Correzione collaborativa di risorse digitali acquisite tramite OCR

Federico Boschetti*

federico.boschetti@yahoo.com

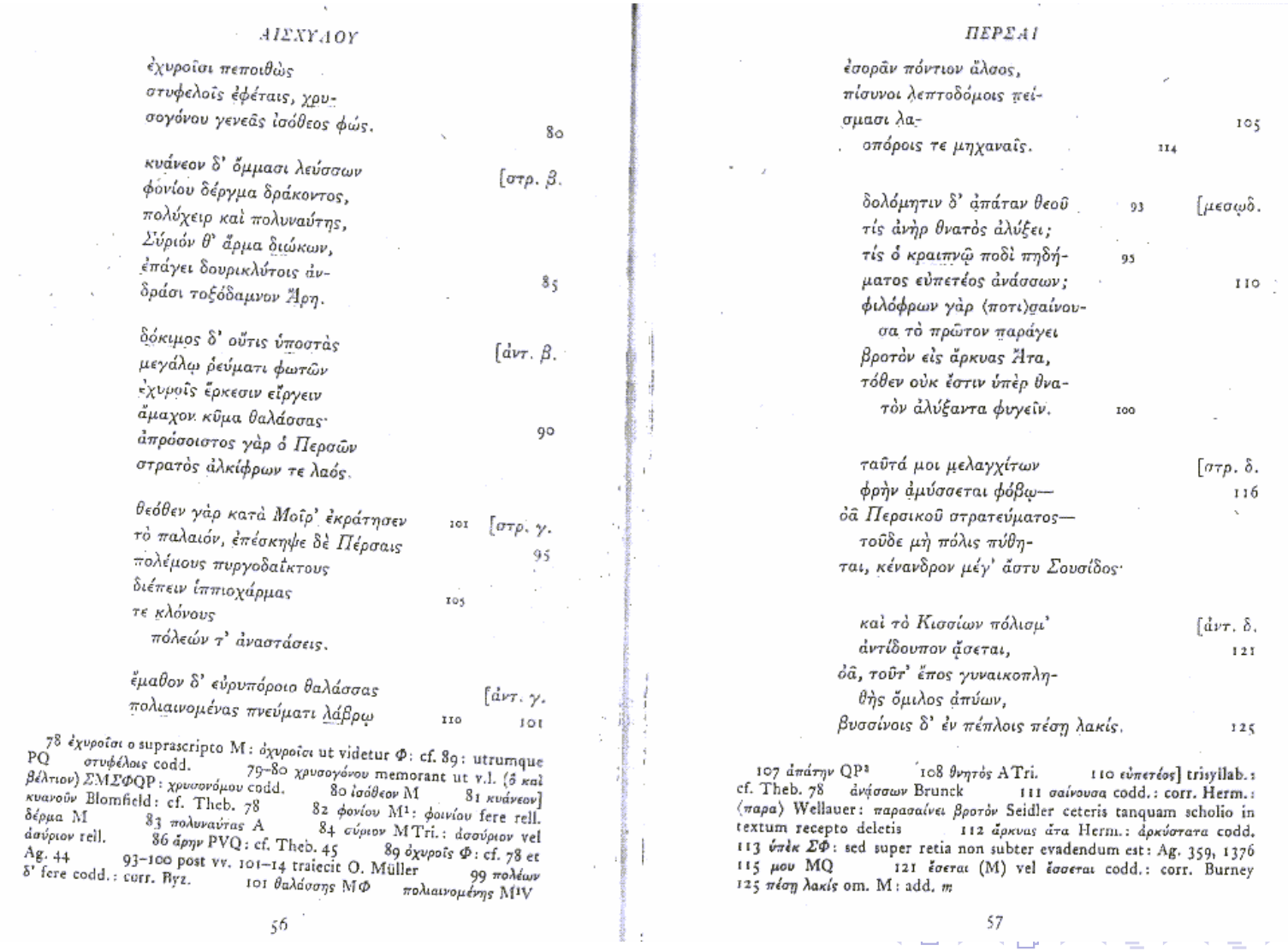
* *CNR-ILC*

Pisa, 13 Giugno 2017

Introduzione

Il miglioramento delle prestazioni dell'OCR ha permesso di spostare l'attenzione dai soli testi digitali alle edizioni digitali. Infatti nella creazione delle prime collezioni digitali, prefazioni, introduzioni, indici, bibliografie, note, apparati critici e varianti testuali presenti in edizioni differenti abitualmente non venivano presi in considerazione.

Edizione critica



ΑΙΣΧΥΛΟΥ

ἐχυροῖσι πεποιθὼς
στυφελοῖς ἐφέταις, χρυ-
σογόνου γενεᾶς ἰσόθεος φώς.

κυάνεον δ' ὄμμασι λεύσσων [στρ. β.
φόνιον δέργμα δράκοντος,
πολύχειρ καὶ πολυναύτης,
Σύριόν θ' ἅρμα διώκων,
ἐπάγει δουρικλύτοις ἀν-
δράσι τοξόδαμνον Ἄρη.

δόκιμος δ' οὔτις ὑποστὰς [ἀντ. β.
μεγάλῳ ῥεύματι φωτῶν
ἐχυροῖς ἔρκεσιν εἴργειν
ἄμαχον κῦμα θαλάσσας·
ἀπρόσοιστος γὰρ ὁ Περσῶν
στρατὸς ἀλκίφρων τε λαός.

θεόθεν γὰρ κατὰ Μοῖρ' ἐκράτησεν [στρ. γ.
τὸ παλαιόν, ἐπέσκηψε δὲ Πέρσαις
πολέμους πυργοδαίκτους
διέπειν ἱππιοχάρμας
τε κλόνους
πόλεών τ' ἀναστάσεις.

ἔμαθον δ' εὐρυπόροιο θαλάσσας [ἀντ. γ.
πολιαινομένας πνεύματι λάβρῳ

78 ἐχυροῖσι o suprascripto M: ὀχυροῖσι ut videtur Φ: cf. 89: utrumque PQ στυφέλοις codd. 79–80 χρυσογόνου memorant ut v.l. (ὃ καὶ βέλτιον) ΣΜΣΦQP: χρυσονόμου codd. 80 ἰσόθεον M 81 κυάνεον] κυανοῦν Blomfield: cf. Theb. 78 82 φονίου M¹: φοινίου fere rell. δέρμα M 83 πολυναύτας A 84 σύριον MTri.: ἀσύριον vel ἀσύριον rell. 86 ἄρην PVQ: cf. Theb. 45 89 ὀχυροῖς Φ: cf. 78 et Ag. 44 93–100 post vv. 101–14 traiecit O. Müller 99 πολέων δ' fere codd.: corr. Byz. 101 θαλάσσης MΦ πολιαινομένης M¹V

56

ΠΕΡΣΑΙ

ἐσορᾶν πόντιον ἅλσος,
πίσυνοι λεπτοδόμοις πεί-
σμασι λα-
οπόροις τε μηχαναῖς.

δολόμητιν δ' ἀπάταν θεοῦ [μεσῳδ.
τίς ἀνὴρ θνατὸς ἀλύξει;
τίς ὁ κραιπνῷ ποδὶ πηδή-
ματος εὐπετέος ἀνάσσων;
φιλόφρων γὰρ (ποτι)σαίνου-
σα τὸ πρῶτον παράγει
βροτὸν εἰς ἄρκυας Ἄτα,
τόθεν οὐκ ἔστιν ὑπὲρ θνα-
τὸν ἀλύξαντα φυγεῖν.

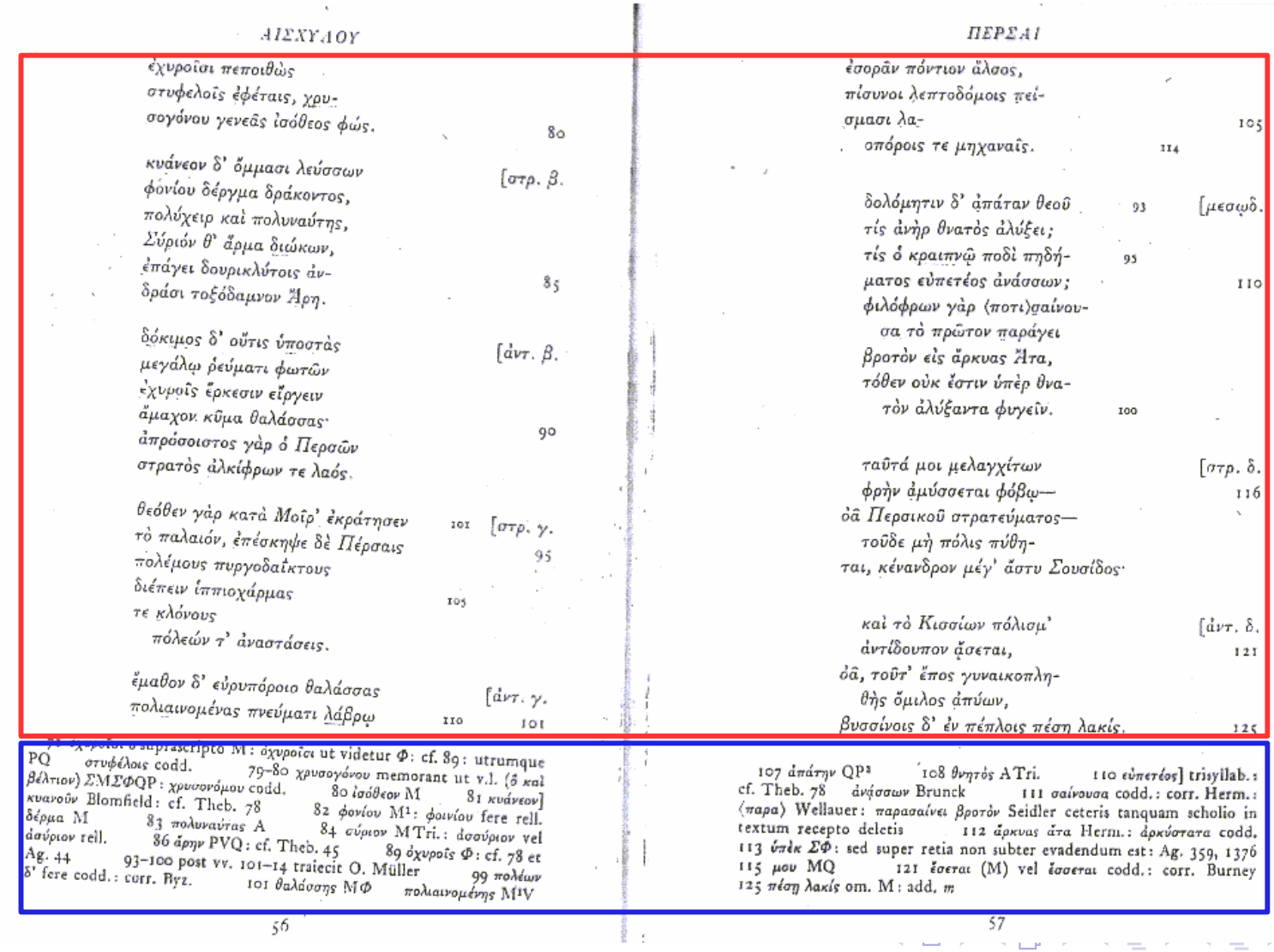
ταῦτά μοι μελαγχίτων [στρ. δ.
φρὴν ἀμύσσεται φόβῳ—
ὰ Περσικοῦ στρατεύματος—
τοῦδε μὴ πόλις πύθη-
ται, κένανδρον μέγ' ἄστυ Σουσίδος·

καὶ τὸ Κισσίων πόλισμ' [ἀντ. δ.
ἀντίδουπον ἄσεται,
ὰᾶ, τοῦτ' ἔπος γυναικοπλη-
θὴς ὅμιλος ἀπύων,
βυσσίνοις δ' ἐν πέπλοις πέσῃ λακίς.

107 ἀπάτην QP² 108 θνητὸς ATri. 110 εὐπετέος] trisyllab.: cf. Theb. 78 ἀνάσσων Brunck 111 σαίνουσα codd.: corr. Herm.: ⟨παρα⟩ Wellauer: παρασαίνει βροτὸν Seidler ceteris tanquam scholio in textum recepto deletis 112 ἄρκυας ἄτα Herm.: ἀρκύστατα codd. 113 ὑπὲκ ΣΦ: sed super retia non subter evadendum est: Ag. 359, 1376 115 μου MQ 121 ἔσεται (M) vel ἔσσεται codd.: corr. Burney 125 πέσῃ λακίς om. M: add. m

57

Edizione critica



ΑΙΣΧΥΛΟΥ

ἐχυροῖσι πεποιθὼς
στυφελοῖς ἐφέταις, χρυ-
σογόνου γενεᾶς ἰσόθεος φώς.

κυάνεον δ' ὄμμασι λεύσσων [στρ. β.
φόνιον δέργμα δράκοντος,
πολύχειρ καὶ πολυναύτης,
Σύριόν θ' ἅρμα διώκων,
ἐπάγει δουρικλύτοις ἀν-
δράσι τοξόδαμνον Ἄρη.

