

# Digitization of written sources

Federico Boschetti

[federico.boschetti@ilc.cnr.it](mailto:federico.boschetti@ilc.cnr.it)

CNR-ILC & VeDPH

June 8, 2022 - University of Pisa

Summer School - Digital Tools for Humanists

# Introduction

# The importance of OCR and HTR

The Optical Character Recognition (OCR) is a bottleneck in many activities that need large quantities of legacy information:

- digital libraries
- corpus linguistics
- digital history
- ...

# The importance of OCR and HTR

Nowadays OCR can perform 99% of accuracy on recent, good quality printed editions and it can reach 98% of accuracy on challenging printed documents

The new field of Handwritten Text Recognition is very promising, so that libraries, universities and other institutions (such as state archives) are planning to acquire the digital text not only from printed documents but also from manuscripts

# **Acquisition and pre-processing of digital images**

# Digital images and digital texts

Scanning is the process of acquiring information from two-dimensional or three-dimensional objects, in order to create digital images

Different operations can be performed on digital images of a document and digital texts:

- crop an arbitrary part
- change brightness and contrast
- compare the high fidelity of the layout and of the figures to the original manuscript or printed edition
- ...
- copy and paste it
- search it
- tokenize it
- count the tokens
- make indexes
- ...

# Scanners

Various kinds of scanners are available, but a simple flatbed scanner can be enough, if the document is not fragile. The coplanarity of the written surface of the document with the moving carriage of the scanner has a high impact on the accuracy of the recognition



# Scanners

Currently documents are acquired also by smartphones, but the quality of the acquisition is poor, compared to a flatbed scanner





# DPI and PPI

DPI means Dots per Inch and PPI means Pixels per Inch. In order to have an accurate OCR, 600 DPI are optimal, but 300 DPI can be acceptable.

# The preservation of the master images and metadata

Along the digital text acquisition workflow one or more image elaborations are required. It is necessary to keep always the original images and possibly to preserve also the metadata related to the necessary transformations and use naming conventions for the files, with minimal metadata about dpi, color, etc.

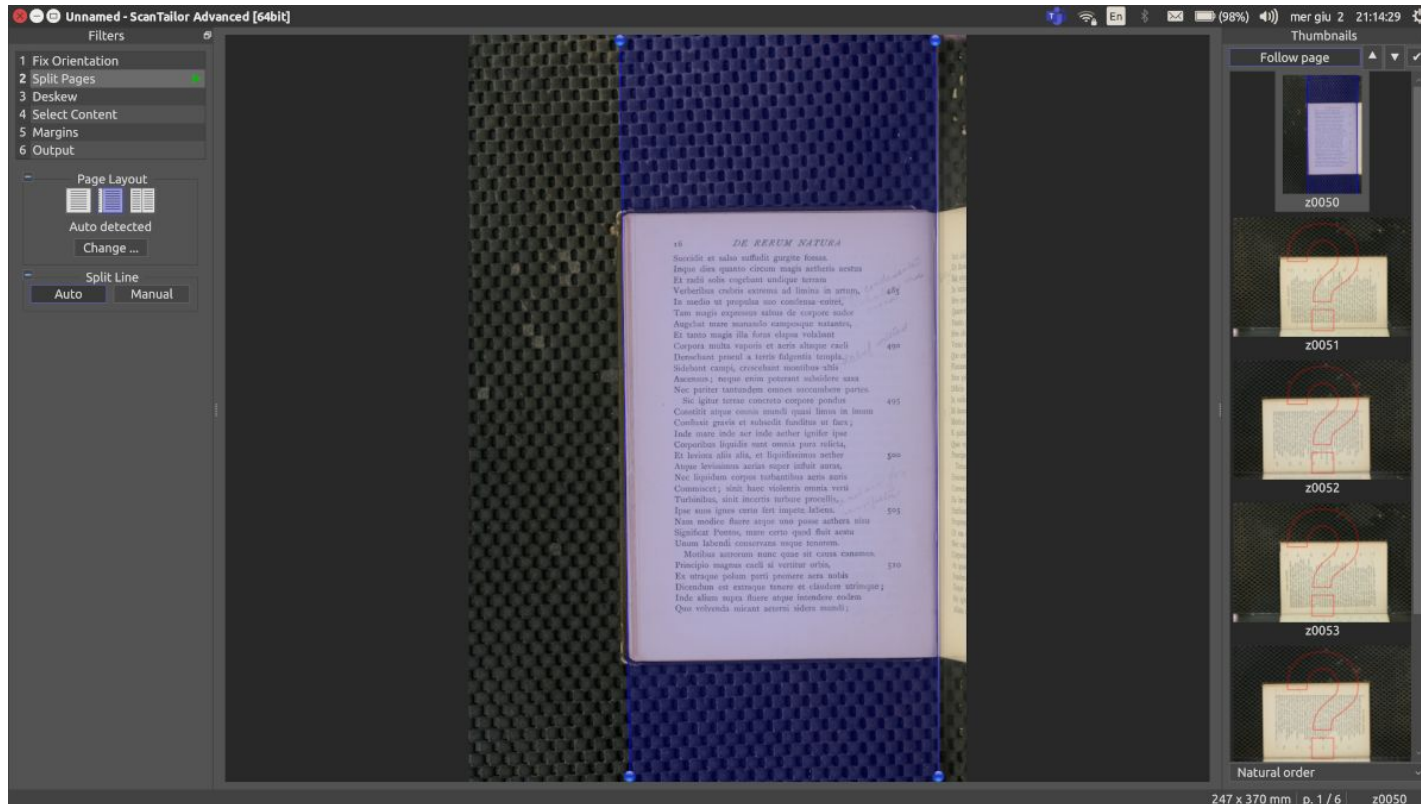
# OCR (or HTR) preprocessing on images

In order to improve the accuracy of OCR or HTR, images must be processed at least with the following operations:

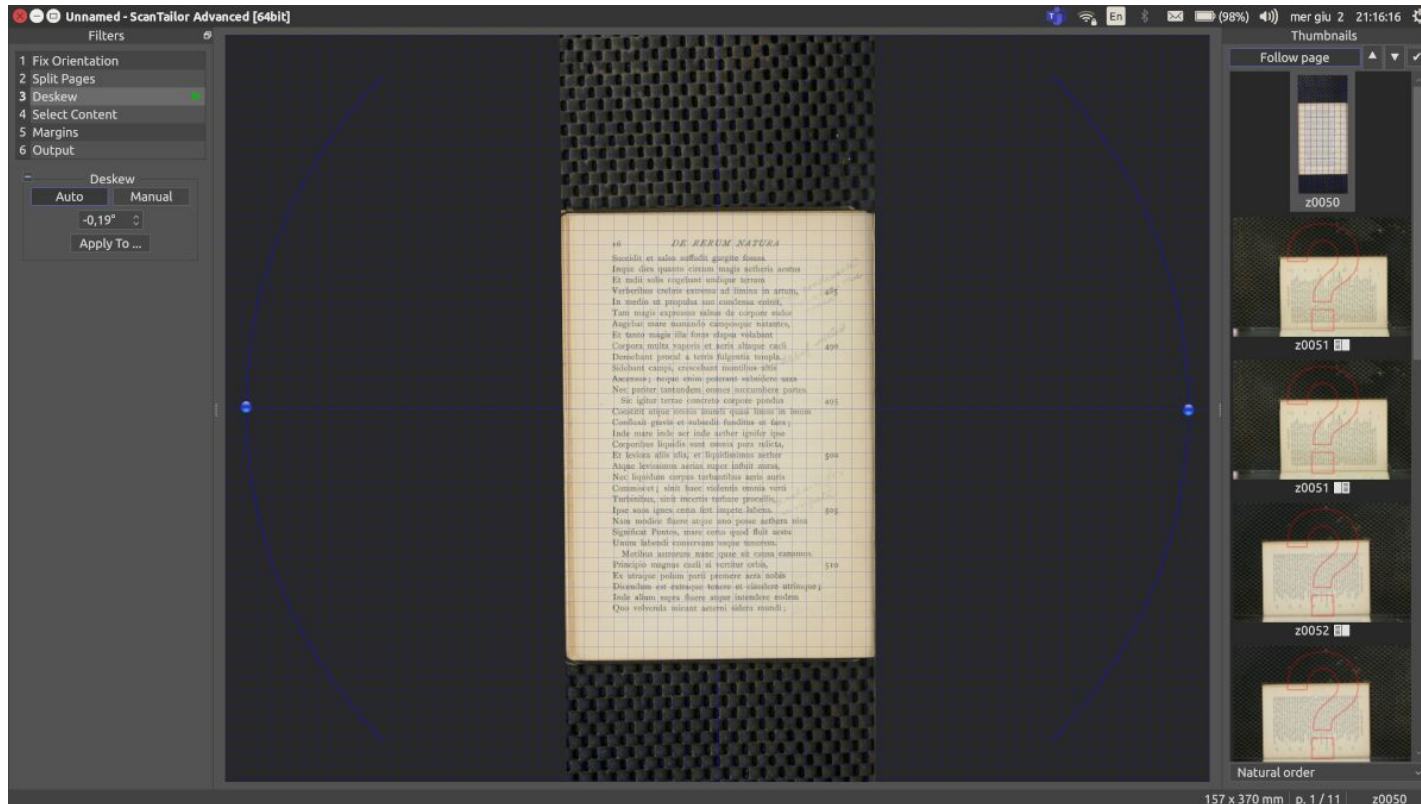
- fixing orientation (if necessary)
- splitting pages (if two pages have been scanned together)
- deskewing (i.e. small rotation)
- selecting content
- adding margins
- change the output resolution (if necessary)
- binarization
- dewarping



# Splitting pages



# Deskewing



# Selecting content

The screenshot displays the ScanTailor Advanced interface. On the left, a 'Filters' sidebar lists steps: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content (highlighted), 5 Margins, and 6 Output. Below this are 'Page Box' and 'Content Box' settings, each with 'Disable', 'Auto', and 'Manual' options. The main workspace shows a scanned page with a red rectangular selection box around the text. A vertical ruler on the right indicates the selection's vertical extent. On the far right, a 'Thumbnails' panel shows a sequence of page thumbnails, with the current page (z0051) highlighted and a red asterisk indicating the selected content area. The status bar at the bottom shows the page dimensions and zoom level: '-75, 73 | 157 x 370 mm | p. 1 / 11 | z0050'.

