Open Source OCR and PDF compression at the Internet Archive

Switching to Open Source Software

Merlijn B.W. Wajer (merlijn@archive.org)

August 14, 2021
Internet Archive

Overview

- Who am I?
- Internet Archive (IA)

Structure of the presentation

- OCR: what is it and what is it used for at IA?
- PDF generation and compression

What is OCR?

Optical Character Recognition: "reading text from photos"

Why do we need OCR?

- Document analysis, exploration and accessibility

What is OCR?

Optical Character Recognition: "reading text from photos"

Why do we need OCR?

- Document analysis, exploration and accessibility

Examples:

- Linking to specific pages (analysis)

What is OCR?

Optical Character Recognition: "reading text from photos"

Why do we need OCR?

- Document analysis, exploration and accessibility

Examples:

- Linking to specific pages (analysis)
- Creation of other downstream formats like plaintext, PDF, ePUB (accessibility)

What is OCR?

Optical Character Recognition: "reading text from photos"

Why do we need OCR?

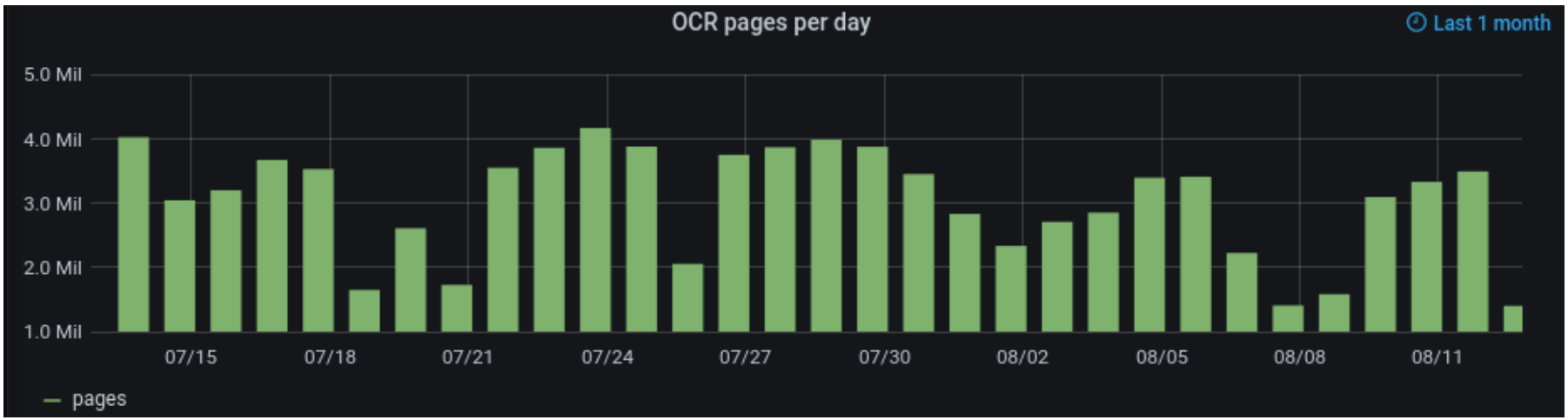
- Document analysis, exploration and accessibility

Examples:

- Linking to specific pages (analysis)
- Creation of other downstream formats like plaintext, PDF, ePUB (accessibility)
- Full text search (exploration)

Scale

- 3+ million pages every day



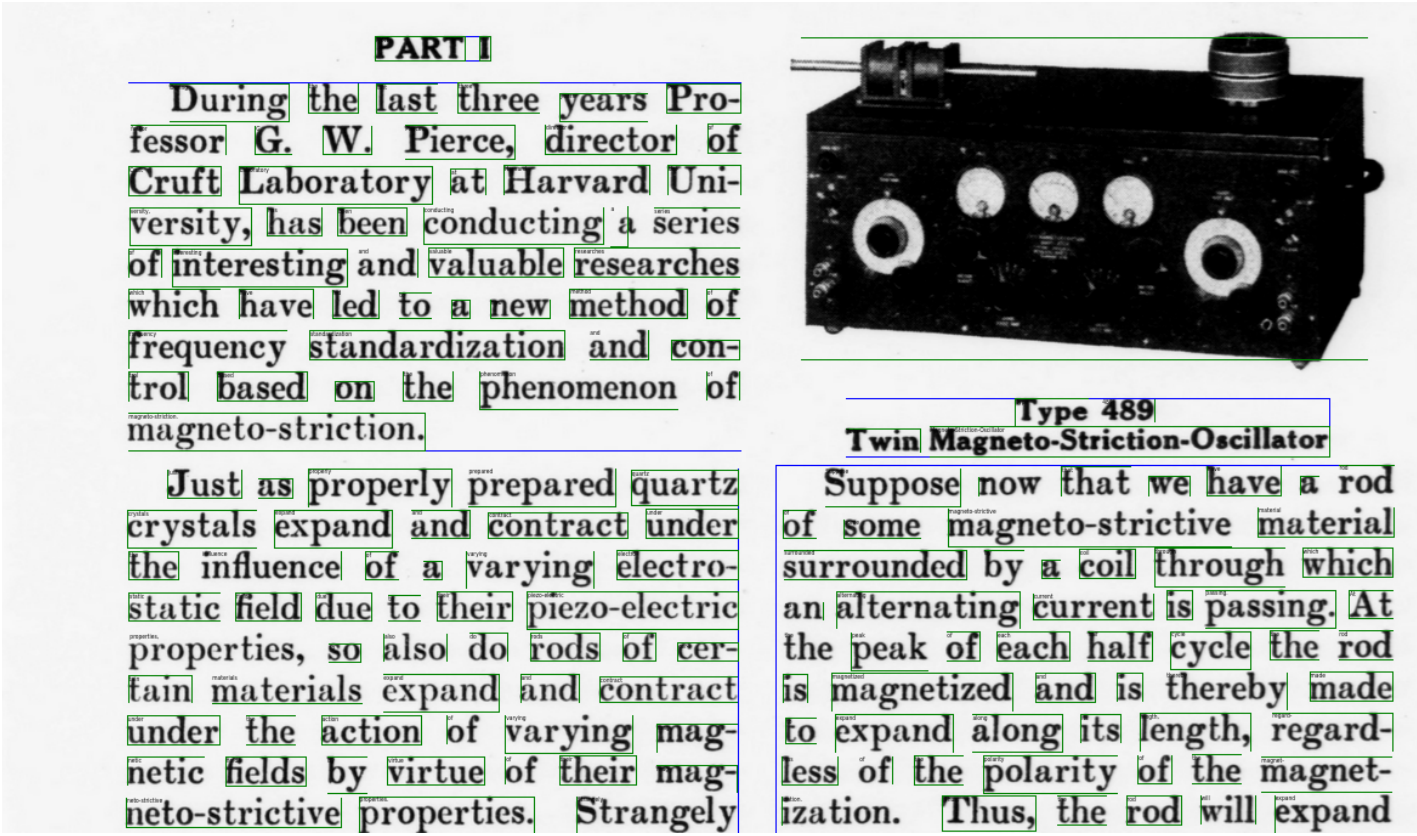
Moving to an OSS stack

Moving away from Abbyy

- Prior experience with Tesseract: used to deskew images in microfilm project
- Available engines: Tesseract, Calamari and ocropy
- File formats: Abbyy XML, ALTO, PAGE XML, hOCR, ...