δόκιμος δ' οὔτις ὑποστὰς [ἀντ. β.
μεγάλῳ ῥεύματι φωτῶν
ἐχυροῖς ἔρκεσιν εἴργειν
ἄμαχον κῦμα θαλάσσας·
ἀπρόσοιστος γὰρ ὁ Περσῶν
στρατὸς ἀλκίφρων τε λαός.

θεόθεν γὰρ κατὰ Μοῖρ' ἐκράτησεν [στρ. γ.
τὸ παλαιόν, ἐπέσκηψε δὲ Πέρσαις
πολέμους πυργοδαΐκτους
διέπειν ἱππιοχάρμας
τε κλόνους
πόλεών τ' ἀναστάσεις.

ἔμαθον δ' εὐρυπόροιο θαλάσσας [ἀντ. γ.
πολιαινομένας πνεύματι λάβρῳ

PQ στυφέλοις codd. 79–80 χρυσογόνου memorant ut v.l. {ὃ καὶ βέλτιον} ΣΜΣΦQP: χρυσονόμου codd. 80 ἰσόθεον M 81 κυάνεον] κυανοῦν Blomfield: cf. Theb. 78 82 φονίον M¹: φοινίου fere rell. δέρμα M 83 πολυναύτας A 84 σύριον MTri.: ἀσύριον vel ἀσύριον rell. 86 ἄρην PVQ: cf. Theb. 45 89 ὀχυροῖς Φ: cf. 78 et Ag. 44 93–100 post vv. 101–14 traiecit O. Müller 99 πολέων δ' fere codd.: corr. Byz. 101 θαλάσσης MΦ πολιαινομένης M¹V

56

ΠΕΡΣΑΙ

ἐσορᾶν πόντιον ἅλσος,
πίσυνοι λεπτοδόμοις πεί-
σμασι λα-
οπόροις τε μηχαναῖς.

δολόμητιν δ' ἀπάταν θεοῦ [μεσῳδ.
τίς ἀνὴρ θνατὸς ἀλύξει;
τίς ὁ κραιπνῷ ποδὶ πηδή-
ματος εὐπετέος ἀνάσσων;
φιλόφρων γὰρ (ποτι)σαίνου-
σα τὸ πρῶτον παράγει
βροτὸν εἰς ἄρκυας Ἄτα,
τόθεν οὐκ ἔστιν ὑπὲρ θνα-
τὸν ἀλύξαντα φυγεῖν.

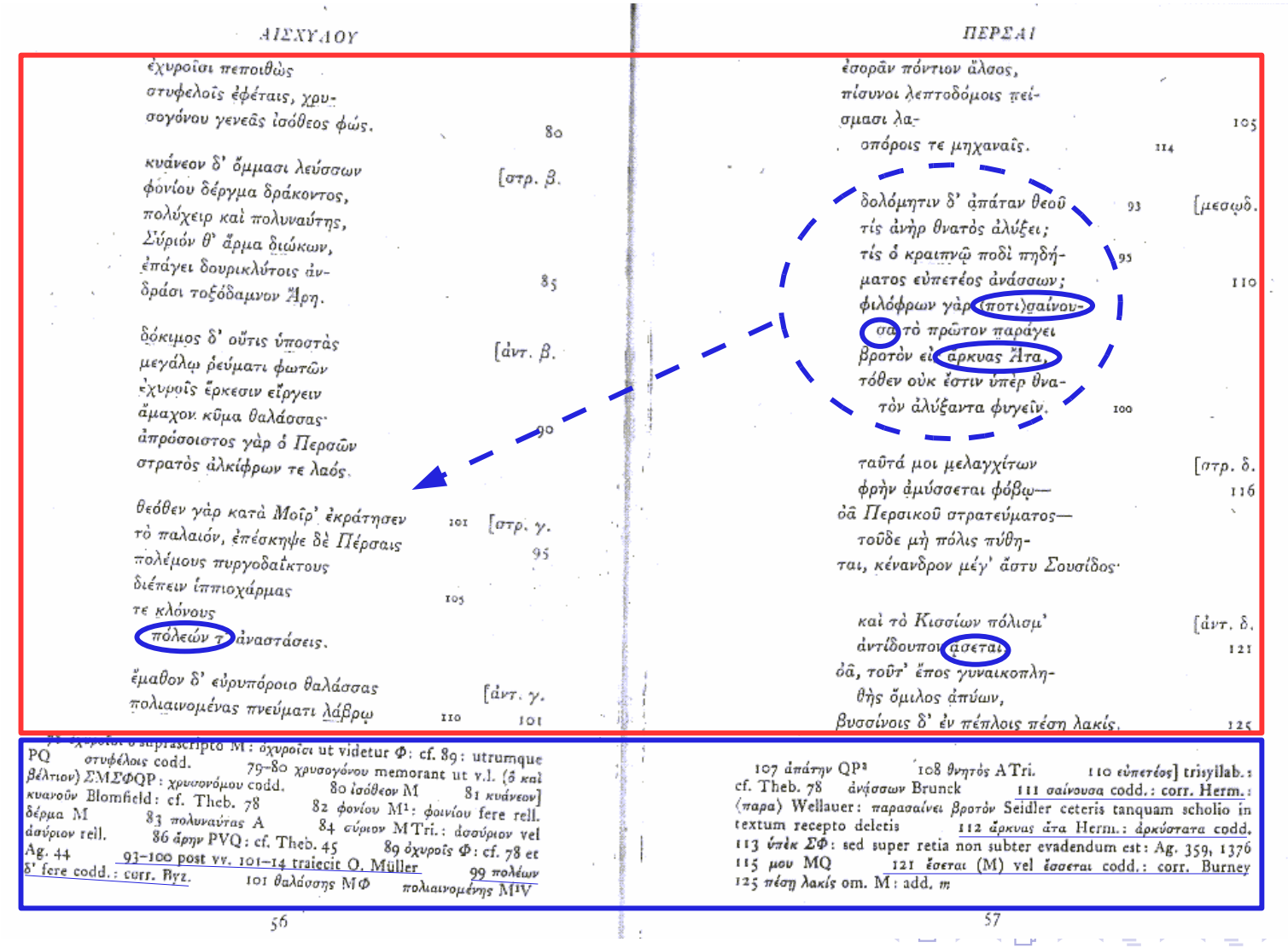
ταῦτά μοι μελαγχίτων [στρ. δ.
φρὴν ἀμύσσεται φόβῳ—
ὀᾶ Περσικοῦ στρατεύματος—
τοῦδε μὴ πόλις πύθη-
ται, κένανδρον μέγ' ἄστυ Σουσίδος·

καὶ τὸ Κισσίων πόλισμ' [ἀντ. δ.
ἀντίδουπον ἄσεται,
ὀᾶ, τοῦτ' ἔπος γυναικοπλη-
θὴς ὅμιλος ἀπύων,
βυσσίνοις δ' ἐν πέπλοις πέση λακίς.

107 ἀπάτην QP³ 108 θνητὸς ATri. 110 εὐπετέος] trisyllab.: cf. Theb. 78 ἀνάσσων Brunck 111 σαίνουσα codd.: corr. Herm.: ⟨παρα⟩ Wellauer: παρασαίνει βροτὸν Seidler ceteris tanquam scholio in textum recepto deletis 112 ἄρκυας ἄτα Herm.: ἀρκύστατα codd. 113 ὑπὲκ ΣΦ: sed super retia non subter evadendum est: Ag. 359, 1376 115 μου MQ 121 ἔσεται (M) vel ἔσσεται codd.: corr. Burney 125 πέσῃ λακίς om. M: add. m

57

Edizione critica

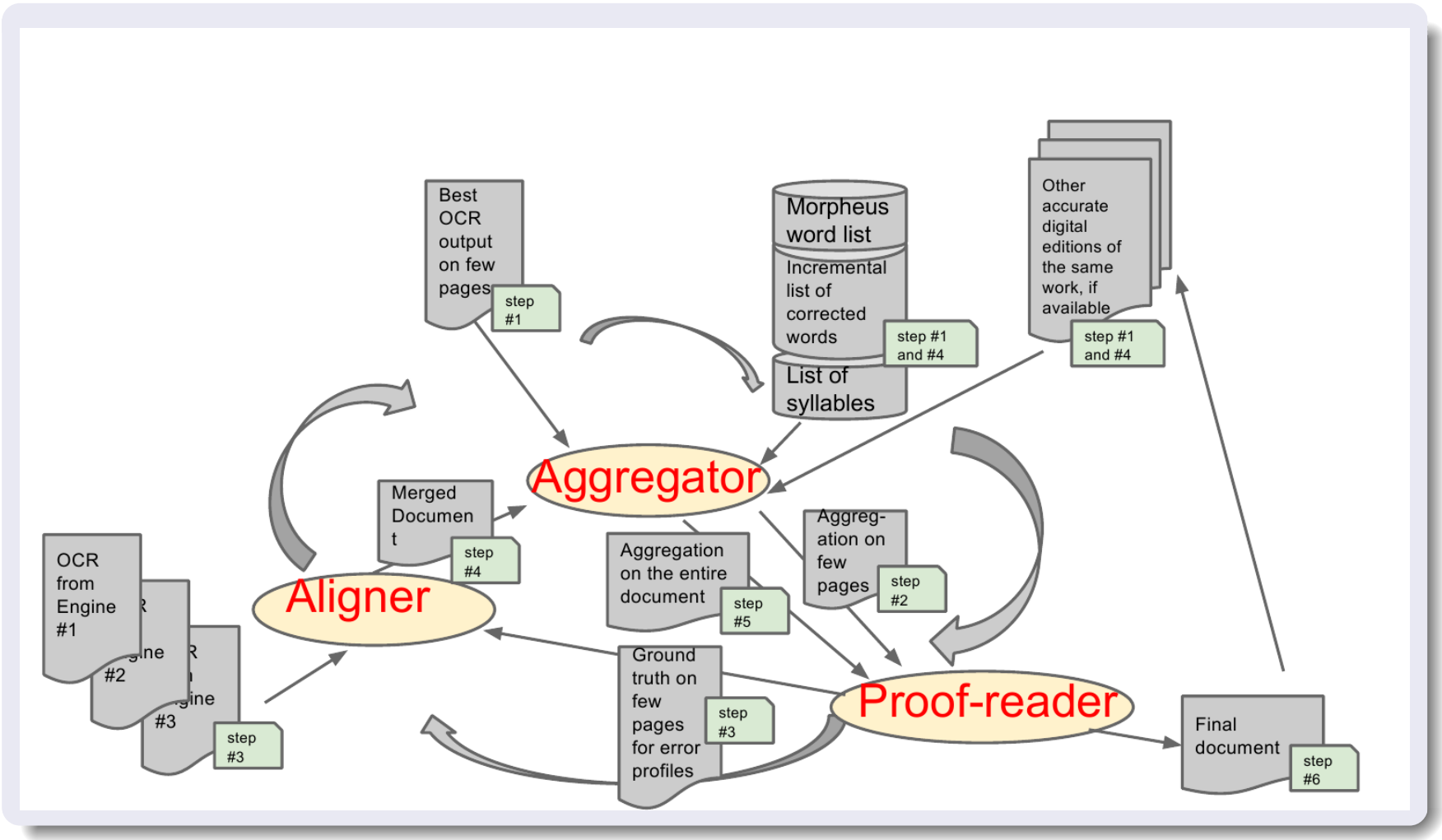


Peculiarità dell'OCR storico

Peculiarità dell'OCR applicato a documenti di interesse storico

- Qualità variabile della carta e degli inchiostri (materie prime usate)
- Qualità variabile dell'impressione a stampa (tecnologia impiegata)
- Stato di conservazione (fattore temporale)
- Varietà linguistiche assenti nei dizionari di riferimento (variabili diacroniche, diatopiche, diastratiche, diafasiche, diamesiche)

OCR Work-flow



Overview

I repertori di immagini online

https://archive.org

"Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more."

https://www.hathitrust.org

"HathiTrust is a partnership of academic and research institutions, offering a collection of millions of titles digitized from libraries around the world."

Altre risorse

- https://www.bsb-muenchen.de
- https://books.google.com
- ...

