PDF and Open Source

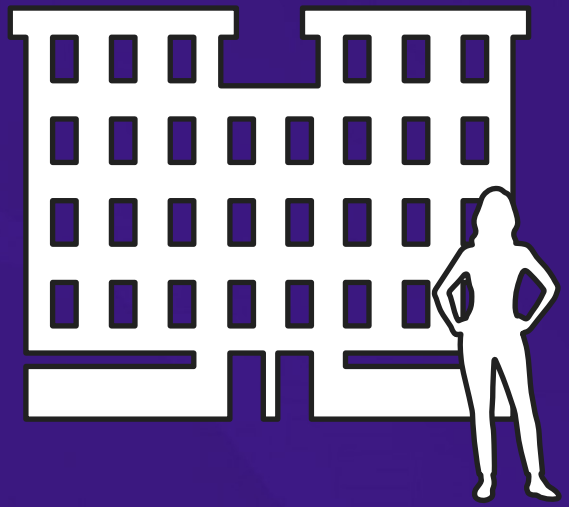
How far can we get?

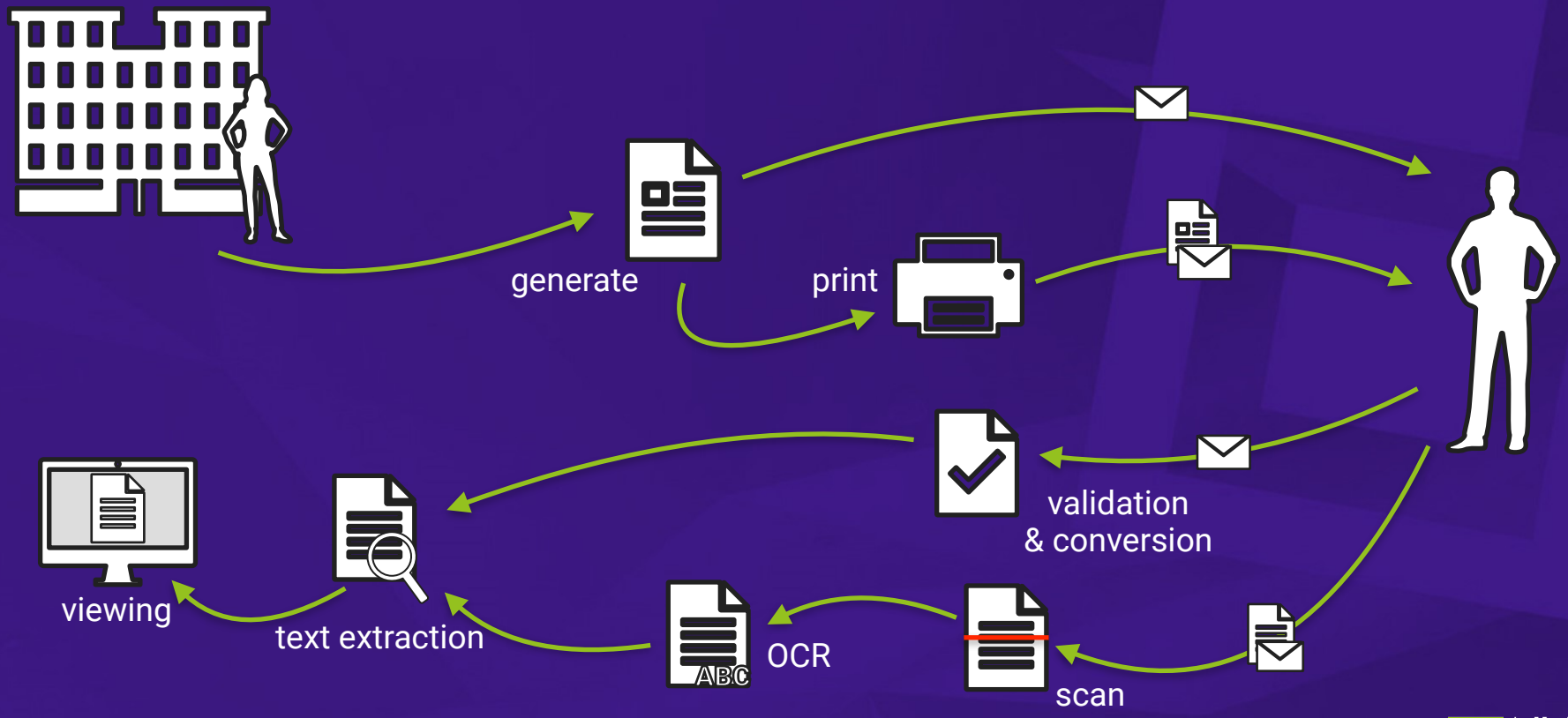
François Fernandès

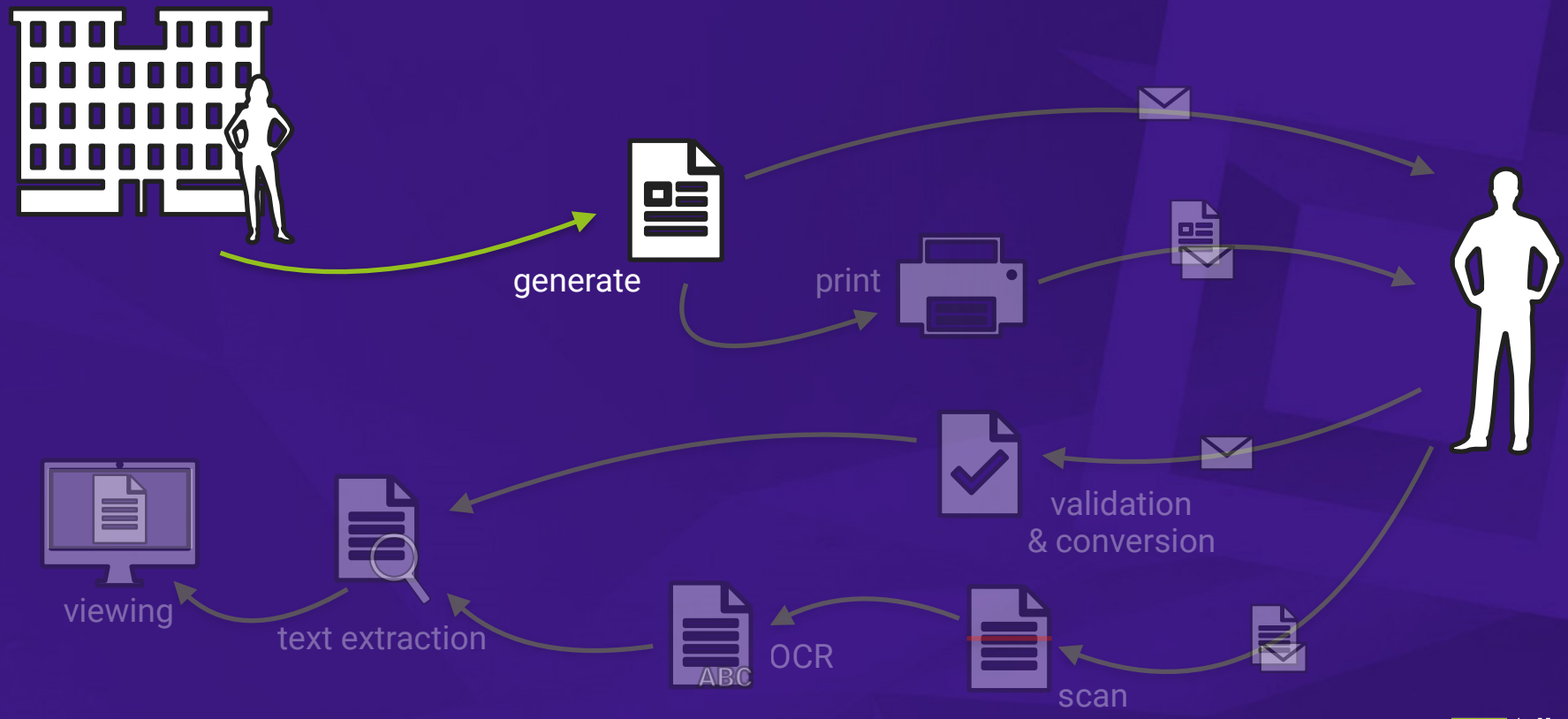
Senior Solution Architect

francois.fernandes@digitalfrontiers.de
@tellme_francois
GitHub.com/fernanfs

digital frontiers




generate
print
validation
& conversion
scan
OCR
ABC
text extraction
viewing


generate
print
validation
& conversion
scan
OCR
ABC
text extraction
viewing

GENERATE

Printing

Office Software

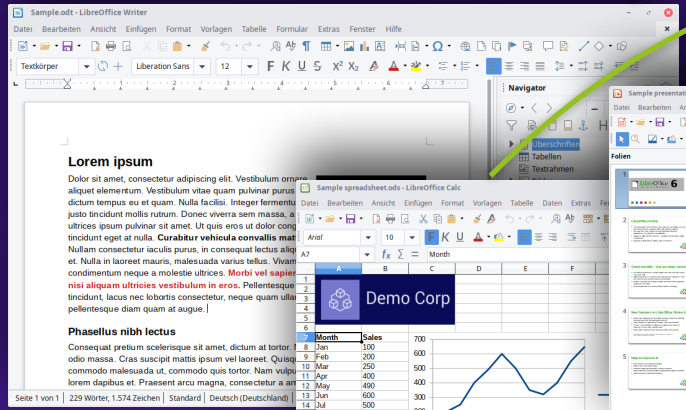
Programmatic Generation

Declarative Generation

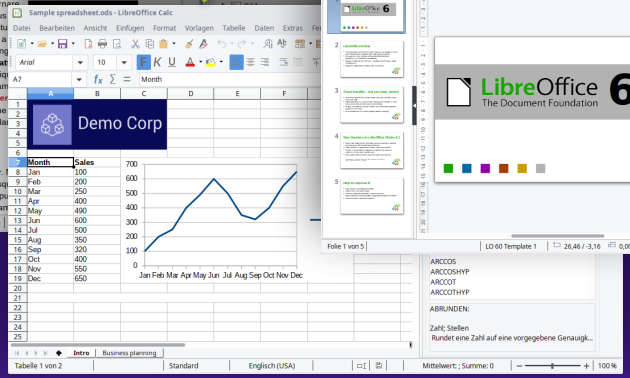
GENERATE

LibreOffice
The Document Foundation

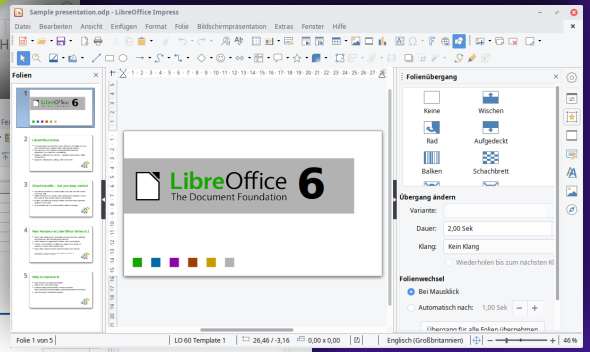
PDF



Word Processing



Spreadsheets



Presentations

digital frontiers

LibreOffice
The Document Foundation

+ mature Office Suite
+ capable of generating PDF/A
+ mostly compatible with other Office products

- some compatibility issues when opening documents of other office products

License
Mozilla Public License, Version 2

URL
https://libreoffice.org/

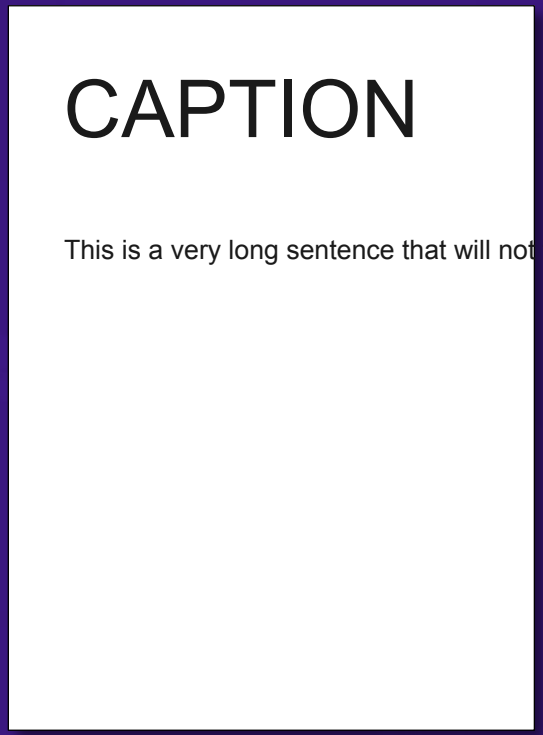
Language
C++

Platforms
Windows, MacOS, Linux

Programmatic vs. Declarative

Programmatic vs. Declarative

Programmatic vs. Declarative

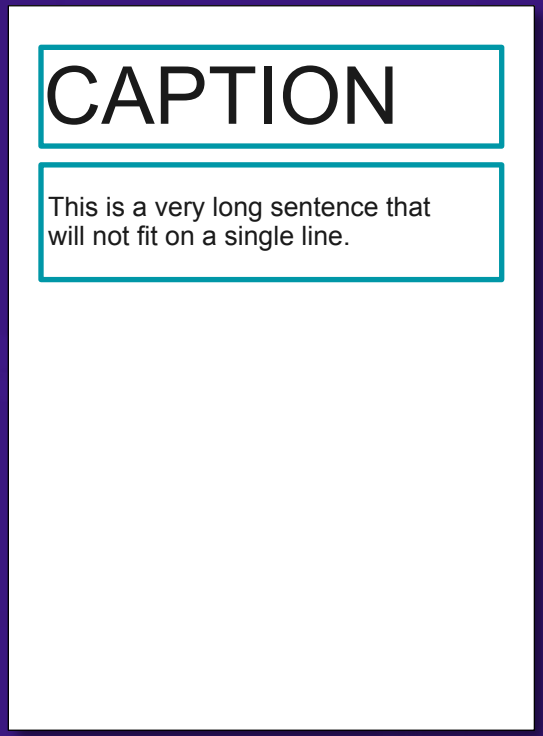



```
SetFont Arial 50
MoveTo 100,200
WriteText CAPTION

