# *Creating Plain Text from a pdf Scan using OCR*

David Joseph Wrisley, Assoc Prof of Digital Humanities, NYU AD
@DJWrisley, NYU Buenos Aires, 22-23 March 2018

جامعة نيويورك أبوظبي

NYU | ABU DHABI

# Outline

-brief video of what we do

-brief discussion of OCR as a problem of computer vision

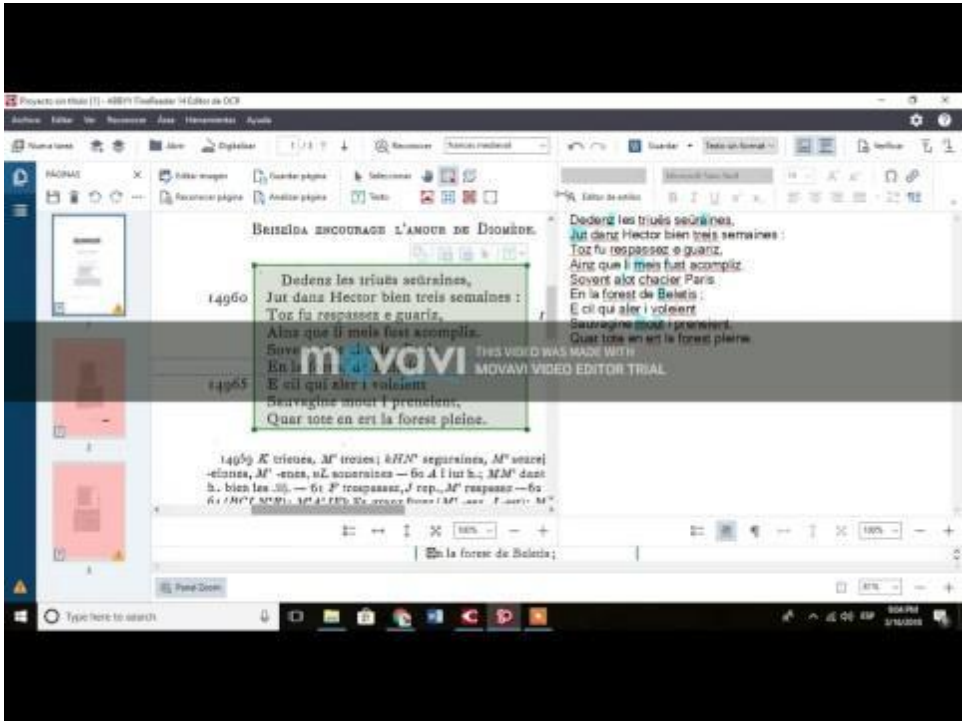
-places where OCR is happening around us (and we don’t know it)

-frontiers of OCR (non-English, in coding environments, historical OCR)