Edizioni ed esemplari

- Online non solo è possibile trovare molteplici edizioni della stessa opera, ma anche molteplici esemplari della stessa edizione
- L'applicazione dell'OCR a più esemplari della stessa edizione e l'allineamento dei risultati possono migliorare l'accuratezza della digitalizzazione
- È buona norma catalogare gli esemplari disponibili online, possibilmente secondo il modello denominato *Functional Requirements for Bibliographic Records* (FRBR) raccomandato dall'*International Federation of Library Associations* (IFLA)

La scansione

In assenza di immagini di alta qualità disponibili online, è necessario procedere alla scansione delle proprie immagini. Se i materiali non sono delicati (edizioni antiche, preziose, facili da rovinare), un'economico scanner piatto può essere sufficiente

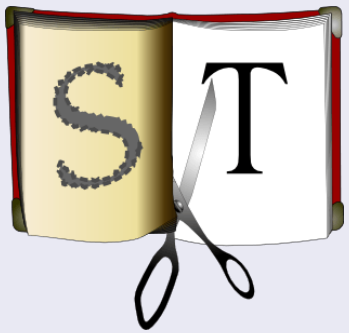
DPI

- La risoluzione delle immagini acquisite dipende dal miglior compromesso fra qualità e costo (costo per gli strumenti di acquisizione e per lo *storage*
- Attualmente, 600 punti per pollice (*dots per inch*: DPI) sono considerati sufficienti per ottenere risultati ottimali con l'OCR
- È buona norma salvare i *master files* con le immagini originali al massimo della risoluzione (anche molto superiore a 600 DPI) e applicare le necessarie trasformazioni a copie delle immagini, eventualmente a risoluzione inferiore

ScanTailor

http://scantailor.org

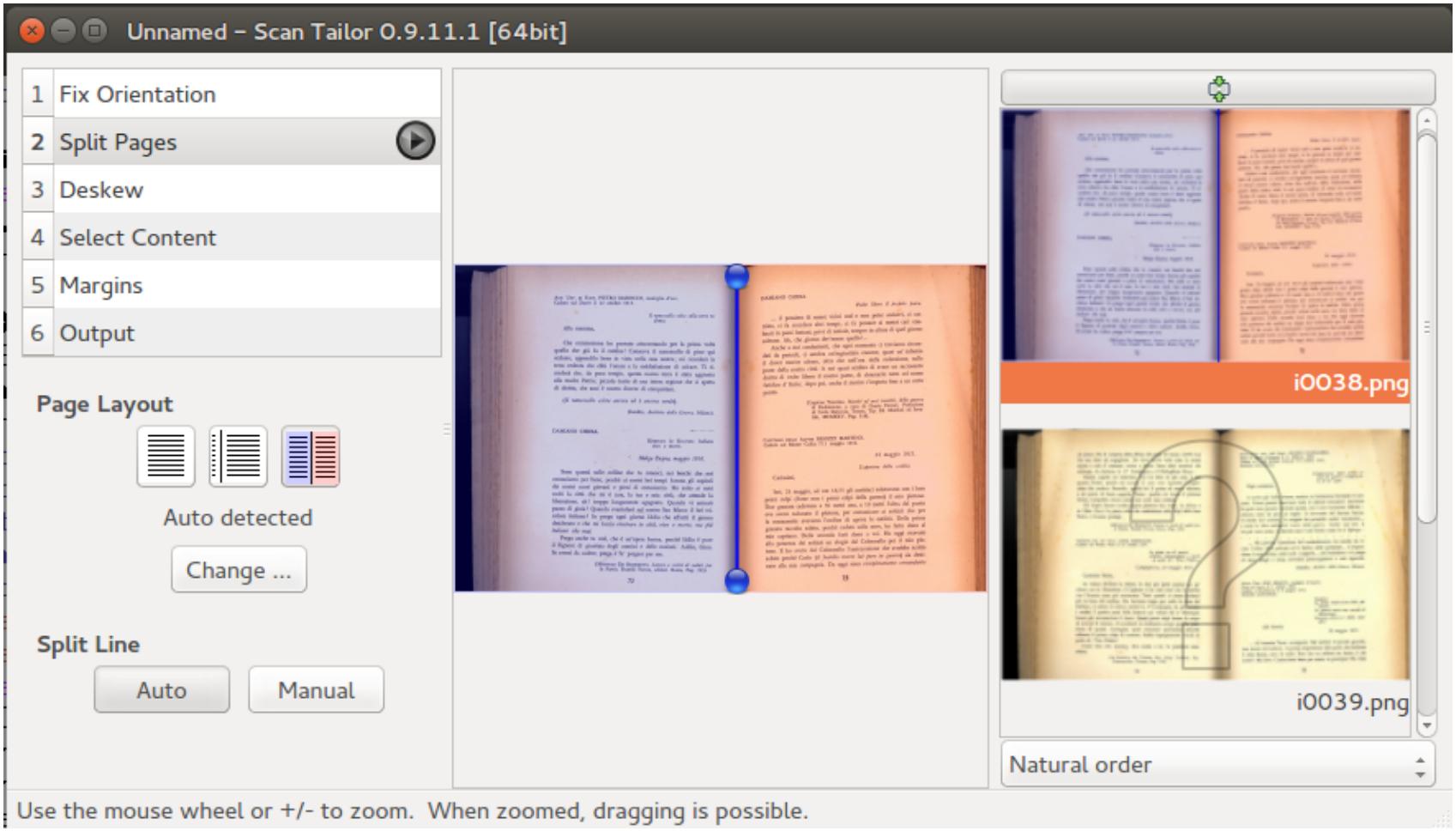
"Scan Tailor is an interactive [...] tool for scanned pages. It performs operations such as page splitting, deskewing, adding/removing borders, and others. You give it raw scans, and you get pages ready to be printed or assembled into a PDF or DJVU file. Scanning, optical character recognition, and assembling multi-page documents are out of scope of this project."



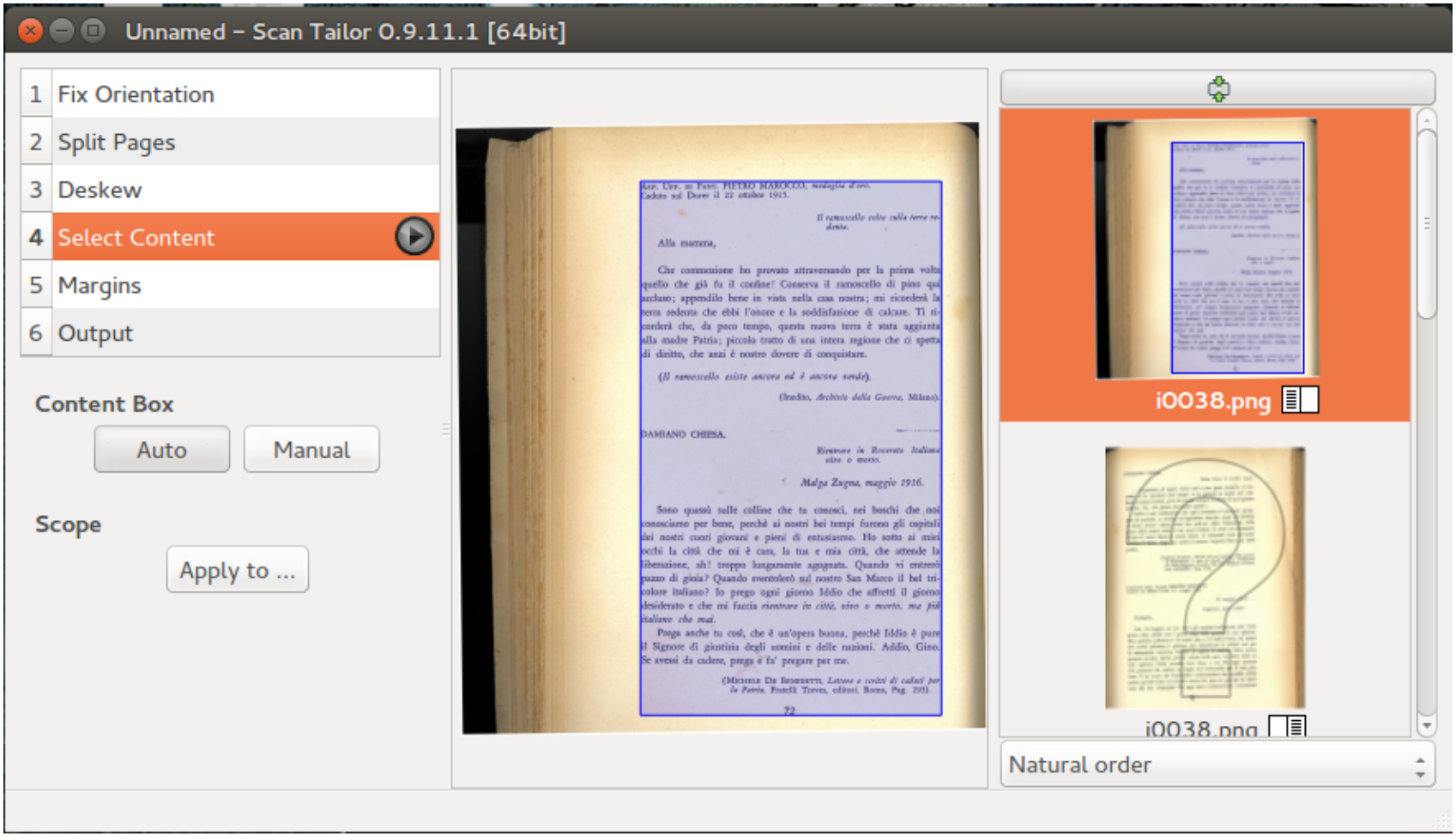
ScanTailor GUI & CLI

ScanTailor può essere usato sia interattivamente, attraverso l'interfaccia grafica, sia tramite linea di comando, per progetti massivi

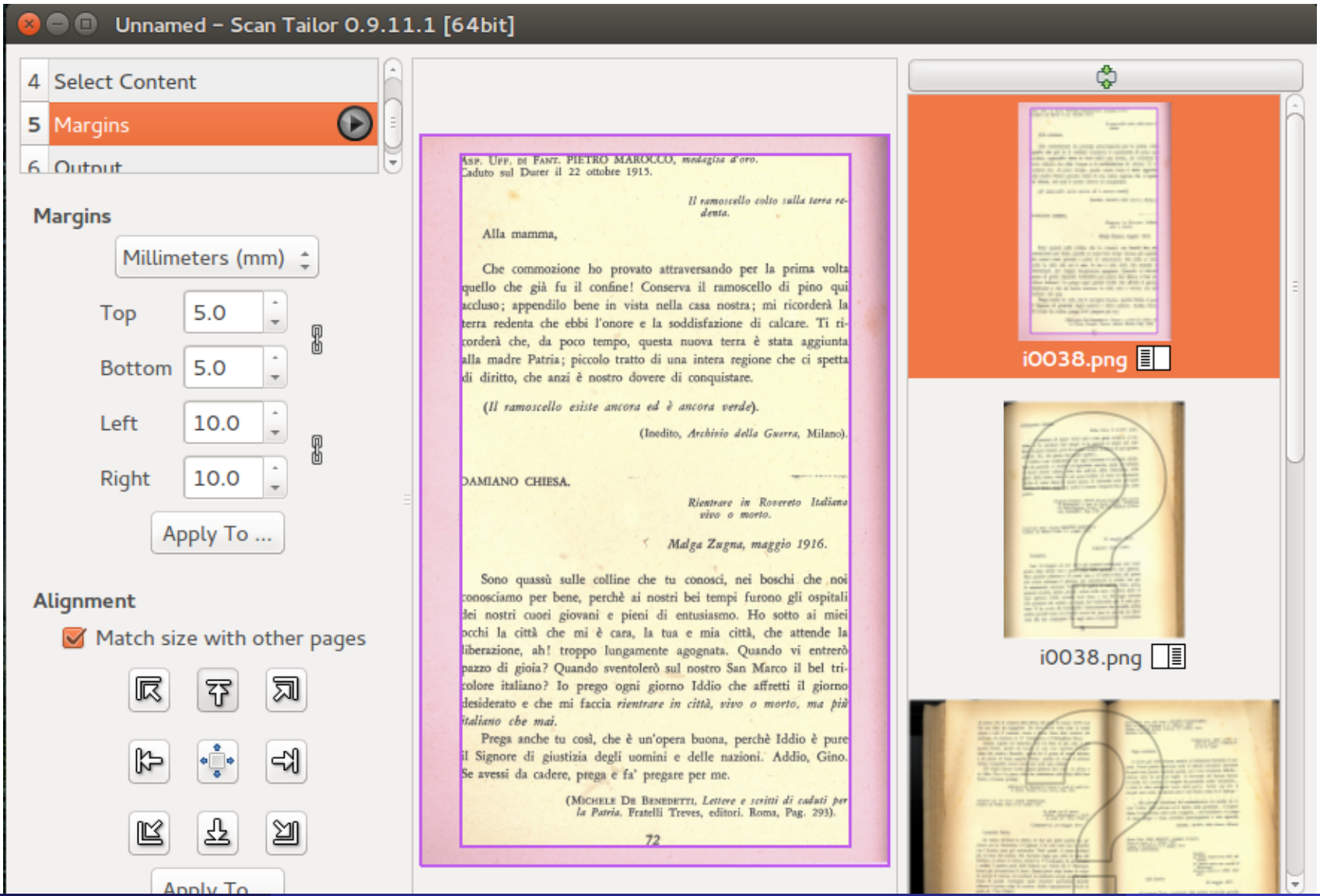
Page Splitting, Content Selection and Margins



