PDF Days Online 2021

Traversing PDF Files

SELECT * FROM pdf.future

Michaël Demey

PDF Days Online 2021

- iText Employee since 2011
  - Research Manager
- Bass Player
- Dad

- @mymilkedeek

iText

iText is a leading technology company active in the digital documents space

Flagship product

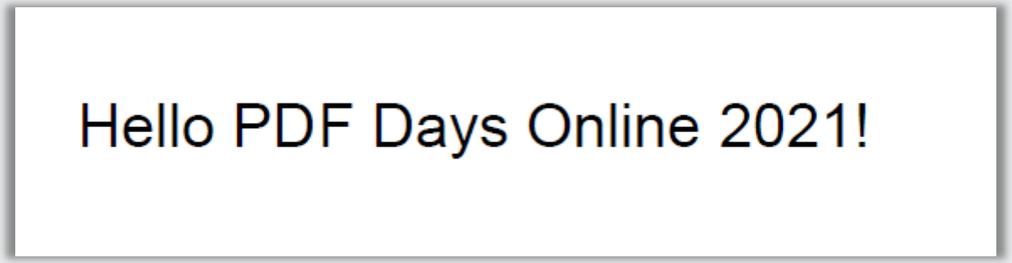
Open-source software libraries to create and manipulate PDF documents in Java and .NET (C#).

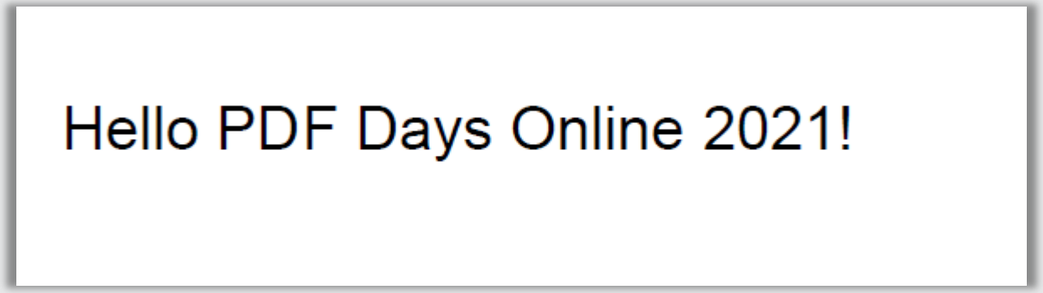
Customers

There are currently millions of iText users, both open source and commercial.

- Software developers, technology vendors, software integrators, but also
- Financial, public, government and health care sectors including many of the Fortune 500 companies.

PDF files can be identical visually…





But they can differ structurally

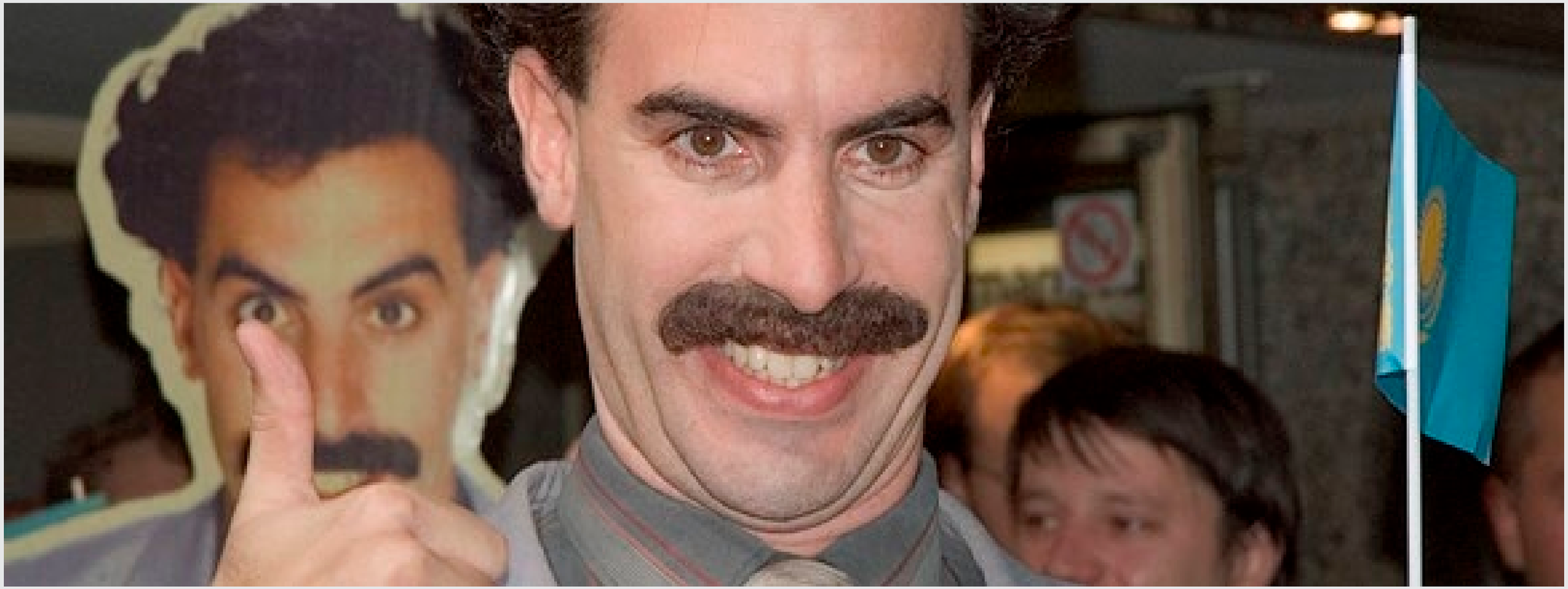
- Content Stream Syntax
- Multiple content streams instead of 1
- Use of XObjects