We picked Tesseract as engine and hOCR as format

What a OCR result looks like: hocrjs



PART I

During the last three years Professor G. W. Pierce, director of Cruft Laboratory at Harvard University, has been conducting a series of interesting and valuable researches which have led to a new method of frequency standardization and control based on the phenomenon of magneto-striction.

Just as properly prepared quartz crystals expand and contract under the influence of a varying electrostatic field due to their piezo-electric properties, so also do rods of certain materials expand and contract under the action of varying magnetic fields by virtue of their magneto-strictive properties. Strangely

Type 489
Twin Magneto-Striction-Oscillator

Suppose now that we have a rod of some magneto-strictive material surrounded by a coil through which an alternating current is passing. At the peak of each half cycle the rod is magnetized and is thereby made to expand along its length, regardless of the polarity of the magnetization. Thus, the rod will expand

- https://archive.org/services/hocr-view/view?identifier=sim_general-radio-experimenter_1928-06_3_1
- https://github.com/kba/hocrjs

PDF generation

By HORATIO W. LAMSON, Engineering Department

PART I

During the last three years Professor G. W. Pierce, director of Cruft Laboratory at Harvard University, has been conducting a series of interesting and valuable researches which have led to a new method of frequency standardization and control based on the phenomenon of magneto-striction.

Just as properly prepared quartz crystals expand and contract under the influence of a varying electrostatic field due to their piezo-electric properties, so also do rods of certain materials expand and contract under the action of varying magnetic fields by virtue of their magneto-strictive properties. Strangely enough, pure iron, and steels which are alloys of iron and carbon, although they are strongly magnetic, show only very feeble magneto-strictive effects. On the other hand, pure nickel, which is only slightly magnetic, gives a strong magneto-strictive response. Alloys of nickel and iron in certain proportions are active, especially those having about 36% nickel and 64% iron, which is the approximate composition of invar

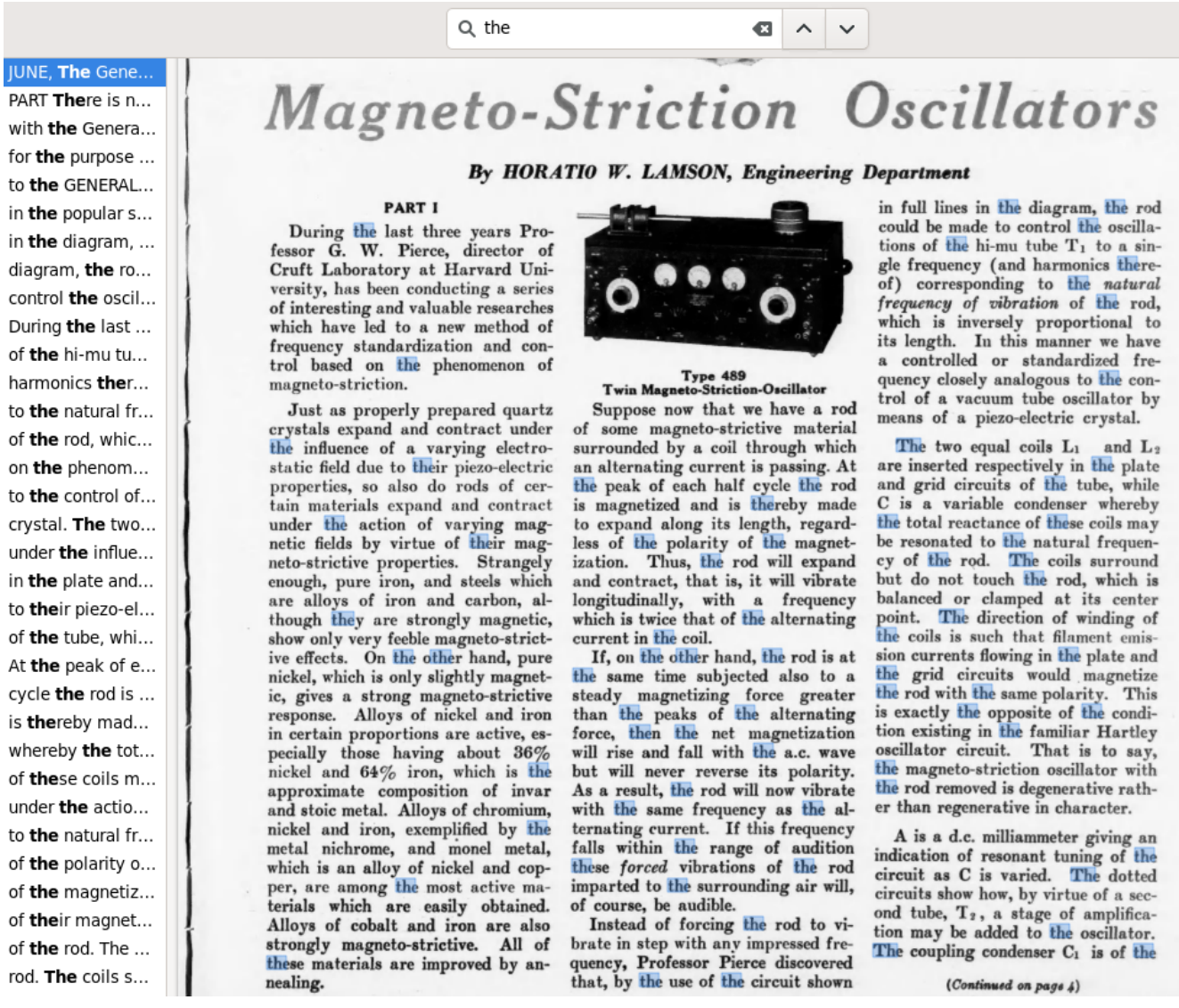
Type 489
Twin Magneto-Striction-Oscillator

Suppose now that we have a rod of some magneto-strictive material surrounded by a coil through which an alternating current is passing. At the peak of each half cycle the rod is magnetized and is thereby made to expand along its length, regardless of the polarity of the magnetization. Thus, the rod will expand and contract, that is, it will vibrate longitudinally, with a frequency which is twice that of the alternating current in the coil.

If, on the other hand, the rod is at the same time subjected also to a steady magnetizing force greater than the peaks of the alternating force, then the net magnetization will rise and fall with the a.c. wave but will never reverse its polarity. As a result, the rod will now vibrate in full lines in the diagram, the rod could be made to control the oscillations of the hi-mu tube T_1 to a single frequency (and harmonics thereof) corresponding to the *natural frequency of vibration* of the rod, which is inversely proportional to its length. In this manner we have a controlled or standardized frequency closely analogous to the control of a vacuum tube oscillator by means of a piezo-electric crystal.