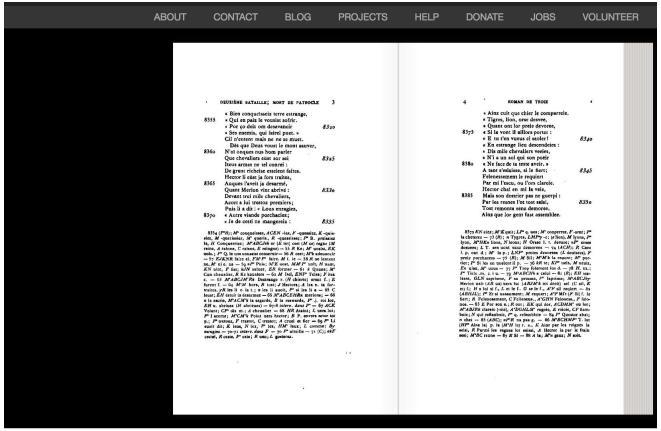
-our hands on with Abbyy FineReader

# Marielle's Video Tutorial (8 min)

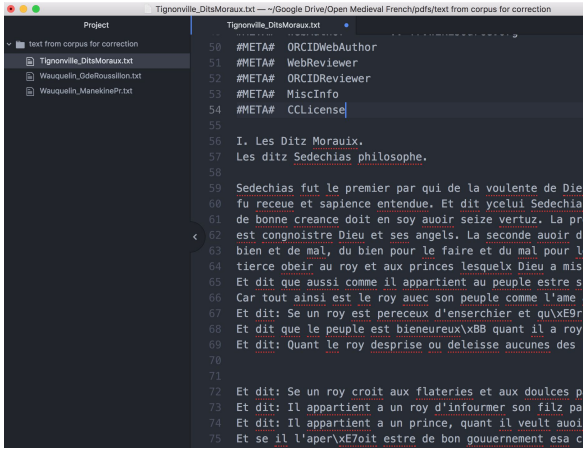
Key points:

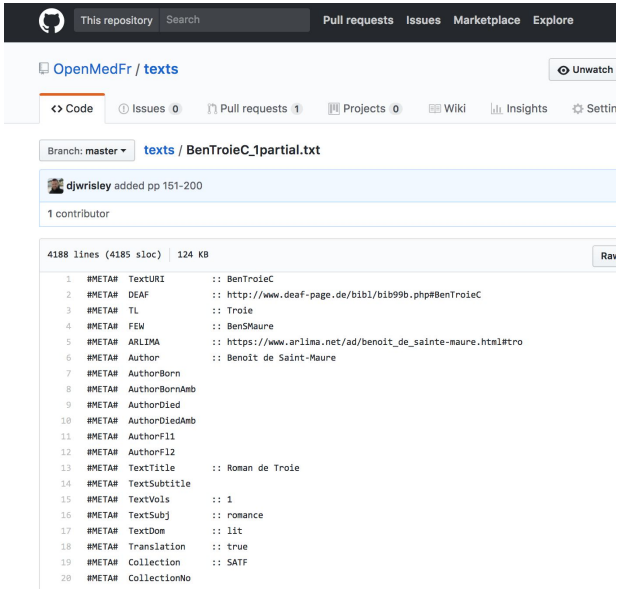


# MAIN GOAL: From pdf scan to txt file to GitHub




Le roman de Troie
by Benoît, de Sainte-More, 12th cent; Constans, Léopold Eugène, 1845-1916, [from old catalog] ed



# OCR (optical character recognition)

## Extracting text from an image of text

**Optical character recognition** (also **optical character reader**, **OCR**) is the conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast) (Wikipedia)

-photo of credit card or a check and information is extracted
-unrecognized photos of pdf documents in Google drive
-reCaptcha
-Google car mapping the streets