Navigating PDF files is easy

PDF Days Online 2021

- PDF files are *neat* data structures
- Everything is always uniformly structured
- No PDF file ever strays from how things should be

NOT



source: WikiMedia Commons https://commons.wikimedia.org/wiki/File:Borat_in_Cologne.jpg

Navigation Is Hard

PDF Days Online 2021

- Complex, different structures *per* file
  - Even within the same vendors output
- Not standardized
  - Each processor does their own thing

Demo Time!

PDF Days Online 2021

- Let’s look at 2 vendors
  - Apache PDF Box
  - iText
- Goal: Navigate PDF using these implementations

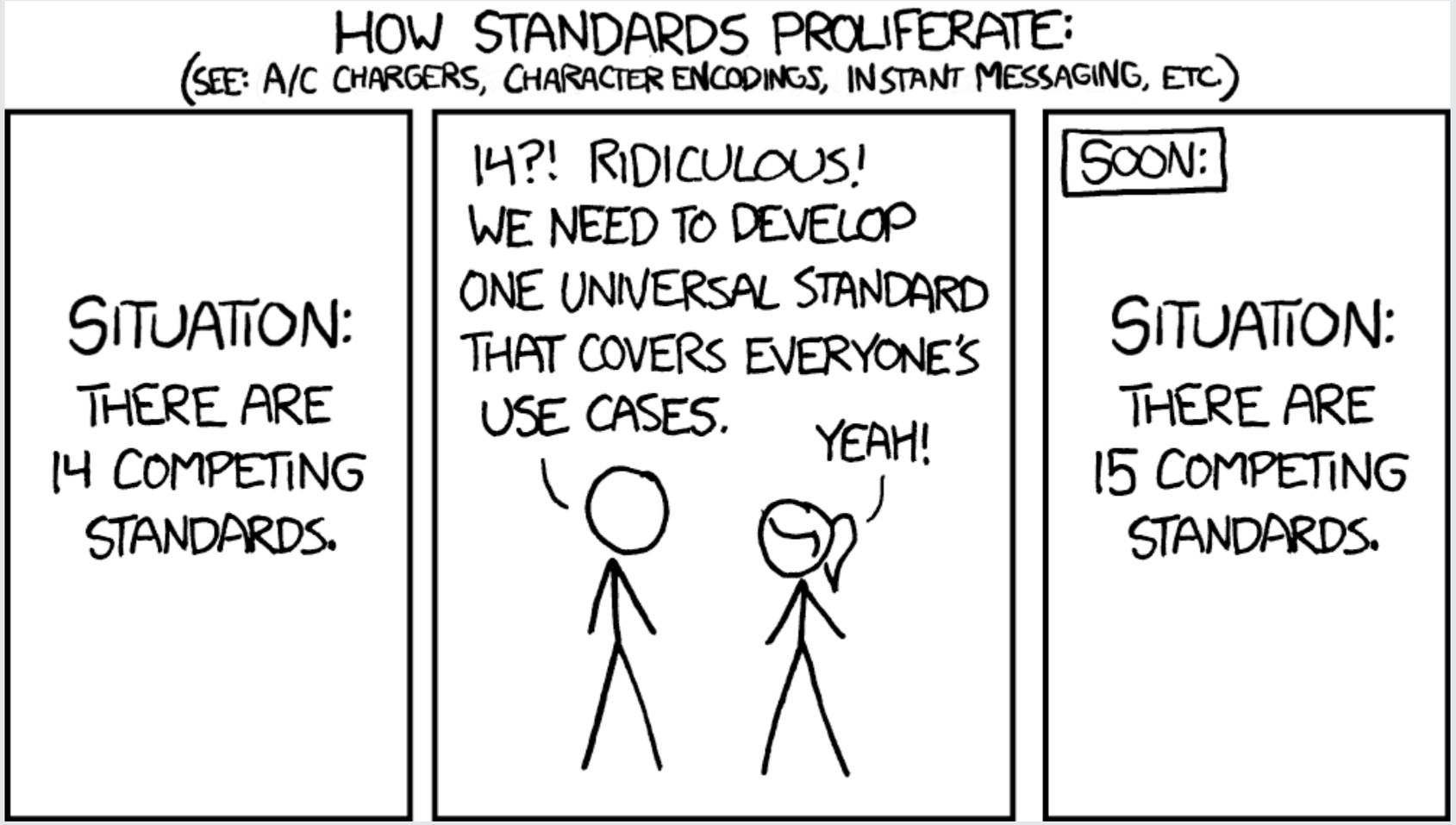
Demo Time!