The two equal coils L_1 and L_2 are inserted respectively in the plate and grid circuits of the tube, while C is a variable condenser whereby the total reactance of these coils may be resonated to the natural frequency of the rod. The coils surround but do not touch the rod, which is balanced or clamped at its center point. The direction of winding of the coils is such that filament emission currents flowing in the plate and the grid circuits would magnetize the rod with the same polarity. This is exactly the opposite of the condition existing in the familiar Hartley oscillator circuit. That is to say, the magneto-striction oscillator with the rod removed is degenerative rath-

PDF text searching



Magneto-Striction Oscillators

By HORATIO W. LAMSON, Engineering Department

PART I

During the last three years Professor G. W. Pierce, director of Cruft Laboratory at Harvard University, has been conducting a series of interesting and valuable researches which have led to a new method of frequency standardization and control based on the phenomenon of magneto-striction.

Just as properly prepared quartz crystals expand and contract under the influence of a varying electrostatic field due to their piezo-electric properties, so also do rods of certain materials expand and contract under the action of varying magnetic fields by virtue of their magneto-strictive properties. Strangely enough, pure iron, and steels which are alloys of iron and carbon, although they are strongly magnetic, show only very feeble magneto-strictive effects. On the other hand, pure nickel, which is only slightly magnetic, gives a strong magneto-strictive response. Alloys of nickel and iron in certain proportions are active, especially those having about 36% nickel and 64% iron, which is the approximate composition of invar and stoic metal. Alloys of chromium, nickel and iron, exemplified by the metal nichrome, and monel metal, which is an alloy of nickel and copper, are among the most active materials which are easily obtained. Alloys of cobalt and iron are also strongly magneto-strictive. All of these materials are improved by annealing.

Type 489
Twin Magneto-Striction-Oscillator

Suppose now that we have a rod of some magneto-strictive material surrounded by a coil through which an alternating current is passing. At the peak of each half cycle the rod is magnetized and is thereby made to expand along its length, regardless of the polarity of the magnetization. Thus, the rod will expand and contract, that is, it will vibrate longitudinally, with a frequency which is twice that of the alternating current in the coil.

If, on the other hand, the rod is at the same time subjected also to a steady magnetizing force greater than the peaks of the alternating force, then the net magnetization will rise and fall with the a.c. wave but will never reverse its polarity. As a result, the rod will now vibrate with the same frequency as the alternating current. If this frequency falls within the range of audition these *forced* vibrations of the rod imparted to the surrounding air will, of course, be audible.

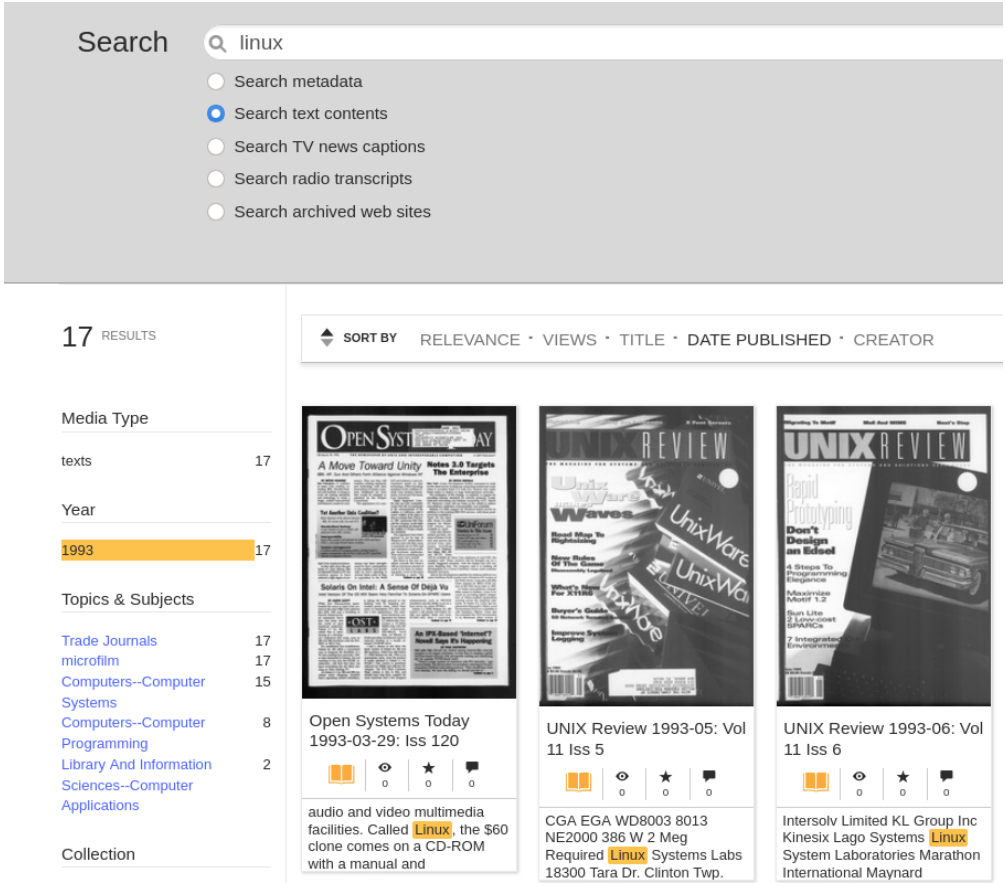
Instead of forcing the rod to vibrate in step with any impressed frequency, Professor Pierce discovered that, by the use of the circuit shown in full lines in the diagram, the rod could be made to control the oscillations of the hi-mu tube T_1 to a single frequency (and harmonics thereof) corresponding to the *natural frequency of vibration* of the rod, which is inversely proportional to its length. In this manner we have a controlled or standardized frequency closely analogous to the control of a vacuum tube oscillator by means of a piezo-electric crystal.

The two equal coils L_1 and L_2 are inserted respectively in the plate and grid circuits of the tube, while C is a variable condenser whereby the total reactance of these coils may be resonated to the natural frequency of the rod. The coils surround but do not touch the rod, which is balanced or clamped at its center point. The direction of winding of the coils is such that filament emission currents flowing in the plate and the grid circuits would magnetize the rod with the same polarity. This is exactly the opposite of the condition existing in the familiar Hartley oscillator circuit. That is to say, the magneto-striction oscillator with the rod removed is degenerative rather than regenerative in character.