SetFont Arial 20
MoveTo 100 300
WriteText This is a very long
          sentence that will not
          fit on a single line.
```

(Extremely-Simplified PDF Instruction Set - ESPIS™)

Programmatic vs. Declarative




```
<Block style=„font:Arial;size:50“>
  CAPTION
</Block>

<Block style=„font:Arial;size:20“>
  This is a very long sentence that
  will not fit on a single line.
</Block>
```

Apache PDFBox

```
try (PDDocument doc = new PDDocument()) {
  PDPage page = new PDPage();
  doc.addPage(page);

  PDFont font = PDType1Font.HELVETICA_BOLD;

  try (PDPageContentStream contents = new PDPageContentStream(doc, page)) {
    contents.beginText();
    contents.setFont(font, 12);
    contents.newLineAtOffset(100, 700);
    contents.showText(message);
    contents.endText();
  }

  doc.save(filename);
}
```

Apache PDFBox

+ mature library
+ capable of generating PDF/A
+ full control of how the document is generated

- steep learning curve

License
Apache License, Version 2

URL
https://pdfbox.apache.org

Language
Java

Platforms
Java

Apache FOP

XSL:FO

eXtensible Stylesheet Language: Formatting Objects

GENERATE

Programmatic vs. Declarative

Apache FOP

XSL:FO

eXtensible Stylesheet Language: Formatting Objects

XML Data

```
<person>
  <name>François Fernandès</name>
  <position>Senior Solution Architect</position>
</person>
```