Use the context menu to enable / disable the content box. Hold Shift to drag a box. Use double-click on content to automatically adjust the content area.

# Adding margins

The screenshot shows the ScanTailor Advanced software interface. The main window displays a page of Latin text from "DE RERUM NATURA" with a pink border indicating the added margins. The text is as follows:

16 DE RERUM NATURA  
Succidit et salso suffudit gurgite fossas.  
Inque dies quanto circum magis aetheris aestus  
Et radii solis cogeabant undique terram  
Verberibus crebris extrema ad limina in artum, 485  
In medio ut propulsa suo condensa coiret,  
Tam magis expressus salsus de corpore sudor  
Augebat mare manando camposque natantes,  
Et tanto magis illa foras elapsa volabant  
Corpora multa vaporis et aeris atque caeli 490  
Densebant procul a terris fulgentia templa.  
Sidebant campi, crescebant montibus altis  
Ascensus; neque enim poterant subsidere saxa  
Nec pariter tantundem omnes succumbere partes.  
Sic igitur terrae concreto corpore pondus 495  
Constitit atque omnis mundi quasi limus in imum  
Confluxit gravis et subsedit funditus ut faex;  
Inde mare inde aer inde aether ignifer ipse  
Corporibus liquidis sunt omnia pura relictia,  
Et leviora aliis alia, et liquidissimus aether 500  
Atque levissimus aeris super influit auras,  
Nec liquidum corpus turbantibus aeris auris  
Commiscet; sinit haec violentis omnia verti  
Turbinibus, sinit incertis turbare procellis,  
Ipse suos ignes certo fert impete labens, 505  
Nam modice fluere atque uno posse aethera nisu  
Significat Pontos, mare certo quod fluit aestu  
Unum labendi conservans usque tenorem.  
Motibus astrorum nunc quae sit caenam.  
Principio magnus caeli si vertitur orbis, 510  
Ex utraque polum parti premere aera nobis  
Dicendum est extraque tenere et claudere utrumque;  
Inde alium supra fluere atque intendere eodem  
Quo volvenda micant aeterni sidera mundi;

The left sidebar shows the "Margins" section with the following settings:

- Auto Margins:
- Top: 5,0
- Bottom: 5,0
- Left: 10,0
- Right: 10,0

The "Alignment" section is set to "Manual" with the "Match size with other page" checkbox checked. The "Guides Help" section provides instructions on how to create, delete, and move guides.

The right sidebar shows a "Thumbnail" gallery with five thumbnails of the page, each with a red question mark and a box around it, indicating that the margins are not yet applied to these thumbnails. The thumbnails are labeled z0050, z0051, z0051, z0052, and z0050.

The bottom status bar shows the page dimensions: -31,45 132 x 195 mm p. 1 / 11 z0050.



# Changing resolution

The screenshot shows the ScanTailor Advanced software interface. The 'Filters' panel on the left has the 'Output Resolution (DPI)' setting circled in red, with a 'Change ...' button below it. The main window displays a scanned page of Latin text from 'DE RERUM NATURA' with line numbers 16, 485, 490, 495, 500, 505, and 510. The right sidebar shows a 'Thumbnails' panel with a 'Follow page' button and a vertical list of thumbnails labeled z0050, z0051, z0052, z0053, and z0050. The bottom status bar shows '-51, 92 | 135 x 200 mm | p. 1 / 6 | z0050'.

# Binarization

The screenshot displays the ScanTailor Advanced interface. On the left, the 'Filters' panel is visible, with the 'Output' filter selected. The 'Mode' is set to 'Black and White', and the 'Threshold' is set to 0. The 'Color operations' section includes 'Reduce noise: 7' and 'Posterize: Level: 4'. The 'Despeckling' section is checked. A red circle highlights the 'Mode' dropdown menu. The central area shows a scanned page of Latin text from 'DE RERUM NATURA' with line numbers 485, 490, 495, 500, 505, and 510. The right side shows a 'Thumbnails' panel with a 'Follow page' button and a vertical stack of thumbnails labeled z0050, z0051, z0052, z0053, and z0050. The bottom status bar indicates '-51, 92 | 135 x 200 mm | p. 1 / 6 | z0050'.

Unnamed - ScanTailor Advanced [64bit]  
Filters

- 1 Fix Orientation
- 2 Split Pages
- 3 Deskew
- 4 Select Content
- 5 Margins
- 6 Output

Output Resolution (DPI)  
600  
Change ...

Mode  
Black and White

Options

- Fill offcut
- Fill margins
- Equalize illumination (B...
- Morphological smoothin...

Threshold  
Method: Otsu  
0

Thinner Thicker

Color operations  
Color segmentation  
R 0 G 0 B 0  
Reduce noise: 7  
Posterize  
Level: 4  
Normalize  
 Force b&w

Apply To ...

Despeckling  
 Despeckle

Apply To ...

16 *DE RERUM NATURA*

Succidit et salso suffudit gurgite fossas.  
Inque dies quanto circum magis aetheris aestus  
Et radii solis cogeabant undique terram 485  
Verberibus crebris extrema ad limina in artum,  
In medio ut propulsa suo condensa coiret,  
Tam magis expressus salsus de corpore sudor  
Augebat mare manando camposque natantes,  
Et tanto magis illa foras elapsa volabant 490  
Corpora multa vaporis et aeris altaque caeli  
Densebant procul a terris fulgentia templa.  
Sidebant campi, crescebant montibus altis  
Ascensus; neque enim poterant subsidere saxa  
Nec pariter tantundem omnes succumbere partes. 495  
Sic igitur terrae concreto corpore pondus  
Constitit atque omnis mundi quasi limus in imum  
Confluxit gravis et subsedit funditus ut faex;  
Inde mare inde aer inde aether ignifer ipse  
Corporibus liquidis sunt omnia pura relicta,  
Et leviora aliis alia, et liquidissimus aether 500  
Atque levissimus aeras super influit auras,  
Nec liquidum corpus turbantibus aeris auris  
Commiscet; sinit haec violentis omnia verti  
Turbinibus, sinit incertis turbare procellis,  
Ipse suos ignes certo fert impete labens. 505  
Nam modice fluere atque uno posse aethera nisu  
Significat Pontos, mare certo quod fluit aestu  
Unum labendi conservans usque tenorem.  
Motibus astrorum nunc quae sit causa canamus.  
Principio magnus caeli si vertitur orbis, 510  
Ex utraque polum parti premere aera nobis  
Dicendum est extraque tenere et claudere utrimque;  
Inde alium supra fluere atque intendere eodem  
Quo volvenda micant aeterni sidera mundi;

Output  
Picture Zones  
Fill Zones  
Dewarping  
Despeckling

Thumbnails  
Follow page

z0050  
z0051  
z0052  
z0053  
z0050

Natural order

-51, 92 | 135 x 200 mm | p. 1 / 6 | z0050

# Despeckling

The screenshot displays the ScanTailor Advanced software interface. On the left, a sidebar contains a list of filters: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content, 5 Margins, and 6 Output. Under the 'Output' filter, the 'Despeckling' section is expanded, and the 'Despeckle' checkbox is checked. A red circle highlights this section. The main workspace shows a scanned page of Latin text from 'DE RERUM NATURA' with a line number '16' in the top left. The text is arranged in columns with line numbers on the right. The software interface includes a top status bar with system icons and a right sidebar with 'Output', 'Picture Zones', 'Fill Zones', 'Dewarping', and 'Despeckling' options. A 'Thumbnails' panel on the far right shows a sequence of image thumbnails labeled z0050 through z0053, with the current page (z0050) highlighted. At the bottom, a footer reads 'Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.' and 'Natural order' is indicated at the bottom right of the thumbnails panel.

Unnamed - ScanTailor Advanced [64bit]

Filters

- 1 Fix Orientation
- 2 Split Pages
- 3 Deskew
- 4 Select Content
- 5 Margins
- 6 Output

Output Resolution (DPI): 600

Mode: Black and White

Options

- Fill offcut
- Fill margins
- Equalize illumination (B)
- Savitzky-Golay smoothing
- Morphological smoothing

Threshold

Method: Otsu

0

Thinner Thicker

Color operations

Color segmentation

R 0 G 0 B 0

Reduce noise: 7

Posterize

Level: 4

Normalize

Force b&w

Apply To ...

Despeckling

Despeckle

Apply To ...