A is a d.c. milliammeter giving an indication of resonant tuning of the circuit as C is varied. The dotted circuits show how, by virtue of a second tube, T_2, a stage of amplification may be added to the oscillator. The coupling condenser C_1 is of the

(Continued on page 4)

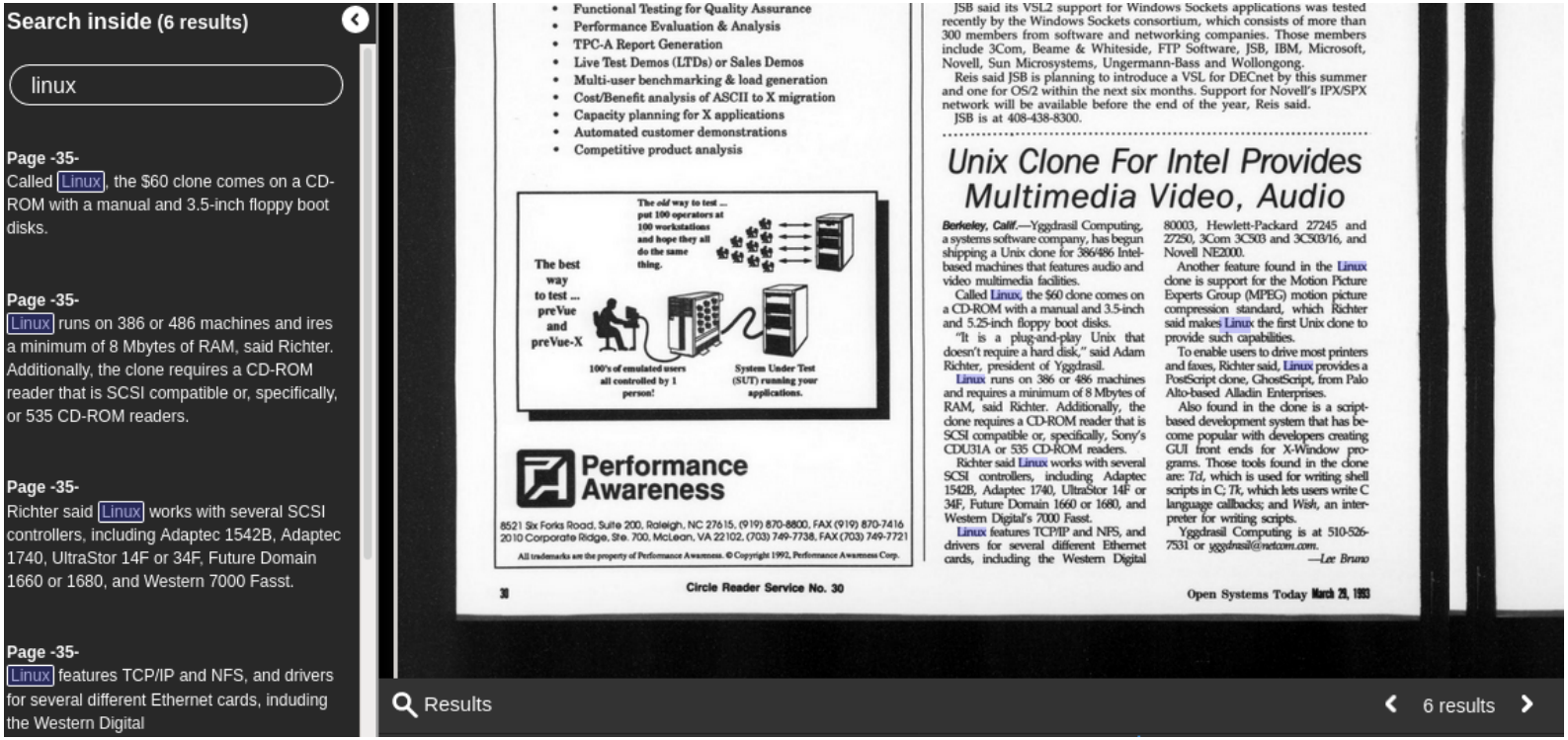
Full text searching in lots of documents



Magazine from March 1993 mentions Linux

First release: September 1991

Searching inside



Link to this issue

Just $60 for a copy of the UNIX clone...

Searching inside: speeding up

Parsing (big) XML documents for every search is slow

Searching inside: speeding up

Parsing (big) XML documents for every search is slow

Solution:

- Preprocessed plain text to compare against elastic search
- page index to map plaintext bytes to XML page byte ranges

Depending on the amount of matches, results can come back within a second

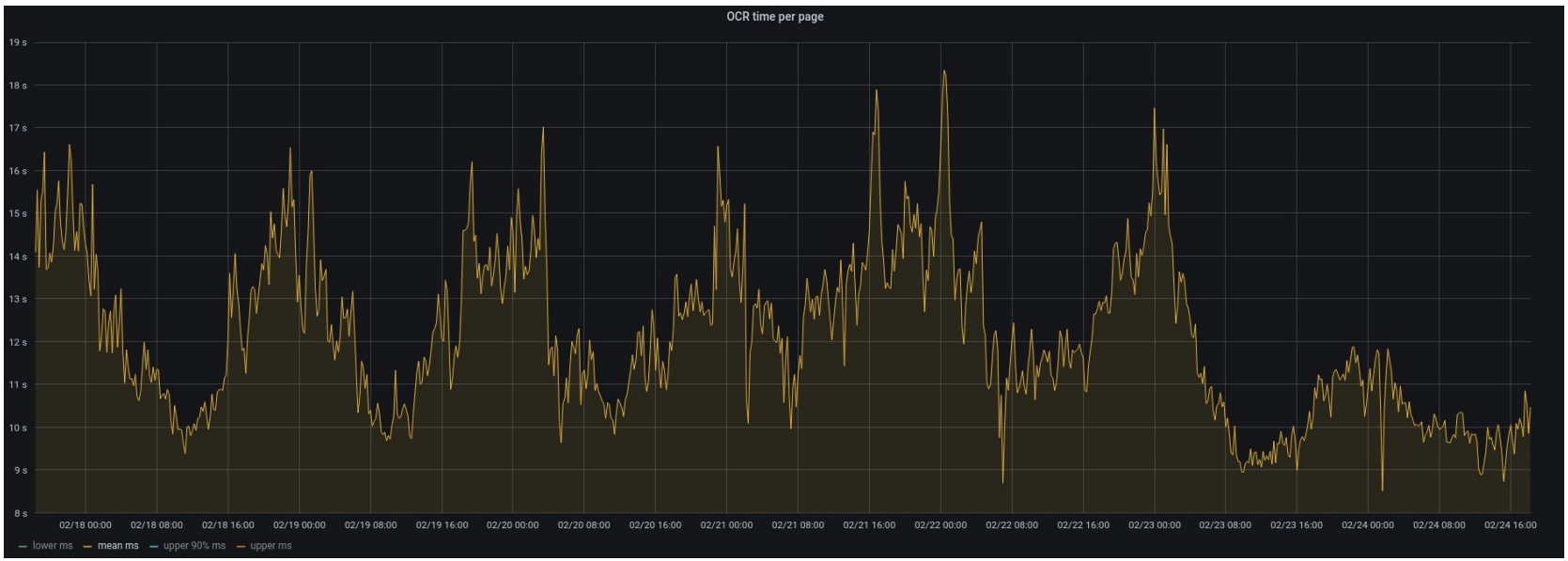
Tesseract

- Developed by HP in 1980s, open sourced, Google 2006
- Actively maintained by community
- Many languages and scripts including: Arabic, CJK, Indian, Fraktur script
- Script and orientation detection
- Version 4 has new recognition engine

List of supported languages

Tesseract 5

Tesseract 4 vs Tesseract 5 (20201231 alpha)



https://twitter.com/brewster_kahle/status/1364742767880990722

OCR module

- Python
- Heuristics for script and language detection
- "Autonomous mode"
- Extensive language and script mapping
- Can convert from Abbyy XML
- Separate, small modules for downstream files
- Custom Tesseract debian repo: https://archive.org/download/tesseract-deb/

Ocr	tesseract 4.1.1
Ocr_detected_lang	de
Ocr_detected_lang_conf	1.0000
Ocr_detected_script	Fraktur
Ocr_detected_script_conf	1.0000
Ocr_module_version	0.0.7
Ocr_parameters	-l deu+Fraktur

Challenges

- Large XML documents
- Quality and quality comparison is hard
- There are a **lot** of languages and scripts out there
- Many edge cases in user uploaded content
- Working on PDF creation/compression in parallel

PDF (Portable Document Format)

- PDF is a well supported document format
- Supports images, vectors, arbitrary text placement
- We create a PDF for every digitised book

Our requirements:

- PDFs with text layer, using our hOCR (tesseract result) files
- MRC (Mixed Raster Content) compression
- PDF/A (Archival standard) compliant, PDF/UA (accessibility)
- OCR and PDF must be separate steps and programs
- Fast and scalable

How much of this can be done with existing FOSS (Free and Open Source Software)?

