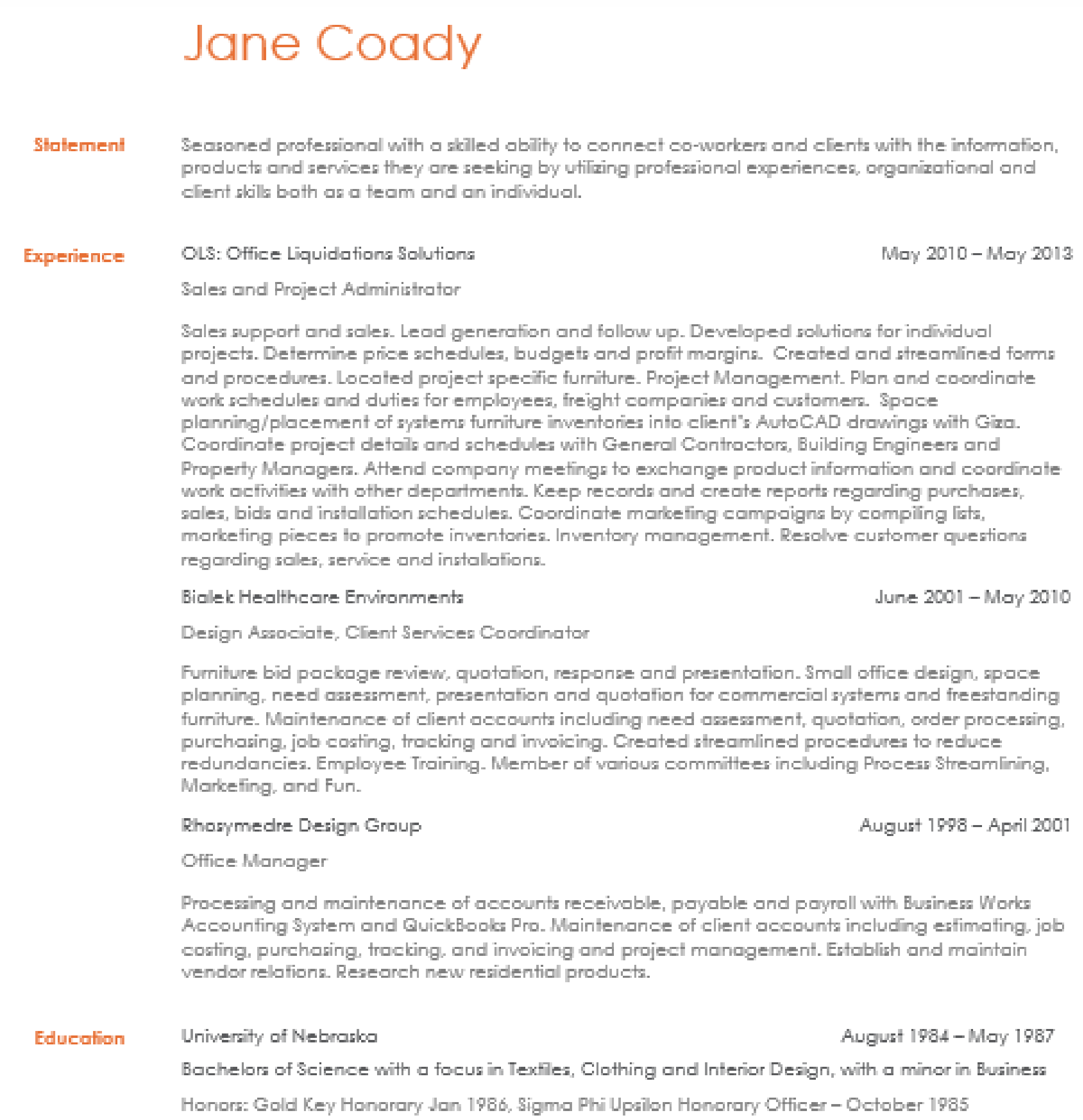


Taking a close look at the forest or open meadows reveals that there are often subtle differences in plant species across a wide landscape. Unique micro-climates, exposure to the sun, soil types, moisture availability, and a variety of other factors influence the types of plant species present in any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

Missing Text (TIKA-1130)