16 DE RERUM NATURA

Succidit et salso suffudit gurgite fossas.  
Inque dies quanto circum magis aetheris aestus  
Et radii solis cogeabant undique terram 485  
Verberibus crebris extrema ad limina in artum,  
In medio ut propulsa suo condensata coiret,  
Tam magis expressus salsus de corpore sudor  
Augebat mare manando camposque natantes,  
Et tanto magis illa foras elapsa volabant 490  
Corpora multa vaporis et aeris altaque caeli  
Densebant procul a terris fulgentia templa.  
Sidebant campi, crescebant montibus altis  
Ascensus; neque enim poterant subsidere saxa  
Nec pariter tantundem omnes succumbere partes. 495  
Sic igitur terrae concreto corpore pondus  
Constitit atque omnis mundi quasi limus in imum  
Confluxit gravis et subsedit funditus ut faex;  
Inde mare inde aer inde aether ignifer ipse  
Corporibus liquidis sunt omnia pura relicta,  
Et leviora aliis alia, et liquidissimus aether 500  
Atque levissimus aëria super influit auras,  
Nec liquidum corpus turbantibus aeris auris  
Commiscet; sinit haec violentis omnia verti  
Turbinibus, sinit incertis turbare procellis,  
Ipse suos ignes certo fert impete labens. 505  
Nam modice fluere atque uno posse aëthera nisu  
Significat Pontos, mare certo quod fluit aestu  
Unum labendi conservans usque tenorem.  
Motibus astrorum nunc quae sit causa canamus.  
Principio magnus caeli si vertitur orbis, 510  
Ex utraque polum parti premere aera nobis  
Dicendum est extraque tenere et claudere utrimque;  
Inde alium supra fluere atque intendere eodem  
Quo volvenda micant aeterni sidera mundi;

Output

Picture Zones

Fill Zones

Dewarping

Despeckling

Thumbnails

Follow page

z0050

z0051

z0052

z0053

Natural order

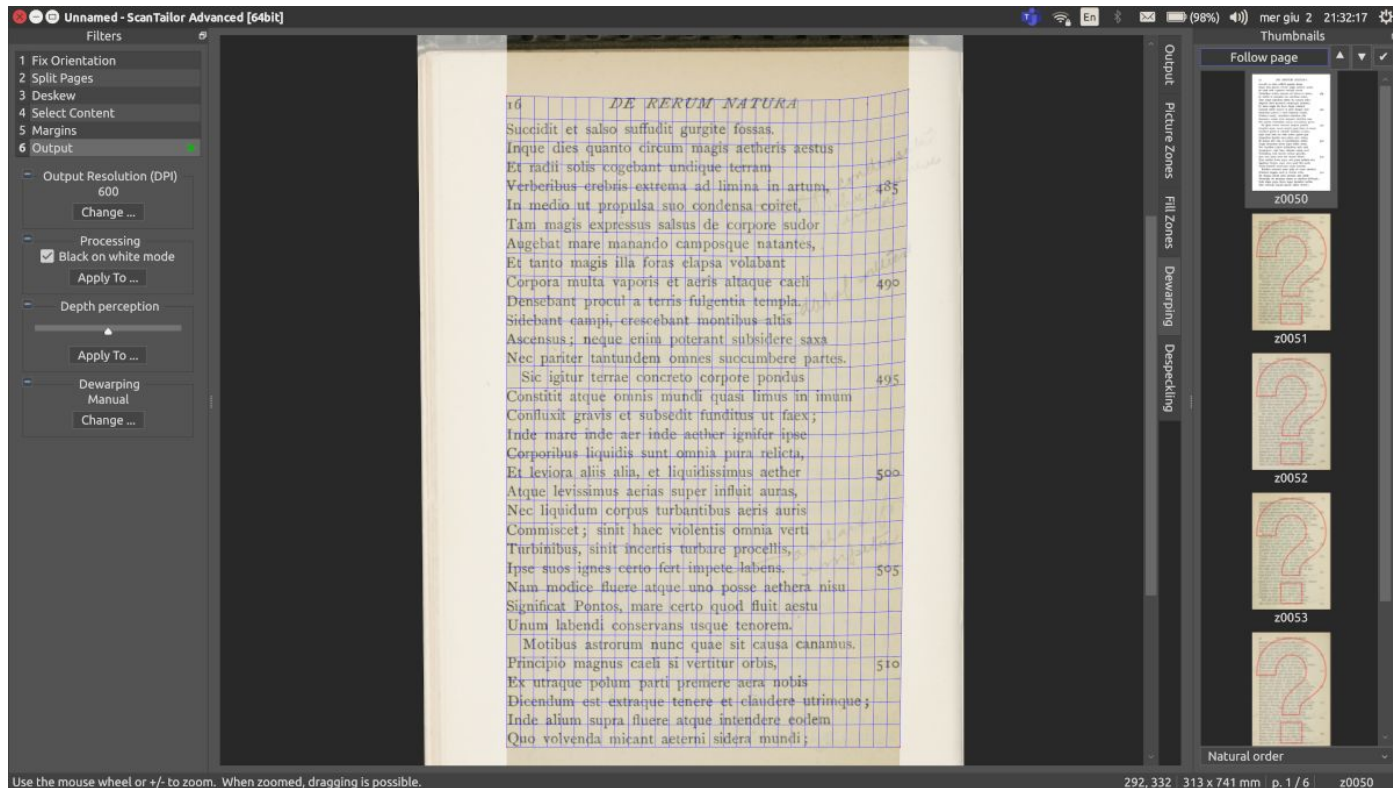
mer giu 2 21:25:10

(98%)

Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

-51,92 135 x 200 mm p. 1 / 6 z0050

# Dewarping



# Manual or automated?

On small projects, this operations usually are performed manually; on massive projects, usually they are performed automatically.

# Exercise

Compare processed and raw images on <https://archive.org>

# **Optical Character Recognition (OCR)**

# Commercial applications

There are many commercial applications for Optical Character Recognition, such as Abbyy FineReader, Adobe Acrobat Pro, etc. (Among many other comparative evaluations, see for example: <https://www.adamenfroy.com/best-ocr-software>)

There are also many solutions online (e.g. a new service on GoogleDrive to extract text from PDF of images)

And finally there are many apps to capture images with a smartphone and convert them into text or searchable PDFs.



# Commercial applications: strength points

The main advantages of commercial software are:

- simple to install on Windows and Mac
- easy to use
- graphical interface

# Commercial applications: weakness points

The main issues of commercial software are:

- necessity to renew the license for new versions
- scalability (when you must pay per page recognized)
- languages and scripts (FineReader has additional packages for old or ancient scripts, such as Fraktur)

# Open source applications for OCR

The most performant open source applications for OCR are:

- Tesseract (<https://github.com/tesseract-ocr/tesseract>)
- OCRopus (<https://github.com/ocropus/ocropy>) and its derivatives, listed below
- Kraken (<http://kraken.re>)
- Calamari (<https://github.com/Calamari-OCR/calamari>)

Another interesting OCR project is

- Gamera

# Open source applications: strength points

The main advantages of these projects are:

- scalability (to process millions of pages)
- scientific research to process challenging documents (endangered languages, ancient languages and scripts, low quality paper and ink, damaged documents)
- support of the community

# Open source applications: weakness points

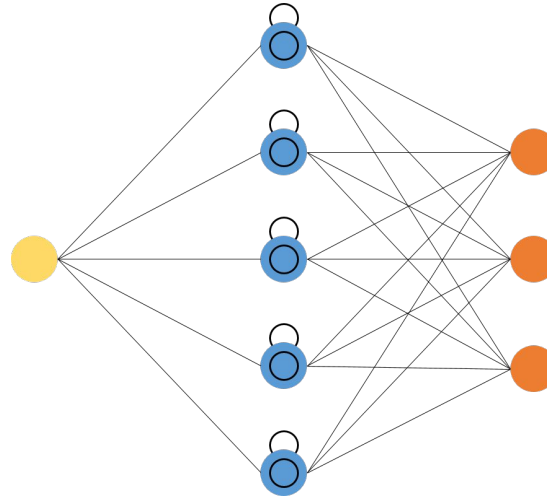
The main issues of these projects are:

- incompatible versions in quick evolution
- no graphical interface (only command line)
- not available for all the Operative Systems

# What is an OCR engine?

In simple words, an OCR engine is a classifier, which assigns a **label** (i.e. a character or a sequence of characters) to an **image region**

For this reason, the most recent OCR engines are based on Neural Networks



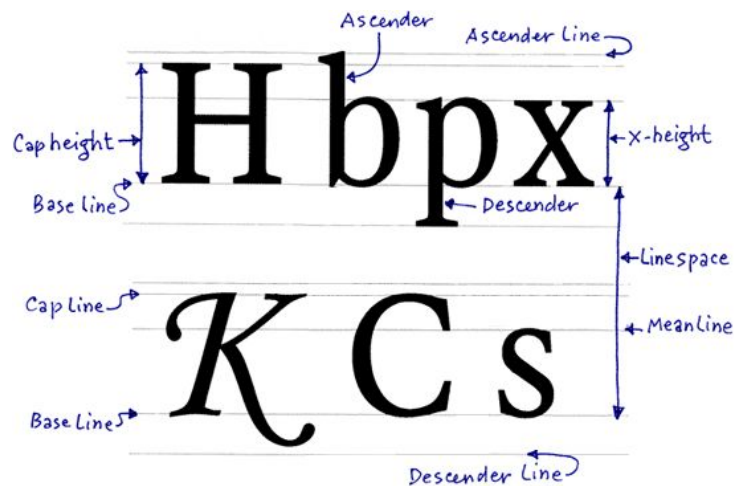
# Layout analysis

In order to assign a label, at the character level, to an image region, regions must be identified by layout analysis and segmentation

The **layout analysis** decomposes the page in its textual and graphical components (e.g. columns of text, illustrations, and tables)

# Segmentation

Textual blocks are hierarchically segmented in lines, words, and characters



*que le processus de paix réussisse*". "Il ne saurait en aucun cas être question de nouvelles concessions palestiniennes", a-t-il pour-

how-ocr-works.com



# Segmentation issues

Bad segmentation causes bad OCR

Factors that must be taken into account:

- avoid artifacts during the image acquisition process, such as page warping (when it is possible, pages should be unbounded!)
- preprocess the images to reduce artifacts
- if an OCR engine makes a bad segmentation, try another one (for example, if Abby FineReader does not satisfy your needs, try tesseract or Kraken and vice versa)

# Trained data sets

Both commercial and open source OCR applications are provided with pre-trained data sets

For this reason, we can perform the optical character recognition on a variety of languages and scripts, without taking care of the training phase

# Training

When the accuracy of the recognition is not satisfactory, it is necessary to train the system

Training is based on an **accurate** association between **text** and **image**

The text that exactly matches the image is called **ground truth**

Some OCR engines, such as tesseract, need a small amount of ground truth, some others on the contrary need a large amount.

