

Digitization of documents

Federico Boschetti

federico.boschetti@ilc.cnr.it

CNR-ILC & VeDPH

June 3, 2021 - University of Pisa

Summerschool - 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

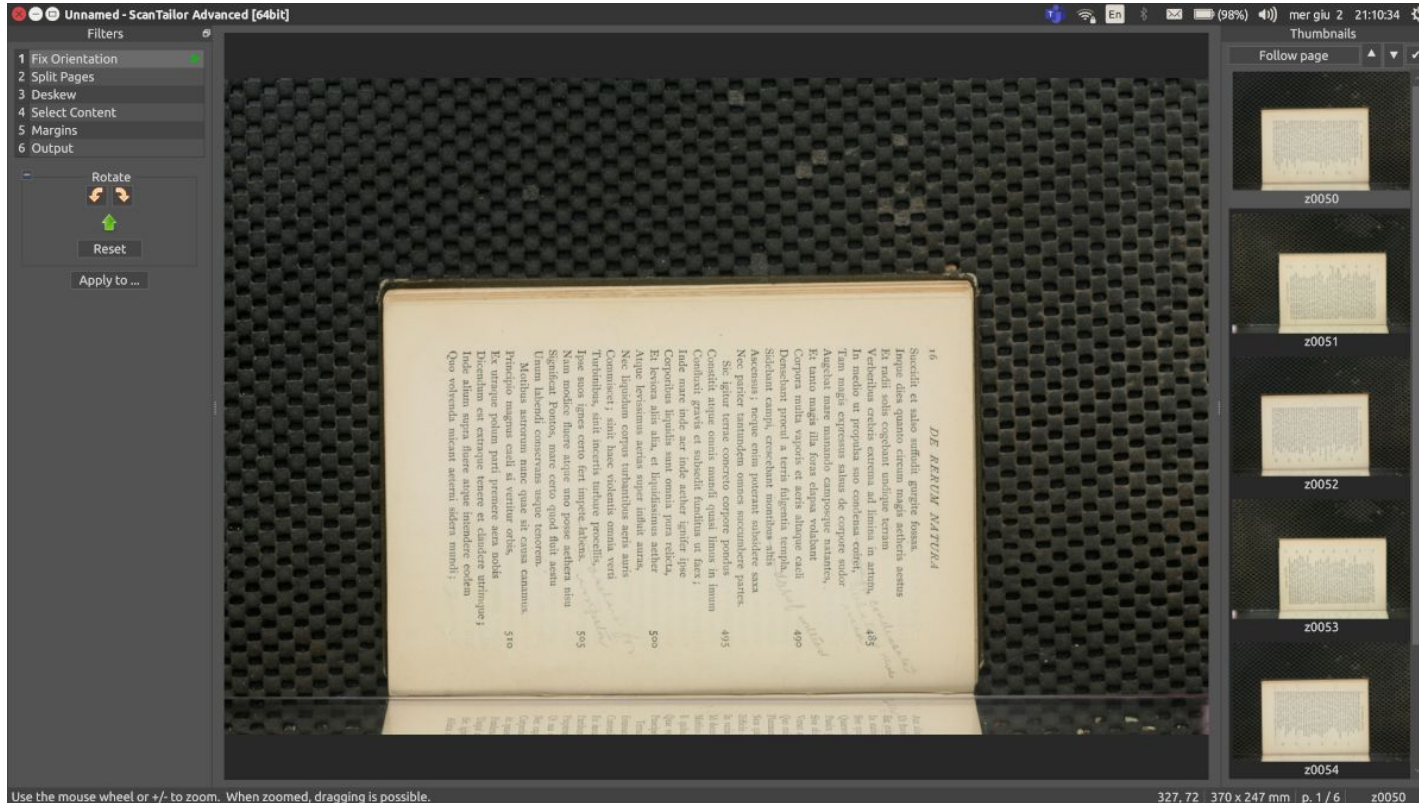
Along the digital text acquisition workflow one or more image elaborations are required. It is necessary to keep always the original images and possibly to preserve also the metadata related to the necessary transformations and use naming conventions for the files, with minimal metadata about dpi, color,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