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 May 2010 – May 2013

Sales and Project Administrator

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 June 2001 – May 2010

Design Associate, Client Services Coordinator

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 August 1998 – April 2001

Office Manager

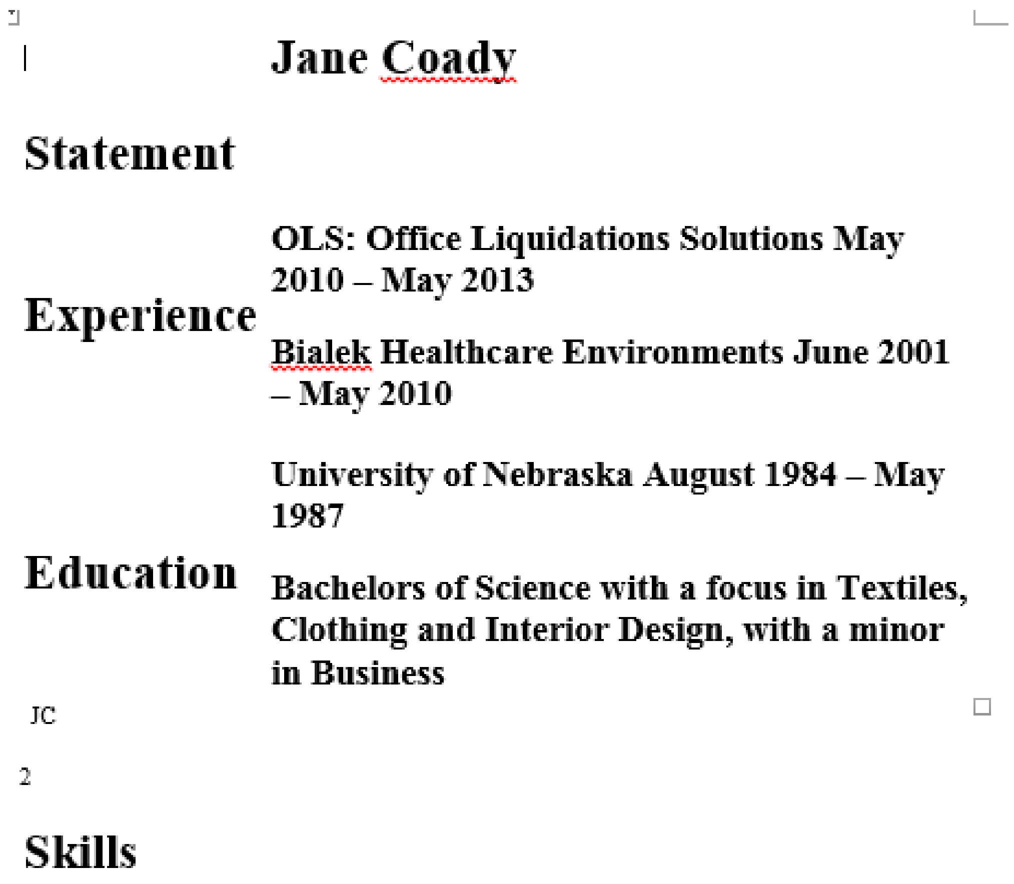
Processing and maintenance of accounts receivable, payable and payroll with Business Works Accounting System and QuickBooks Pro. Maintenance of client accounts including estimating, job costing, purchasing, tracking, and invoicing and project management. Establish and maintain vendor relations. Research new residential products.

Education

University of Nebraska August 1984 – May 1987

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



Jane Coady

Statement

Experience

OLS: Office Liquidations Solutions May 2010 – May 2013

Bialek Healthcare Environments June 2001 – May 2010

Education

University of Nebraska August 1984 – May 1987

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

JC

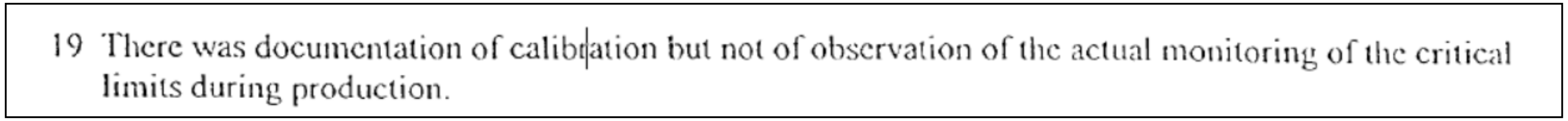
2

Skills

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

When Things Go Not as Well as They Might with Content Extraction – OCR

Image:

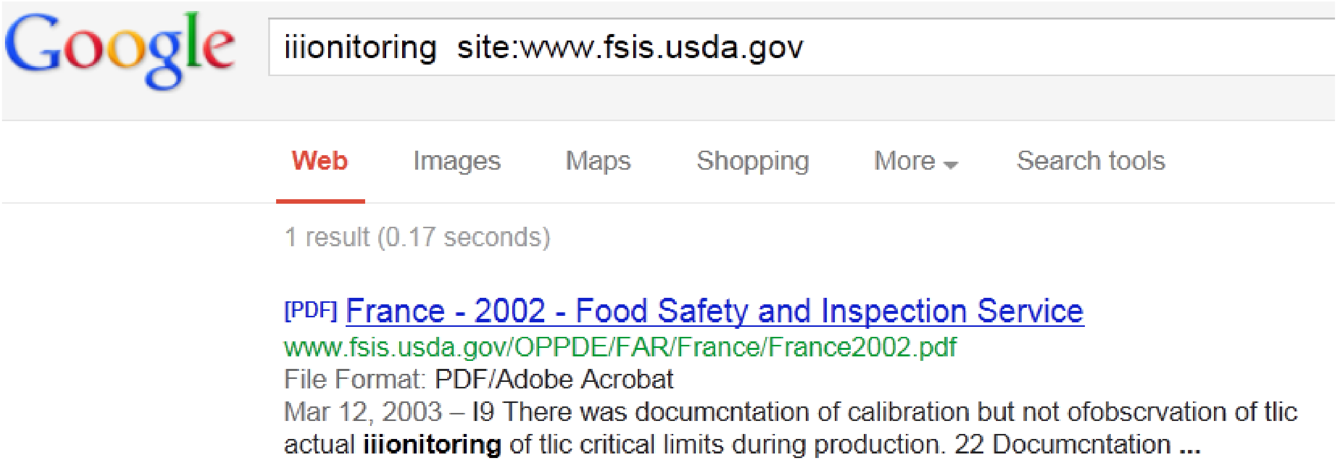


19 There was documentation of calibration but not of observation of the actual monitoring of the critical limits during production.

Text Extracted:

I9 There was documcntation of calibration but not ofobscrvation of tlic actual iiionitoring of tlic critical limits during production.

Search Results:



iiionitoring site:www.fsis.usda.gov

1 result (0.17 seconds)

[PDF] France - 2002 - Food Safety and Inspection Service
www.fsis.usda.gov/OPPDE/FAR/France/France2002.pdf
File Format: PDF/Adobe Acrobat
Mar 12, 2003 – I9 There was documcntation of calibration but not ofobscrvation of tlic actual **iiionitoring** of tlic critical limits during production. 22 Documcntation ...

Take-away

- **If you don’t evaluate content extraction…**

You don’t know
what you can’t find

TIKA-1302: The Dream

- **Motivation**
  - All of the above
  - We have only roughly 1,000 test files in unit tests in Apache POI, Apache PDFBox and Apache Tika
  - POI/PDFBox/Tika mistakenly made me a committer
- **Run Tika on much larger corpus nightly/weekly**
- **Automatically recognize regressions**