# Training: the case of tesseract

The screenshot shows the jTessBoxEditorFX application window. The title bar reads "jTessBoxEditorFX - vie.times.exp0.tif". The menu bar includes "File", "Edit", "Settings", "Tools", and "Help". Below the menu bar, there are tabs for "Trainer", "Box Editor", and "TIFF/Box Generator". The "Box Editor" tab is active, showing a toolbar with "Open", "Save", "Reload", "Merge", "Split", "Insert", and "Delete". A "Character" field is set to "x 0".

The main area is divided into two panes. The left pane contains a table with the following columns: "Box Coordinates", "Char", "X", "Y", "Width", and "Hei...". The table lists 28 rows of data, each representing a character and its bounding box coordinates.

Box Coordinates	Char	X	Y	Width	Hei...
1	a	101	116	15	16
2	A	125	108	25	24
3	à	161	108	15	24
4	À	185	100	25	32
5	â	221	108	15	24
6	Â	245	100	25	32
7	ã	281	110	15	22
8	Ã	305	102	25	30
9	á	341	108	15	24
10	Á	365	100	25	32
11	ø	401	116	15	21
12	A	425	108	25	29
13	ä	461	109	15	23
14	Ä	485	101	25	31
15	å	521	100	15	32
16	Å	545	100	25	32
17	ä	581	100	15	32
18	Ä	605	100	25	32
19	ä	641	101	15	31
20	Ä	665	100	25	32
21	ä	701	100	15	32
22	Ä	725	100	25	32
23	ä	761	109	15	28
24	Ä	785	104	25	34
25	ä	821	108	15	24
26	Ä	844	100	26	32
27	ä	881	100	15	32
28	Ä	904	100	26	32

The right pane shows a preview of the text "gguangschoolsoisuthan..." with blue bounding boxes around each character. The text is displayed in a serif font. The preview area includes a scroll bar on the right and a "Find" button at the bottom.

# Performing OCR

```
tesseract -l <language(s)> <image> <output without suffix>
```

```
tesseract -l ita+lat img001.tiff doc001
```

(training data are available here: <https://github.com/tesseract-ocr/tessdata>)

tesseract -l <language(s)> <image> <output without suffix> hocr

**tesseract -l ita+lat img001.tiff doc001 hocr**

```
4 <html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
5 <head>
6 <title></title>
7 <meta http-equiv="Content-Type" content="text/html;charset=utf-8"/>
8 <meta name="ocr-system" content="tesseract 5.0.0-alpha-20210401-94-ga968" />
9 <meta name="ocr-capabilities" content="ocr_page ocr_carea ocr_par ocr_line ocrx_word ocrp_wconf"/>
10 </head>
11 <body>
12 <div class="ocr_page" id="page_1" title="image "z0053.tif"; bbox 0 0 12181 19222; ppageno 0">
13 <div class="ocr_carea" id="block_1_1" title="bbox 3910 479 11495 887">
14 <p class="ocr_par" id="par_1_1" lang="ita" title="bbox 3910 479 11495 887">
15 <span class="ocr_line" id="line_1_1" title="bbox 3910 479 11495 887; baseline -0.006 -81; x_size 400; x_descenders
16 93; x_ascenders 120">
17 <span class="ocrx_word" id="word_1_1" title="bbox 3910 500 5473 805; x_wconf 68">LIBER</span>
18 <span class="ocrx_word" id="word_1_2" title="bbox 5838 479 8127 887; x_wconf 51" lang="lat">QUINTUS.</span>
19 <span class="ocrx_word" id="word_1_3" title="bbox 11121 576 11227 763; x_wconf 95" lang="lat">I</span>
20 <span class="ocrx_word" id="word_1_4" title="bbox 11311 573 11495 854; x_wconf 95" lang="lat">9</span>
21 </span>
22 </p>
23 <div class="ocr_carea" id="block_1_2" title="bbox 402 1167 11496 18639">
24 <p class="ocr_par" id="par_1_2" lang="lat" title="bbox 485 1167 11496 6397">
25 <span class="ocr_line" id="line_1_2" title="bbox 525 1167 10060 1598; baseline -0.004 -88; x_size 407; x_descenders
26 85; x_ascenders 117">
27 <span class="ocrx_word" id="word_1_5" title="bbox 525 1182 3279 1598; x_wconf 89">Quandoquidem</span>
28 <span class="ocrx_word" id="word_1_6" title="bbox 3515 1198 4706 1505; x_wconf 89">claram</span>
29 <span class="ocrx_word" id="word_1_7" title="bbox 4944 1185 6378 1593; x_wconf 89">speciem</span>
30 <span class="ocrx_word" id="word_1_8" title="bbox 6610 1241 8492 1583; x_wconf 89">certamque</span>
```

# **Early printed editions and Handwritten Text Recognition (HTR)**

# Early printed editions and manuscripts

Early printed editions and manuscripts are challenging:

- complex and/or irregular layout
- abbreviations
- ligatures
- irregular letters





# How to train Kraken

Tutorial:

<http://kraken.re/training.html#training>

Training ancient Greek, early editions:

[https://github.com/pharos-alexandria/ocr-greek\\_cursive/blob/91d72606e2a60593e5eccafe14e6c98493a90ce7/README.md](https://github.com/pharos-alexandria/ocr-greek_cursive/blob/91d72606e2a60593e5eccafe14e6c98493a90ce7/README.md)



# Improving OCR and HTR

# Accuracy

OCR is evaluated according to the **accuracy**, a measure that is expressed by the following formula

$$\text{matches} / (\text{matches} + \text{mismatches} + \text{adds} + \text{dels})$$

according to the general formula

$$\text{TP} + \text{TN} / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

matches are the agreement between the OCR result and the ground truth

# Techniques to improve OCR and HTR

OCR and HTR can be improved by **postprocessing**

A couple of strategies are worthy of attention:

- alignment of multiple and independent OCR engines with efficient selecting criteria
- alignment to different editions of the same text, with criteria to distinguish between OCR errors to be corrected and genuine variants

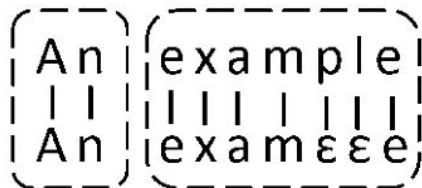
# Alignment

<https://link.springer.com/article/10.1007/s10032-020-00359-9>

**GT:** An example

**OCR:** An exam e

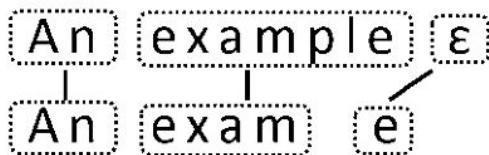
Char alignment first:



# char errors: 2

# word errors: 1

Direct word alignment:



# word errors: 2

# Exercise

Try to align two sequences of characters

[https://bioboot.github.io/bimm143\\_W20/class-material/nw](https://bioboot.github.io/bimm143_W20/class-material/nw)



**Manual correction**



WIKISOURCE

- Pagina principale
- Portali tematici
- Un testo a caso
- Un indice a caso
- Un autore a caso
- Una pagina a caso
- Ultime modifiche

Comunità

- Aiuto
- Portale Comunità
- Bar
- Progetti tematici
- Fai una donazione
- Contatti

Strumenti

- Puntano qui
- Modifiche correlate
- Cerca un file
- Pagine speciali
- Informazioni pagina
- » Crop/Tooi (Ritaglio immagine)

Stampa/esporta

- Scarica ePub
- Scarica MOBI
- Scarica PDF
- Altri formati

In altre lingue

Strumenti per la rilettura (Aiuto)

- Trova & sostituisci
- Elimina riga 1 Alt+5
- Aggiusta paragrafi Alt+6
- PostOCR Alt+7
- Unisci linee Alt+8
- AutoBt

< > Pagina [Discussione](#) [Immagine](#) [مناقشة](#)

Federico boschetti [discussione](#) [preferenze](#) [beta](#) [osservati speciali](#) [contributi](#) [esci](#)

[Leggi](#) [Modifica](#) [Cronologia](#) [Altro](#)