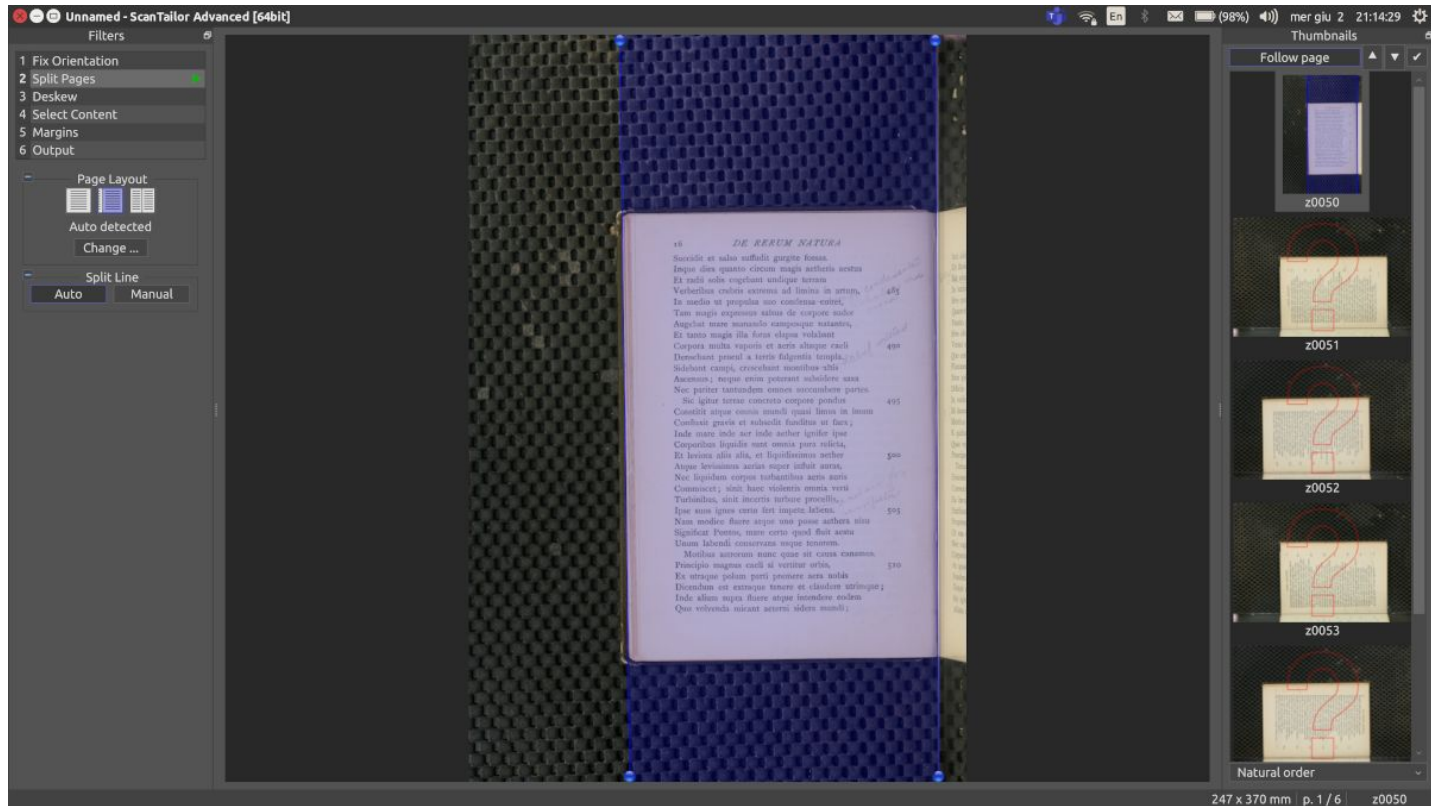
Fixing orientation



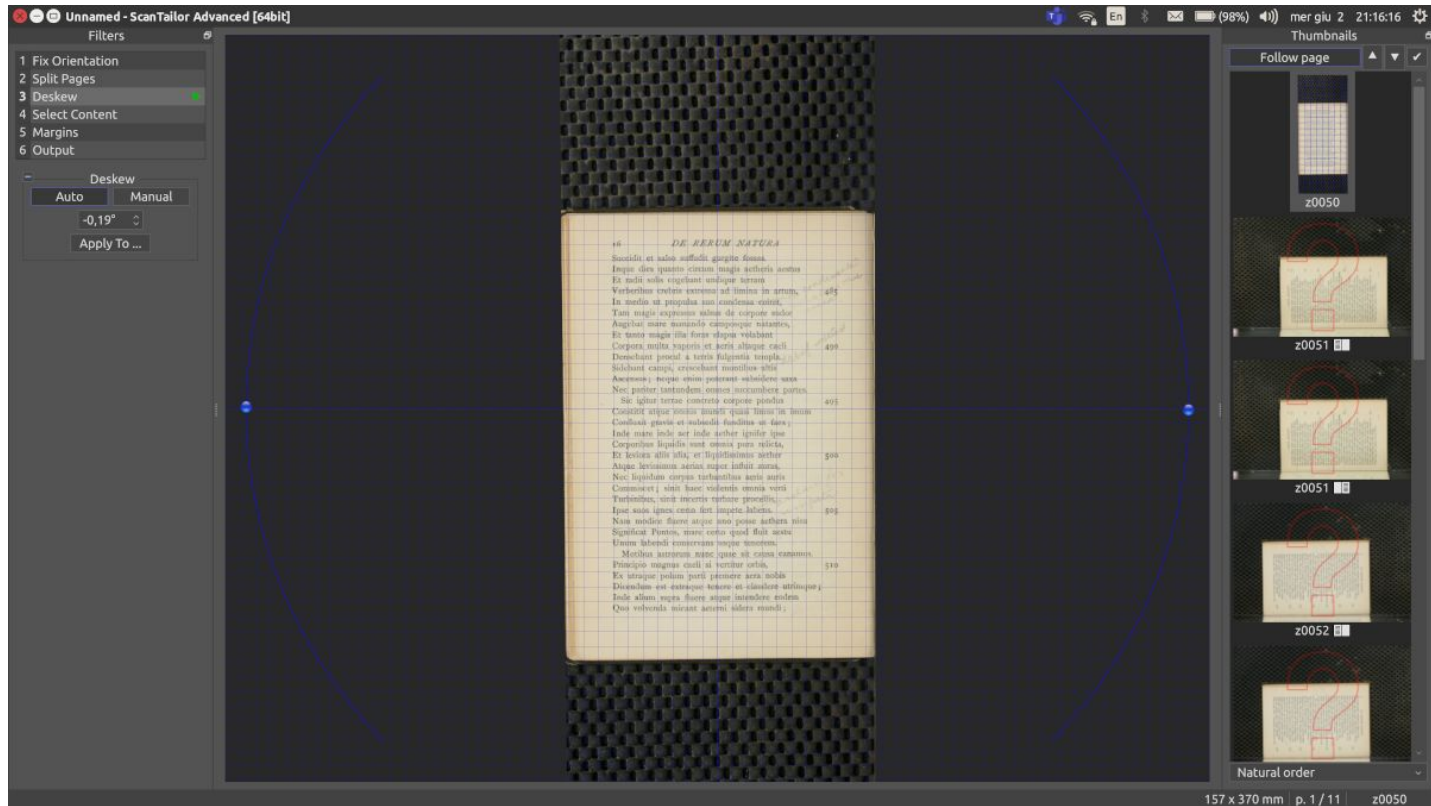
Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

