

# *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



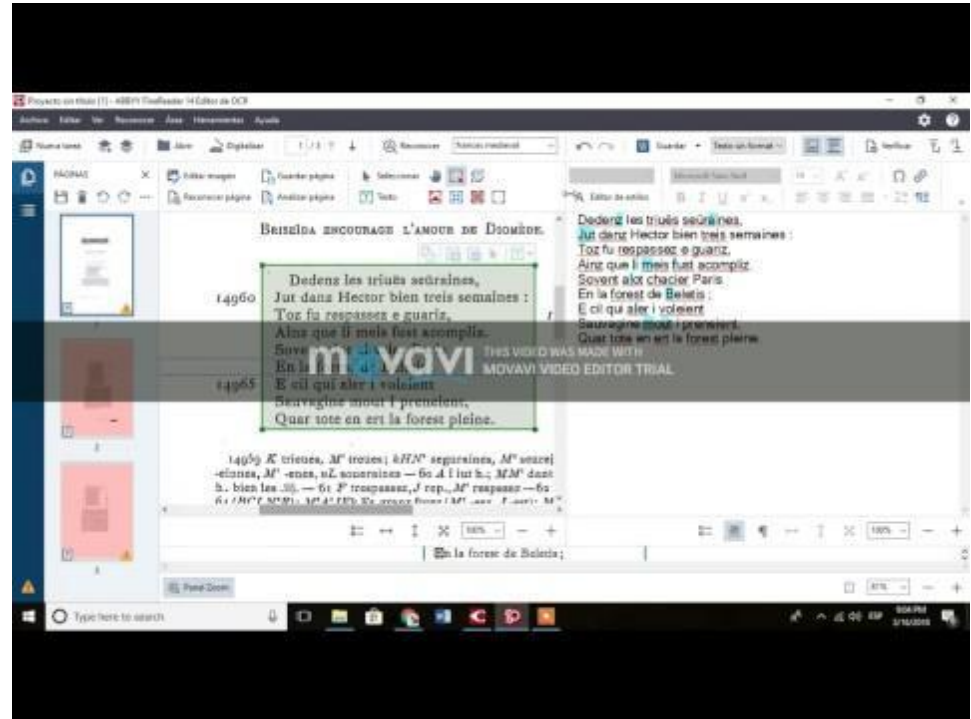
NYU | ABU DHABI

# Outline

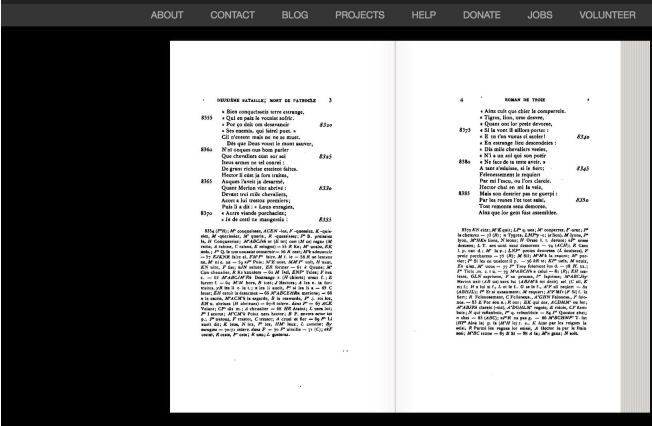
- brief video of what we do
- brief discussion of OCR as a problem of computer vision
- places where OCR is happening around us (and we don't know it)
- frontiers of OCR (non-English, in coding environments, historical OCR)
- our hands on with Abbyy FineReader

# Marielle's Video Tutorial (8 min)

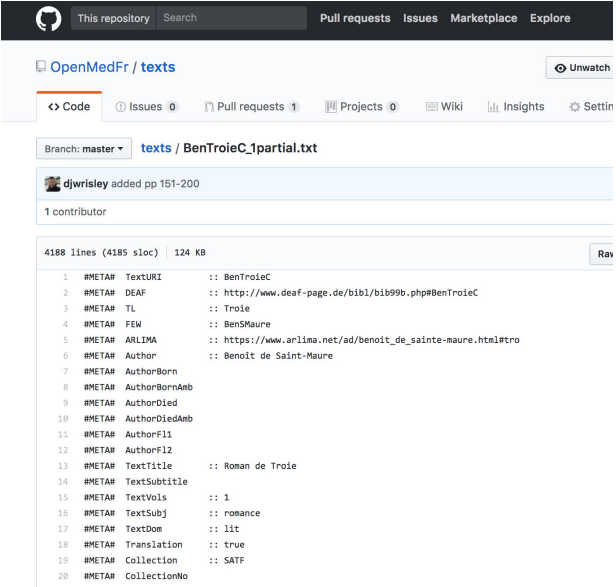
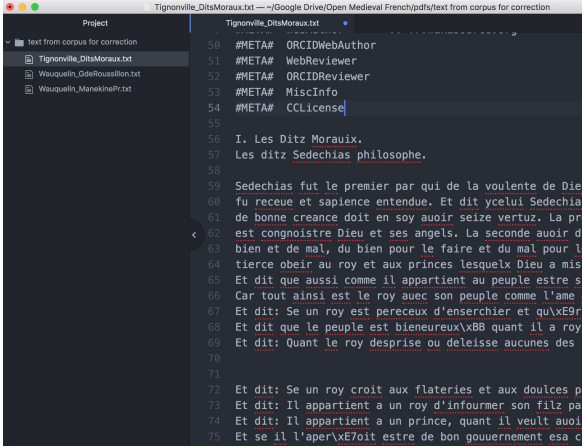
Key points:



# 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



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

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk

Read | Edit | View history

Search Wikipedia

## 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+ <sup>[1]</sup>	Any printed font	Text, hOCR, <sup>[2]</sup> PDF, others with different user interfaces <sup>[3]</sup> or the API	Created by Hewlett-Packard; under further development by Google <sup>[4]</sup>
Readiris	1986	16	?	Proprietary	?	Yes	Yes	?	?	?	Yes	100+ <sup>[5]</sup>	?	?	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 <sup>[6]</sup> 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-Sc <sup>[7]</sup>	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.
ABBY FineReader	1989	14	2017-01-25	Proprietary	Yes	Yes	Yes	Yes	Yes	C/C++	Yes	192 <sup>[8]</sup>	?	DOC, DOCX, XLS, XLSX, PPTX, RTF, PDF, HTML, CSV, TXT, ODT, DjVu, EPUB, FB2 <sup>[9]</sup>	ABBY also supplies SDKs for embedded and mobile devices. Professional, Corporate and Site License Editions for Windows, Express Edition for Mac. <sup>[10]</sup>

- 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,  
30 E moi reconuis a seignor.  
A moi servir met ton porpens,  
Tut[e] ta force e tot tun sens,  
Adam aime, e lui tien chier:  
Il est marid et tu sa mullier;  
35 A lui soies tot tens encline,  
Nen issir de sa discipline;  
Lui serf e aim[e] par bon coraje;  
Car ço est droiz de mariage.  
Se tu le fais bon[e] adjutoire,

```
Pages_from_G
"L
sa and! xu m un. m
m Mu m. mm
y" m m. nm. huln
lu m mm; m a..."
au x m "mm .
.\ mm m, m. m Mm
me] h [W . m ŷln ....
AL... .m. a m m. nm.
n a mm ,L h. .. mm.'
y. A m m W m. mm
m m «c n mm."
un .m . "mm 11m m cm,
nu M M W ae. nm
s. l. 1. a. mg "Mm
en :n n mm un m m EH"

z"

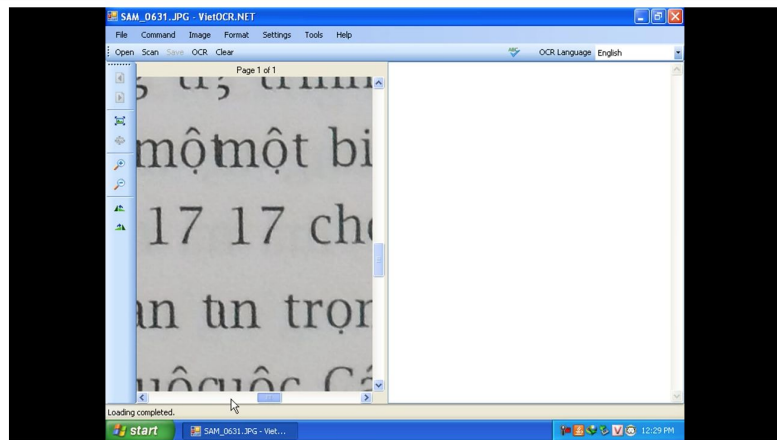
m w Mm . m M".
y. me.» man 1k mm m
'm :an'um! . Mm
y. . Mm" E . mm,

45 w m. m, W M. m.
nc mm m. m m u
1. m. pmm. x. m... m.
```



# OCR in languages other than English

https://so



VietOCR

8,424 views



**Nguyễn Trường Long**

Published on Aug 12, 2011

SUBSCRIBE

Nhận dạng ký tự quang học (tiếng Anh: Optical Character Recognition, viết tắt là OCR), là loại phần mềm máy tính được tạo ra để chuyển các hình ảnh của chữ viết tay hoặc chữ đánh máy (thường được quét bằng máy scanner) thành các văn bản tài liệu. OCR được hình thành từ