## Modifica di Pagina:Tragedie di Eschilo (Romagnoli) I.djvu/129

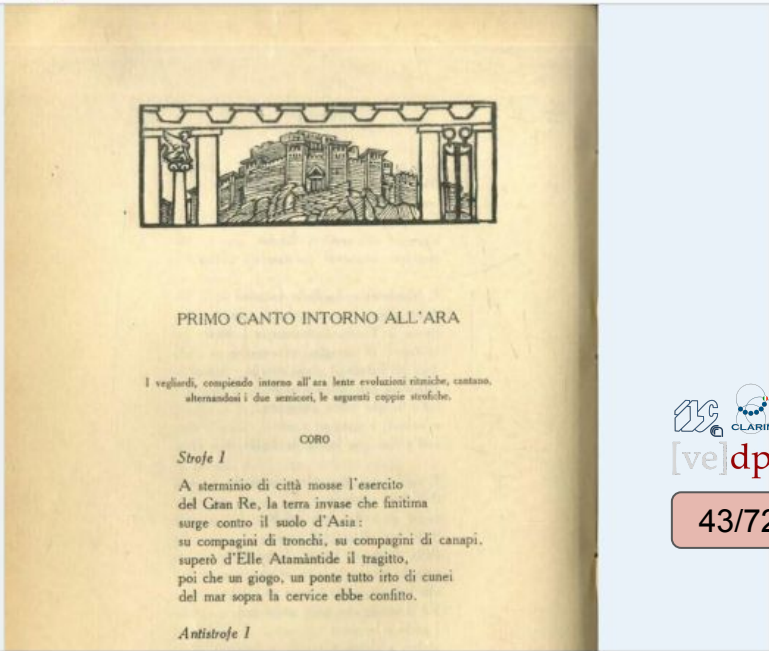
[G](#) [C](#) [∞](#) [🖨](#) [🔍](#) [OCR](#) [✍](#) [> Avanzate](#) [> Caratteri speciali](#) [> Aiuto](#) [> Zoom/Altro](#) [Template usati \(clicca per info\)](#) [Ct](#) [Nota separata](#) [Sc](#) [Smaller](#) [Vc](#)

```
Intestazione (non inclusa):

Corpo della pagina (da includere):
<poem>
[[file:Tragedie di Eschilo (Romagnoli) I-25.png|400px|center]]
{{C|v=2|t=2|PRIMO CANTO INTORNO ALL'ARA}}
{{vc|{{smaller|I vegliardi, compiendo intorno all'ara lente evoluzioni ritmiche, cantano, alternandosi i due senicori,
le seguenti coppie strofiche.}}}}
{{vc|{{sc|coro}}}}
''Strofe I''
A sterminio di città mosse l'esercito
del Gran Re, la terra invase che finitima
surge contro il suolo d'Asia:
su compagni di tronchi, su compagni di canapi,
superò d'Elle Atamantide il tragitto{{Nota separata|Pagina:Tragedie di Eschilo (Romagnoli) I.djvu/351|14}},
poi che un giogo, un ponte tutto irto di cunei
del mar sopra la cervice ebbe confitto.

''Antistrofe I''
Il Signore dei frequenti asiatici popoli
furioso, da due bande spinte d'uomini
una greggia innumerevole
su la terra dei nemici, qua pedoni, là dal pelago.
</poem>
```

PIÙ di pagina (non inclusa)



43/72





## Aiuto:Stato di Avanzamento del Lavoro

Aiuto: Stato di Avanzamento del Lavoro

Manuale ► Guida del percorso di qualità dei testi ► **Stato di Avanzamento del Lavoro**


Lo **Stato di Avanzamento del Lavoro (SAL)** è il livello che indica la qualità dei testi che hanno intrapreso il percorso di qualità di Wikisource.

Nel namespace pagina [ modifica ]





	<b>SAL 25%</b>	<i>predefinito</i>
	<b>SAL 50%</b>	La pagina è problematica
	<b>SAL 75%</b>	La pagina è stata trascritta e formattata
	<b>SAL 100%</b>	La pagina è stata riletta da un utente diverso da quello che ha portato la pagina al SAL 75%

Il SAL 00% si usa per segnalare:

- pagine vuote
- pagine che contengono testo (es.: pubblicità di altri volumi della collana, *ex libris*, oppure i "giudizi della critica") o immagini (es.: timbri o etichette della biblioteca o del proprietario) che non fanno parte dell'opera in senso stretto
- pagine in lingue diverse dall'italiano (vedi {{fwpage}})

	<b>SAL 00%</b>	La pagina non necessita di trascrizione
---	----------------	---

Nel namespace indice [ modifica ]

	<b>SAL 25%</b>	<i>predefinito</i>
	<b>SAL 50%</b>	Tutte le pagine hanno raggiunto o superato il SAL 50% (escluse le pagine SAL 00%)
	<b>SAL 75%</b>	Tutte le pagine hanno raggiunto o superato il SAL 75% (escluse le pagine SAL 00%)
	<b>SAL 100%</b>	Tutte le pagine hanno raggiunto il SAL 100% (escluse le pagine SAL 00%)

# Commerce Numérique

commerce 9 Page 80: Show Image Status: ★★★★★ Save

Index  
OCR

— 80 —  
— 80 —

suite creusé des lacs, songé à capter les eaux, à faire  
suite creusé des lacs, songé à capter les eaux, à faire

établir des barrages, sans quoi (on pouvait faire le  
établir des barrages, sans quoi (on pouvait faire le

calcul) en trente-deux heures plus une goutte d'eau.  
calcul) en trente-deux heures plus une goutte d'eau.

Mei is déposé un projet tendant à augmenter la

TEI

# Lace (http://heml.mta.ca/lace)

Lace: Visualizing, Editing and Searching Polylingual OCR Results    Latest Edits    Search    FAQ    Editing Guide    About

Aristotle (1829), Aristotelis De generatione animalium libri quinque

-20   -5   Previous   13   Next   +5   +20

98%

Zone Type    Line Mode    Clear Zones

— ▲ —      7

ἑταίρα. διὰ καὶ ἐν τῇ ὁμίλῃ ἡ σύνταξις γίνεται τῶν σκελῶν τὸ τε γὰρ ἔργων νευρώδης καὶ ἡ φύσις τῶν σκελῶν νευρώδης. ὥστ' ἐπεὶ τὸτ' ἔκ ἀδεχεται ἔχει, ἀνάγκη καὶ ἔρχει ἢ μὴ ἔχει ἢ μὴ ἑταίρ' ἔχει· τοὺς γὰρ ἔχουσιν ἢ αὐτῇ ὅτις ἀμφοτέρων αὐτῶν. ἐπεὶ δὲ τοῖς γε τοῖς ἔρχεται ἔχουσιν ἔξω διὰ τῆς κινήσεως θερμοκρασίαν τῆ αἰδέου προέρχεται τὸ σπέρμα συναβρασιῶν, ἀλλ' ἔχει ὡς ἔτιμον ἐν πύθι θηῶν, ὥσπερ τοῖς ἰχθύσι. πάντα δ' ἔχει τὰ ζῴοτα τῶν ἔρχεται ἐν τῷ πρῶτον ἢ ἔξω, πλὴν ἔχει· ἔτος δὲ πρὸς τῇ ἰσοφύμῳ, διὰ τὴν αἰτὴν αἰτίαν δι' ἧπερ καὶ οἱ ἄρνες ταχύνονται, γὰρ ἀναγκαῖον γίνεσθαι τὸν συνδυασμὸν αὐτῶν· οὐ γὰρ ὥσπερ τὰ τετραπόδα ἐπὶ τὰ πρῶτ' ἐπιβαίνει, ἀλλ' ὄρθα μίγνεται διὰ τὰς ἀνάσεις. δι' ἣν μὲν ἐν αἰτίαν ἔχουσιν τὰ ἔχοντα ἔρχεται, ἄρεται, καὶ δι' ἣν αἰτίαν τὰ μὲν ἔξω τὰ δ' ἐντός. ἴσα δὲ μὴ ἔχει, καθάπερ εἴρηται, διὰ τε τὸ μὴ εἶναι τὸ ἀναγκαῖον μόνον οὐκ ἔχει τοῦτο τὸ μέρος, καὶ διὰ τὸ ἀναγκαῖον εἶναι ταχύν γίνεσθαι τὴν ἔχοντα· ταυταὶ δ' ἐστὶν ἡ τῶν ἰχθύων φύσις καὶ ἡ τῶν ἄρνων. εἰ μὲν γὰρ ἰχθύεις ἐχρῶνται παραπίπτουσαι καὶ ἀπολύουσαι ταχέως, ὥσπερ γὰρ ἐπὶ τῶν ἀνθρώπων καὶ πάντων τῶν ζῴων ἀνάγκη κατασχέσθαι τὸ πνεῦμα πρῶτον τὴν γυνή· τοῦτο δ' ἐκείναις συμβαίνει μὴ δεχόμεναι τὴν θαλάσσης, εἰσὶ δὲ εὐφραστοὶ τοῦτο μὴ ποιούσαι. ἕκαστος δὲ ἐν τῷ συνδυασμῷ τὸ σπέρμα πέττει αὐτῆς, ὥσπερ τὰ πᾶσα καὶ ζῴοτα, ἀλλ' ὑπὸ τῆς ὥρας τὸ σπέρμα πεπεμαμένον ἄβασιν ἔχουσιν, ὥστε μὴ ἐν τῷ ὄργασμῳ ἀλλήλων πεύει.