XSL Stylesheet

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0"
      xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
      xmlns:fo="http://www.w3.org/1999/XSL/Format">
  <xsl:output method="xml" indent="yes"/>
  <xsl:template match="/">
    <fo:root>
      <fo:layout-master-set>
        <fo:simple-page-master master-name="A4-portrait"
              page-height="29.7cm" page-width="21.0cm" margin="2cm">
          <fo:region-body/>
        </fo:simple-page-master>
      </fo:layout-master-set>
      <fo:page-sequence master-reference="A4-portrait">
        <fo:flow flow-name="xsl-region-body">
          <fo:block>
            Hello, <xsl:value-of select="person/name"/>!
          </fo:block>
        </fo:flow>
      </fo:page-sequence>
    </fo:root>
  </xsl:template>
</xsl:stylesheet>
```

FO Document

```
<?xml version="1.0" encoding="utf-8"?>
<fo:root xmlns:fo="http://www.w3.org/1999/XSL/Format">
  <fo:layout-master-set>
    <fo:simple-page-master master-name="A4-portrait"
          page-height="29.7cm" page-width="21.0cm" margin="2cm">
      <fo:region-body/>
    </fo:simple-page-master>
  </fo:layout-master-set>
  <fo:page-sequence master-reference="A4-portrait">
    <fo:flow flow-name="xsl-region-body">
      <fo:block>
        Hello, François Fernandès!
      </fo:block>
    </fo:flow>
  </fo:page-sequence>
</fo:root>
```

XSLT Processor

Hello, François Fernandès!

digital frontiers

Apache FOP

+ based on the XSL:FO Standard
+ capable of generating PDF/A and PDF/X

- moderate community activity
- XSL:FO not fully supported
- XSL:FO has a steep learning curve

License
Apache License, Version 2

URL
https://xmlgraphics.apache.org/fop/

Language
Java

Platforms
Java

```
//Initialize PDF writer
PdfWriter writer = new PdfWriter(dest);

//Initialize PDF document
PdfDocument pdf = new PdfDocument(writer);

// Initialize document
Document document = new Document(pdf);

//Add paragraph to the document
document.add(new Paragraph(message));

//Close document
document.close();
```

iText

+ extremely mature library
+ supporting way more than only declarative generation
+ commercial support available
+ available for Java as well as the .net ecosystem

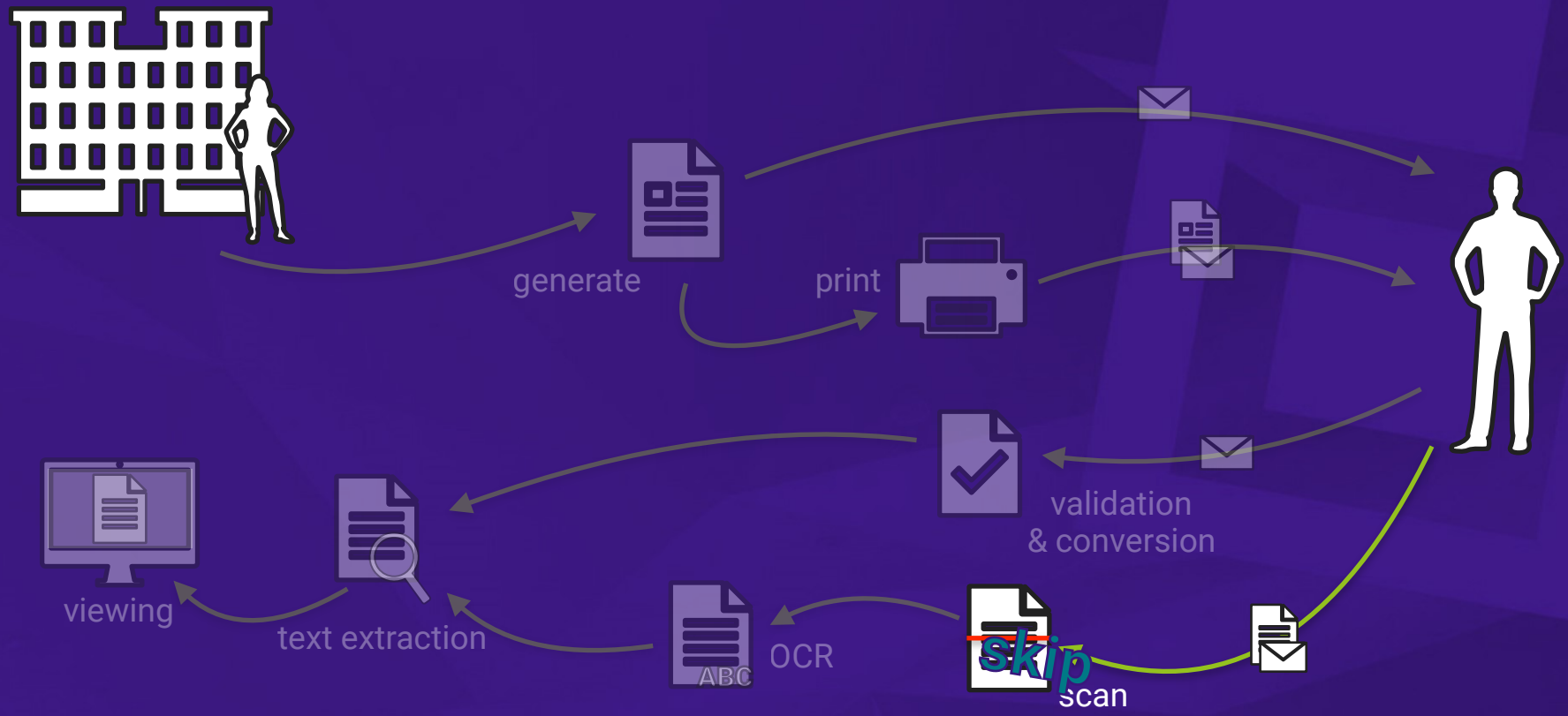
- restrictive license

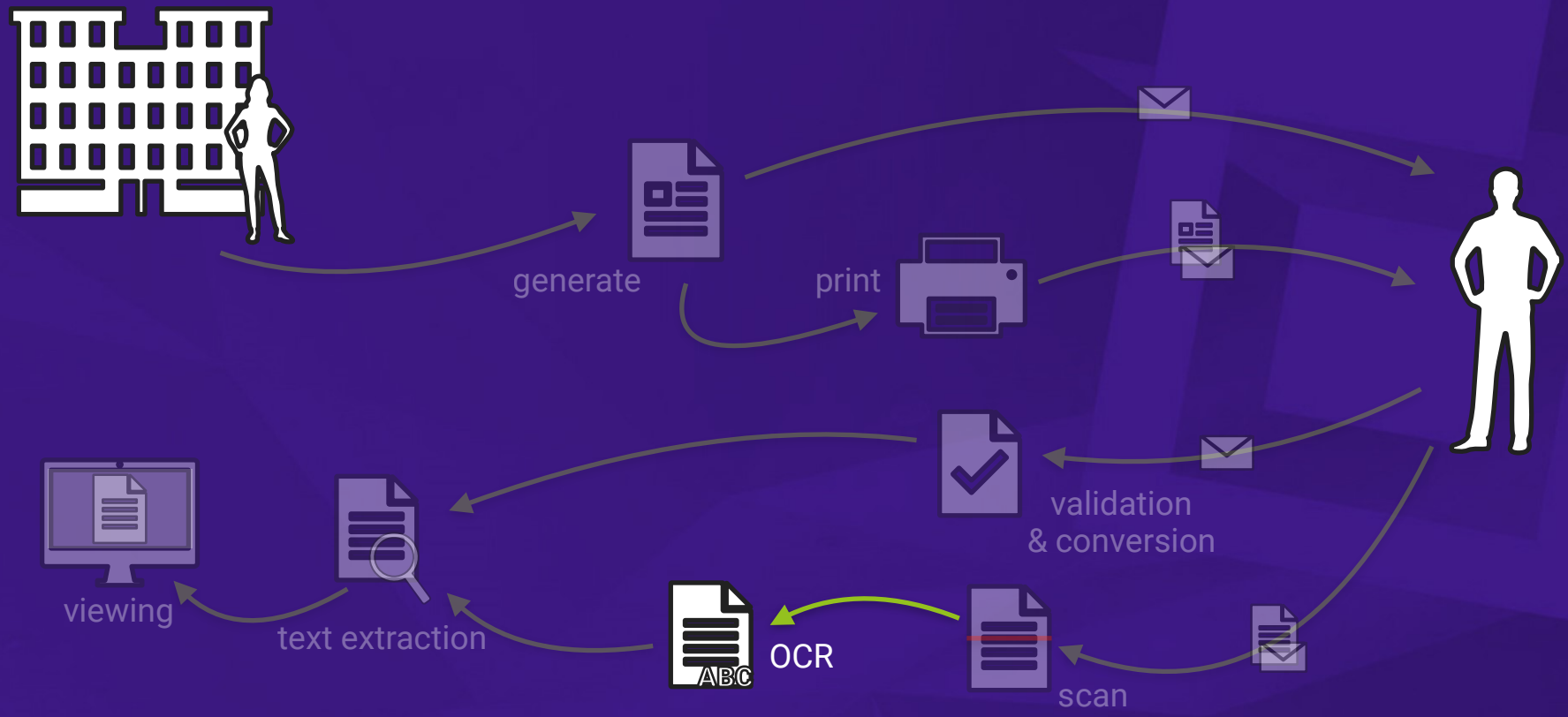
License
GNU Affero General Public License or Commercial License

URL
https://itextpdf.com

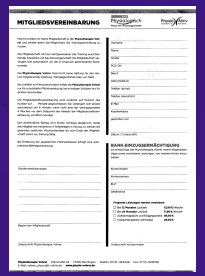
Language
Java, C#

Platforms
Java, .net


generate
print
validation
& conversion
viewing
text extraction
ABC
OCR
skip
scan


generate
print
validation
& conversion
scan
OCR
ABC
text extraction
viewing

Tesseract OCR



Scan.tiff