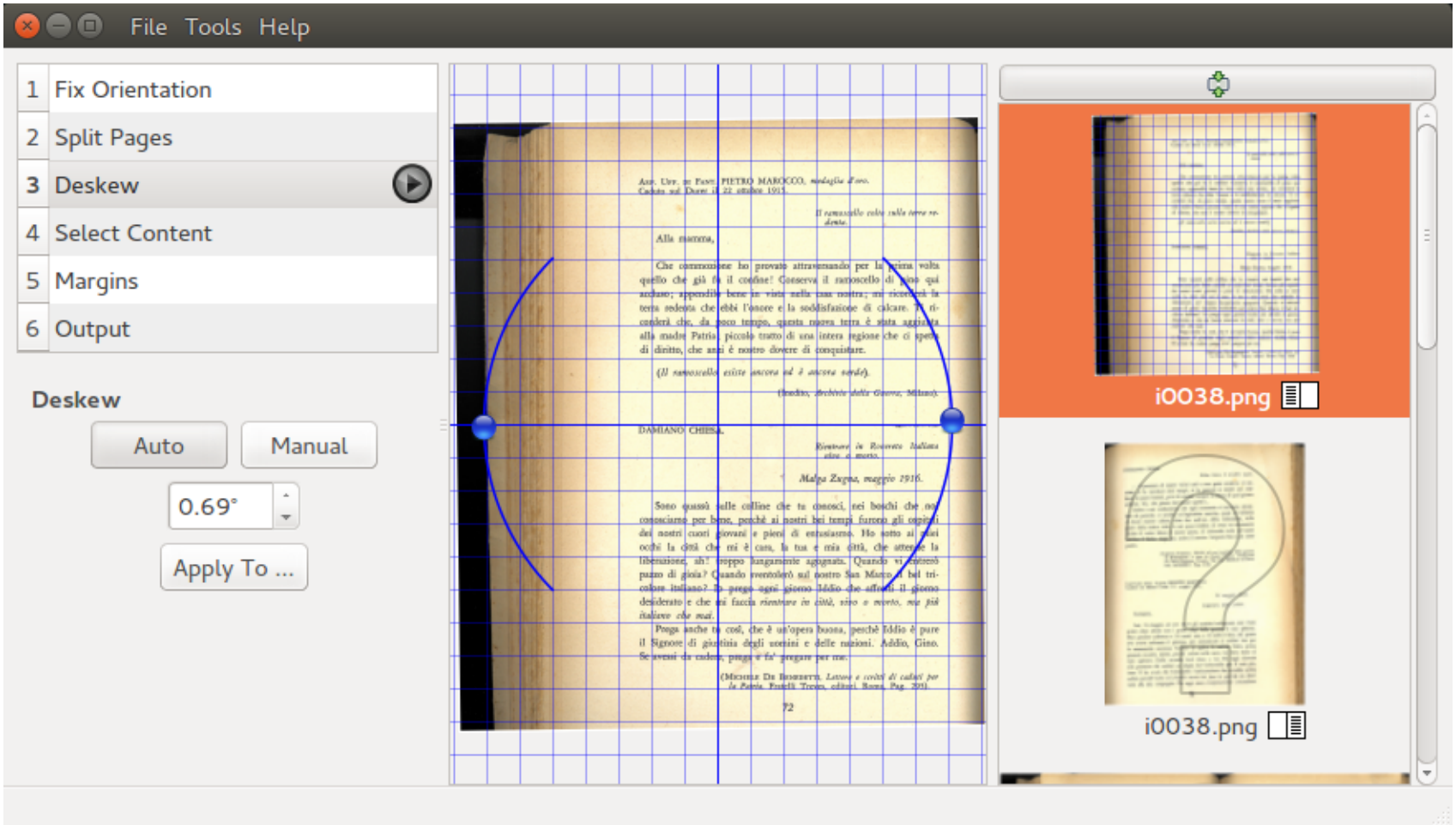
Page Splitting, Content Selection and Margins



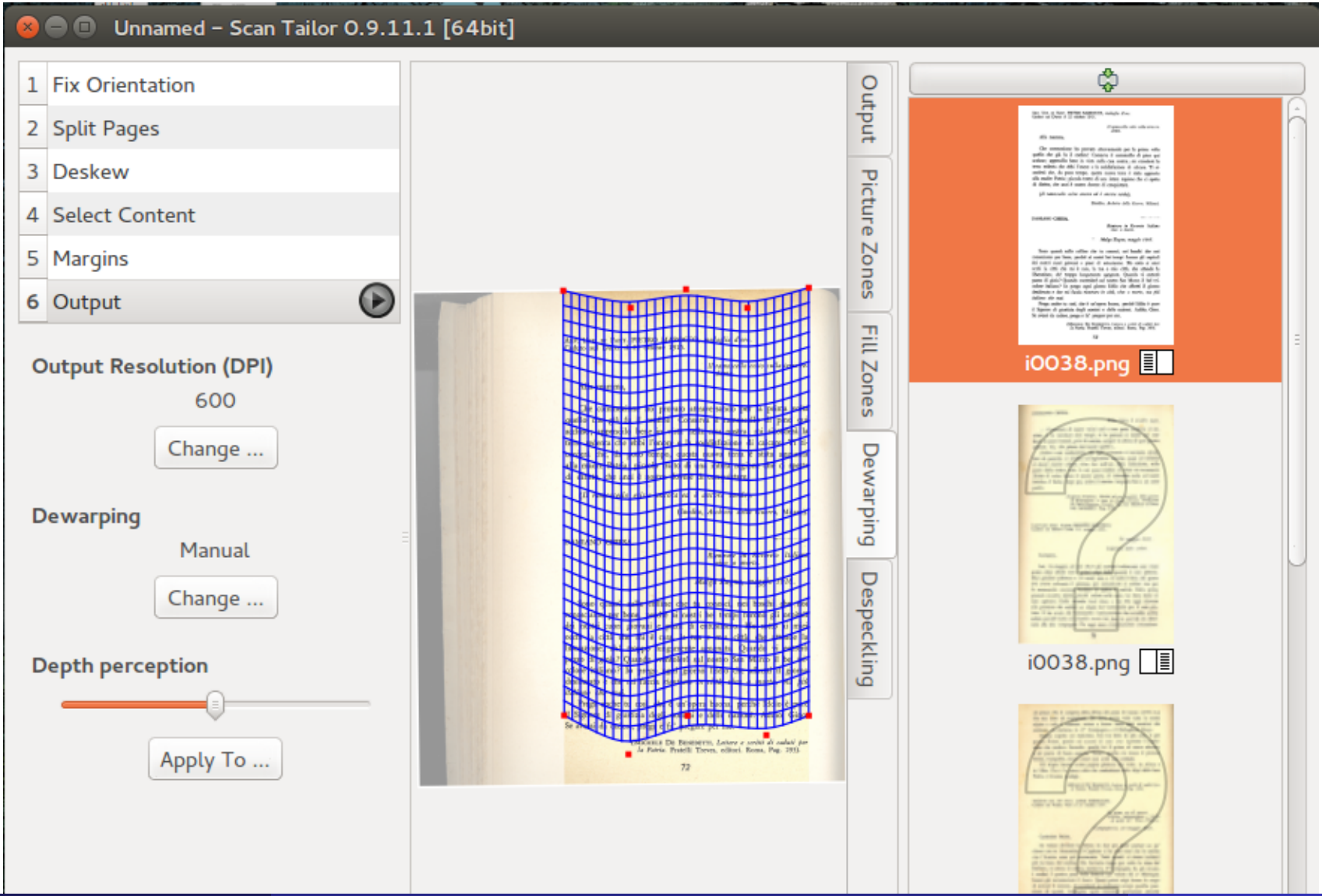
Page Splitting, Content Selection and Margins



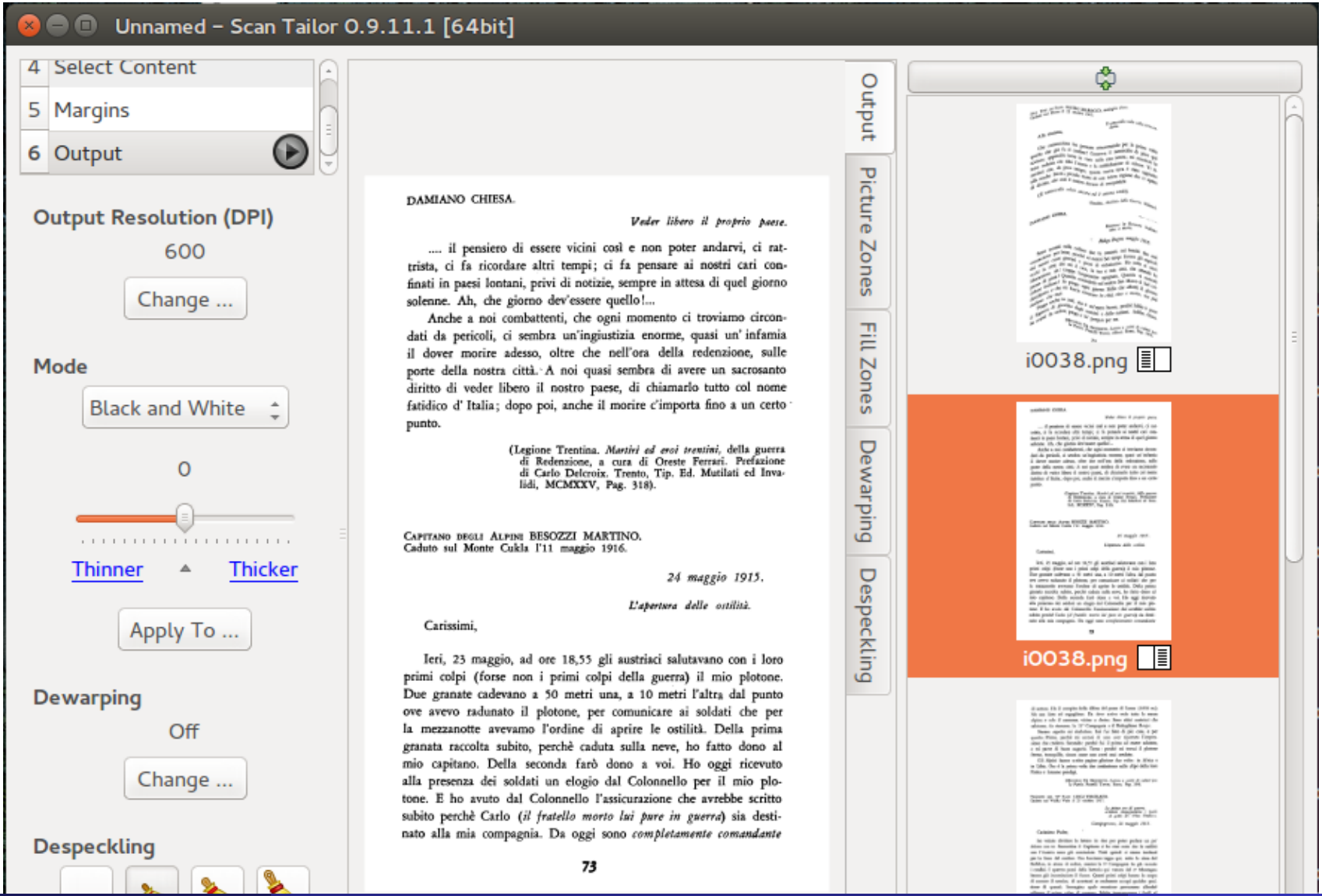
Deskewing


File Tools Help
1 Fix Orientation
2 Split Pages
3 Deskew
4 Select Content
5 Margins
6 Output
Deskew
Auto
Manual
0.69°
Apply To ...
i0038.png
i0038.png

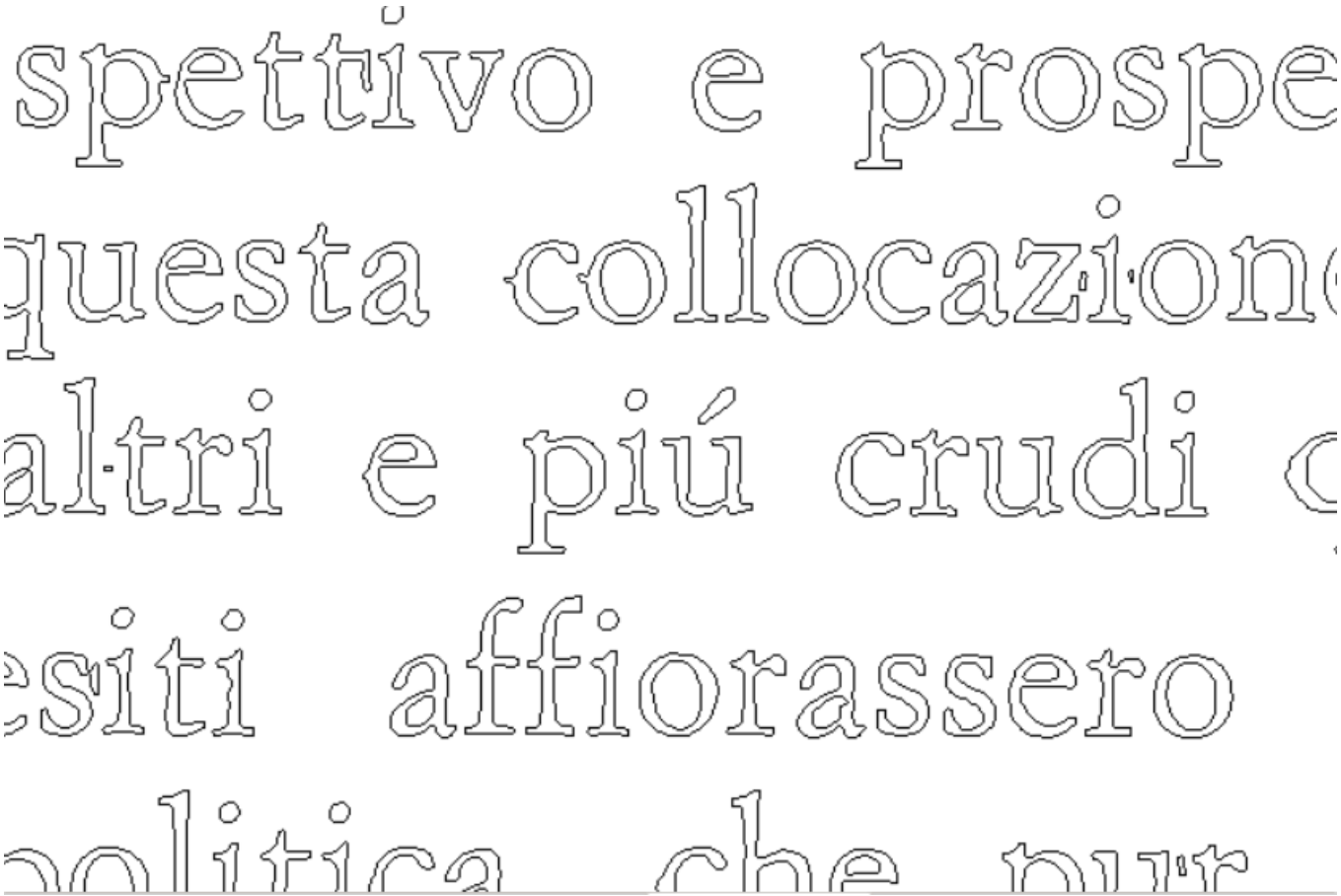
Dewarping


Unnamed – Scan Tailor 0.9.11.1 [64bit]
1 Fix Orientation
2 Split Pages
3 Deskew
4 Select Content
5 Margins
6 Output
Output Resolution (DPI)
600
Change ...
Dewarping
Manual
Change ...
Depth perception
Apply To ...
Output
Picture Zones
Fill Zones
Dewarping
Despeckling
i0038.png
i0038.png