A —

ἑταίρα. διὰ καὶ ἐν τῇ ὁμίλῃ ἡ σύνταξις γίνεται τῶν σκελῶν τὸ τε γὰρ ἔργων νευρώδης καὶ ἡ φύσις τῶν σκελῶν νευρώδης. ὥστ' ἐπὶ τούτ' οὐκ ἀδεχεται ἔχει, ἀνάγκη καὶ ἔρχεται ἢ μὴ ἔχει ἢ μὴ ἑταίρ' ἔχει· τοὺς γὰρ ἔχουσιν ἢ αὐτῇ ὅτις ἀμφοτέρων αὐτῶν. ἐπεὶ δὲ τοῖς γε τοῖς ἔρχεται ἔχουσιν ἔξω διὰ τῆς κινήσεως θερμοκρασίαν τοῦ αἰδέου προέρχεται τὸ σπέρμα συναβρασιῶν, ἀλλ' οὐκ ὡς ἔτιμον ἂν εἴθεθαι θεῶν, ὥσπερ τοῖς ἰχθύσι. πάντα δ' ἔχει τὰ ζῴοτα τοῖς ἔρχεται ἐν τῷ πρῶτον ἢ ἔξω, πλὴν ἔχοντα οὐκ ἔχει πρὸς τῇ ἰσοφύμῳ, διὰ τὴν αἰτὴν αἰτίαν δι' ἧπερ καὶ οἱ ἄρνες ταχύνονται, γὰρ ἀναγκαῖον γίνεσθαι τὸν συνδυασμὸν αὐτῶν· οὐ γὰρ ὥσπερ τὰ τετραπόδα ἐπὶ τὰ πρῶτ' ἐπιβαίνει, ἀλλ' ὄρθα μίγνεται διὰ τὰς ἀνάσεις. δι' ἣν μὲν οὖν αἰτίαν ἔχουσιν τὰ ἔχοντα ἔρχεται, εἴρηται, καὶ δι' ἣν αἰτίαν τὰ μὲν ἔξω τὰ δ' ἐντός. ἴσα δὲ μὴ ἔχει, καθάπερ εἴρηται, διὰ τε τὸ μὴ εἶναι τὸ ἀναγκαῖον μόνον οὐκ ἔχει τοῦτο τὸ μέρος, καὶ διὰ τὸ ἀναγκαῖον εἶναι ταχέως γίνεσθαι τὴν ὄρθαν· ταυταὶ δ' ἐστὶν ἡ τῶν ἰχθύων φύσις καὶ ἡ τῶν ἄρνων. οἱ μὲν γὰρ ἰχθύεις ὄρθουσαι παραπίπτουσαι καὶ ἀπολύουσαι ταχέως, ὥσπερ γὰρ ἐπὶ τῶν ἀνθρώπων καὶ πάντων τῶν ζῴων ἀνάγκη κατασχέσθαι τὸ πνεῦμα πρῶτον τὴν γυνή· τοῦτο δ' ἐκείναις συμβαίνει μὴ δεχόμεναι τὴν θαλάσσης, εἰσὶ δὲ εὐφραστοὶ τοῦτο μὴ ποιούσαι. οὐκ ἔστι δὲ ἐν τῷ



**HTR**

# From OCR to HTR

It is necessary to continue improving OCR accuracy, because it is the real bottleneck for computational linguistics and digital humanities

Currently OCR is very satisfactory on modern languages and documents with a simple layout, but it is challenging on early documents and old or ancient scripts

HTR is emerging with very promising results

# HTR: new challenges

High quality digital images, image enhancement techniques (preprocessing), larger amounts of training data, and better HTR engines (processing) are crucial to increase the performance of handwritten text recognition systems, but also textual and linguistic analysis applied to the HTR output (**post-processing**) can be taken into account



# Processing: segmentation (regions)

Il *Paraclytus* di Lelio Manfredi, composto fra il 1515 e il 1520 alla corte di Ferrara

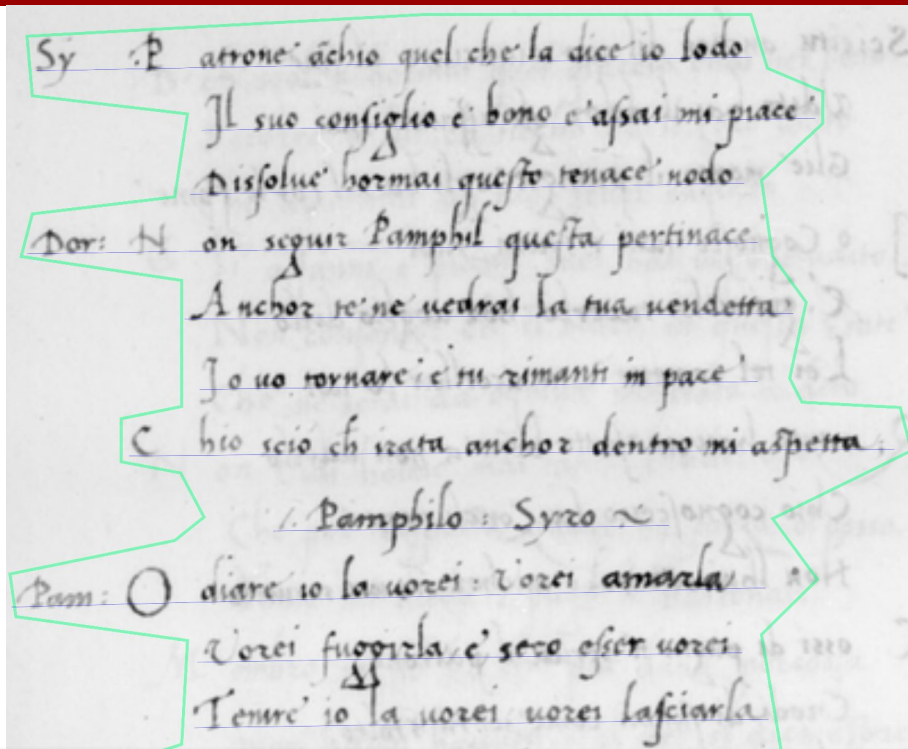


image from gallica.bnf.fr

# Processing: segmentation (baselines and masks)

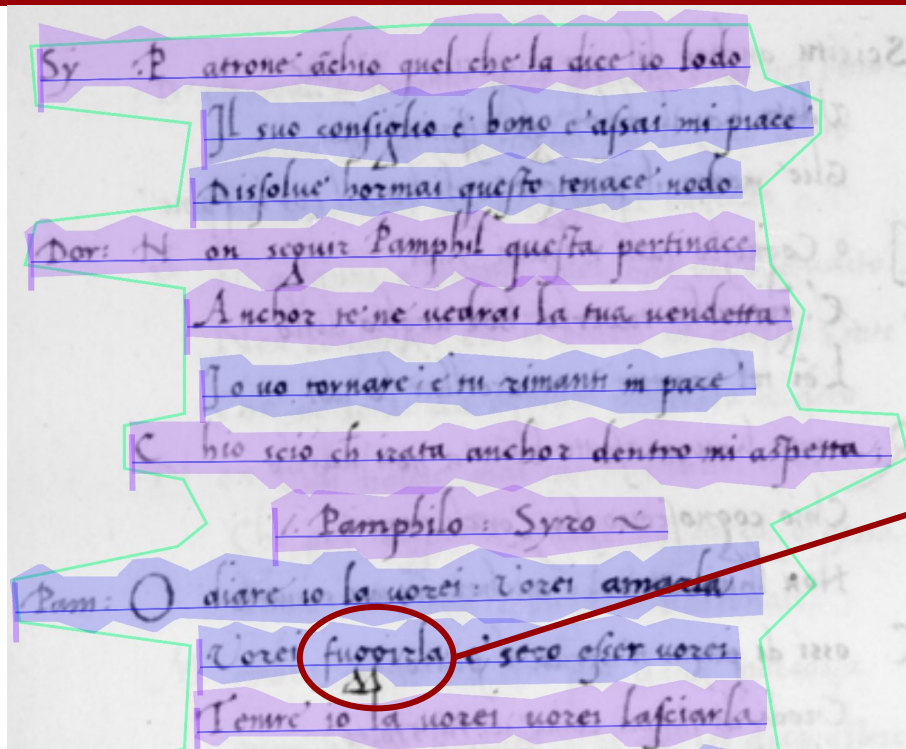


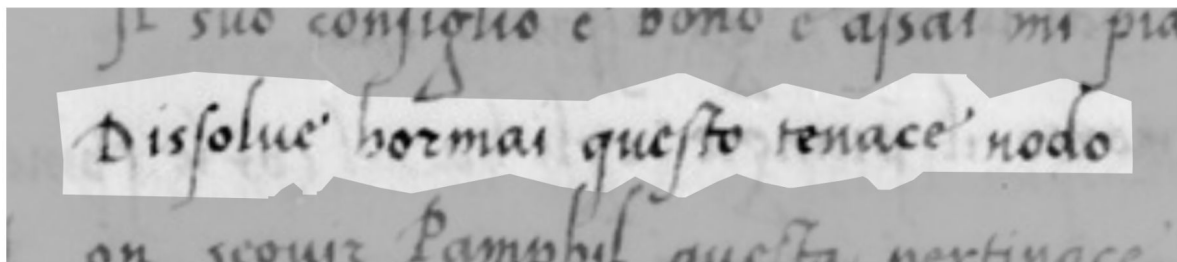
image from gallica.bnf.fr

# Post-processing goals

- **detection and classification of potential errors**
  - working with non-standard varieties of language is challenging, because it is harder to distinguish between orthographic variants and HTR errors
- **self-corrections**
  - the most likely errors can be automatically substituted by their most probable corrections, according to the specific linguistic, metrical and stylistic context
- **clues and suggestions for the proofreaders**
  - proofreaders, especially when they are students or volunteers, are highly facilitated by highlights in different colors (according to the type of potential error) and lists of suggestions