# OCR as part of code

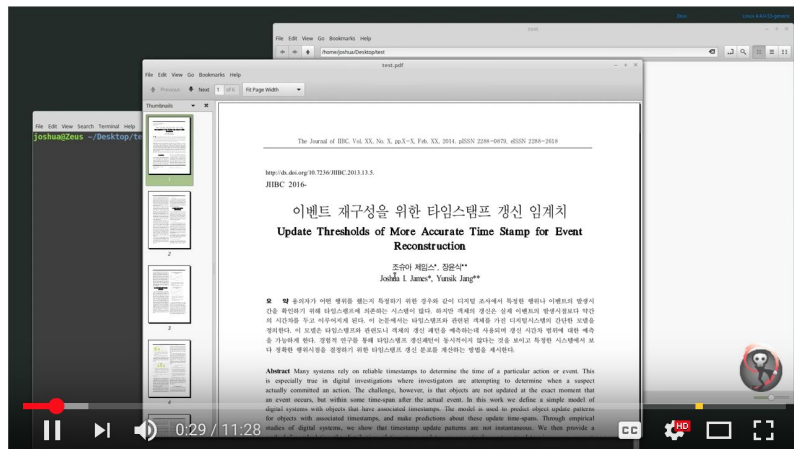
Extracting I

-tesseract i  
(batch) “un



) on Ubuntu Linux

or as part of your code for serial



Using Tesseract-OCR to extract text from images

28,739 views

143 8 SHARE



DFIR.Science

Published on Apr 14, 2017

SUBSCRIBE 1.9K

In this video we use tesseract-ocr to extract text from images in English and Korean. Optical character recognition is useful in cases of data hiding or simple embedded PDF. For OCR using tesseract, we must first convert PDF documents to high-resolution images.

@DJWrisley NYU Buenos Aires, 22-23 March 10

# Historical OCR



The Berkeley NLP Group

[Overview](#) [Publications](#) [Software](#) [Tutorials](#) [Members](#) [Comics](#)

## Ocular Historical Document Recognition System

### Overview

Ocular is a state-of-the-art historical OCR system described in the following papers:

[Improved Typesetting Models for Historical OCR \[PDF\]](#)

Taylor Berg-Kirkpatrick and Dan Klein.

ACL 2014.

[Unsupervised Transcription of Historical Documents \[PDF\]](#)

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein.

ACL 2013.

Ocular can recognize collections of documents that use historical fonts. The system is unsupervised: you don't need document images that are labeled with human transcriptions in order to learn a particular historical font. Instead, Ocular learns the font directly, straight from the set of input document images you want transcribed.

**EMOP**  
EARLY MODERN OCR PROJECT

Home About News OCR Instruction Presentations Publications

### Cobre

#### Digital Books at Texas A&M

**BOOKS**  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...

**ABOUT**  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...

**COLLECTIONS**  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...  
1800-1850: The new edition of commentaries, printed for sale... with an exact collation of...

Cobre is a robust image comparison environment, presenting versions of texts in filmstrip view along side each other and collating these images of different texts while allowing users to adjust the collation.

Cobre was originally created for Primeros Libros, but has been modified for the Early Modern OCR Project (eMOP) to add the following features:

- Add the possibility for transcription of pages on the Annotations window:

Transkribus Register Login

## Transcribe. Collaborate. Share...

...and benefit from cutting edge research in Handwritten Text Recognition!

Download version 1.4 Download version 1.4 for Mac Wiki - How-to guide (pdf)

### Scholars

Are you transcribing historical documents? Handwritten or printed, from the middle ages or from the 20th century? Would you like to do this in a highly standardized, flexible and reliable way? And do you appreciate to get support from automated tools such as Handwritten Text Recognition and Layout Analysis?

[View details >](#)

### Archives

Are you responsible for large collections of handwritten and printed documents? Do you believe that digitisation paves the way to realise new opportunities to access, enrich and explore archival material? And are you open to involve humanities scholars and volunteers so that they can work with these documents in an effective way - producing data which can also be integrated in your repository?

[View details >](#)

### Volunteers

Are historical letters, postcards, manuscripts or medical documents fascinating for you? Do you enjoy deciphering handwriting - this wonderful feeling when you can read something which may be hidden to most other people? And do you believe that everyone can make a valuable contribution to scholarship and science?

[View details >](#)

### Scientists

Are you a computer scientist and working in the field of computer vision, document analysis, pattern recognition, natural language processing or a related field? You are seeking interesting documents from 1000 years of handwriting, printing and publishing? And you would really enjoy to get some reference data in a well-subdocumented format, such as PAGE?

[View details >](#)

### Humanities Scholars

TRANSCRIBUS enables you to:

- register in the platform
- download an expert tool specifically designed for your needs
- upload your own documents/images
- manage your own private collection (no one else has access to your documents)
- segment the images into blocks, line/baseline and words with the support of layout analysis tools

- transcribe text in any language and with any character set (load your own virtual keyboard)
- export your documents at every time in several formats such as TEI, RTF, PDF, XML

Now, this sounds interesting, but it starts to get really exciting if you consider that:

- once you have properly transcribed e.g. 100 images you may inform us and we will train an HTR engine from the Computational...

Get started!

- Register
- Download Transkribus
- Try out some test documents with automatically produced full-text in the TranskribusCloud collection



# Our exercise: Abby 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....