tika-eval

Available in Apache Tika 1.15

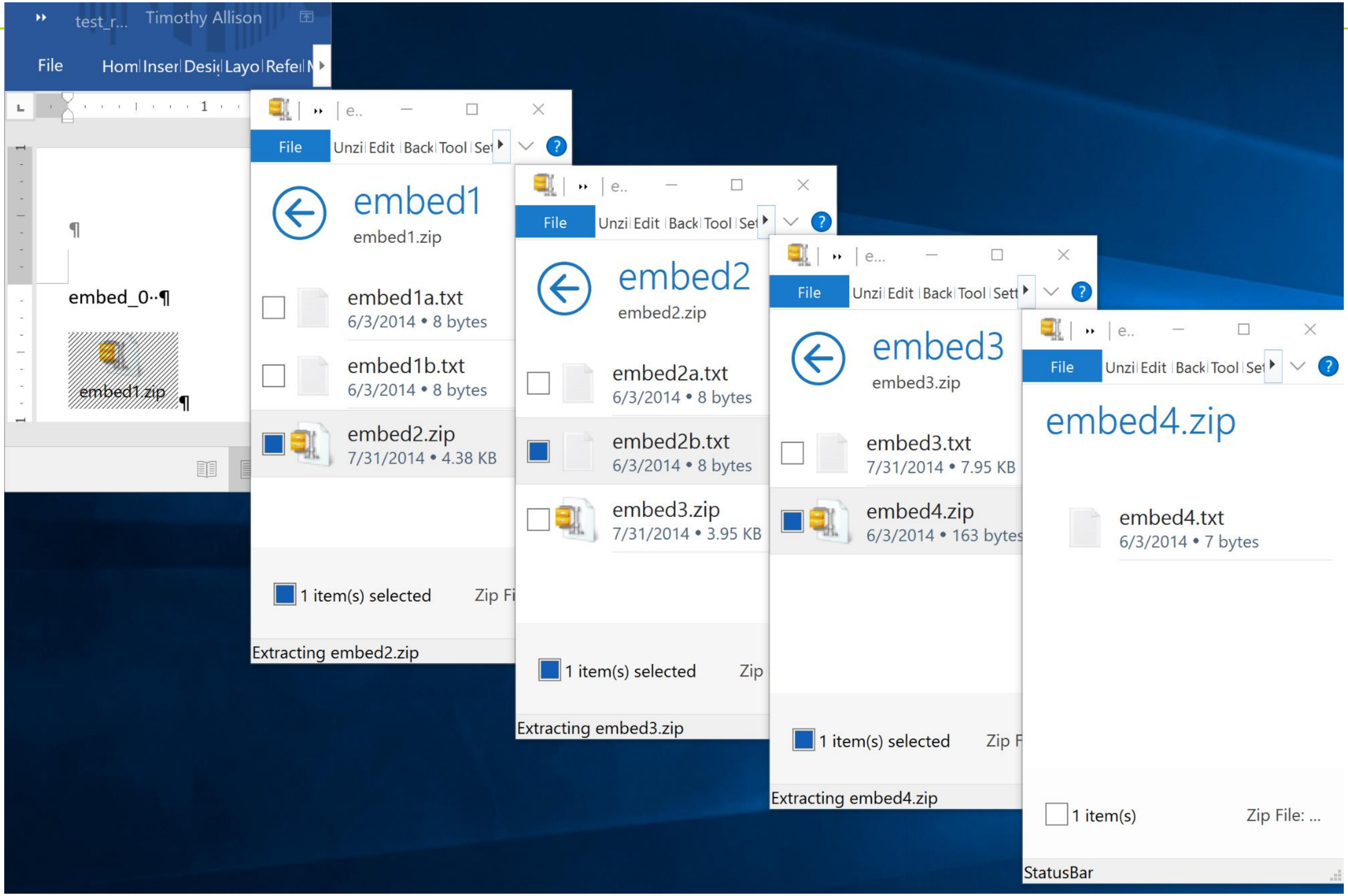
High-level overview

- **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

Definitions

- **“original documents” or “container documents” – the original binary documents from which you’d like to extract text, whether or not they actually have attachments.**
- **“embedded documents” – any document contained within another document, including those that only ever exist as embedded docs: `emf/wmf/xmp/xfa`.**
- **“extract” – .txt or .json representation of the extracted text/metadata.**
  - tika-eval was designed for .json
    - `RecursiveParserWrapper` via API
    - (`–J`) for tika-app
    - `/rmeta` for tika-server
  - tika-eval can handle .txt files – details on our wiki

Why the `RecursiveParserWrapper`?



embed1
embed1.zip

□ embed1a.txt
6/3/2014 • 8 bytes

□ embed1b.txt
6/3/2014 • 8 bytes

■ embed2.zip
7/31/2014 • 4.38 KB

■ 1 item(s) selected Zip Fi

Extracting embed2.zip

embed2
embed2.zip

□ embed2a.txt
6/3/2014 • 8 bytes

■ embed2b.txt
6/3/2014 • 8 bytes

□ embed3.zip
7/31/2014 • 3.95 KB

■ 1 item(s) selected Zip

Extracting embed3.zip

embed3
embed3.zip

□ embed3.txt
7/31/2014 • 7.95 KB

■ embed4.zip
6/3/2014 • 163 bytes

■ 1 item(s) selected Zip F

Extracting embed4.zip

embed4.zip

embed4.txt
6/3/2014 • 7 bytes

□ 1 item(s) Zip File: ...