327,72 370 x 247 mm p. 1 / 6 z0050

Splitting pages



Deskewing



Selecting content

The screenshot displays the ScanTailor Advanced interface. On the left, a 'Filters' panel 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' sections, each with 'Disable', 'Auto', and 'Manual' options. The main workspace shows a scanned page with a red rectangular selection box around the text. The text is Latin, starting with 'DE REBUS NATURA' and 'Societas et vobis sufficere gurgite fossa.' On the right, a 'Thumbnails' panel shows a vertical list of page thumbnails. The top thumbnail is labeled 'z0050' and has a red asterisk. Below it are thumbnails for 'z0051' and 'z0052', each with a red selection box. The bottom thumbnail is labeled 'z0050' and has a red asterisk. The status bar at the bottom indicates '-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 displays the ScanTailor Advanced interface. On the left, a sidebar contains a list of processing steps: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content, 5 Margins (highlighted), and 6 Output. The 'Margins' section is expanded, showing 'Auto Margins' checked and manual settings for Top (5,0), Bottom (5,0), Left (10,0), and Right (10,0). Below this, the 'Alignment' section is set to 'Manual' with various alignment icons. The 'Guides Help' section provides instructions on how to create, delete, and move guides. The main workspace shows a scanned page of Latin text from 'DE RERUM NATURA' with a pink border and numerical guides (485, 490, 495, 500, 505, 510) on the right side. A 'Follow page' panel on the right shows a vertical stack of thumbnails for pages z0050, z0051, z0051, z0052, and z0050, with 'Natural order' selected at the bottom. The status bar at the bottom indicates the page is at -31,45, 132 x 195 mm, page 1 of 11, and thumbnail z0050.

Filters

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

Margins

Auto Margins

Top 5,0

Bottom 5,0

Left 10,0

Right 10,0

Apply To ...

Alignment

Match size with other page

Mode: Manual

Apply To ...

Guides Help

- Right-click to create/remove guides from the context menu called.
- Right-click on a guide to delete that guide from the context menu called.
- **SHIFT+LMB** - drag the guide under the cursor.
- **SHIFT+Ctrl+LMB** on the content rectangle - drag the page content. Hold **SHIFT** pressed to restrict moving along the horizontal axis only or **Ctrl** for the vertical one. Hold **SHIFT+Ctrl** for usual dragging.
- **Double-click** on content - automatically attach that content to the nearest guide. Hold **SHIFT** pressed to not attach content to the nearest guide.

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 coeret,
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 caena canamus.
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;

Follow page

z0050

z0051

z0051

z0052

z0050

Natural order

Resize margins by dragging any of the solid lines.

-31,45 132 x 195 mm p. 1 / 11 z0050

Changing resolution

The screenshot displays the ScanTailor Advanced software interface. On the left, a sidebar contains various processing filters. The 'Output Resolution (DPI)' filter is highlighted with a red circle, showing a value of 600 and a 'Change ...' button. Below it, the 'Mode' is set to 'Black and White', and several 'Options' are checked, including 'Fill offcut', 'Fill margins', 'Equalize illumination (B)', 'Savitzky-Golay smoothing', and 'Morphological smoothing'. The 'Threshold' is set to 0, and 'Color operations' are also visible. The main window shows a scanned page of Latin text from 'DE RERUM NATURA' with line numbers 485, 490, 495, 500, 505, and 510. On the right, a 'Thumbnails' panel shows a sequence of processed images labeled z0050 through z0053. The status bar at the bottom indicates the page is at -51, 92, 135 x 200 mm, p. 1 / 6, with a zoom level of z0050.