```
tesseract Scan.tiff out -l deu+eng
```

MITGLIEDSVEREINBARUNG

Physiologisch Physi) Aktiv

Hiermit erkläre ich meine Mitgliedschaft in der Physiotherapie Vohrer und erhalte damit die Möglichkeit, die Trainingseinrichtung zu nutzen.

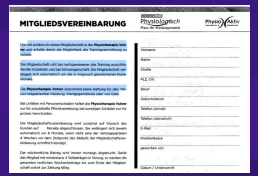
out.txt

```
tesseract Scan.tiff out -l deu+eng hocr
```

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<body>
 <div class='ocr_page' id='page_1' title='image "Sample_Scan.tiff";
  <div class='ocr_carea' id='block_1_1' title="bbox 37 62 394 82">
   <p class='ocr_par' id='par_1_1' lang='deu' title="bbox 37 62 394
    <span class='ocr_line' id='line_1_1' title="bbox 37 62 394 82;
     <span class='ocrx_word' id='word_1_1' title='bbox 37 62 394 82;
    </span>
   </p>
  </div>
```

out.hocr

```
tesseract Scan.tiff out -l deu+eng pdf
```



out.pdf

digital frontiers

Tesseract OCR

+ mature project
+ active community
+ used (and extended) by Google
+ generates searchable PDF

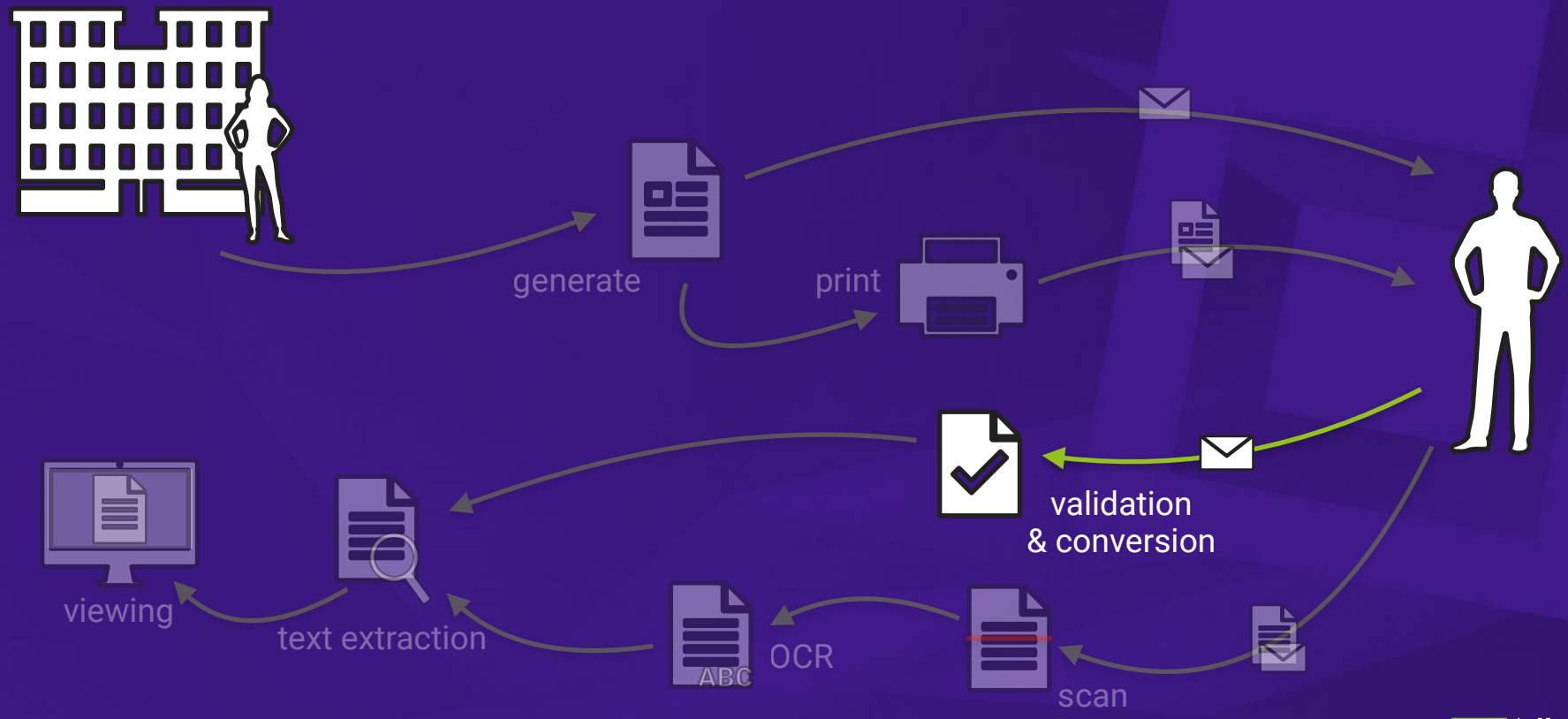
- only accepts images as input

License
Apache License, Version 2.0

URL
https://github.com/tesseract-ocr/tesseract

Language
C++

Platforms
MacOS, Windows, Linux


generate
print
validation
& conversion
viewing
text extraction
ABC
OCR
scan

veraPDF

+ active community
+ supported by the PDF Association
+ very extensive coverage of PDF/A-1, PDF/A-2 and PDF/A-3

- only validates PDF/A, not PDF in general

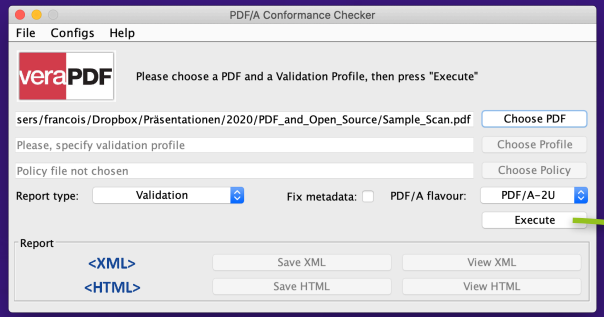
License
GNU General Public License, Version 2 or Mozilla Public License, Version 2