source: WikiMedia Commons https://commons.wikimedia.org/wiki/File:Hands-Fingers-Crossed.jpg

To Summarize

- Things could be better across vendors
- And across files

We Need a New Standard



source: xkcd.com

Prototypes

- Word of caution:
  - The following are prototypes developed by a single vendor (me). They don’t set the rules that might be (yet).

Two Approaches

- High Level
  - PDFQL
- Low Level
  - PDFPath

PDFPath

- Navigate PDF Tree using XPath like syntax

XPath:

```
/wikimedia/projects/project/editions/*[2]
```

XML document:

```
<?xml version="1.0" encoding="utf-8"?>
<wikimedia>
 <projects>
  <project name="Wikipedia" launch="2001-01-05">
   <editions>
    <edition language="English">en.wikipedia.org</edition>
    <edition language="German">de.wikipedia.org</edition>
    <edition language="French">fr.wikipedia.org</edition>
    <edition language="Polish">pl.wikipedia.org</edition>
   </editions>
  </project>
  <project name="Wiktionary" launch="2002-12-12">
   <editions>
    <edition language="English">en.wiktionary.org</edition>
    <edition language="French">fr.wiktionary.org</edition>
    <edition language="Vietnamese">vi.wiktionary.org</edition>
    <edition language="Trukish">tr.wiktionary.org</edition>
   </editions>
  </project>
 </projects>
</wikimedia>
```

source: WikiMedia Common https://commons.wikimedia.org/wiki/File:XPath_example.svg

PDFPath Demo Time!

PDFPath (Dis)Advantages

PDF Days Online 2021

- Cross vendor implementation
  - “copy paste” across implementations
  - One PDFPath query should work everywhere
- Reduces boilerplate code

- The Query is still PDF File specific
  - Possibly even specific to a given revision of the file

PDFQL

- Works on a “higher” level than PDFPath
- High Level Objects as shortcuts/entry points

PDFQL Demo Time!

PDFQL (Dis)Advantages

- Portability
  - Across vendors, across PDF files
- More abstract
  - Requires less knowledge of the PDF specification
- Less focused, when compared to PDFPath

To Summarize

- Navigating PDF is hard
- Trying to bring two solutions to the table
  - PDFPath
  - PDFQL

Thank you!

PDF Days Online 2021

- Any questions?