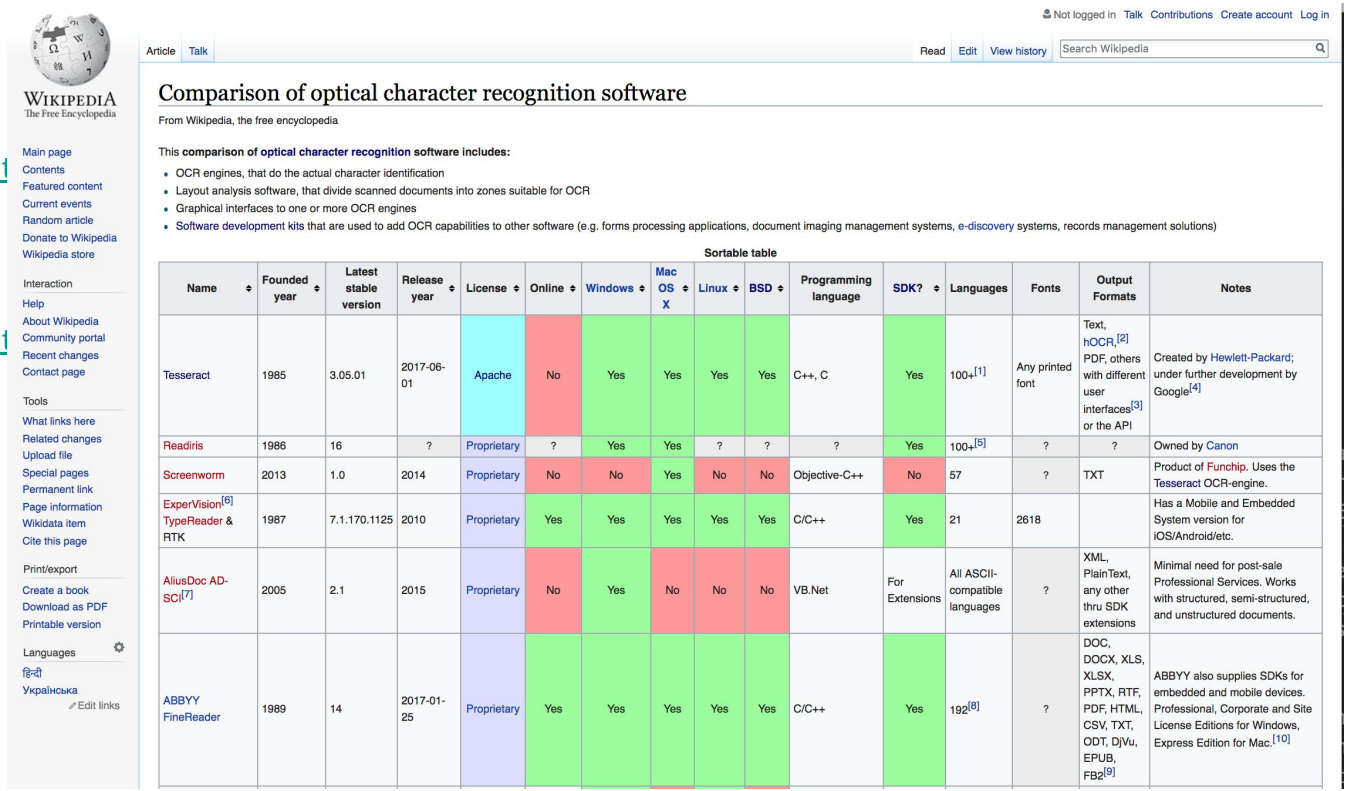
# What is the Difference between and Image and Machine-Readable Text?

- a pdf image is only a set of pixels, not understandable except to the human eye as letters and words and sentences
- a searchable pdf has some OCR done, but if you extract it you will find errors (try this some time in Adobe)
- In order to make the text quality better, our OCR process takes scanned editions and we will follow these basic steps
    - build a “dictionary” from French (one time only)
    - manually choose text blocks
    - read with some “pattern training”
    - some further read through for correction

# Comparison of OCR tools

ht

ht



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
हिन्दी
Українська
Edit links

## Comparison of optical character recognition software

From Wikipedia, the free encyclopedia

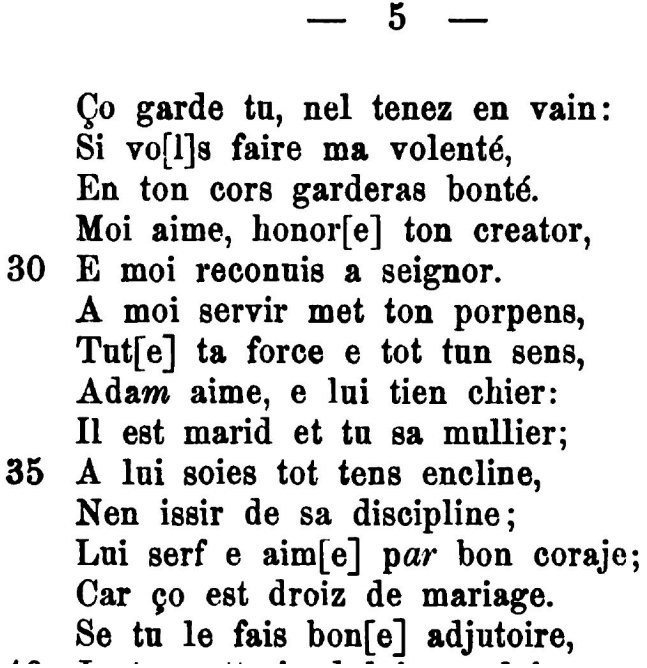
This **comparison of optical character recognition** software includes:

- OCR engines, that do the actual character identification
- Layout analysis software, that divide scanned documents into zones suitable for OCR
- Graphical interfaces to one or more OCR engines
- Software development kits that are used to add OCR capabilities to other software (e.g. forms processing applications, document imaging management systems, e-discovery systems, records management solutions)

Sortable table

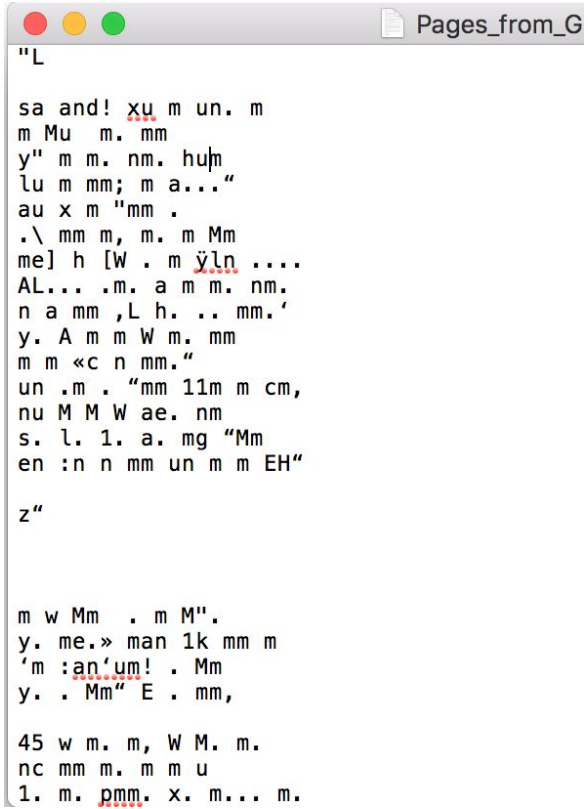
| Name | Founded year | Latest stable version | Release year | License | Online | Windows | Mac OS X | Linux | BSD | Programming language | SDK? | Languages | Fonts | Output Formats | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tesseract | 1985 | 3.05.01 | 2017-06-01 | Apache | No | Yes | Yes | Yes | Yes | C++, C | Yes | 100+[1] | Any printed font | Text, hOCR,[2] PDF, others with different user interfaces[3] or the API | Created by Hewlett-Packard; under further development by Google[4] |
| Readiris | 1986 | 16 | ? | Proprietary | ? | Yes | Yes | ? | ? | ? | Yes | 100+[5] | ? | ? | Owned by Canon |
| Screenworm | 2013 | 1.0 | 2014 | Proprietary | No | No | Yes | No | No | Objective-C++ | No | 57 | ? | TXT | Product of Funchip. Uses the Tesseract OCR-engine. |
| ExperVision[6] TypeReader & RTK | 1987 | 7.1.170.1125 | 2010 | Proprietary | Yes | Yes | Yes | Yes | Yes | C/C++ | Yes | 21 | 2618 | | Has a Mobile and Embedded System version for iOS/Android/etc. |
| AliusDoc AD-SCI[7] | 2005 | 2.1 | 2015 | Proprietary | No | Yes | No | No | No | VB.Net | For Extensions | All ASCII-compatible languages | ? | XML, PlainText, any other thru SDK extensions | Minimal need for post-sale Professional Services. Works with structured, semi-structured, and unstructured documents. |
| ABBYY FineReader | 1989 | 14 | 2017-01-25 | Proprietary | Yes | Yes | Yes | Yes | Yes | C/C++ | Yes | 192[8] | ? | DOC, DOCX, XLS, XLSX, PPTX, RTF, PDF, HTML, CSV, TXT, ODT, DjVu, EPUB, FB2[9] | ABBYY also supplies SDKs for embedded and mobile devices. Professional, Corporate and Site License Editions for Windows, Express Edition for Mac.[10] |