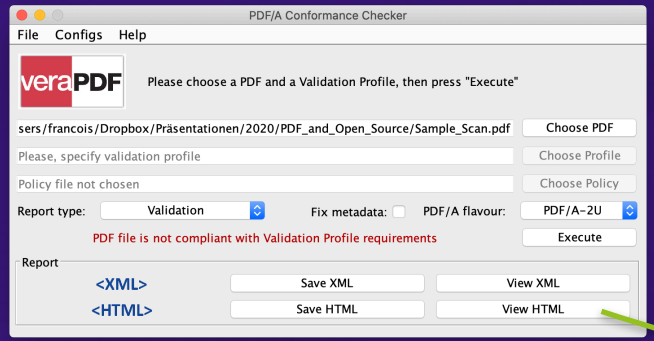
URL
https://verapdf.org

Language
Java

Platforms
MacOS, Windows, Linux

VALIDATION & CONVERSION - veraPDF

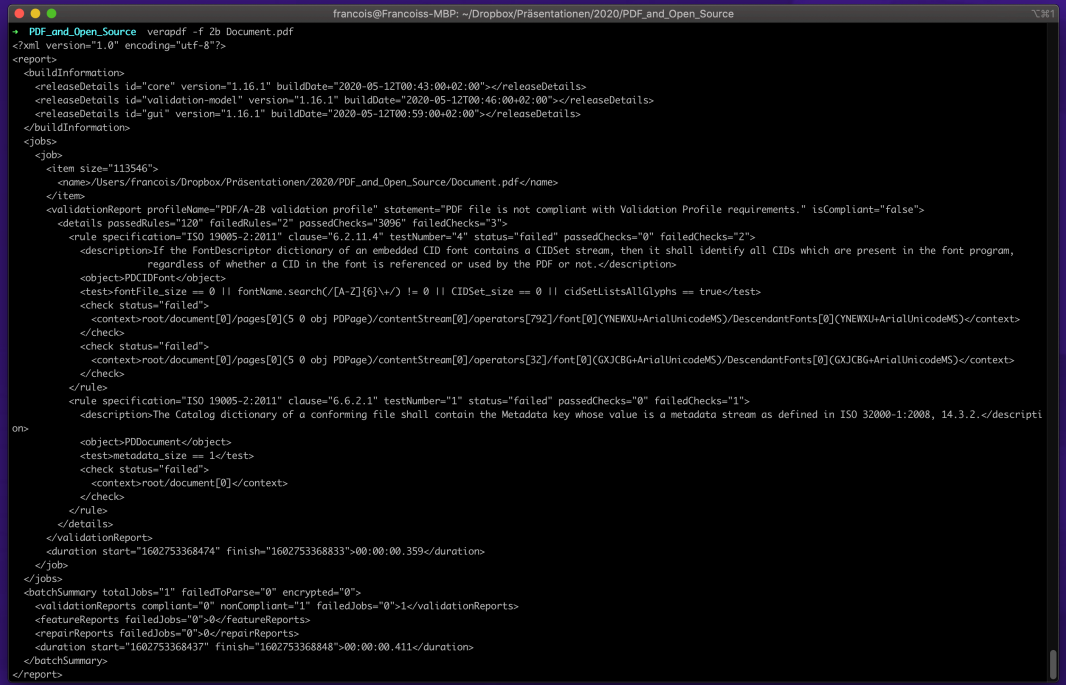




digital frontiers

VALIDATION & CONVERSION - veraPDF

```
~/PDF_and_Open_Source # verapdf -f 2b Document.pdf
```



