Scrubber: An open source compilation to protect journalistic sources

Ethan Gregory Dodge
@egd_io
egd@truthandtransparency.org

THE TRUTH & TRANSPARENCY FOUNDATION

Who am I?

- Born and raised Mormon
- Digital forensics professional
- Journalist
- Co-Founder of the Truth & Transparency Foundation
- Craft beer enthusiast

THE TRUTH & TRANSPARENCY FOUNDATION

Truth & Transparency Foundation

A nonprofit, investigative newsroom dedicated to empowering the disenfranchised by promoting transparency within religious institutions.

THE TRUTH & TRANSPARENCY FOUNDATION

THE TRUTH & TRANSPARENCY FOUNDATION

THE TRUTH & TRANSPARENCY FOUNDATION

Problem

- I was spending a lot of time cleaning, compressing, and optimizing PDFs.
- Largely manual, slowly automated what I could.

THE TRUTH & TRANSPARENCY FOUNDATION

Solution: Scrubber

- Leverages PDF Redact Tools, OCRmyPDF, and QPDF to clean, OCR, and linearize.
- Bash script tying them all together

THE TRUTH & TRANSPARENCY FOUNDATION

Solution: Scrubber

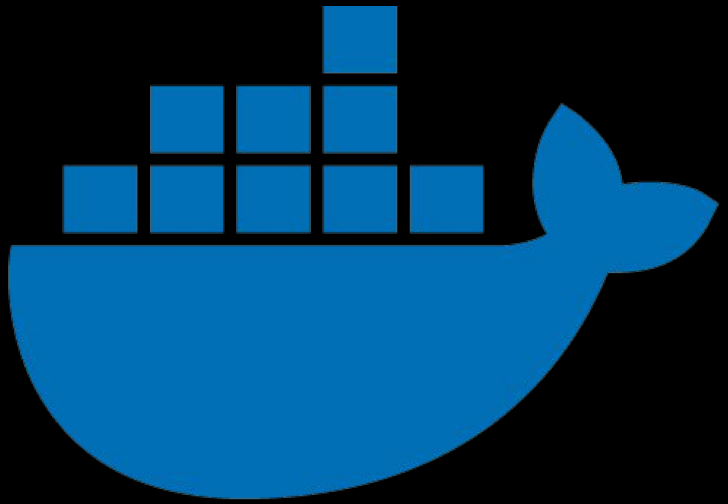
- Turns every page into png using PDF Redact Tools
  - Optionally can redact individual images
- Combines them all again into a PDF, effectively stripping any embedded image data.

THE TRUTH & TRANSPARENCY FOUNDATION

Solution: Scrubber

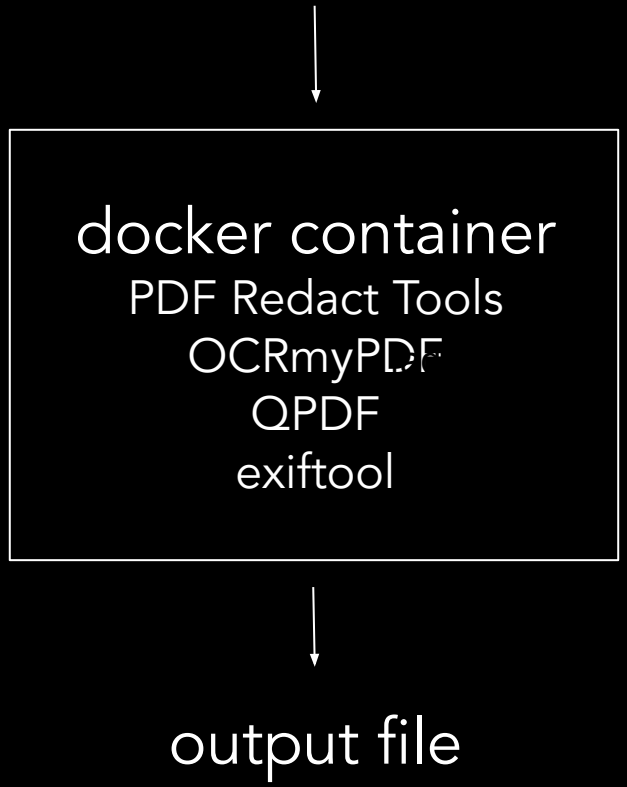
- Adds the text layer using OCRmyPDF
  - OCRmyPDF also handle compression
- QPDF linearizes it for optimum web hosting
- All the exif data created in the process is then removed by exif tool

THE TRUTH & TRANSPARENCY FOUNDATION



docker

THE TRUTH & TRANSPARENCY FOUNDATION



THE TRUTH & TRANSPARENCY FOUNDATION

Benefits to Scrubber

- Can run on any OS thanks to Docker
  - Pull from Docker Hub or build locally
- Can handle large PDFs (if your machine can)
- Provides a one stop tool for cleaning and optimization
- Supports “batches” of PDFs

THE TRUTH & TRANSPARENCY FOUNDATION

https://github.com/truthandtransparency/scrubber

THE TRUTH & TRANSPARENCY FOUNDATION

Scrubber: An open source compilation to protect journalistic sources

Ethan Gregory Dodge
@egd_io
egd@truthandtransparency.org

THE TRUTH & TRANSPARENCY FOUNDATION