Classic XHTML

```
<?xml version="1.0" encoding="UTF-8"?>
<meta name="Content-Type" .../>
…
<p>embed_0  </p>
<p><div class="embedded" id="rId7"/>
<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"
    ...
  },
  {
    "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.

Yes, the current GUI is a bunch of xlsx files! Please help on TIKA-1334!

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

Workflow – StartDB

- **Start db:**

```
java –jar tika-eval.jar StartDB
```

- **Open browser to localhost:8082**
- **Select db (full path!):**
  - jdbc:h2:/C:/data/my_db

- **Notes on db structure: https://wiki.apache.org/tika/TikaEvalDbDesign**

Reports (Profile)

- **Metadata – count of metadata values**
- **Attachments – counts**
- **Mimes – mime counts for containers and embedded docs**
- **Exceptions**
  - Counts by type (e.g. password vs. actual exception)
  - Counts by mime
  - Counts by normalized stacktrace
  - All stack traces
- **Content**
  - Language id
  - Word count
  - Common words count
  - Word length stats
  - Page count

Reports (Compare)*

- **Metadata – comparison counts A to B**
- **Attachments – comparison counts A to B**
- **Mimes**
  - Comparison mime counts for containers and embedded docs
  - Counts of mime changes `mimeA->mimeB`
- **Exceptions**
  - Comparisons of counts by mime
  - Counts by mime
  - Counts by normalized stacktrace
  - All stack traces
- **Content**
  - Language id
  - Word count
  - Word length stats
  - Page count

Includes Profile data for both A and B and then also some comparison reports

Content – “Common words” and their Utility in Profile

- **Top 20k most common words per language in Wikipedia*, ****
  - Require > 3 letters for non-CJK
  - Remove common html markup terms, e.g. “body”, “table”
- **To find PDFs that are mostly image only:**
  - `number of words/number of pages`
- **To find very corrupt text:**
  - `number of common words/number of alphabetic words`

* Many thanks, Apache Lucene!

** Metric was recommended by Tilman Hausherr

Content Comparisons

- **Similarity metrics between A and B**
  - `how many words in common/total number of words` (with counts normalized to 0/1 per doc)