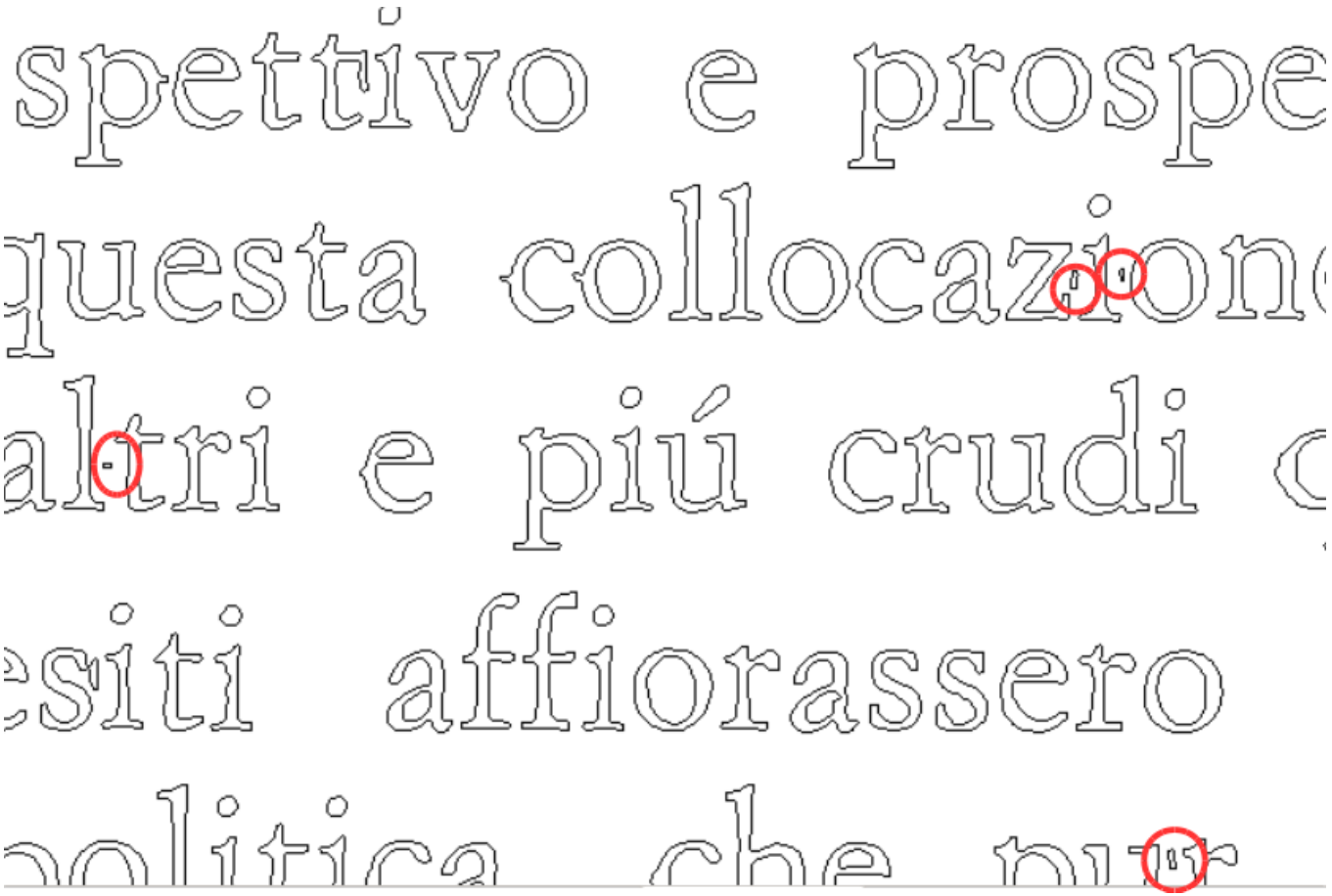
Binarization



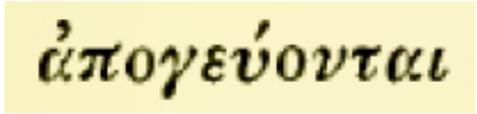
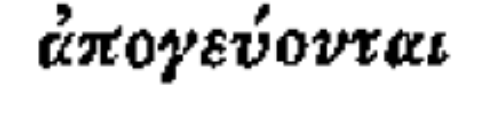
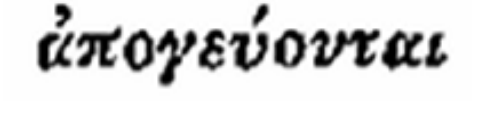
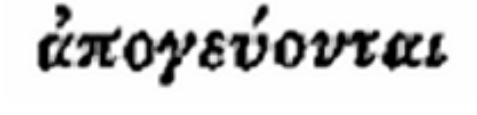
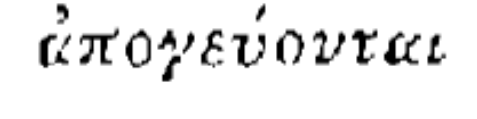
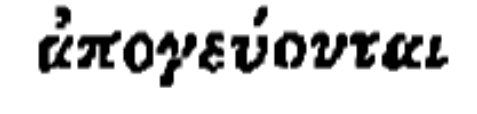
Despeckling



Despeckling



Ulteriori operazioni

original		ἀπογεύονται	CHARSEQ
threshold		a3πογεύοΜαι	CHARSEQ
scale		δπογεδονται	GREEKSEQ
erosion		ùno211eúουιαι	CHARSEQ
blur		δπογεύοπαι	GREEKSEQ
dilation		ἀπο.''εύοιιται	CHARSEQ
erosion		ἀπο'ι'εύονται	GREEKSEQ
erosion		ἀπογεύονται	WORD

Creare i propri script

```
import numpy as np
import cv2, sys, os
in_img=sys.argv[1]
out_img=in_img[0:-4]+"_clean.png"
im=cv2.imread(in_img)
gray = cv2.cvtColor(im,cv2.COLOR_BGR2GRAY)
grayout=gray.copy()
blur = cv2.GaussianBlur(gray,(5,5),1)
thresh = cv2.adaptiveThreshold(blur,255,1,1,11,2)
img,contours,hierchy = cv2.findContours(thresh,cv2.RETR_TREE,cv2.CHAIN_APPROX_NONE)
wcnt = 0
for item in contours:
    area =cv2.contourArea(item)
    [x,y,w,h] = cv2.boundingRect(item)
    if area<sys.argv[2]:
        cv2.drawContours(grayout,contours,wcnt,255,-1)
    wcnt = wcnt + 1
grayout=cv2.GaussianBlur(grayout,(5,5),1)
cv2.imwrite(out_img,grayout)
```

Overview

Abbyy Fine Reader

https://www.abbyy.com

- FineReader è una delle applicazioni commerciali dalle prestazioni più alte per l'OCR.
- FineReader è in grado di compiere complesse analisi del *layout* e di riconoscere testi multilingui.
- È possibile addestrare FineReader a riconoscere nuovi caratteri, associando l'immagine dei glifi ai corrispondenti codici Unicode.

Ideatech Anagnostis

http://anagnostis.soft112.com

- Anagnostis è un'applicazione commerciale capace di riconoscere, anche senza addestramento, il Greco antico. L'addestramento (training) può migliorare ulteriormente le prestazioni.
- Spiriti e accenti sono trattati separatamente dal corpo del carattere, migliorando la precisione del sistema di riconoscimento.

Tesseract

https://github.com/tesseract-ocr

- Tesseract è attualmente una delle applicazioni *open source* per l'OCR che danno risultati più accurati.
- Tesseract necessita un addestramento *ad hoc* per riconoscere il Greco politonico (o qualsiasi nuovo *set* di caratteri). Il riconoscimento di *set* misti di caratteri dà risultati accettabili.
- Il formato dell'*output* è solo testo oppure xhtml arricchito con un *microformat* (hOCR) che registra le posizioni delle parole (o anche dei singoli caratteri) sull'immagine della pagina.

OCRopus / ocropy

https://github.com/tmbdev/ocropy

- Anche OCRopus è *open source* e dà prestazioni molto alte.
- OCRopus è in continua evoluzione. Attualmente è sviluppato in python (ocropy).
- OCRopus utilizza reti neurali.

Training

rmordeo nõ momordi cum o:ſed cũ e memordi dicebant.Q.Ennius in ſatyris meũ

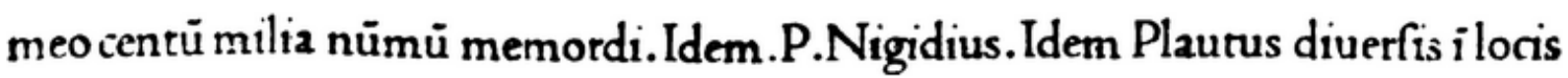
rmordeo nom momordi cum o:sed cum e memordi dicebant.Q.Ennius in satyris meum

(x) 269: tort-016/010024.bin.png

inqt non eſt:at ſi me canis memorderit.Laberius in gallis de integro patrimonio

inqt non est:at si me canis memorderit.Laberius in gallis de integro patrimonio

(x) 270: tort-016/010025.bin.png



meo centũ milia nũmũ memordi.Idem.P.Nigidius.Idem Plautus diuerſis ĩ locis

meo centum milia nummum memordi Idem P.Nigidius Idem Plautus diuersis im locis

Gamera / Rigaudon

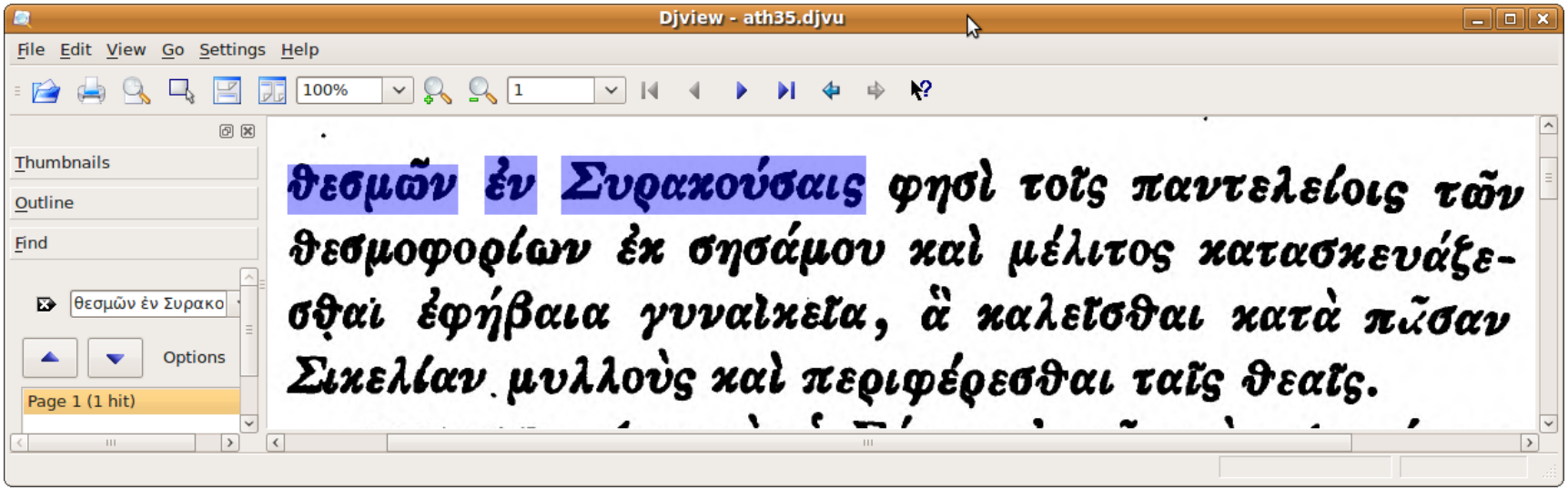
http://gamera.informatik.hsnr.de/addons/ocr4gamera

- Anche Gamera, come Tesseract e OCRopus, è *open source* e dà prestazioni molto alte.
- Bruce Robertson ha ottimizzato Gamera e ha creato dei moduli *ad hoc* per il Greco: *Rigaudon* (`https://github.com/brobertson/rigaudon`).