Binarization

The screenshot displays the ScanTailor Advanced interface. On the left, the 'Filters' panel is active, showing a list of processing steps: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content, 5 Margins, and 6 Output. The 'Output' filter is selected, and its settings are visible: Output Resolution (DPI) is 600, Mode is 'Black and White', and 'Force b&w' is checked. A red circle highlights the 'Mode' dropdown menu. Below the 'Output' filter, there are sections for 'Options' (with checkboxes for 'Fill offcut', 'Fill margins', 'Equalize illumination (B)', and 'Morphological smoothing'), 'Threshold' (Method: Otsu, value: 0), 'Color operations' (with sliders for 'Reduce noise' and 'Posterize'), and 'Despeckling' (with a checked 'Despeckle' option). The main workspace shows a page of Latin text from 'DE RERUM NATURA' with line numbers 485, 490, 495, 500, 505, and 510. The text is being processed into a binary (black and white) format. On the right, a 'Thumbnails' panel shows a sequence of images: the original page, followed by several versions of the page with red question marks overlaid, indicating areas of concern or error during the binarization process. The status bar at the bottom indicates the page is at -51, 92, 135 x 200 mm, p. 1 / 6, with a zoom level of 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 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 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

Natural order

mer giu 2 21:25:10
98%

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

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

Despeckling

The screenshot displays the ScanTailor Advanced interface. On the left, the 'Filters' panel is active, showing a list of processing steps: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content, 5 Margins, and 6 Output. Under the 'Output' section, the 'Despeckling' filter is selected, and the 'Despeckle' checkbox is checked. A red circle highlights the 'Despeckling' section. The main workspace shows a scanned page of Latin text from 'DE RERUM NATURA' with a large red question mark overlaid on the center, indicating a detected speckle or artifact. The page number '16' is visible in the top left corner of the document. On the right, the 'Output' panel shows a vertical stack of thumbnails for the processed page, labeled 'z0050' through 'z0053'. The status bar at the bottom indicates the page is at 92% zoom, 135 x 200 mm, page 1 of 6, and zoomed to 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)
- 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 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 relictia,
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

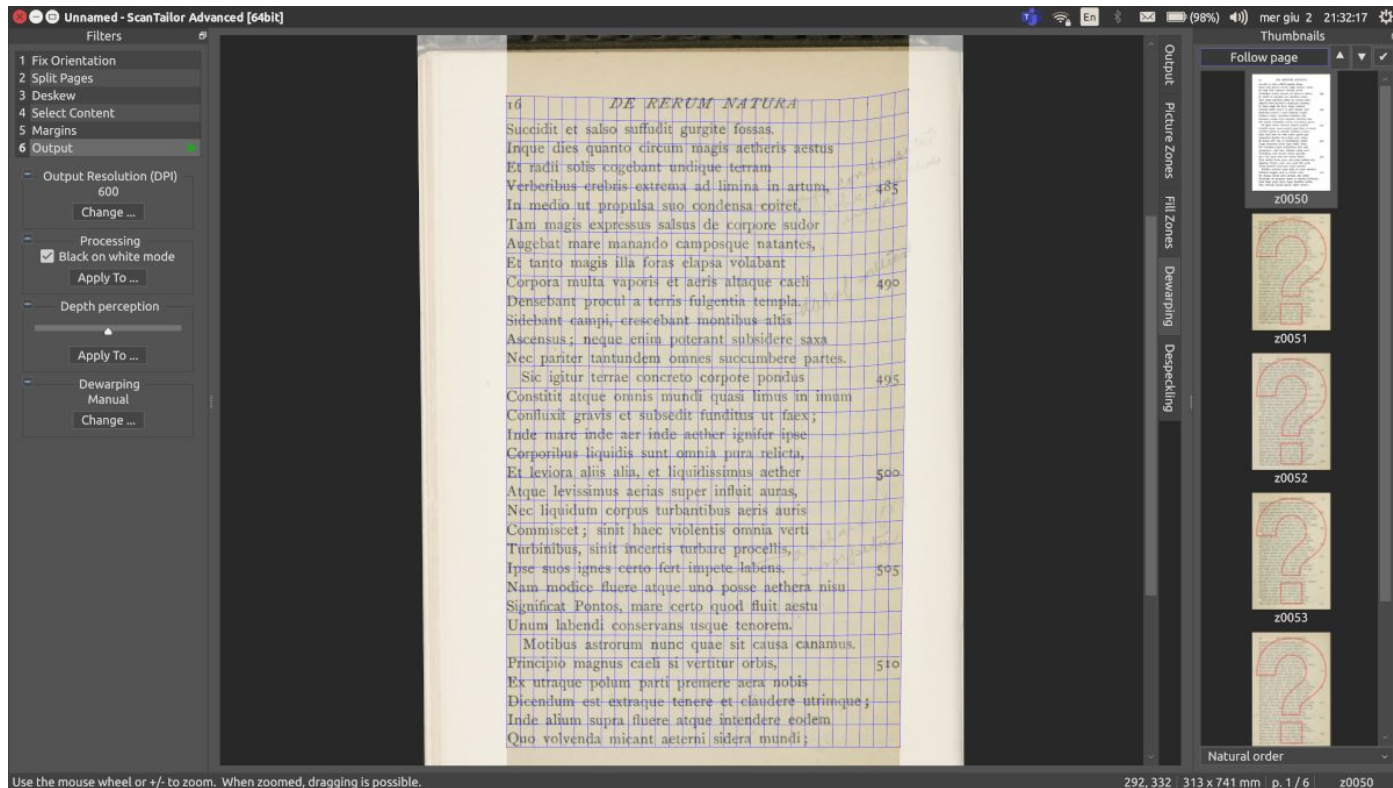
Follow page

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, these 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