PDF generation with FOSS

- Tesseract can generate PDFs with text layers (no compression, pdfs generated at OCR time)
- paperwork and ocrmypdf do not compress well, also perform OCR at the same time
- pymupdf is a powerful python library to create and modify PDFs

PDF generation with FOSS

- Tesseract can generate PDFs with text layers (no compression, pdfs generated at OCR time)
- paperwork and ocrmypdf do not compress well, also perform OCR at the same time
- pymupdf is a powerful python library to create and modify PDFs

Our solution: two-step PDF generation

- Port Tesseract PDF generation to Python
- Modify PDF in-place with pymupdf, insert MRC compressed images

MRC compression: The concept

MRC decomposes an image into a background, foreground and mask

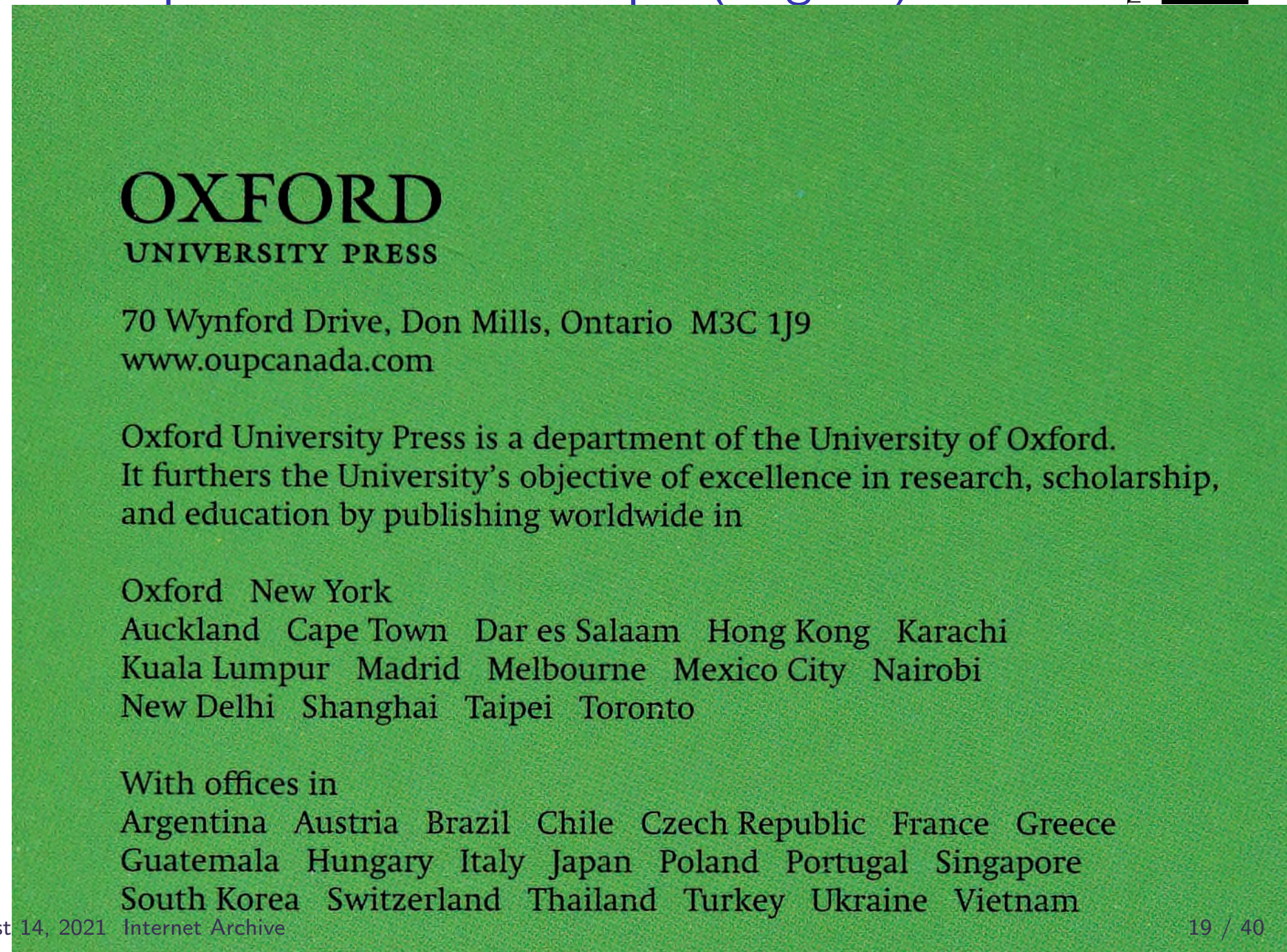
- Encode background + foreground with JPEG2000
- Encode mask using JBIG2 or CCITT (bi-tonal compression)
- Optionally downscale background image

MRC compression: The concept

MRC decomposes an image into a background, foreground and mask

- Encode background + foreground with JPEG2000
- Encode mask using JBIG2 or CCITT (bi-tonal compression)
- Optionally downscale background image
- High (500x) compression ratio for uncompressed images (A 27MB image can be turned into a 52kB PDF with text layer)
- Compression rate of 10x for highly compressed (JPEG2000) images
- Quality/compression ratio can be tweaked

MRC compression: Book example (original)



OXFORD
UNIVERSITY PRESS

70 Wynford Drive, Don Mills, Ontario M3C 1J9
www.oupcanada.com

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

MRC compression: Book example (mrc compressed)