Valutazione dell'accuratezza

$$accuratezza = \frac{corrispondenze}{corrispondenze + substituzioni + inserzioni + cancellazioni}$$

Mappatura del testo sull'immagine



hOCR

```
<p class='ocr_par' id='par_1_7' lang='fra' title="bbox 797 1884 3144 2194">
   <span class='ocr_line' id='line_1_8' title="bbox 802 1884 3144 2032; baseline -0 -27; x_size 114;
         x_descenders 31; x_ascenders 28">
      <span class='ocrx_word' id='word_1_49' title='bbox 802 1884 893 2005; x_wconf 83'>Il</span>
      <span class='ocrx_word' id='word_1_50' title='bbox 940 1950 1043 2006; x_wconf 91'>en</span>
      <span class='ocrx_word' id='word_1_51' title='bbox 1083 1926 1360 2032; x_wconf 69'>prend</span>
      <span class='ocrx_word' id='word_1_52' title='bbox 1408 1952 1529 2007; x_wconf 88'>un</span>
      <span class='ocrx_word' id='word_1_53' title='bbox 1569 1929 1796 2031; x_wconf 59'>pour</span>
      <span class='ocrx_word' id='word_1_54' title='bbox 1845 1917 2001 2016; x_wconf 89'>soi,</span>
      <span class='ocrx_word' id='word_1_55' title='bbox 2056 1917 2259 1998; x_wconf 89'>dont</span>
      <span class='ocrx_word' id='word_1_56' title='bbox 2308 1916 2392 1999; x_wconf 90'>la</span>
      <span class='ocrx_word' id='word_1_57' title='bbox 2444 1943 2684 1999; x_wconf 90'>corne</span>
   </span>
</p>
```

Allineamento

Esempio di *progressive multiple sequence alignment*

```
 ἱερᾶς Μέμφιδος ἄρχων¶μέγας Ἀρ°σά°μης, τάς τ° ὠγυγίους¶θήβας ἐφέπων Ἀριόμαρδος,¶
 ἱερᾶς ΜέμφιδOς ἄρχων¶μέγας Ἀρ°σά°μης, τάς τ΄ ὠγυγίους¶θήβας ἐφέπων Ἀριόμαρδος,¶
 1 ρᾶς Μέμφιδος ἀρχων¶μΕγας Αρ εάςμης, τάς π’ ὠγιπέους¶θηβας φσέκων Αριόμαρδος,¶

καὶ ἐλειοβάται ναῶν ἐρέται, ^0¶δεινοὶ πλῆθός τ’ ἀνάριθμοι.¶ἁβροδιαίτων δ’ ἑπεται
καὶ ἐλειοβάται ναῶν ἐρέται,°°°¶δεινοὶ πλῆθός τ΄ ἀνάριθμοι.¶ἁβροδιαίτων δ΄ ἕπεται
και°°λειοβάται ναῶν Ιρέται, 40¶δεινοὶ πλῆθός τ’ ἀνάριOμοι.¶ἁρροδιαίτων δ’ ἑκεται

 .^Αυδῶν¶ὅχλος, οἵτ’ ἐπίπαν ἠπειρογενὲς°¶κατέχουσιν ἔ°θνος, τους Μιτρογαθὴς¶Ἀρκτ
 °°Λυδῶν¶ὅχλος, οἵτ΄ ἐπίπαν ἠπειρογενἔς°¶κατέχOυσιν ἔ°θνOς, τOὺς Μιτρογαθῆς¶Ἀρκτ
 °°Λυδῶν¶ὀχΙος, οιτ’ ἐπΠα’’ ἠπε ρογενΙς·¶κατἐχουσιν ε’ἄνος, τοὺς Μιτροκα°ης¶Ἀρ°

εύς τ° αγαθός, βασιλῆς δίοποι, ^¶χαἰ πολύχρυσοι °Σάρδεις ἐπόχους¶πολλοῖς ἄρμασιν
εύς τ’ ἀγσθός, βασιλῆς δίσποι,°°¶χαἰ πολύχρυσOι °Σάρδεις ἔπόχους¶πολλοῖς ἄρμασιν
ἐυς τ’ αγααός, βασιλῆς δ°οπο`` α¶αα’ πολ°σΕυσοι -σάρδαις ἐπόχους¶πολλοΤς ἄρμασιν

 ἐ°ξορμῶσιν,¶δίρρυμά τε καὶ τρίρρυμα τέλη¶φοβερὰν ὅ°ψιν προςιδέσθαι.¶στεῦται δ’
 ἐ°ξOρμῶσιν,¶δίρρυμά τε καὶ τρίρρυμα τέλη¶φOβερὰν ὅ°ψιν προςιδέσθαι·¶στεῦται δ΄
 ἐ’ζορμῶσιν,¶δΙρρυμά τι κα’ πρέρρυμα τἐλη¶φορερὰν ὀ’ψιν προςιδέσθαι.¶στέυτσι δ’
```

Selezione

ἄ λ λ ο ς δ ' ε κ ε ί ν ο υ _ π α ῖ ς τ ό δ ' έ _ ρ γ ο ν η ν υ σ ε ν .	FineReader
ἄ λ λ ο ς δ ' ἐ κ ε ί ν ο υ * π α ῖ ς τ ό δ ' ἔ ' ρ γ ο ν ἤ ν υ σ ε ν	OCRopus
; λ λ ο ς _ ό ἐ χ ε ; τ ο υ _ . κ α ˜ ς τ ό δ _ _ _ P ϒ ο ≫ ἤ ν ν σ ι ν .	Anagnostis
ἄ λ λ ο ς δ ' ἐ κ ε ί ν ο υ _ π α ῖ ς τ ό δ ' ἔ _ ρ γ ο ν ἤ ν υ σ ε ν .	Result

Spell-checking

Output di FineReader	RegEx per tutte le applic. OCR	Suggerimenti dello spell-checker	Risultati
ἐξερήμωσεν	ἐξερήἑ?[μι]ωσεν	ἐξερήμωσε, ἐξερήμωσέ, ἐξηρήμωσεν	ἐξερήμωσεν
ωπασεν	[ωοὦ]π[αο]σ[εό]ν	ὥπασεν, ὥπασέν, σπάσεν	ὥπασεν
εν᾽	[εἔ]ν᾽	ἐν, ἐν᾽ … ἕν᾽ (34. elemento)	ἕν᾽
επάσης	ε?ἀ?πάσης	πάσης, πάσῃς … ἁπάσης (11. elemento)	ἁπάσης
ἐὐθυντἠριον	[εἐ][ὐυ]θυντ[ἠή]ριον	εὐθυντήριον, εὐθυντήριόν, εὐθυντῆρι	εὐθυντήριον
πρώτος	πρ[ώῶ]τος	πρῶτος, πρῶτός, πρωτὸς	πρῶτος
Κύρος	[ΚΧΗ][ύῦῖ]ρος	Κῦρος, Κῦρός, Κύπρος	Κῦρος
εθηκε	[εἔ]θηκε	ἔθηκε, ἔθεκέ, θῆκε	ἔθηκε
Δυδῶν	[ΔΛ]υδῶν	Δυῶν, Διδῶν … Λυδῶν (6. item)	Λυδῶν
λάὸν	λ[αά][òὸ]ν	λαὸν, λαόν, Λάιόν	λαὸν
ἦλασεν	[ἦἤ]λασ[ετ]?ν	ἤλασεν, ἤλασέν, ἤασεν	ἤλασεν
ευφρων	ε?ι?[υὄ]φρωο?ν	ἐύφρων, Εὔφρων, **εὔφρων** (corretto)	ευφρων

Suggerimenti correttamente accettati

Output di FineReader

ωπασεν

RegEx per tutte le applic. OCR

[ωοώ]π[αο]σ[εό]ν

Suggerimenti dello spell-checker

ώπασεν, ώπασέν, σπάσεν

Risultato

ώπασεν

Suggerimenti correttamente rifiutati

Output di FineReader

ἐξερήμωσεν

RegEx per tutte le applic. OCR

ἐξερήἑ?[μι]ωσεν

Suggerimenti dello spell-checker

ἐξερήμωσε, ἐξερήμωσέ, ἐξηρήμωσεν

Risultato

ἐξερήμωσεν

Suggerimenti scorrettamente rifiutati

Output di FineReader

ευφρων

RegEx per tutte le applic. OCR

ε?ι?[ῠŏ]φρωο?ν

Suggerimenti dello spell-checker

ἐύφρων, Εὔφρων, **εὔφρων** (corretto)

Risultato

ευφρων

Overview

Collaborativo / Cooperativo

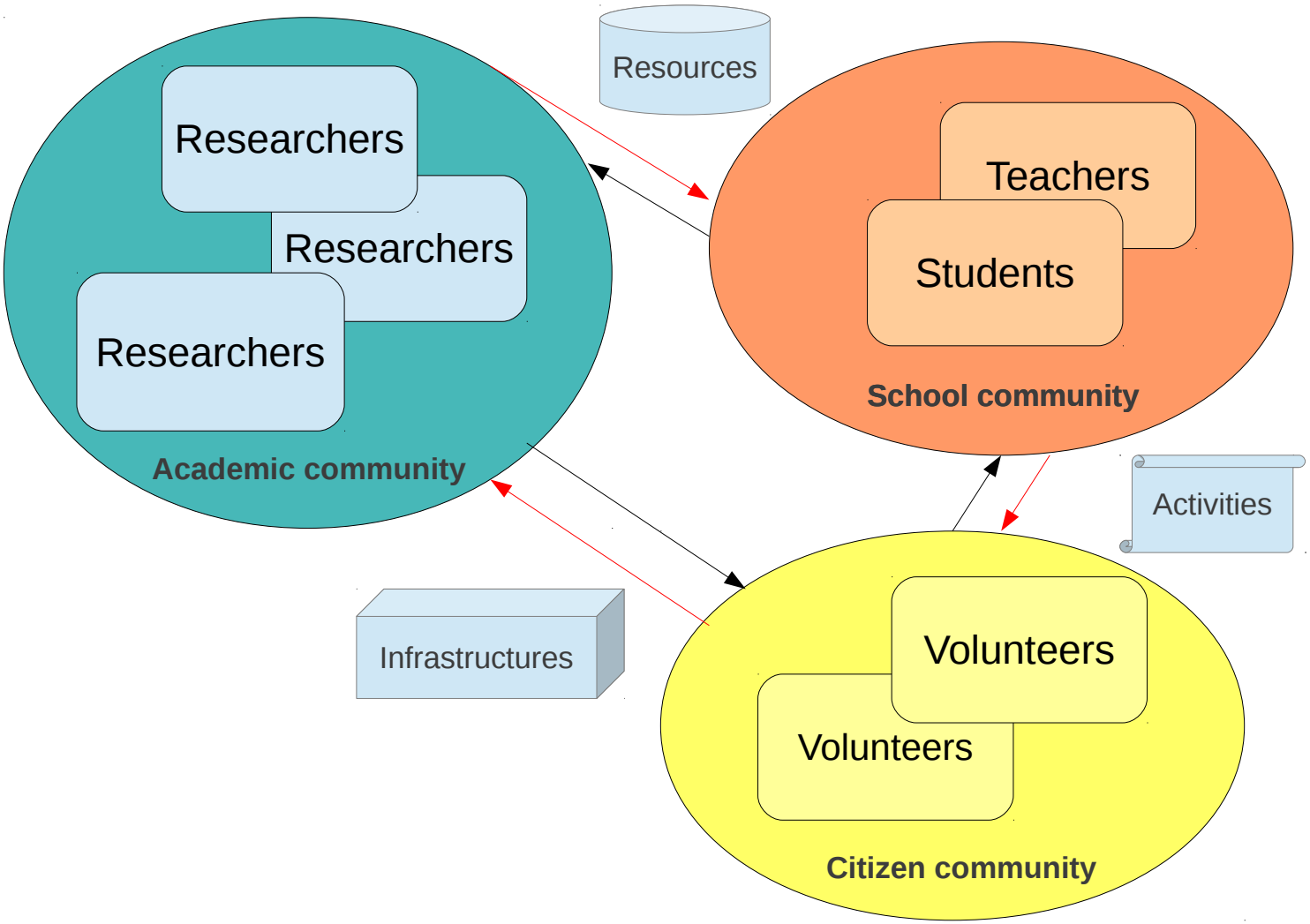
Collaborazione

La collaborazione implica l'interazione di più soggetti o di più gruppi con un obiettivo comune. Il prodotto ottenuto è frutto di negoziazioni fra le parti e costituisce il raggiungimento di tale obiettivo.