# http://www.free-ocr.com/



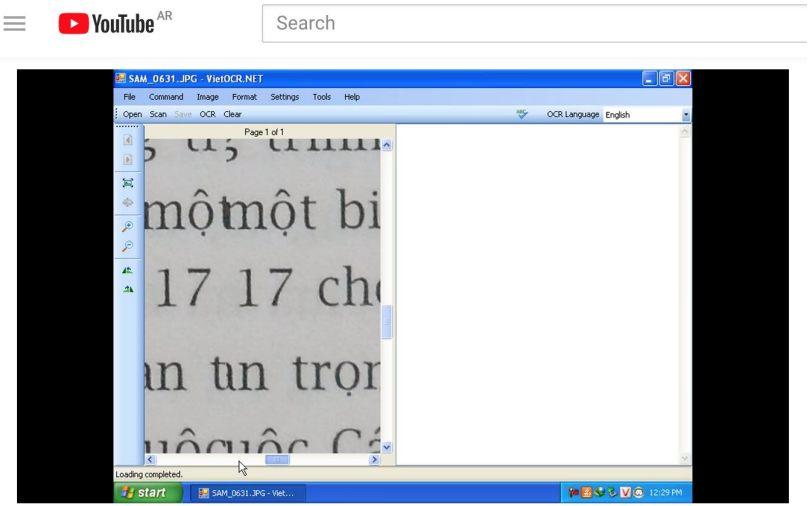
— 5 —

Ço garde tu, nel tenez en vain:
Si vo[l]s faire ma volenté,
En ton cors garderas bonté.
Moi aime, honor[e] ton creator,
E moi reconuis a seignor.
A moi servir met ton porpens,
Tut[e] ta force e tot tun sens,
Ada*m* aime, e lui tien chier:
Il est marid et tu sa mullier;
A lui soies tot tens encline,
Nen issir de sa discipline;
Lui serf e aim[e] p*ar* bon coraje;
Car ço est droiz de mariage.
Se tu le fais bon[e] adjutoire,