# Three different types of potential errors in a row

Line #4

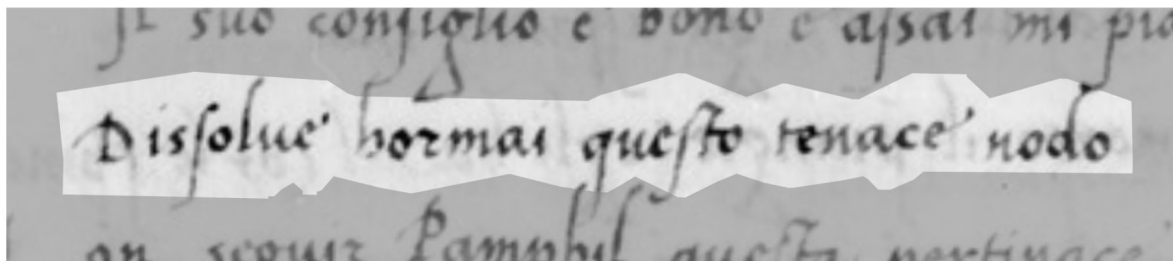


Diefolue hormai queft tenace nodo

by (eScriptorium) on Tue Jan 18 2022 08:36:11 GMT+0100

# Three different types of potential errors in a row

Line #4



Diefolue hormai queft tenace nodo

by (eScriptorium) on Tue Jan 18 2022 08:36:11 GMT+0100

wrong form

false negative

wrong form in helpful context

# Detection strategies

## For each token

- check in the list of forms previously corrected by hand
- check in the list of forms extracted by corpora of *similar* texts
- check in the list of all available forms
- in case of elision, verify if the next word starts by a vowel
- in case of articles and pronouns, verify if they agree with the following noun phrase (noun, adjective+noun...)
- in case of poems, verify the rhymes
- classify previously unattested forms in **potential words** (i.e. sequences of characters that respect phonetic and orthographic rules) and **nonwords** (i.e. random sequences of characters)

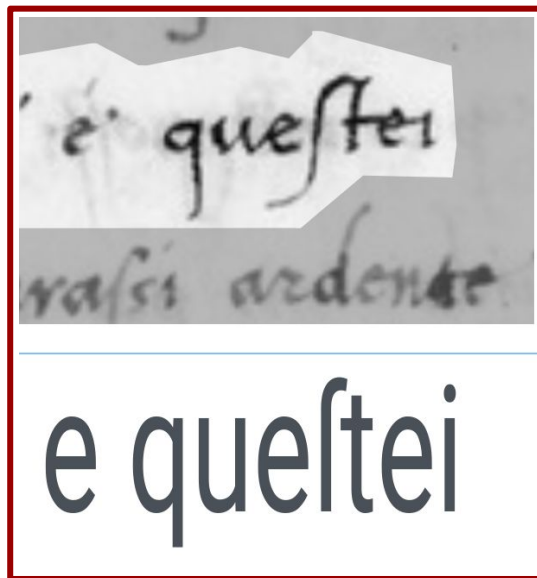
# Repertories of attested historical forms

## Repertories of attested historical forms

- are built on corpora of literary and documentary sources
- contain much more inflected forms than the lists used to create the most popular spell-checkers, which are based on the standard language
- may contain
  - number of occurrences (total or divided by subcorpora)
  - part of speech and morphological traits
  - diachronic information (range between first and last attestation)
  - genre information (depending on the metadata of the original corpora)
  - peculiar use by authors

# Repertories of attested historical forms: an example

*questei* is an ancient variant of *costei*  
(=this woman) that a standard  
spell-checker rejects





# Variant spellings

Manuscripts contain multiple spellings of the same forms, due to

- evolution of orthographic rules (e.g. principii, principij, principî, principi)
- concurrent spellings in the same manuscript

Different spellings may be

- attested (**true positives**)
- previously unattested but inferrable (**borderline false negatives**)
- previously unattested and unpredictable (**false negatives**)

# Variant spellings: examples

## Uses of H

- initial h before vowel
  - hoggi
  - hormai ( but horamai)
- h after c
  - focho
  - secho
- h after g
  - rogho (but fogho...)
  - vegggho

- h after l
  - alhora
- h after r
  - perhó
  - trarhá

## Proclitics fused with the following word

- article
  - lamor
- preposition
  - detá
- pronoun
  - mha
  - lhavea

# Agreement

In Italian articles and demonstrative, possessive and indefinite adjectives

- are very frequent words
- are inflected and agree with the head of the noun phrase

Whenever

- we can detect a sequence constituted by  
DET ADJ\* NOUN ADJ\* (e.g. *la bella chioma, la chioma dorata, la sua bella chioma dorata*)
- we know the morphological traits (i.e. gender and number) of determiners, nouns and adjectives

we can check their agreement and detect inconsistencies

# Agreement: examples

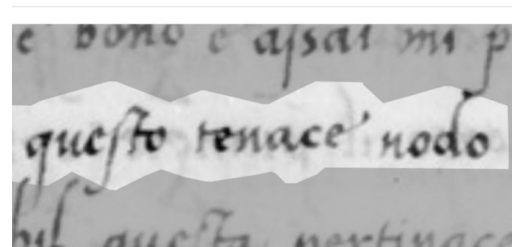
The most probable candidates for **quest** are

- quest' [NP]-----[mf][sp]- (?=[aeiou])
- questa [NP]-----fs-
- queste [NP]-----fp-
- questi [NP]-----mp-
- questo [NP]-----ms-

but the phonetic constraint and the agreement with

- tenace [A]-----[mf]s=
- nodo [N]-----ms-

narrow the choice to **questo**



quest tenace nodo

wrong form in helpful context

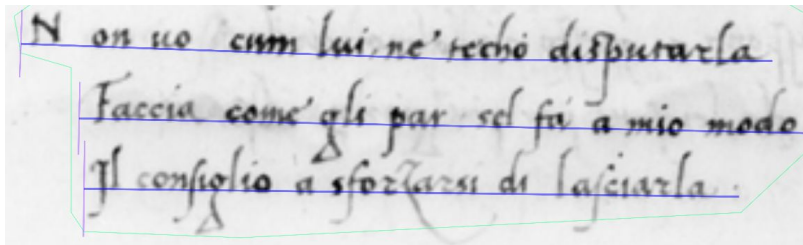
# Rhyme

Poems can exploit the mutual information provided by rhymes, especially with fixed metrical schemes, such as

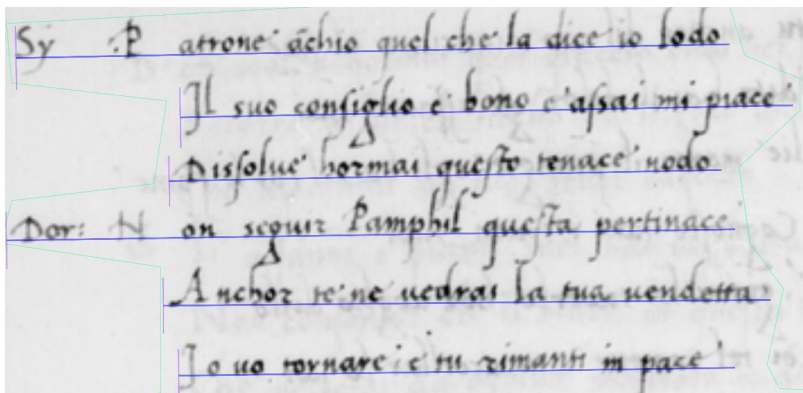
- AA
- ABA BCB CDC ...
- ABAB CDCD ...
- ABBA CDDC ...
- ABABABCC DEDEDEFF ...

Even when we cannot identify which words have the stress on the fourth- or third-last syllable, we can always check the identity of the last vowels with an accent or of the sequences of characters from the second-last vowel to the word end

# Rhyme: examples



N on uo cum lui ne' techo disputarla  
Faccia come' gli par sel fu' a mio modo  
Il consiglio a sforzarsi di lasciarla.



Sy :P atrone' a'chio quel che la dice io lodo  
Il suo consiglio e' bono e' assai mi piace'  
Disfolue' hormai questo tenace' uodo  
Dor: N on seguir Pamphil questa pertinace'  
Anchor re' ne uedrai la tua uendetta  
Io uo tornare' e' tu rimanti in pace'

...

modo

...

for**do** → fodo ← lodo

(manual adjustment is required)

pi**ace**

nodo

pertina**ae** → pertin**ace**

(self-correction is enough)

...

pace

# Vocabulary (+morphology+orthography) restraint

- too large repertoires of inflected forms are deceiving
- **theological treatises, notarial documents, Renaissance poetry** are a few examples of texts with very specific vocabularies
- focus on diachronic and diatopic variants of morphology and orthography close to your case study
- use a spell-checkers based on inflected forms extracted by corpora of texts similar to your case study

# Vocabulary restraint: Leone Orsini

Leone Orsini,  
*Canzoniere*,  
c.a 1564

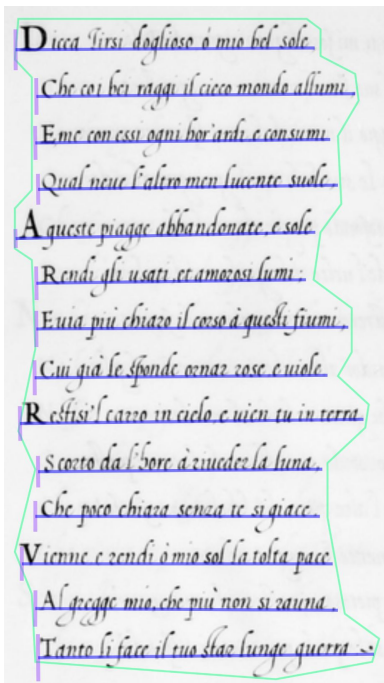


image from gallica.bnf.fr

Dicea **Tirsi** doglioso ò mio bel sole,  
Che coi bei raggi il cieco mondo allumi,  
E me con **essi** ogni hor ardi, e consumi  
Qual neue l'altro men lucente suole,  
A queste piagge abbandonate, e sole  
Rendi gli usati, et amorosi lumi;  
E uia più chiaro il corso à questi fiumi,  
Cui già le **sponde** ornar rose, e **uiole**.  
**Restisi'**l carro in cielo, e uien tu in terra  
Scorto da l'hore à riueder la luna,  
Che poco chiara senza te si giace.  
**Vienne**, e rendi o mio sol la tolta pace  
Al gregge mio, che più non si **rauna**,  
Tanto si face il tuo star lunge guerra :~

## LEGENDA

same in Petrach  
in P. with different spelling  
in P. with different  
inflection  
not attested in P.