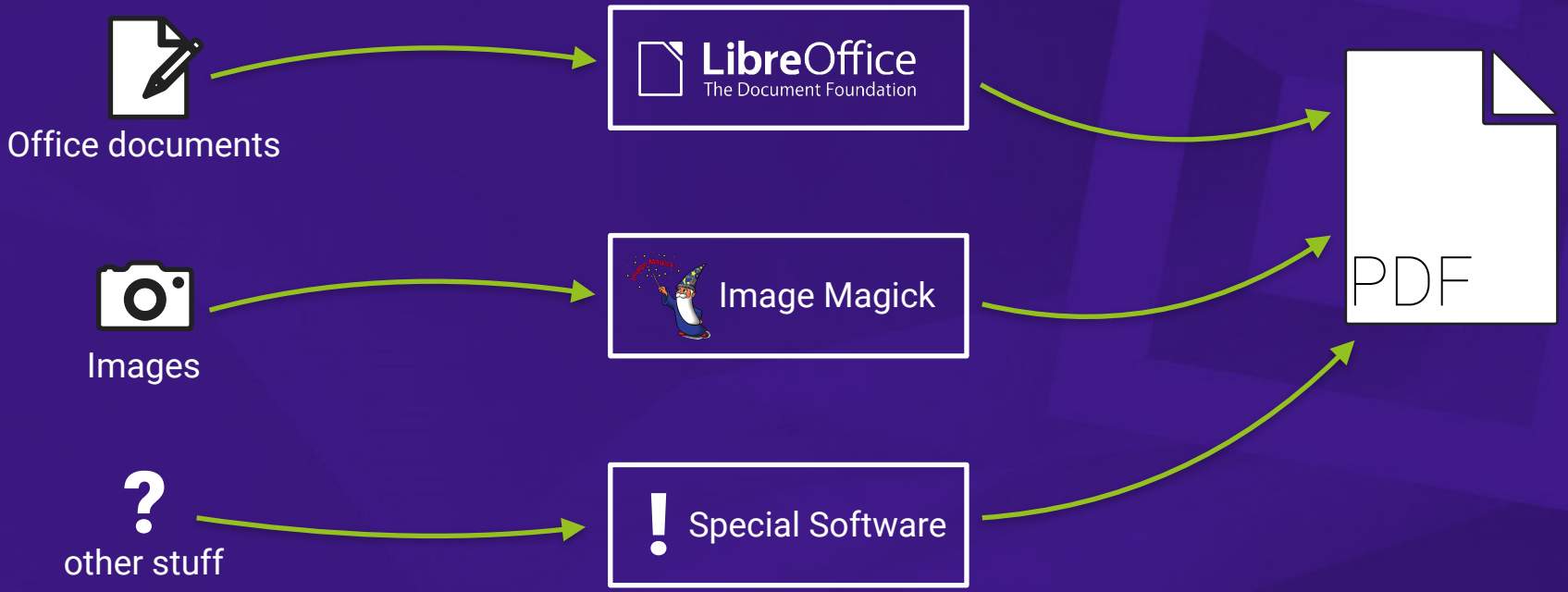
digital frontiers

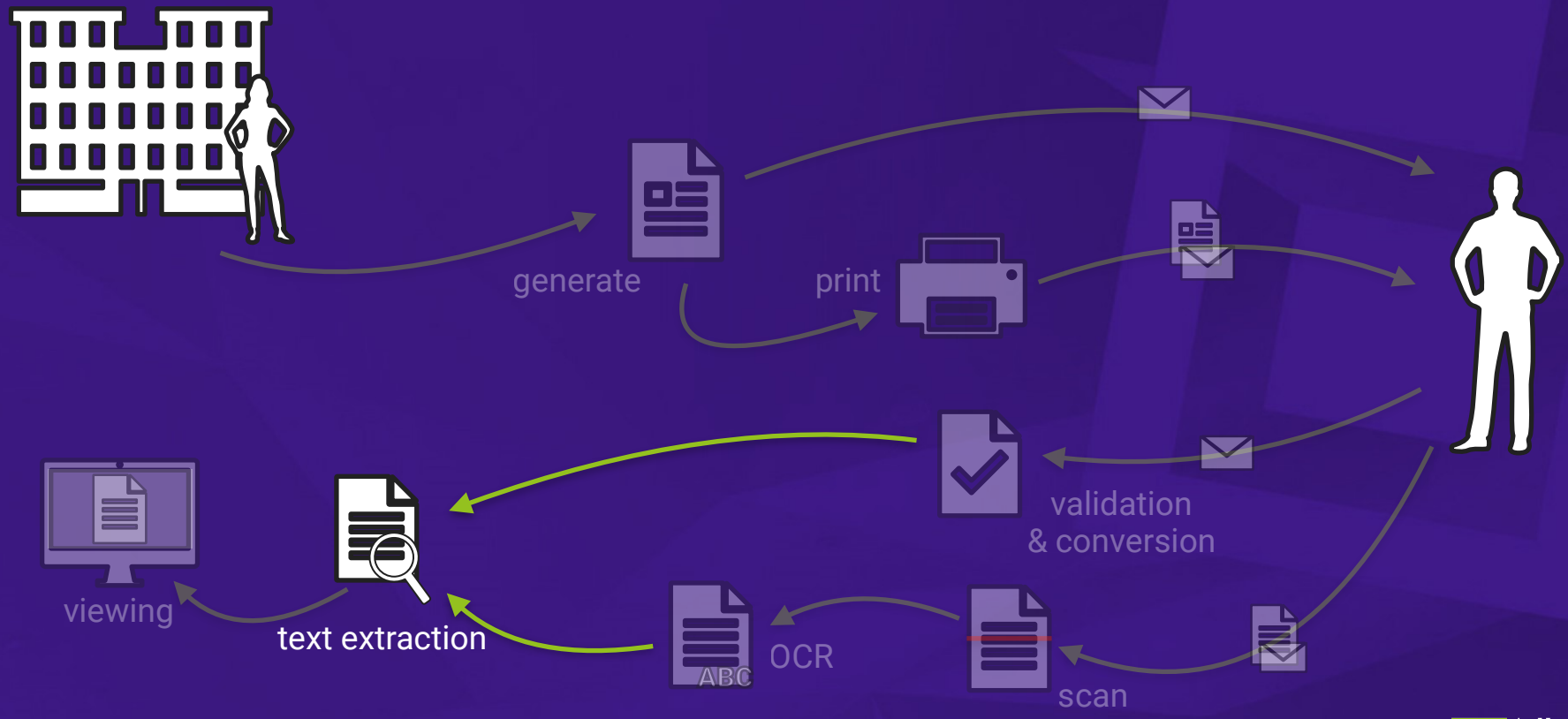
What is the best open source conversion solution?

it depends

none

it depends


Office documents
LibreOffice
The Document Foundation
Images
Image Magick
PDF
?
other stuff
!
Special Software


generate
print
validation
& conversion
scan
OCR
ABC
text extraction
viewing

Apache Tika

He certainly knows way more than me

Great Library!

October, 20, 2020 16:00

Evaluating Text Extraction at Scale

Case Study from Apache Tika

Presenter: Tim Allison
Language: English

Accessibility | Archiving

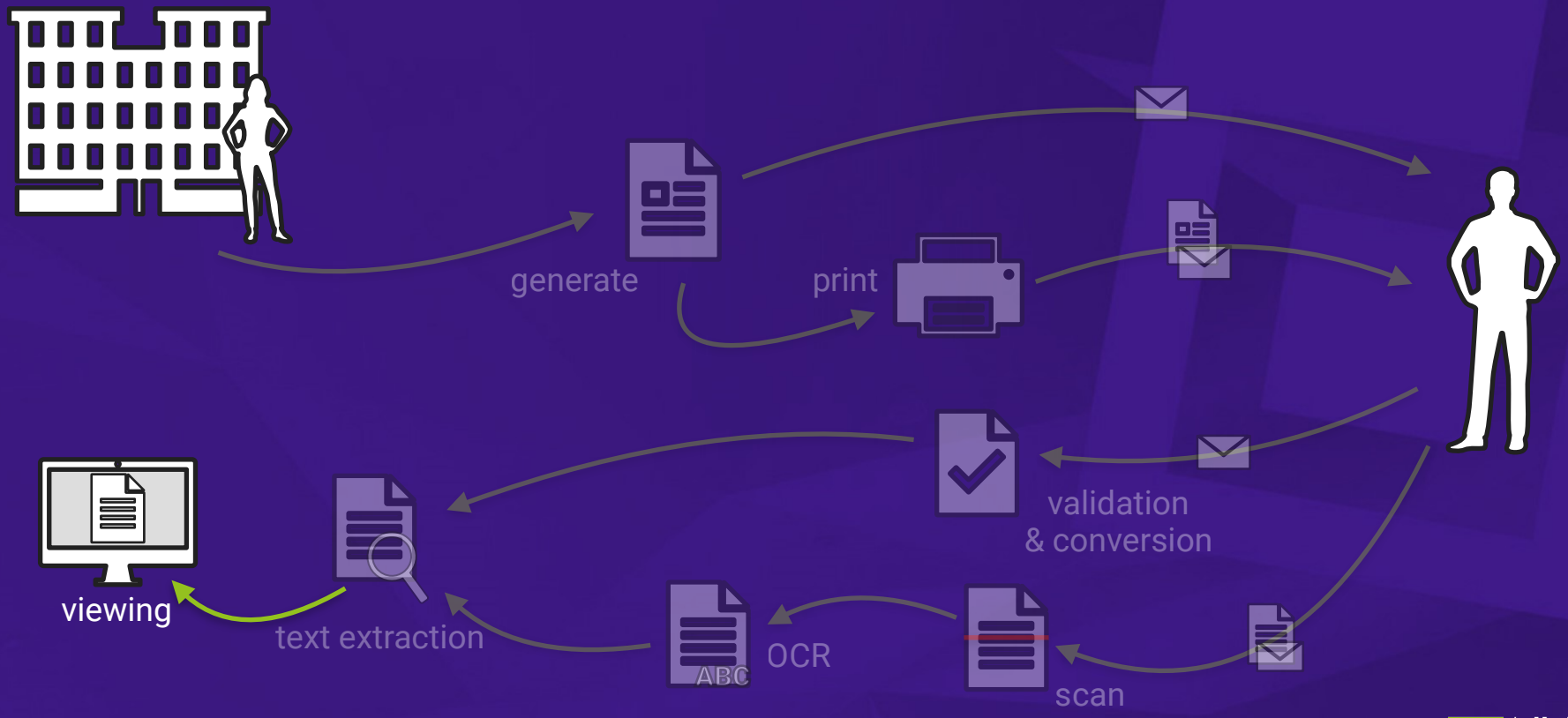
Description

Apache Tika is widely used as a critical enabling technology for search in Apache Solr and other search systems. This open source library performs text and metadata extraction from numerous file formats, including PDF via an integration with Apache PDFBox. As we all know, when something goes wrong with text extraction, the reliability of search and other natural language processing (NLP) applications is greatly hindered.

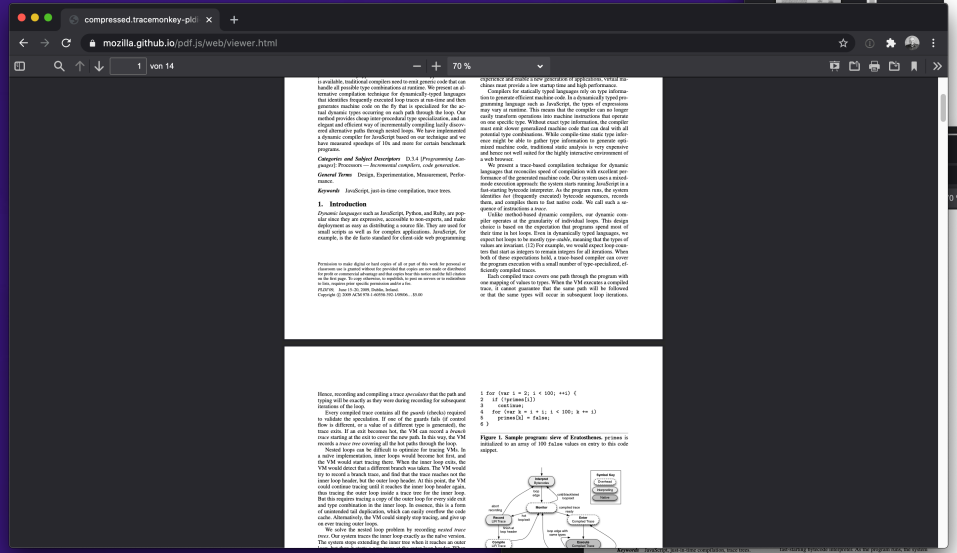
Over the last 5 years, the Tika project has gathered and published a large corpus of files (https://corpora.tika.apache.org/base/docs), and we have developed an evaluation module (tika-eval) and methodology to identify regressions in text extraction and areas for improvement in our parsers.