  - `how many words in common/total number of words` (with actual counts)
- **Improvement in “common words”**
  - `number of Common Words in B – number of Common Words in A`
  - Per mime

Content Comparison Example – Junk -> Better Text

- **File: commoncrawl2/KF/KFGBFTGT47L5JXJVHUL23EIB6SIMMM7C**

	Tika 1.14	Tika 1.15-SNAPSHOT
Unique Tokens	786	156
Total Tokens	1603	272
LangId	zh-cn	de
Common Words	0	116
Alphabetic Tokens	1603	250
Top N Tokens	摇敔: 18 | 獴档: 14 | 略獴: 14 | m: 11 | 杮湥: 11 | 瑶摇: 11 | 畲杮: 11 | 档湥: 10 | 搠敩: 9 | 敤浨: 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

- **File: govdocs1/519/519086.doc**

	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 | iabstract: 2 | idata: 2
Common Words/Alphabetic Tokens	7498/13472 = 56%	7409/13587 = 55%

Overlap: 95.5%

Increase in Common Words: -89

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!)**

TIKA-1302, “The Ticket is Grown; the Dream is Gone”

- Without ground truth, humans need to interpret differences
- This only makes building a gui more important!!! (TIKA-1334)
- Collaborative tagging? As a human reviews diffs, flag document as hopeless or a given extraction as “great”, “awful” (Again, thanks to Tilman Hausherr)
- Dream of TIKA-1302 ran into reality, but we’re better than where we were…

To conclude

- **Text extraction is critical to many of our projects**
- **Please evaluate – you don’t know what you can’t find!**
- **Please use tika-eval if it suits your needs**
- **Join the Apache Tika community and its evaluation efforts!**

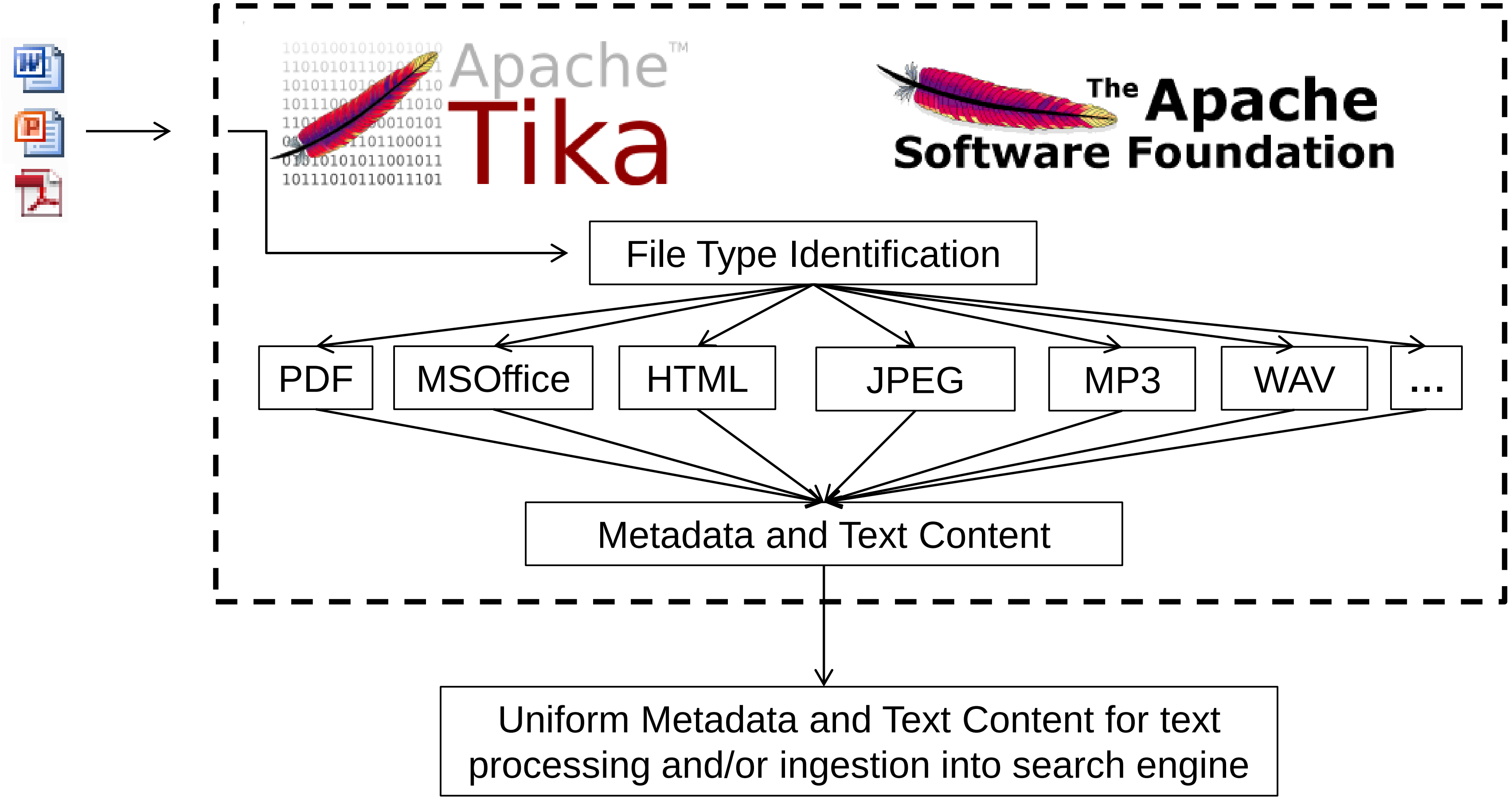
- **Email: tallison@apache.org**
- **Twitter: @_tallison**

Some Resources

- **Nick Burch's talk on Tika** **http://events.linuxfoundation.org/sites/events/files/slides/WhatsNewWithApacheTika_2.pdf**
- **tika-eval wiki – https://wiki.apache.org/tika/TikaEval**
- **Fellow traveler – Ryan Bauman's "Automatic evaluation of OCR"** **https://ryanfb.github.io/etc/2015/03/16/automatic_evaluation_of_ocr_quality.html**

Extras

Apache Tika



Apache™ Tika
The Apache Software Foundation
File Type Identification
PDF
MSOffice
HTML
JPEG
MP3
WAV
...
Metadata and Text Content
Uniform Metadata and Text Content for text processing and/or ingestion into search engine