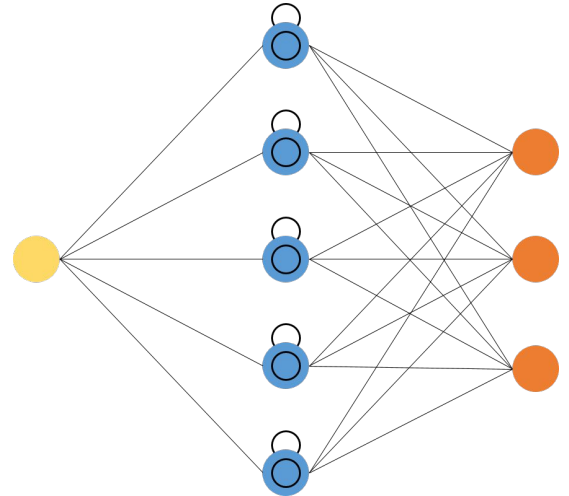
The main issues of these projects are:

- 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



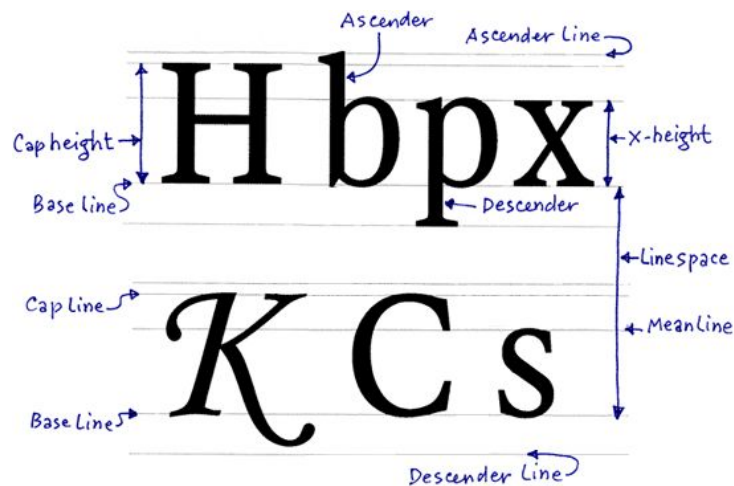
Layout analysis

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-

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 Abbyy 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". The main interface has a toolbar with buttons for "Open", "Save", "Reload", "Merge", "Split", "Insert", and "Delete". Below the toolbar are three tabs: "Box Coordinates", "Box Data", and "Box View". The "Box Coordinates" tab is active, displaying a table with columns "Char", "X", "Y", "Width", and "Hei...". The table contains 28 rows of data, with the last row partially cut off. To the right of the table is a preview window showing the text "gguangschoolsoisuthan..." with blue bounding boxes overlaid on each character. The preview window also has a "Character" field and a "Find" button. The status bar at the bottom indicates "Page: 1/1".

	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

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>)

hocr

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>
24 <div class='ocr_carea' id='block_1_2' title="bbox 402 1167 11496 18639">
25 <p class='ocr_par' id='par_1_2' lang='lat' title="bbox 485 1167 11496 6397">
26 <span class='ocr_line' id='line_1_2' title="bbox 525 1167 10060 1598; baseline -0.004 -88; x_size 407; x_descenders
27 85; x_ascenders 117">
28 <span class='ocrx_word' id='word_1_5' title='bbox 525 1182 3279 1598; x_wconf 89'>Quandoquidem</span>
29 <span class='ocrx_word' id='word_1_6' title='bbox 3515 1198 4706 1505; x_wconf 89'>claram</span>
30 <span class='ocrx_word' id='word_1_7' title='bbox 4944 1185 6378 1593; x_wconf 89'>speciem</span>
31 <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

Kraken on early printed editions

Λογ. ρημλ. ΕΙΣ ΤΟΝ ΤΙΜΙΟΝ ΚΑΙ ΖΩΟΠΟΙΟΝ ΣΤΑΥ-
ρον· καὶ εἰς τὸ, Σὺ εἶ ὁ ἐρχόμενος, ἢ ἕτερον προσδοκῶμεν; καὶ εἰς
τὸν τυφλὸν καὶ μογιάλων· καὶ εἰς τὸ ῥητὸν τοῦ προφήτου Ἀμβακούμ,
Κύριε, εἰσακήκου τὴν ἀκοήν σου, καὶ ἐφοβήθην.