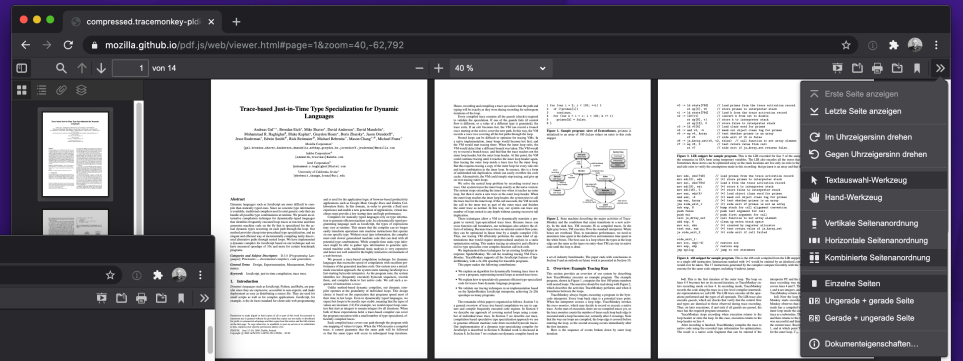
This talk offers an overview of Tika's publicly available regression corpus as well as the tika-eval module. We'll discuss ways of scaling its NLP/language-modeling based metrics to identify potential mojibake, corrupt text and/or bad OCR at scale. These techniques have applicability for PDF parser developers, search system integrators and for those interested in archiving and accessibility.