# OCR in languages other than English

https://sou

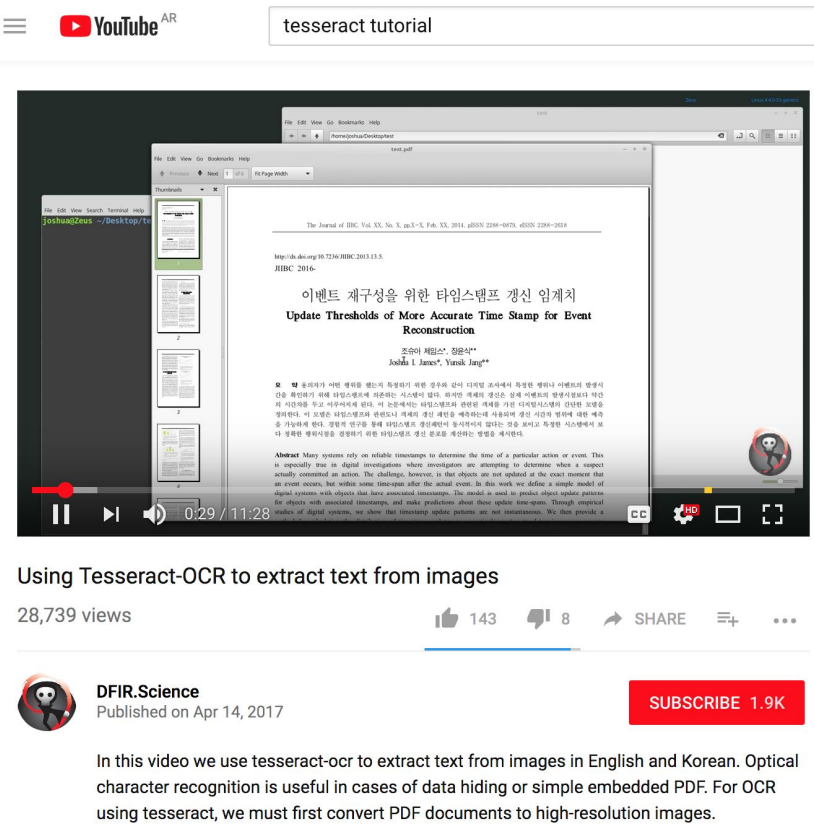


# OCR as part of code

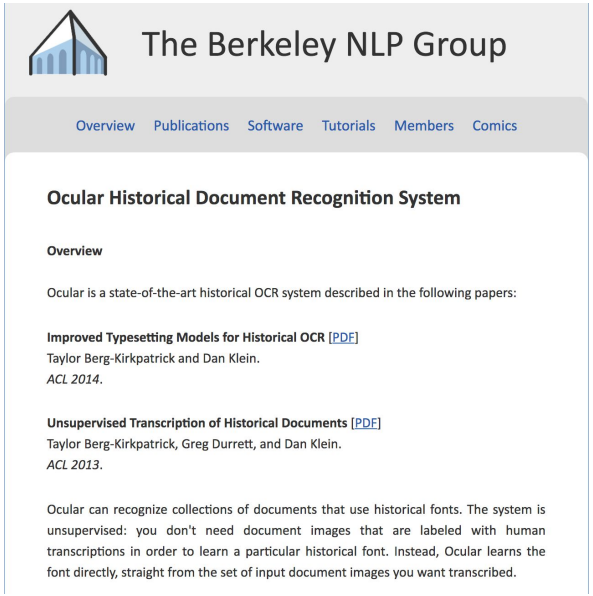
Extracting I ) on Ubuntu Linux

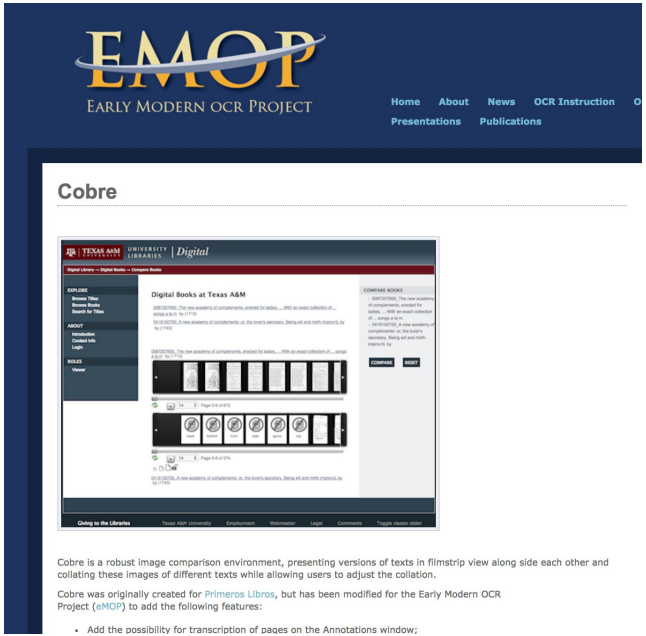
-tesseract i or as part of your code for serial

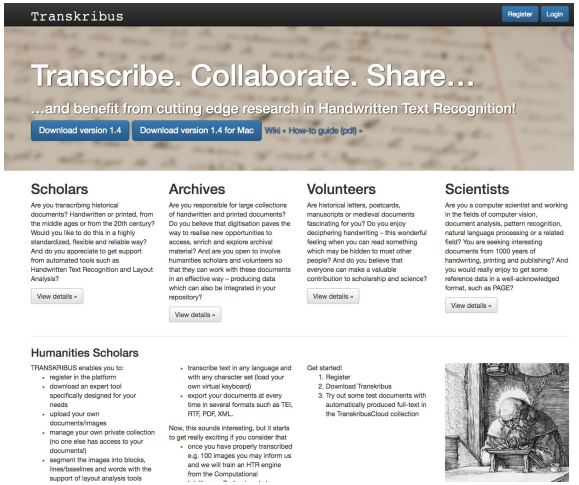
(batch) “un



# Historical OCR



# Our exercise: Abbyy Finereader

-working not with historical handwriting, but historical print publications

-not free (+/- $150), but works very well for a beginner or a lab of non-experts

-licensed to a specific machine

-offers pattern training in the PC version

-allows us to have a simple workflow

# After the OCR… correction at scale

Project Gutenberg - uses systems of Distributed Proofreaders

ECCO/EEBO TCP -  uses collective correction tool TYPEWRIGHT

# Thanks to:

Marielle del Carmen Caballero Ng (NYU Abu Dhabi, class of 2021)

And now we go hands on….