ΚΑΙ ΑΛΩΣ ἡμῶν ἐς σφῶος ἡ πρωτοσημαμένη γλώσσα τῆ σταυροῦ ἴσα ἀκτι-
νας ἐπέδιδυξεν. ἐκέρυξε σταυρῶν, ἔφωτον μὲν τῆ λέξει, μέγα δὲ τῆ
ἐνεργείᾳ· ἐκέρυξε σταυρῶν, ἢ γὰρ ὡς ἀπίστοι, ἢ πρὸς αἰσθητικῶν
ἐκέρυξε σταυρῶν, οἱ ἐπίσημοι Ἰουδαῖοι, ἐπὶ τὸν κόσμον πρωτοσημαμένην
ἐπέδειξεν αὐτῶν ἢ ἀκτίνας. ἀκτίνας γὰρ τῶν σταυρῶν ἠσπάζονται ἀπὸ
ἡμεῶν ἡμῶν σταυρῶν ἢ ἡμῶν σταυρῶν οὐκ ἀπλῶς, ἀλλὰ ἐκ πρωτοσημαμένην
καὶ ἄλλοις ἐκφάνασαι. ἢ ὅπως, ἀκτίνας. πρωτοσημαμένην ὁ σταυρὸς ἢ κόσμος, ἐπὶ τῶν
φωσφορικῶν πᾶσι ἀνέδειξε. τῶν σταυρῶν αἱ ἀκτίνας εἰς τὸν κόσμον ἡμῶν κερύττεινται, ἀλλὰ
ἐπὶ τῶν ἔργων ἐρμηνεύονται. οὐ γὰρ εἰς τὸν σταυρὸν γλώσσῃς πειθούσης ἐπὶ λέξεως ἐρμη-
νεύουσας, ἀλλὰ ψυχῆς ἐργασιμότητος ἐπὶ δικαιοσύνης ἐπιτελούσης. ἀφ' οὗ σταυροῦ ἐπά-
γη, οὐ μῆσοι νηστεύουσιν· ἀφ' οὗ σταυροῦ ἐπάγη, οὐ πόνοι κήρυκες τῆς εὐσεβείας ἐπὶ τῆς σω-
φροσύνης ἐγένοντο· ἀφ' οὗ ὁ σταυρὸς, πάλαια διακρίσθη, ἢ ἀδικίαι κήρυκες τῆς εὐσεβείας
ἐδείχθησαν. ἐπίσημοι ἐπὶ τῶν σταυρῶν ἀπὸ τῶν ἀδελφῶν ἡμῶν κερύττεινται, ἢ τῶν ῥίζων.
ἢ ἀφ' οὗ σταυροῦ ἡμῶν σταυρῶν τεκοῦσα παρθένος, κτίκουσα οὐ νόμον φύσεως, ἀλλὰ
δυνάμει ἐκ ἐνεργείας τῆς φύσεως. μηδὲν ἀπαίτη ἐπὶ τῆς παρθένου ἢ ἀν-
δρα, ἐπὶ ἀπατηθῆσθαι ἐπὶ τοῦ Ἀδάμ τὴν γυναῖκα. ἐάν γὰρ λέγη, πῶς ἐγέννησεν ἡ παρ-
θένης ἀνευ ἀνδρός; ἐρώ σοι καγά, πῶς ἡ Εὐα προήλθεν ἐκ τοῦ Ἀδάμ ἀνευ γυναικός; καὶ
τί δεῖ σώματι σῶμα παραβάλλειν; τίς δ' οὗτο μαχόμενος Ἰουδαῖος; ἀισχυθέντο οἱ ἀναγι-
νώσκων, οὐδὲν ἐπιγινώσκουσι. ἀμφιβάλλει πῶς ἔτεκεν ἡ παρθένος; ἐρμηνεύσθαι πῶς ἔτεκε ἢ 25
τὸ φαινόμενον πῶς ἔτεκεν ἡ Ἰουδαῖα Χριστός. ἐρμηνεύει Παῦλος, ὅτι ἡ πάντα ἐκείνη
ἐπὶ τῶν ῥίζων, φασὶν, ἐκ τῶν ἀδικημάτων ἀποκαθάρσις πάντας, ἢ ἡ πᾶσα