OXFORD
UNIVERSITY PRESS

70 Wynford Drive, Don Mills, Ontario M3C 1J9
www.oupcanada.com

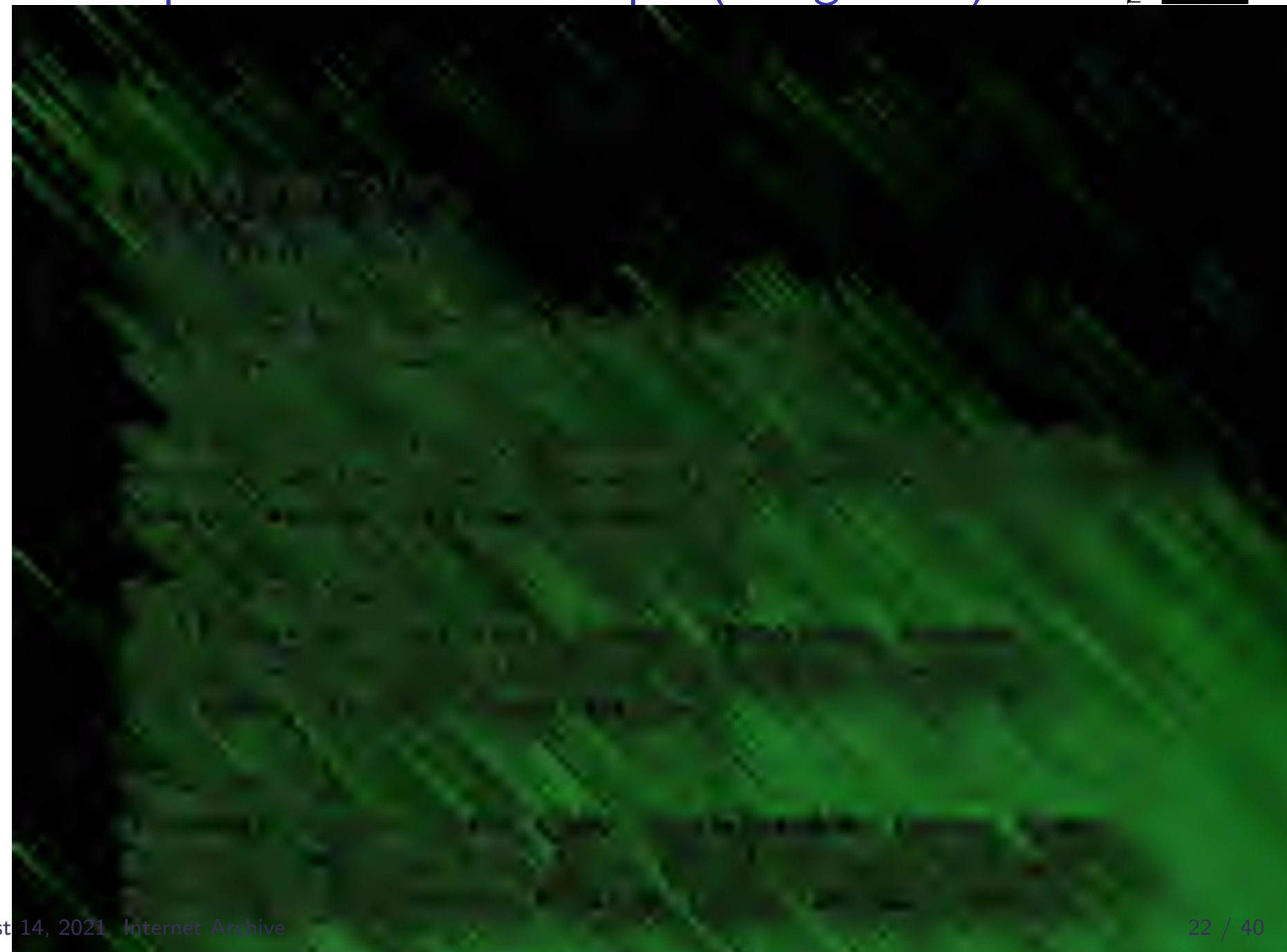
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

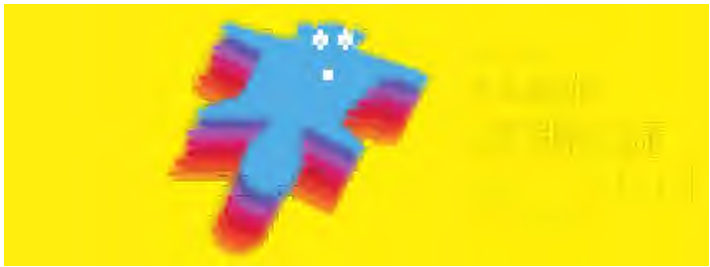
MRC compression: Book example (background)

MRC compression: Book example (foreground)



MRC compression: Book example (mask)

MRC compression: Initlab Logo example



(Background)

(Foreground)

(Alpha Mask)

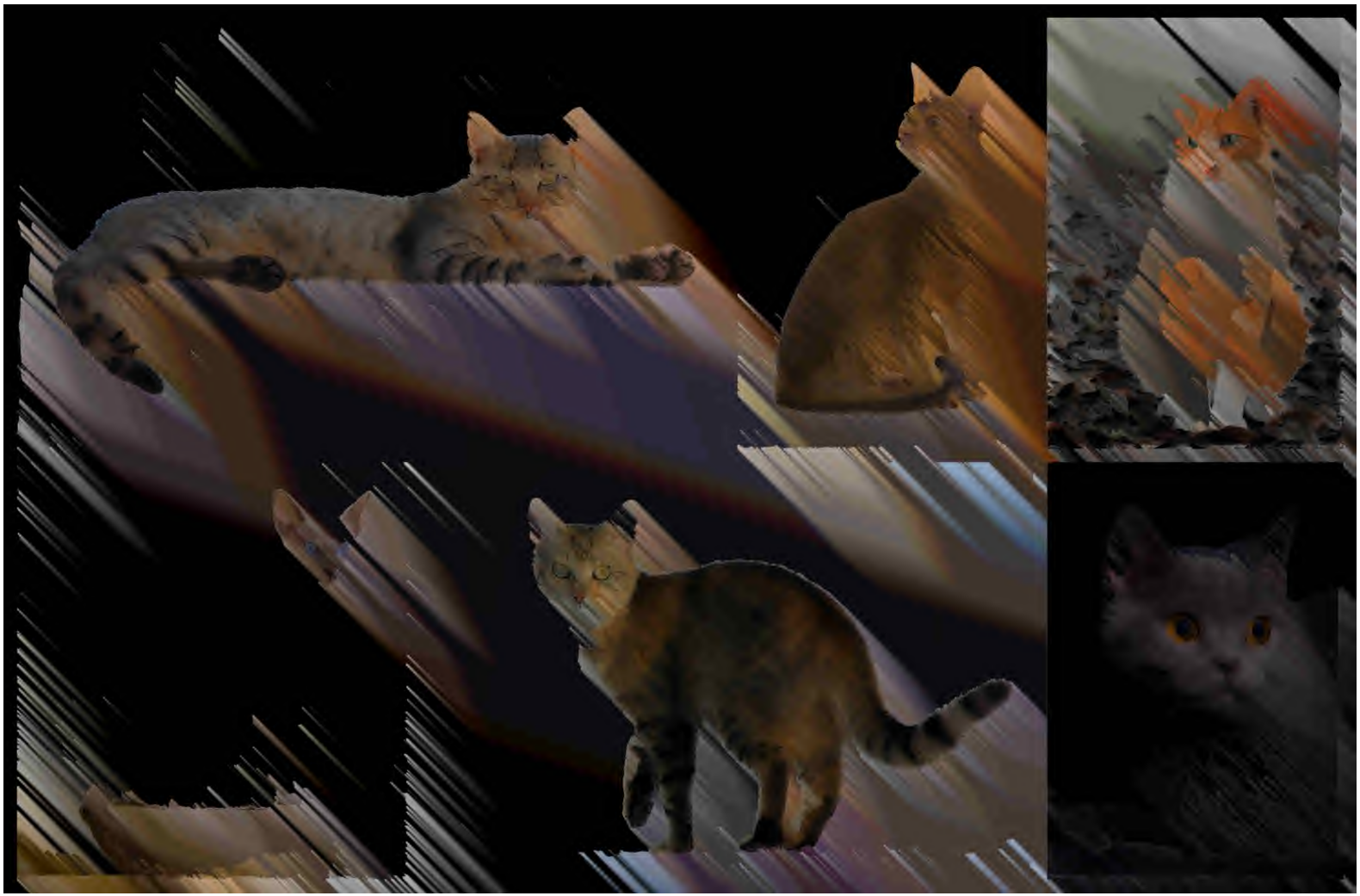
(Result)