Cooperazione

La cooperazione implica la possibilità di suddividere un obiettivo in parti modulari, in modo che i sottoprodotti siano autonomi e riusabili da terze parti per scopi non necessariamente previsti dagli autori originari

Coinvolgere comunità differenti



WikiSource

WIKISOURCE

English
The Free Library
765,000+ pages

Français
La bibliothèque libre
1,209,000+ pages

Русский
Свободная библиотека
293,000+ статей

Español
La biblioteca libre
102,000+ páginas

Deutsch
Die freie Quellensammlung
349,000+ Seiten

Italiano
La biblioteca libera
106,000+ pagine

Português
A biblioteca livre
28,000+ páginas

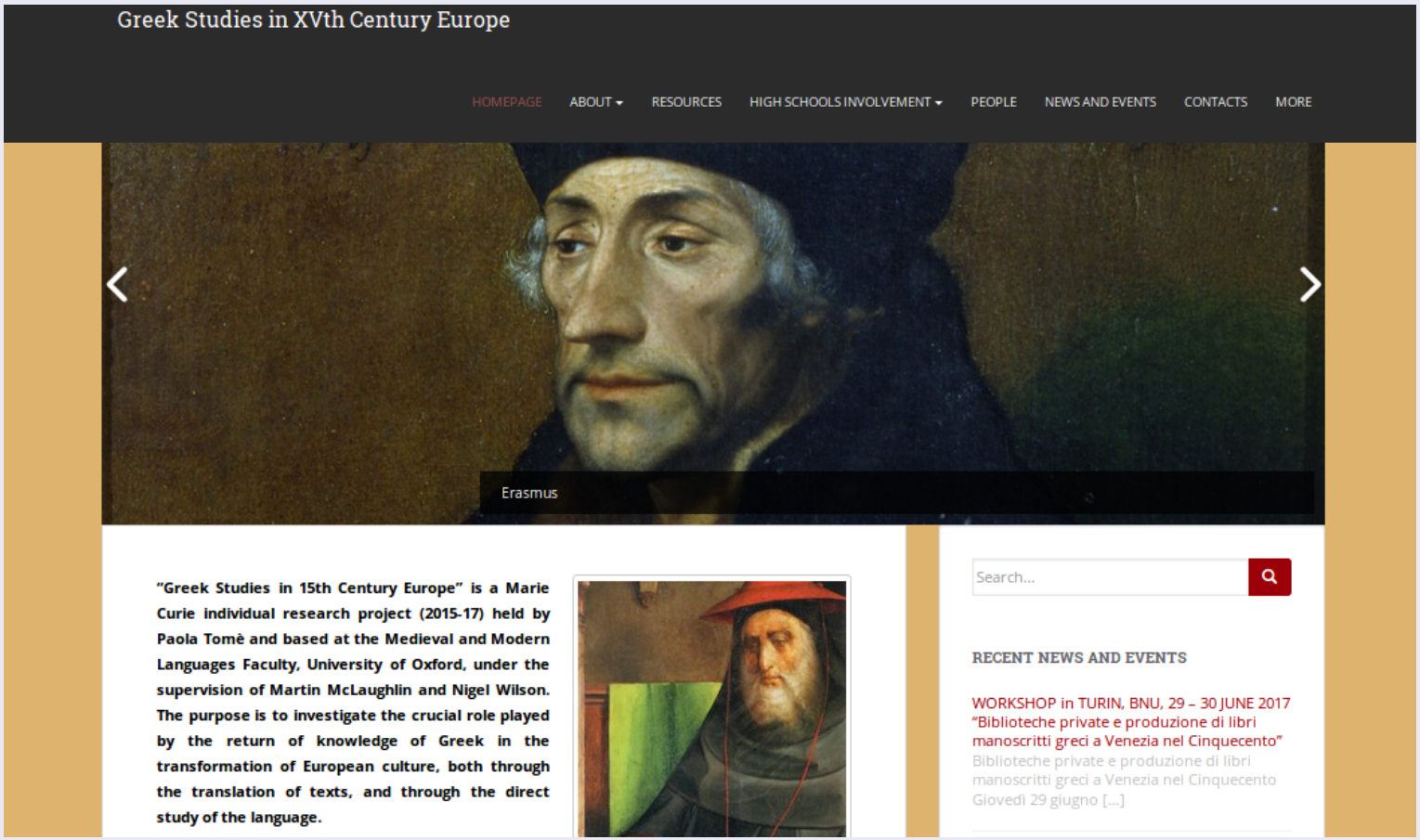
עברית
הספריה החופשית
141,000+ מאמרים

Polski
wolna biblioteka

العربية
المكتبة الحرة

Lo studio del greco nell'Europa del XV secolo
Progetto Marie Curie di Paola Tomè

http://greek15century.mml.ox.ac.uk



Orthographia di Giovanni Tortelli

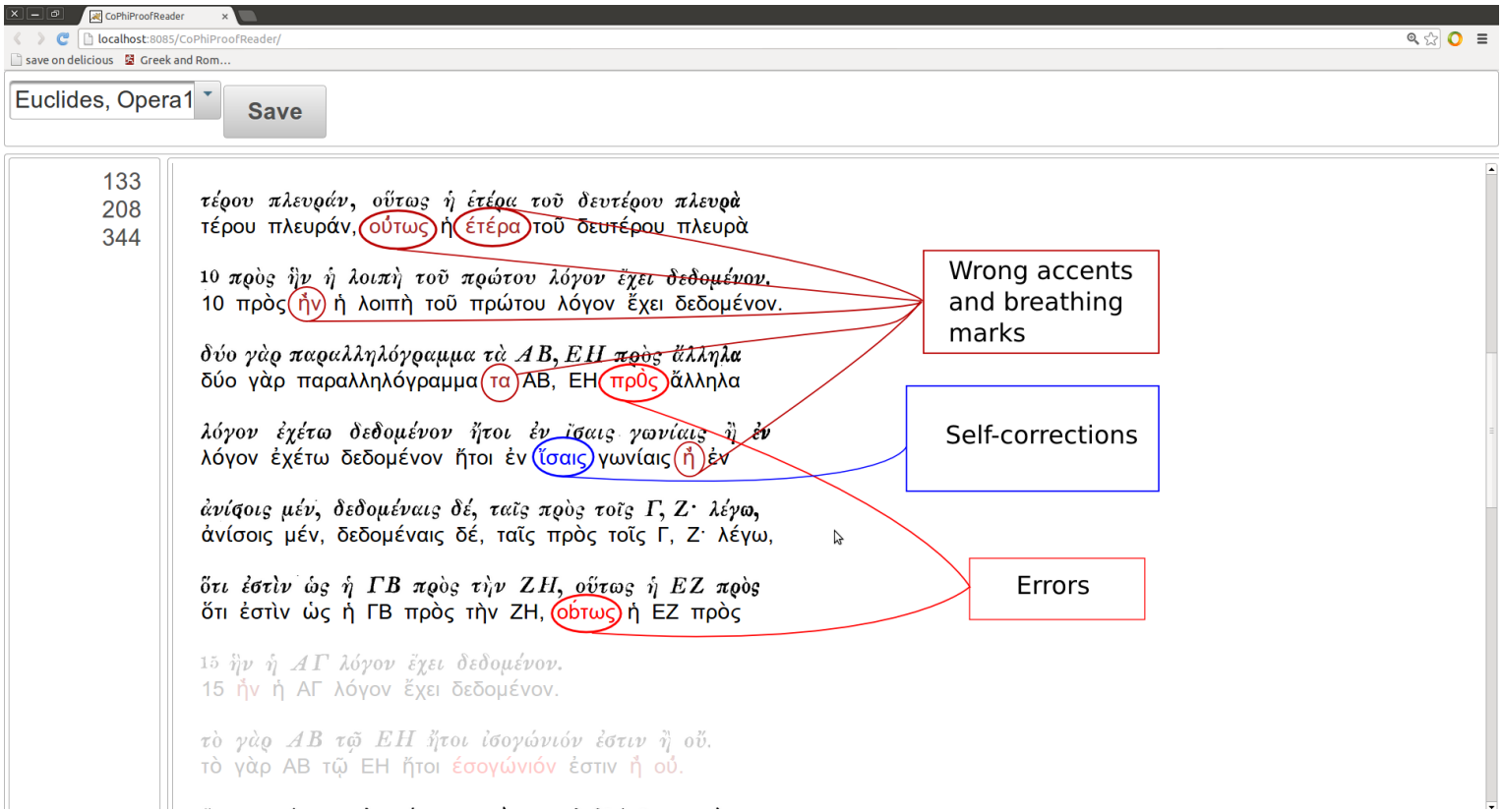
IOANNIS TORTELLII ARRETINI COMMENTARIORVM GRAM
MATICORVM DE ORTHOGRAPHIA DICTIONVM E GRAECIS
TRACTARVM PROOEMIVM INCIPIT AD SANCTISSIMVM
PATREM NICOLAVM QVINTVM PONTIFICEM MAXIMVM.

OEPERAM OLIM BEATISSIME
pater Nicolae. v. ſumme pontifex com/
mentaria quædam grammatica condere:
qbus omnem litterariam antiquitatem
& orthographiæ rationē cū opportunis
hiſtoriis pro poetarum declaratione cō/
nectere conabar: pfuturus ſane pro mea
uirili ſtudioſis linguæ latinæ. interim
ab aliis ſtudiis negociiſq; familiaribus
interceptus illa prorſus relinquere: at/
que longo tempore abiicere uiſus ſum.
Sed nuper cum apud Alatrium campa/
niæ oppidum ex aeris romani moleſtia
ſeceſſiſſē: ea abſoluere quorundam ami/
corum rogatu conatus ſum: ac ſic quoq;
ad calcem uix uſque perduxi opus magnum uarium & diffuſum: atque illud tuæ
ſanctitati a qua uelut fonte omnia mea bona fluxerunt: dedicare conſtitui. non

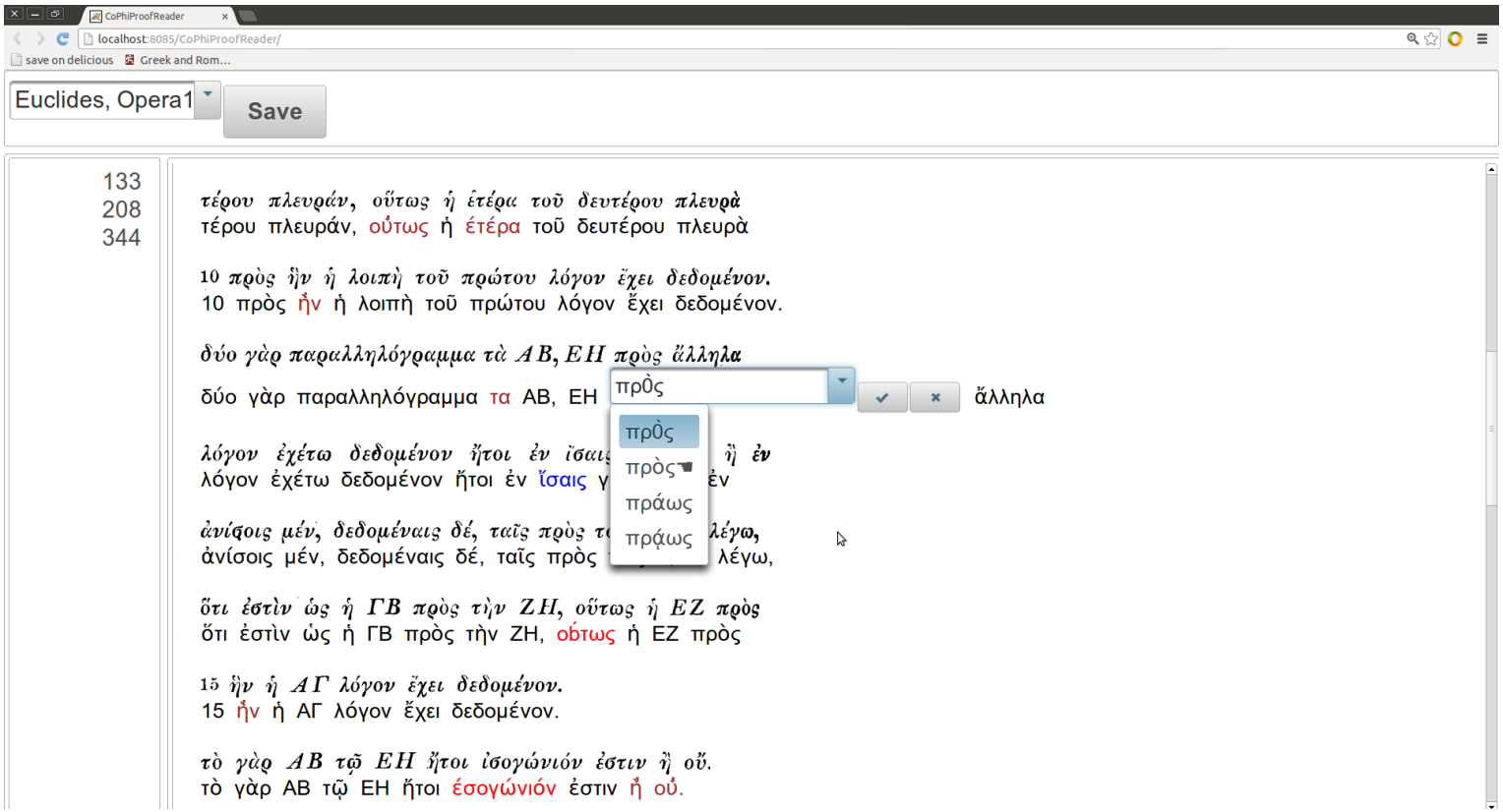
Sciogliere o non sciogliere le abbreviazioni?