898 ΧΡΥΣΟΣΤΟΜΟΥ Πανγκυρ. ἀμφιβάλω.
Λογ. ρημλ. ΕΙΣ ΤΟΝ ΤΙΜΙΟΝ ΚΑΙ ΖΩΟΠΟΙΟΝ ΣΤΑΥΡ-
ρον· καὶ εἰς τὸ, Σὺ εἶ ὁ ἐρχόμενος, ἢ ἕτερον προσδοκῶμεν; καὶ εἰς
τὸν τυφλὸν καὶ μογιάλων· καὶ εἰς τὸ ῥητὸν τοῦ προφήτου Ἀμβακούμ,
Κύριε, εἰσακήκου τὴν ἀκοήν σου, καὶ ἐφοβήθην.
Ἀλως ἡμῶν καὶ σφῶος ἡ πρωτοσημαμένη γλώσσα τοῦ σταυροῦ τὰς ἀκτι-
νας ἐπέδειξεν. ἐκέρυξε σταυρῶν, τὸν τυφλὸν μὲν τῆ λέξει, μέγα δὲ τῆ
ἐνεργείᾳ· ἐκέρυξε σταυρῶν, ὃν γὰρ ὡς ἀπίστοι, καὶ τρέμουσι δαίμονες·
ἐκέρυξε σταυρῶν, ὃν ἐπίηξαν οἱ Ἰουδαῖοι, καὶ ὁ κόσμος προσεκύνησεν· 10
ἐδείξεν αὐτοῦ καὶ ἀκτίνας. ἀκτίνας γὰρ τοῦ σταυροῦ ὑποτίθεται παρ-
θενίαν λαμβύσαν· καὶ λάμβυσαν οὐκ ἀπλῶς, ἀλλ' ἐκ προνοίας τοῦ
κάλου ἐκφάνασαν. καὶ ὅπως, ἀκούει. πορευόμενα παρέλαβεν ὁ σταυρὸς τὸν κόσμον, καὶ σω-
φρονούντα πᾶσιν ἀνέδειξε. τοῦ σταυροῦ αἱ ἀκτίνας οὐκ ἀπὸ τῶν λόγων κερύττεινται, ἀλλὰ
διὰ τῶν ἔργων ἐρμηνεύονται. οὐ χρεια ἔχει σταυρὸς γλώσσῃς πειθούσης καὶ λέξεως ἐρμη- 15
νεύουσας, ἀλλὰ ψυχῆς εὐγνωμονούσης καὶ ἔργα δικαιοσύνης ἐπιτελούσης. ἀφ' οὗ σταυροῦ ἐπά-
γη, οἱ μῆσοι νηστεύουσιν· ἀφ' οὗ σταυροῦ ἐπάγη, οἱ πόνοι κήρυκες τῆς εὐσεβείας καὶ τῆς σω-
φροσύνης ἐγένοντο· ἀφ' οὗ σταυροῦ, τελῶνα εὐαγγελιστοί, καὶ δίκται κήρυκες τῆς εὐσεβείας
ἐδείχθησαν. ἐπίσημοι δὲ καὶ ὁ προσητευσάμενος λόγος τὴν μνημῆν τοῦ σταυροῦ, καὶ τὴν ῥίζαν.
ῥίζα δὲ σταυροῦ ἡ παρθένια, ἡ τὸν παθόντα τεκοῦσα παρθένος, κτίκουσα οὐ νόμον φύσεως, ἀλλὰ 20
δυνάμει καὶ ἐνεργείᾳ τοῦ τεχνίτου τῆς φύσεως. μὴ οὐκ ἀπαίτη ἐπὶ τῆς παρθένου τὸν ἀν-
δρα, ἐπὶ ἀπατηθῆσθαι ἐπὶ τοῦ Ἀδάμ τὴν γυναῖκα. ἐάν γὰρ λέγη, πῶς ἐγέννησεν ἡ παρ-
θένης ἀνευ ἀνδρός; ἐρώ σοι καγά, πῶς ἡ Εὐα προήλθεν ἐκ τοῦ Ἀδάμ ἀνευ γυναικός; καὶ
τί δεῖ σώματι σῶμα παραβάλλειν; τίς δ' οὗτο μαχόμενος Ἰουδαῖος; ἀισχυθέντο οἱ ἀναγι-
νώσκων, οὐδὲν ἐπιγινώσκουσι. ἀμφιβάλλει πῶς ἔτεκεν ἡ παρθένος; ἐρμηνεύσθαι πῶς ἔτεκε ἢ 25
πέτρα τοῦ ὕδατος, οὐκ ἔχουσα ὑποκειμένην φλέβα, οὐ ποταμὸν ὑπεστοροεμένον, οὐ ῥίζαν ὑ-

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

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

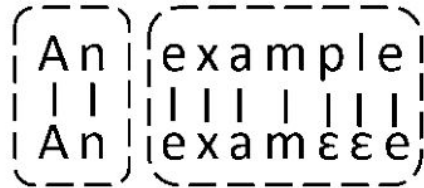
Alignment

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

GT: An example

OCR: An exam e

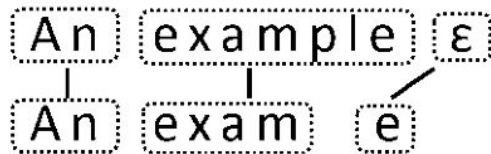
Char alignment first:



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

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
Carica 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
FestaOCR 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 (sta incluso):

```
<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 asiaci 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)



PRIMO CANTO INTORNO ALL'ARA

I vegliardi, compiendo intorno all'ara lente evoluzioni ritmiche, cantano, alternandosi i due senicori, le seguenti coppie strofiche.

CORO

Strofe 1

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,
poi che un giogo, un ponte tutto irto di cunei
del mar sopra la cervice ebbe confitto.

Antistrofe 1





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]





	SAL 25%	<i>predefinito</i>
	SAL 50%	La pagina è problematica
	SAL 75%	La pagina è stata trascritta e formattata
	SAL 100%	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 {{fwpag}})

	SAL 00%	La pagina non necessita di trascrizione
---	----------------	---

Nel namespace indice [modifica]

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

ἑταίῳ. διὰ καὶ ἐν τῇ ὁμίλῳ ἡ σύνταξις γίνεται τῶν σκε-
λῶν· τὸ τε γὰρ ἔργον νευρώδης καὶ ἡ φύσις τῶν σκελῶν
νευρώδης. ὥστ' ἐπεὶ τὸτ' ἐκ ἐδέχεται ἔχει, ἀνάγκη καὶ
ἔρχει ἢ μὴ ἔχει ἢ μὴ ἑταίῳ ἔρχει· τοὺς γὰρ ἔρχουσι ἢ
αὐτῇ ὅτις ἀμφοτέρων αὐτῶν. ἐπεὶ δὲ τοῖς γε τοῖς ἔρχει ἔχουσι
ἔξω διὰ τῆς κινήσεως θερμοκρασίαν τῶν αἰσῶν προέρχεται τὸ
σπέρμα συναβροσθῆναι, ἀλλ' ἔχει ὡς ἔτιμον ἐν πύλῳ θηῶν,
ὥσπερ τοῖς ἰχθύσι. πάντα δ' ἔχει τὰ ζῴοντα τοῖς ἔρχει
ἐν τῷ πρῶτον ἢ ἔξω, πλὴν ἰχθύε· ἔτι δὲ πρὸς τῇ ἰσοφεί-
μῳ, διὰ τὴν αἰτῆν αἰτίαν δι' ἧπερ καὶ οἱ ἄνθρωποι ταχύν-
ονται γὰρ ἀναγκαῖον γίνεσθαι τὸν συνδυασμὸν αὐτῶν· οὐ γὰρ
ὥσπερ τὰ τετραπόδα ἐπὶ τὰ πρῶτῃ ἐπιβαίνει, ἀλλ' ὀρθοὶ
μύνυται διὰ τὸς ἀνάσθας. δὲ ἦν μὲν οὖν αἰτίαν ἔχουσι τὰ
ἔχοντα ἔρχει, εἴρηται, καὶ δὲ ἦν αἰτίαν τὰ μὲν ἔξω τὰ
δ' ἐντός. ὅσα 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 ἔχουσαν·
τοιούτῃ δ' ἔστιν ἡ τῶν ἰχθύων φύσις καὶ ἡ τῶν ἔρχει. εἰ
μὲν γὰρ ἰχθύε ἔρχονται παραπίπτουτες καὶ ἀπολύονται
ταχέως, ὥσπερ γὰρ ἐπὶ τῶν ἀνθρώπων καὶ πάντων τῶν
τούτων ἀνάγκη κατασχέσθαι τὸ πνεῦμα πρῶτον τὴν γο-
νὴν· τοῦτο δ' ἐκείναις συμβαίνει μὴ δεχόμεναι τὴν θαλάσ-
σαν, εἰσὶ δὲ εὐφρόντοι τοῦτο μὴ ποιούτες. ἕκον δὲ ἐν τῷ
συνδυασμῷ τὸ σπέρμα πέττει αὐτίς, ὥσπερ τὰ πτεῖλα καὶ
ζῴοντα, ἀλλ' ὑπὸ τῆς ὥρας τὸ σπέρμα πεπεμαμένον ἀ-
βῶν ἔχουσι, ὥστε μὴ ἐν τῷ ὄργασμῳ ἀλλήλων πεύει.

A —

ἑταίῳ. διὰ καὶ ἐν τῇ ὁμίλῳ ἡ σύνταξις γίνεται τῶν σκε-
λῶν· τὸ τε γὰρ ἔργον νευρώδης καὶ ἡ φύσις τῶν σκελῶν
νευρώδης. ὥστ' ἐπεὶ τὸτ' ἐκ ἐδέχεται ἔχει, ἀνάγκη καὶ
ἔρχει ἢ μὴ ἔχει ἢ μὴ ἑταίῳ ἔχει· τοὺς γὰρ ἔχουσι ἢ
αὐτῇ ὅτις ἀμφοτέρων αὐτῶν. ἐπεὶ δὲ τοῖς γε τοῖς ἔρχει ἔχουσι ἢ
ἔξω διὰ τῆς κινήσεως θερμοκρασίαν τῶν αἰσῶν προέρχεται τὸ
σπέρμα συναβροσθῆναι, ἀλλ' οὐκ ὡς ἔτιμον ἐν πύλῳ θηῶν,
ὥσπερ τοῖς ἰχθύσι. πάντα δ' ἔχει τὰ ζῴοντα τοῖς ἔρχει
ἐν τῷ πρῶτον ἢ ἔξω, πλὴν ἰχθύε· ἔτι δὲ πρὸς τῇ ἰσοφεί-
μῳ, διὰ τὴν αἰτῆν αἰτίαν δι' ἧπερ καὶ οἱ ἄνθρωποι ταχύν-
ονται γὰρ ἀναγκαῖον γίνεσθαι τὸν συνδυασμὸν αὐτῶν· οὐ γὰρ
ὥσπερ τὰ τετραπόδα ἐπὶ τὰ πρῶτῃ ἐπιβαίνει, ἀλλ' ὀρθοὶ
μύνυται διὰ τὸς ἀνάσθας. δὲ ἦν μὲν οὖν αἰτίαν ἔχουσι τὰ
ἔχοντα ἔρχει, εἴρηται, καὶ δὲ ἦν αἰτίαν τὰ μὲν ἔξω τὰ
δ' ἐντός. ὅσα 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 ἔχουσαν·
τοιούτῃ δ' ἔστιν ἡ τῶν ἰχθύων φύσις καὶ ἡ τῶν ἔρχει. οἱ
μὲν γὰρ ἰχθύε ὀρθοὶ παραπίπτουτες καὶ ἀπολύονται
ταχέως, ὥσπερ γὰρ ἐπὶ τῶν ἀνθρώπων καὶ πάντων τῶν
τοιούτων ἀνάγκη κατασχέσθαι τὸ πνεῦμα πρῶτον τὴν γο-
νὴν· τοῦτο δ' ἐκείναις συμβαίνει μὴ δεχόμεναι τὴν θαλάσ-
σαν, εἰσὶ δὲ εὐφρόντοι τοῦτο μὴ ποιούτες. ὅσων δὲ ἐν τῷ

Conclusion

OCR and 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