MRC compression: Cat example (original) - 11MB

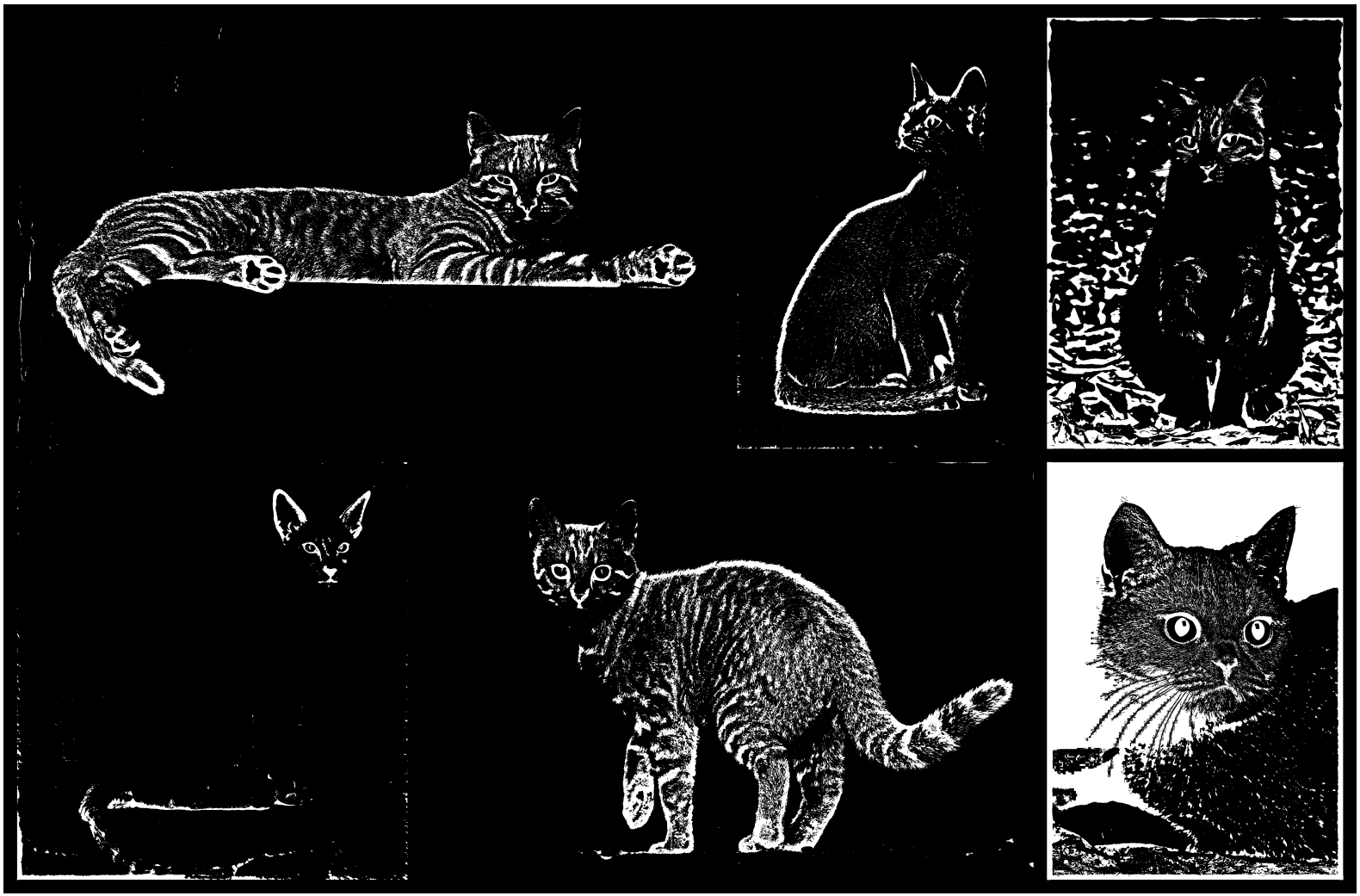
MRC compression: Cat example (mrc compressed) - 244kB

MRC compression: Cat example (background)

MRC compression: Cat example (foreground)



MRC compression: Cat example (mask)



MRC compression: the algorithm

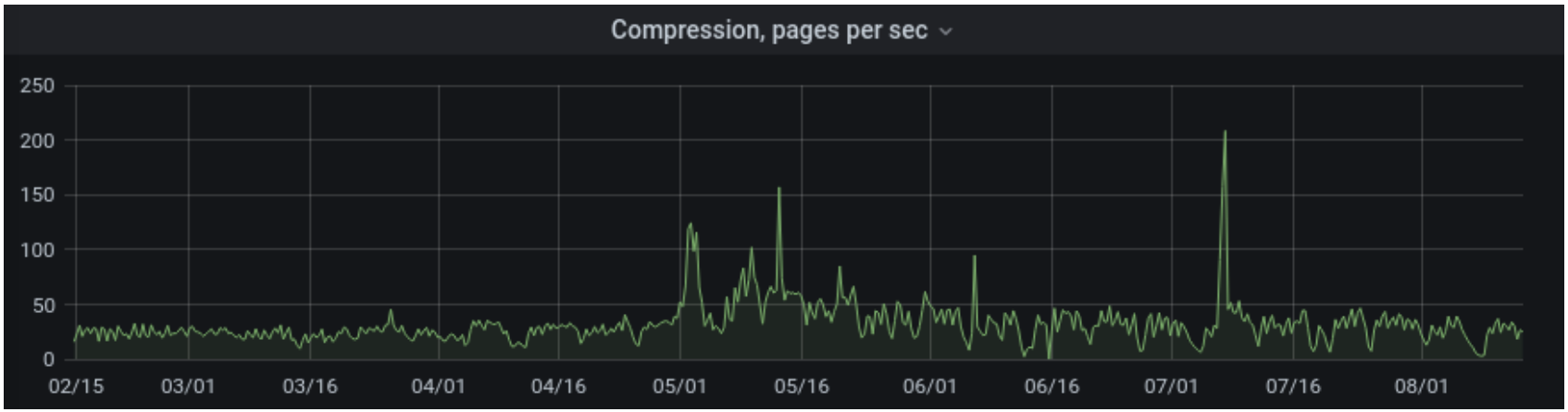
- Mask generation is specialised binarisation (sauvola), plus OCR-based binarisation of regions with text
- Foreground and background are generated by including/excluding pixels that are part of the mask, followed by a step to optimise the image for compressibility

Both these parts are written in Cython for speed, highly optimised versions were later contributed by Bas Weelinck.