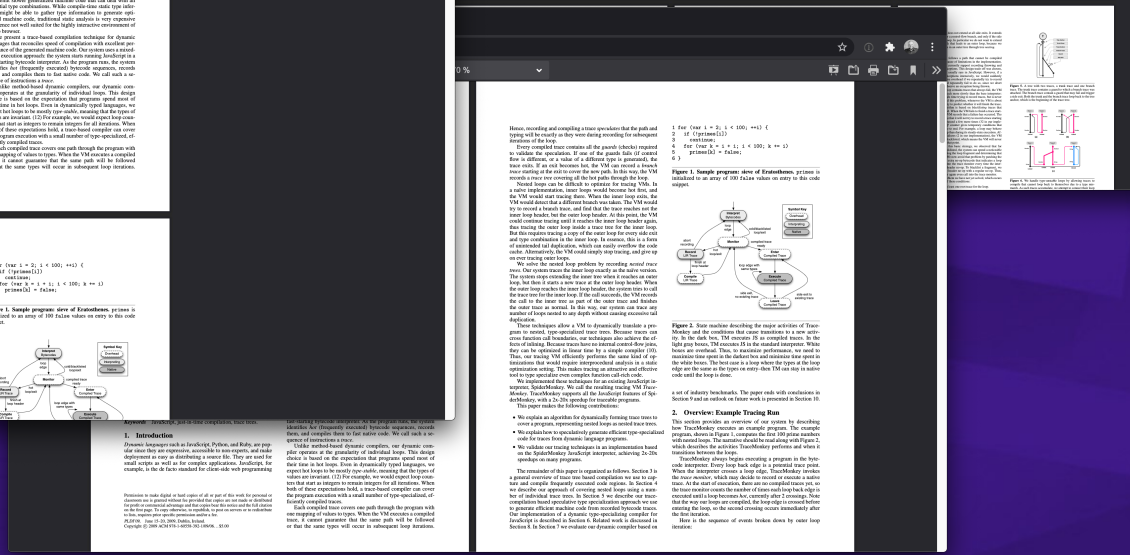
digital frontiers


generate
print
validation
& conversion
viewing
text extraction
ABC
OCR
scan

pdf.js







pdf.js

+ active community
+ used in Firefox

- struggles with some advanced PDF features
- limited by Java Script

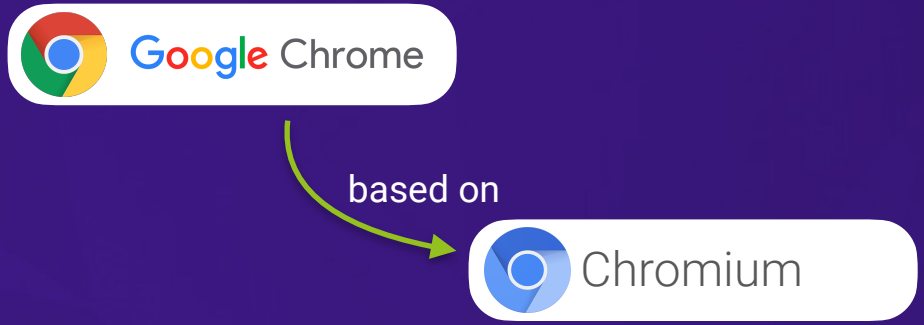
License
Apache License, Version 2

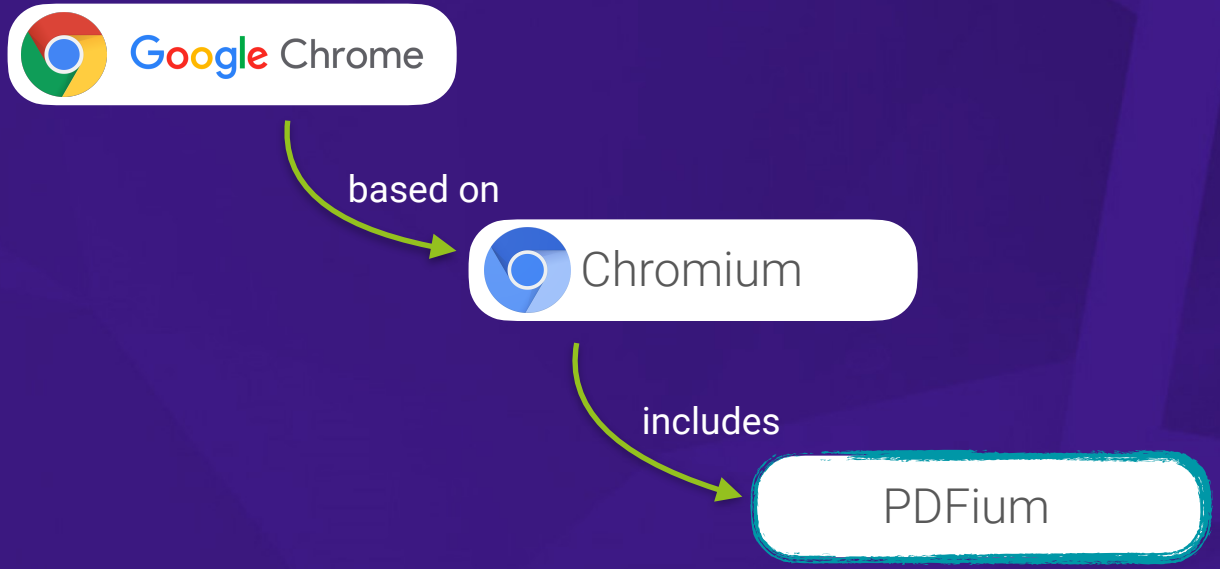
URL
https://mozilla.github.io/pdf.js/

Language
Java Script

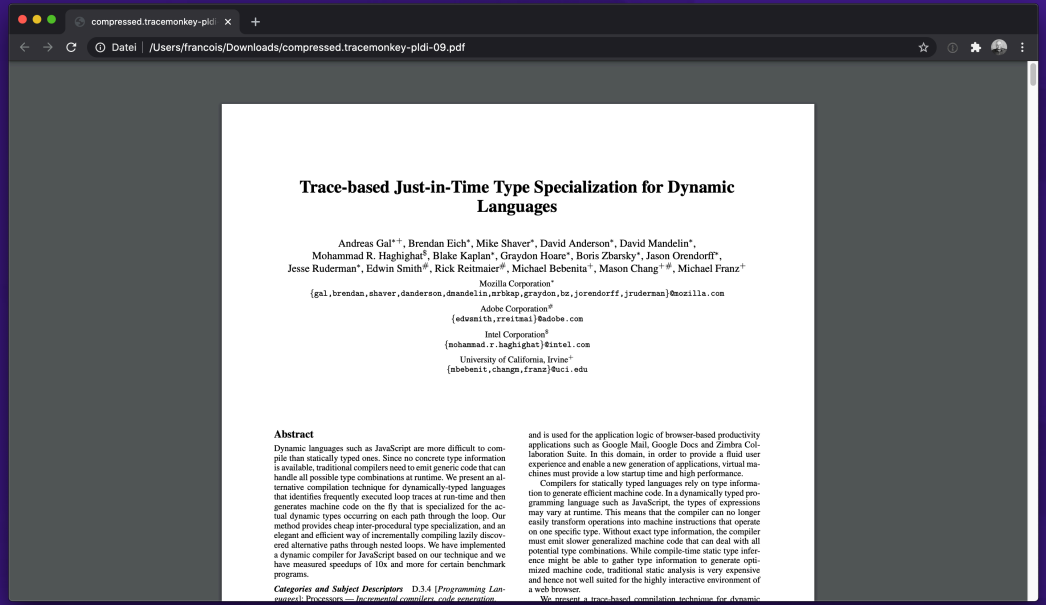
Platforms
any modern Browser


Google Chrome


Google Chrome
based on
Chromium


Google Chrome
based on
Chromium
includes
PDFium

PDFium



Trace-based Just-in-Time Type Specialization for Dynamic Languages

Andreas Gal[+*], Brendan Eich[*], Mike Shaver[*], David Anderson[*], David Mandelin[*], Mohammad R. Haghighat[§], Blake Kaplan[*], Graydon Hoare[*], Boris Zbarsky[*], Jason Orendorff[*], Jesse Ruderman[*], Edwin Smith[#], Rick Reitmaier[#], Michael Bebenita[+], Mason Chang[+#], Michael Franz[+]

Mozilla Corporation[*]
{gal,brendan,shaver,danderson,dmandelin,mrbkap,graydon,bz,jorendorff,jruderman}@mozilla.com

Adobe Corporation[#]
{edwsmith,rreitmai}@adobe.com

Intel Corporation[§]
{mohammad.r.haghighat}@intel.com

University of California, Irvine[+]
{mbebenit,changm,franz}@uci.edu

Abstract

Dynamic languages such as JavaScript are more difficult to compile than statically typed ones. Since no concrete type information is available, traditional compilers need to emit generic code that can handle all possible type combinations at runtime. We present an alternative compilation technique for dynamically-typed languages that identifies frequently executed loop traces at run-time and then generates machine code on the fly that is specialized for the actual dynamic types occurring on each path through the loop. Our method provides cheap inter-procedural type specialization, and an elegant and efficient way of incrementally compiling lazily discovered alternative paths through nested loops. We have implemented a dynamic compiler for JavaScript based on our technique and we have measured speedups of 10x and more for certain benchmark programs.

Categories and Subject Descriptors D.3.4 [*Programming Languages*]: Processors — *Incremental compilers, code generation*

and is used for the application logic of browser-based productivity applications such as Google Mail, Google Docs and Zimbra Collaboration Suite. In this domain, in order to provide a fluid user experience and enable a new generation of applications, virtual machines must provide a low startup time and high performance.

Compilers for statically typed languages rely on type information to generate efficient machine code. In a dynamically typed programming language such as JavaScript, the types of expressions may vary at runtime. This means that the compiler can no longer easily transform operations into machine instructions that operate on one specific type. Without exact type information, the compiler must emit slower generalized machine code that can deal with all potential type combinations. While compile-time static type inference might be able to gather type information to generate optimized machine code, traditional static analysis is very expensive and hence not well suited for the highly interactive environment of a web browser.

We present a trace-based compilation technique for dynamic

digital frontiers

PDFium

+ active community
+ widely available
+ PDFium is available as a library

- not many customization options in Chrom{e,ium}

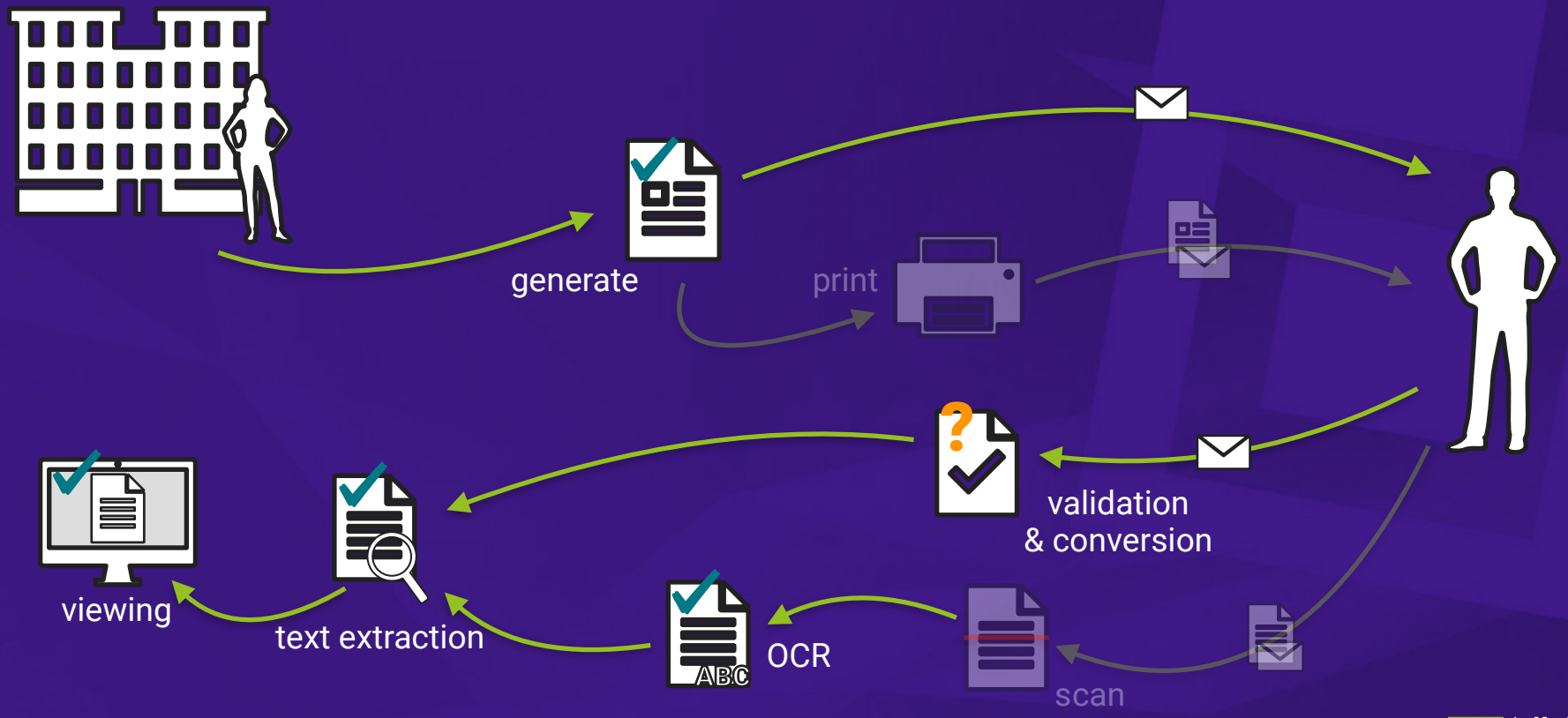
License
BSD-Style

URL
http://pdfium.org

Language
C++

Platforms
MacOS, Windows, Linux

Recap


generate
print
validation
& conversion
viewing
text extraction
OCR
ABC
scan

What is your next PDF project?

François Fernandès
Senior Solution Architect
@tellme_francois
github.com/fernanfs

https://blog.digitalfrontiers.de
https://www.digitalfrontiers.de

#WeAreHiring

digital frontiers