- In fase di training i ricercatori hanno trascritto accuratamente, carattere per carattere, il testo rispettando le abbreviazioni originarie, con l'impiego dei caratteri Unicode raccomandati dalla *Medieval Unicode Font Initiative* (MUFI)
- In fase di correzione cooperativa gli studenti hanno sciolto le abbreviazioni

CoPhi Proof-reader per il greco antico



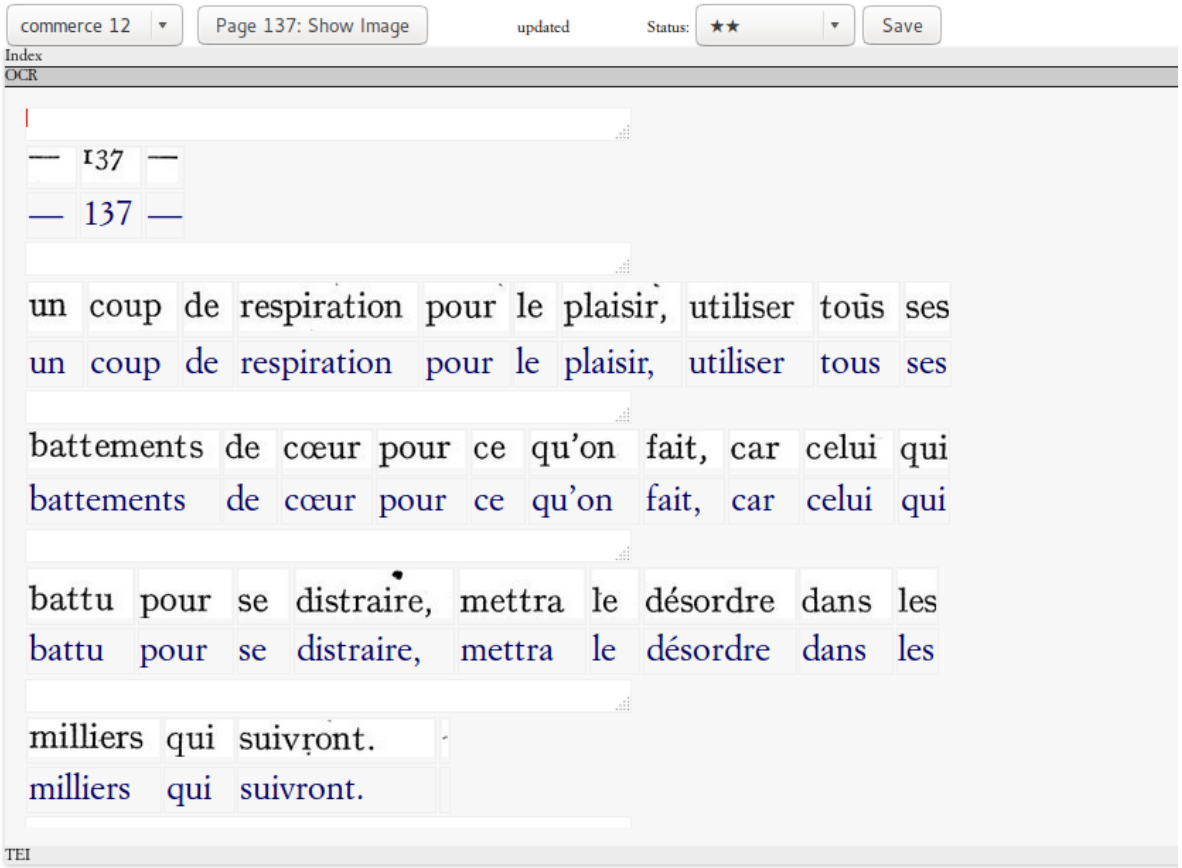
Lista dei suggerimenti



CoPhi Proof-reader per Commerce Online

Progetto coordinato da Antonietta Sanna, Università di Pisa

Digitalizzazione di 29 volumi della rivista letteraria *Commerce*



Reintegrare la formattazione

TEI

(re)generate plain text

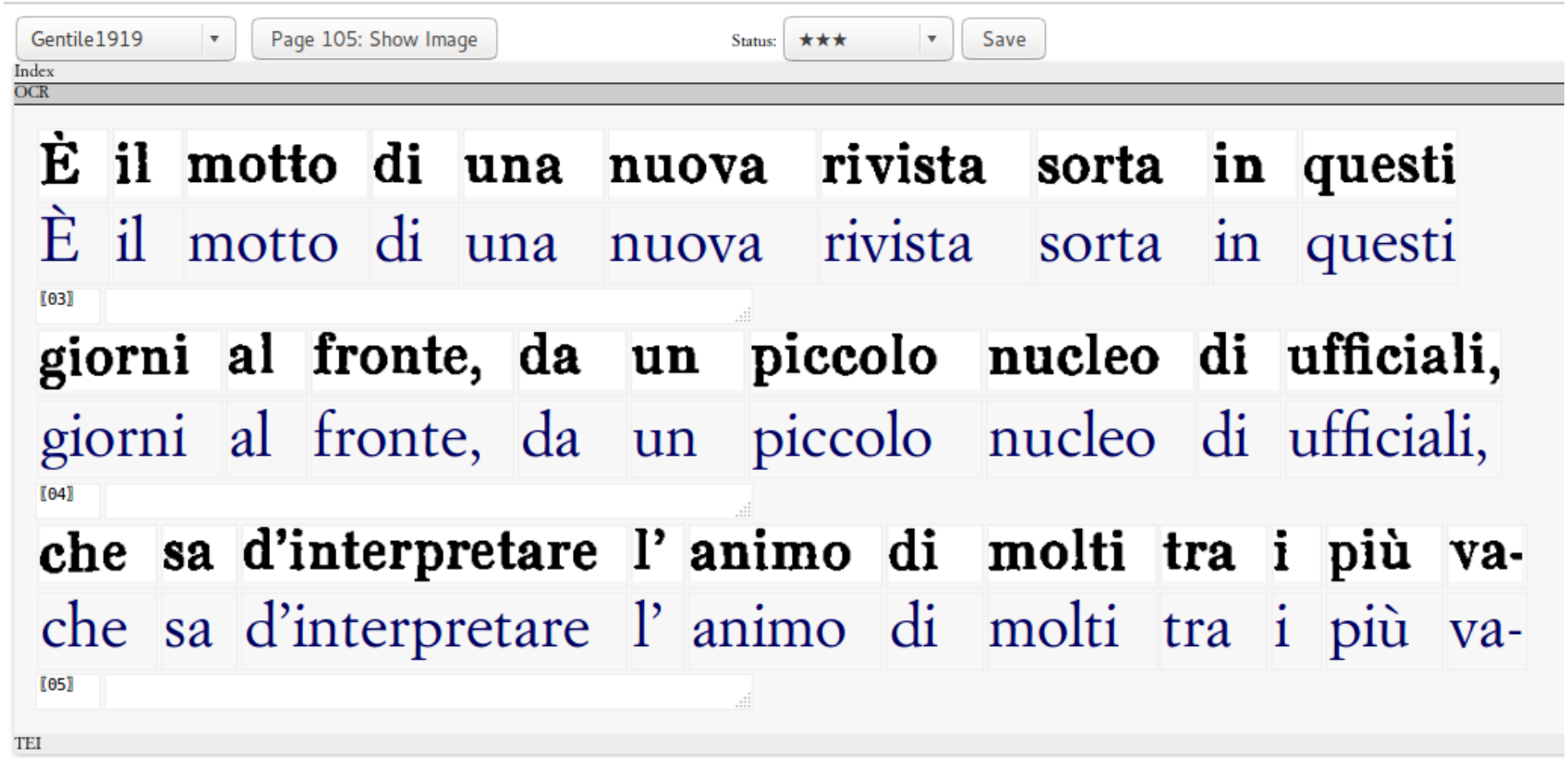
```
1  <fw>— 15 —</fw>
2  véritable d'une hiérarchie fondée sur la rareté. — Je
3  m'amuse parfois d'une image <hi>physique</hi> de nos cœurs,
4  qui sont faits intimement  d'une énorme injustice et
5  d'une petite justice combinées. J'imagine qu'il y a
6  dans chacun de nous un atome important entre nos
7  atomes, et constitué par deux <hi>grains d'énergie</hi> qui
8  voudraient bien se séparer. Ce sont des énergies con-
9  tradictoires mais indivisibles. La nature les a jointes
10 pour toujours, quoique furieusement ennemies. L'une
11 est l'éternel mouvement  d'un gros <hi>électron positif</hi>, et
12 ce mouvement inépuisable engendre une suite de sons
13 graves où l'oreille intérieure distingue sans nulle peine
14 une profonde phrase monotone: <hi>Il n'y a que moi. Il</hi>
15 <hi>n'y a que moi. I'ln'y a que moi, moi, moi...</hi> Quant au
16 petit électron radicalement négatif, il crie à l'extrême
17 de l'aigu, et perce et reperce de la sorte la plus cruelle
18 le thème égotiste de' l'autre : <hi>Oui, mais il ya un tel...</hi>
19 <hi>Oui, mais il y a un, tel... Tel, tel, tel.</hi> Et tel autre !...
20 Car le nom change assez souvent...<_p/>
21 <p_/>Bizarre royaume où tentes les belles choses qui s'y
22 produisent sont  une amère nourriture pour toutes les
```

Il progetto *Voci della Grande Guerra*

http://www.vocidellagrandeguerra.it

Voci della Grande Guerra è un'iniziativa scientifica e culturale promossa dall'Università di Pisa (capofila) in collaborazione con l'Istituto di Linguistica Computazionale "A. Zampolli" del Consiglio Nazionale delle Ricerche (ILC-CNR), l'Università di Siena e l'Accademia della Crusca.
Finanziata dalla Presidenza del Consiglio dei Ministri nell'ambito dell'avviso pubblico per la Commemorazione del Centenario della Grande Guerra, ha l'obiettivo di preservare e diffondere le memorie della Prima Guerra Mondiale attraverso la costruzione e la pubblicazione online di un corpus digitale di testi rappresentativi delle diverse modalità di sentire e raccontare l'Italia in guerra da parte dei suoi protagonisti.

CoPhi Proof-reader per *Voci della Grande Guerra*



Migliorare l'ergonomia

Cos'è l'eye-tracker?

L'*eye tracker* è uno strumento per tracciare il movimento e i punti di fissazione degli occhi su aree di interesse

Esperimenti con l'eye-tracker

Con Barbara Balbi, Flavia De Simone e Vincenzo Broscritto dell'Università "Suor Orsola Benincasa" stiamo conducendo esperimenti con l'*eye tracker* per studiare le diverse strategie di correzione di soggetti non addestrati e di soggetti esperti

Conclusione

- La scelta delle tecniche e degli strumenti di OCR deve essere valutata caso per caso
- Il pre-processing delle immagini con script creati appositamente allo scopo può aumentare notevolmente l'accuratezza
- Il lavoro di correzione può essere ripartito fra diversi soggetti o fra diversi gruppi di lavoro

Domande aperte

- È un modello esportabile?
- Chi dovrebbe guidare progetti di questo tipo?
- Come coinvolgere i ricercatori?
- Come coinvolgere i volontari?
- Come coinvolgere gli insegnanti?
- Come motivare gli studenti?

Grazie per l'attenzione

Bibliografia


S. Feng, R. Manmatha: A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books. JCDL 2006, 109–118 (2006)


W.B. Lund, E.K. Ringger: Improving Optical Character Recognition through Efficient Multiple System Alignment, JCDL (2009)


M. Reynaert: Non-interactive OCR Post-correction for Giga-Scale Digitization Projects. A. Gelbukh (ed.): CICLing 2008, LNCS 4919, 617–630 (2008)


M. Reynaert: All, and only, the Errors: more Complete and Consistent Spelling and OCR-Error Correction Evaluation. 6th International Conference on Language Resources and Evaluation 2008, 1867–1872 (2008)


C. Ringlstetter, K. Schulz, S. Mihov, K. Louka: The same is not the same - postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. 8th International Conference on Document Analysis and Recognition, 1, 406–410 (2005)


M. Spencer, C. Howe: Collating texts using progressive multiple alignment. Computer and the Humanities, 37, 1, 97–109 (2003)


G. Stewart, G. Crane, A. Babeu: A New Generation of Textual Corpora. JCDL 2007, 356–365 (2007)


L. Zhuang, X. Zhu: An OCR Post-processing Approach Based on Multi-knowledge. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, 346–352 (2005)