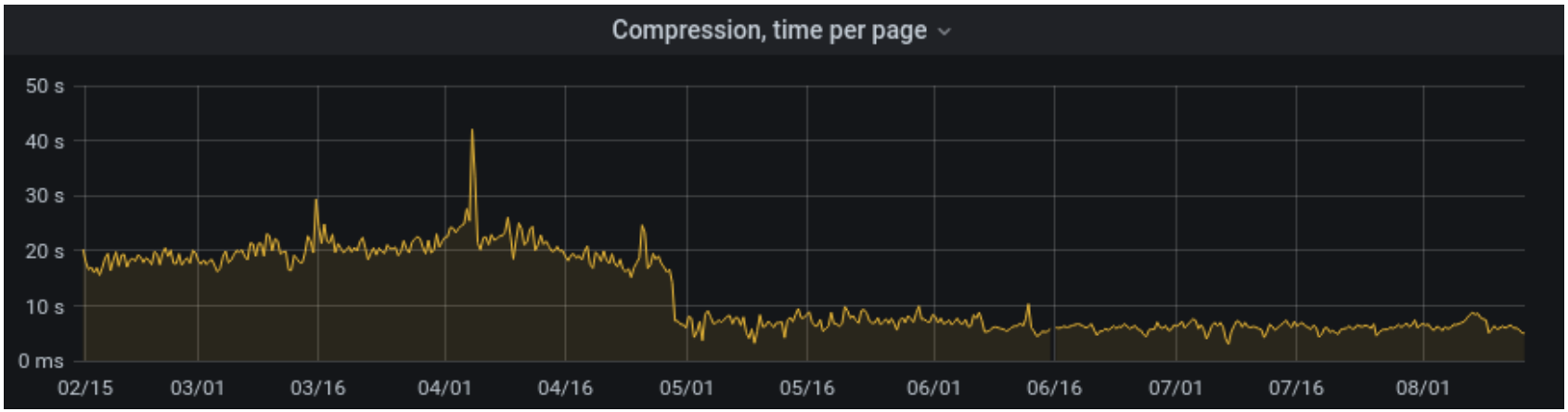
PDF generation and compression: a recap

- Tesseract creates a hOCR file per image, these are combined into a single file
- A PDF with just a text layer is produced using the Python port of the Tesseract PDF renderer
- Images are MRC compressed and inserted into the PDF
- Finally, the PDF is made PDF/A compliant

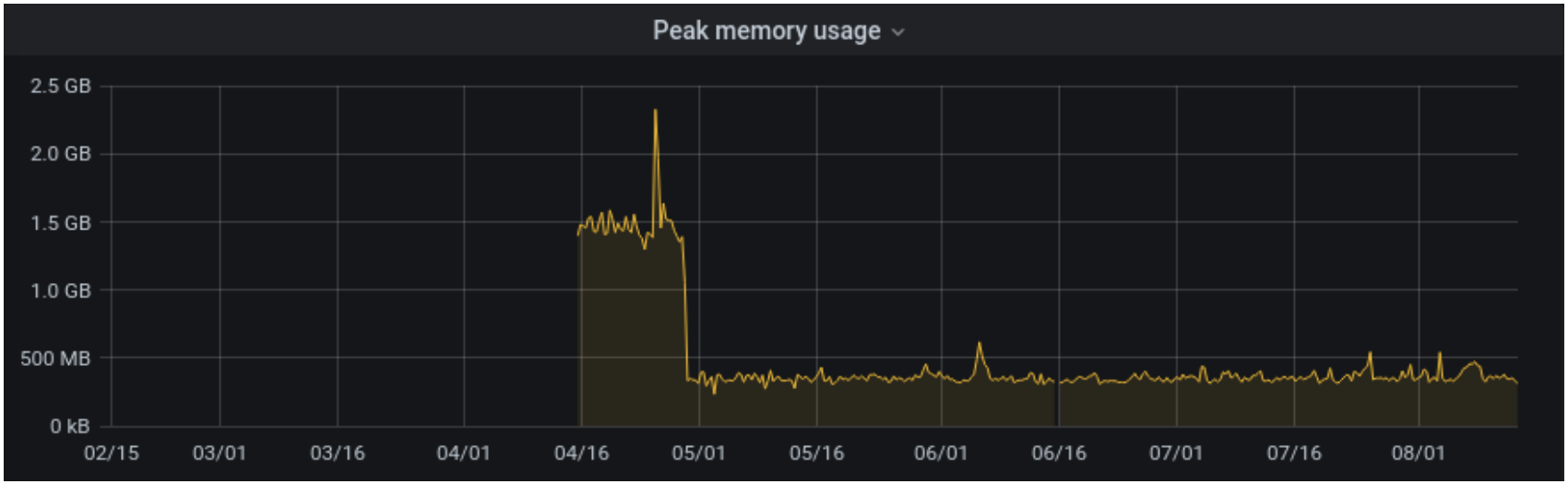
PDF generation: pages per second



PDF generation: time per page



PDF generation: peak memory usage



Source code and Documentation

All code (except for Python-port of Tesseract PDF generation) is AGPLv3.

- https://git.archive.org/www/tesseract
- https://github.com/internetarchive/archive-hocr-tools
- https://github.com/internetarchive/archive-pdf-tools
- https://archive.org/services/docs/api/ocr.html
- https://archive.org/services/docs/api/pdf.html
- https://archive.org/~merlijn/archive-hocr-tools/index.html

Community and Collaboration

- OCR-D project
- Tesseract developers
- mupdf and PyMuPDF
- Slack #ocr-g channel - for all who are interested (drop me an email)

Future work

- MRC foreground and background generation and optimisation could use more improvements (further remove background blur/shadow around text)
- MRC compression without hOCR
- Recoding / compression sophisticated existing PDFs
- Image and photo detection (Tesseract supports it)
- Working on creating an open access data set for OCR training
- Way for users to submit OCR corrections to archive.org?

Summary

- OCR and PDF document processing at the Internet Archive is now based on free software
- We contributed MRC-compression written in Python, usable as a library, AGPL-v3 licensed
- The stack processes millions of pages every day (to date the new PDF software has produced over 4 million PDFs)
- We're happy to collaborate to improve the software

Summary

- OCR and PDF document processing at the Internet Archive is now based on free software
- We contributed MRC-compression written in Python, usable as a library, AGPL-v3 licensed
- The stack processes millions of pages every day (to date the new PDF software has produced over 4 million PDFs)
- We're happy to collaborate to improve the software

Questions?

Feedback

Scan this QR code to give feedback

Attribution

- Cat image https://commons.wikimedia.org/wiki/File:Cat_poster_1.jpg