# Multiple texts

In several cases we transcribe multiple manuscripts of the same work or we have access to previous (sometimes normalized) editions of our manuscripts

In these cases previous transcriptions (published or validated by an accurate proof-reading) of the same or of similar manuscripts become the **collation base** for our transcription, which can be **aligned** and **merged** with the HTR output

# HTR errors vs real variants

In order to avoid **contamination**, it is crucial to distinguish HTR errors from real variants

- agreement between the collation base (CB) and the HTR output reinforces the automated recognition
- disagreement is due
  - **to a real variant**
    - **correctly recognized by HTR** ←highlight it as a possible variant (true words very different from the CB, e.g. “biondi capelli” vs “crini dorati”)
    - **recognized by HTR with errors** ←highlight it as a possible variant with errors (non-words or pseudowords very different from the CB, e.g. “biordi capcli” vs “crini dorati”)
  - **to an artifact generated by HTR**
    - **that can be self-corrected with a high degree of confidence** ←self-correct with the word(s) in the CB but highlight it for manual check (non-words or pseudowords very close to the words in the CB, e.g. “cnini poiati” vs “crini dorati”)
    - **that needs human intervention** ←highlight it as a possible error (true words very close to the words in the CB, e.g. “canini orati” or “crini indorati” vs “crini dorati”)

# K-Centres

# CLARIN Knowledge Centres

## Knowledge Centres

### CLARIN Knowledge Infrastructure

CLARIN Knowledge Centres (K-centres) are a cornerstone of the CLARIN Knowledge Infrastructure (KI), one of the main components ensuring a continuous transfer of knowledge between all players involved in the construction, operation and use of the infrastructure. The mission of the CLARIN KI is to ensure that the available knowledge and expertise does not exist as a fragmented collection of unconnected bits and pieces, but is made accessible in an organised way to both the CLARIN community and the social sciences and humanities research community more widely.

### The Role of K-Centres

The focus of CLARIN is on language resources (in all modalities, from all regions and with any topical orientation) and K-centres serve researchers and educators from any discipline where language plays one of its many roles, ranging from object of study, a means of communication or expression, a means to store and extract information, object of learning or teaching activities, to training source for data-driven analytics, and many others. K-centres share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure and can be mostly found in CLARIN countries, but also exist elsewhere, and they all have a virtual presence.

### Areas of K-Centre Expertise

K-centres all have their own specific areas of expertise, which can belong to many different categories, such as

The Knowledge Centres of CLARIN can be contacted through their **Help Desks**

Homepage:

<https://www.clarin.eu/content/knowledge-centres>

# IMPACT

IMPACT is focused on  
**digitization**

## Tools & Resources

SEARCH OVER

0

TOOLS FOR TEXT DIGITISATION



### CitAttest

Attesting Word Forms in Dictionary Citations. With this **tool**, occurrences of a headword of a historical dictionary are automatically marked in the quotations belonging to that headword in the dictionary.

MORE TOOLS ➔

LEXICA FOR

0

DIFFERENT LANGUAGES



### Diasearch – Diachronic corpus search service

Diasearch is an online service which enables users to perform linguistically enriched queries on a collection of historical texts.

MORE LEXICA ➔

ACCESS

0

IMAGE AND GROUND TRUTH



### VIDA DE LAZARILLO DE TORMES

High quality images with **ground truth** associated from the collection of Biblioteca Nacional de España.

MORE RESOURCES ➔

Homepage:

<https://www.digitisation.eu>

# DiPText-KC

## DiPText-KC

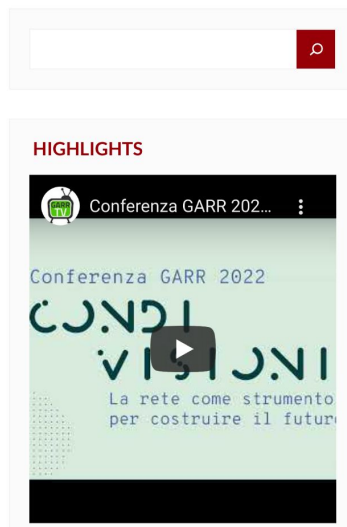
### CLARIN Knowledge Centre for Digital and Public Textual Scholarship

DiPText-KC offers expertise on methods, data, instruments and technologies relevant in the field of Philological and Literary Studies, History, Art History and Cultural Heritage.

Its actions aim at:

- sharing information with scholars and students about the state of the art in digital scholarly editing and text annotation through domain-specific languages;
- supporting scholars and students in the creation and publication of digital scholarly editions and resources;
- organizing training activities (for instance webinars, workshops and summer schools).

DiPText-KC is one of the Centres of [CLARIN-IT](#), the Italian node of [CLARIN](#) (Common Language Resources and Technology Infrastructure), a digital infrastructure of pan-European interest identified by [ESFRI](#) (European Strategy Forum on Research Infrastructures) and classified as a Landmark Research Infrastructure for the Social Sciences and Humanities (ESFRI Landmarks SSH RI).



The Digital and Public Textual Scholarship Knowledge Centre is focused on **digital philology**

Homepage:

<https://diptext-kc.clarin-it.it>

# DiPText-KC

The screenshot shows the DiPText-KC website with a navigation menu and several content sections. The navigation menu includes: CLARIN CENTRE, DiPText-KC, ABOUT, PARTNERS, PEOPLE, KNOWLEDGE, HELPDESK, EVENTS, NEWS, and CONTACT. The main content area is divided into three columns. The left column contains a list of 'Consortia, Associations, Centers' and 'Training' opportunities. The middle column features a 'BREAKING NEWS' section with three items. The right column displays a video player with a title in Italian: 'La Rete come strumento per costruire il futuro'.

**CLARIN CENTRE** **DiPText-KC** ABOUT PARTNERS PEOPLE KNOWLEDGE HELPDESK EVENTS NEWS CONTACT

**Consortia, Associations, Centers**

- Consortia
  - [Unicode Consortium](#)
  - [TEI Consortium](#)
- National and International Associations
  - [ADHO](#)
  - [EADH](#)
  - [AIUCD](#)
- DH Centres and Labs
  - Italy
    - [CIRCSE](#)
    - [LabCD](#)

**Training**

- Summer Schools
  - Venice Centre of Digital and Public Humanities, Department of Humanities, Ca' Foscari University of Venice
    - [Summer Camp 2020](#)
  - University of Pisa
    - [Digital Tools for Humanists 2022](#)  
(past editions: [2021](#) | [2020](#) | [2019](#) | [2018](#) | [2017](#))
  - [Digitising, Cataloguing, Searching and Sharing the Medieval and Early-Modern Image](#)
- Seminars / Webinars
  - [VeDPH Seminars in Digital and Public Humanities – January-May 2022](#)  
(past editions: [October 2019 – May 2020](#) | [September 2020 – December 2021](#))
  - [Humanities Horizons – History, Hacktivism and Genetic Criticism, Solstice Seminar in DPH](#)
  - [The Public Staging of Gender in Shakespearean Theatre Discussion with Pamela Allen Brown](#)

**Digital Libraries**

- Zotero Collections
  - [DiPText-KC Library](#)
  - [CLARIN Library](#)

**BREAKING NEWS**

- › Fourth Appointment of the Workshop Cycle  
“Digital Philology meets Computational Linguistics”
- › Third Appointment of the Workshop Cycle  
“Digital Philology meets Computational Linguistics”
- › **Concluded the First Cycle of the Permanent Seminar Series “A bridge between two worlds”**
- › CNR-ILC CoPhILab @ GARR 2022
- › Second Appointment of the Workshop Cycle  
“Digital Philology meets Computational Linguistics”

La Rete come strumento  
per costruire il futuro

**CNR-ILC CoPhILab @ GARR 2022**  
*The Collaborative and Cooperative Philology Lab (CoPhILab, CNR-ILC): data, applications, services and infrastructures (5:50:11-5:59:00)*

The Digital and Public Textual Scholarship Knowledge Centre keeps you informed on Consortia, Associations, Centres, Training Schools, and Digital Libraries relevant for digital philologists

<https://diptext-kc.clarin-it.it>



70/72

# Conclusion



# Conclusion

- better images and better HTR systems are crucial, but linguistic post-processing can be helpful to improve accuracy
- do not work with a document, work with a library!
- linguistic, metrical and stylistic information increase the confidence at word, phrase and line level
- many linguistic resources are available through the research infrastructure CLARIN
- CLARIN Knowledge Centres can help you to find relevant information on digitization

## Shared folder

<https://bit.ly/3xhBqk1>

# References

Boschetti, F. 2018. *Copisti digitali e filologi computazionali*, Roma.

Grossi, A.M. 2013. *Leone Orsini e il manoscritto “Italien 1535” della Bibliothèque nationale de France*, University of Toronto, PhD Thesis.

Kiessling, B., Tissot, R., Stokes, P. and Stökl Ben Ezra, D. 2019. *eScriptorium: An Open Source Platform for Historical Document Analysis*. In *International Conference on Document Analysis and Recognition Workshops (ICDARW) 2019*, pp. 19-19, doi: 10.1109/ICDARW.2019.10032.

---