

Εισαγωγή στην Ψηφιακή Επεξεργασία Εγγράφων



Βασίλης Γάτος

Διευθυντής Ερευνών



Εργαστήριο Υπολογιστικής Ευφυΐας
Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών
Εθνικό Κέντρο Έρευνας Φυσικών Επιστημών
«Δημόκριτος»

Αγία Παρασκευή, Αθήνα

<http://www.iit.demokritos.gr/~bgat>



Institute of Informatics and Telecommunications – NCSR “DEMOKRITOS”

Εισαγωγή στην Ψηφιακή Επεξεργασία Εγγράφων

- Σκοπός: η ανάπτυξη εφαρμογών για την ανάλυση και την ανάγνωση των ιστορικών εγγράφων



- Τυπωμένα και χειρόγραφα έγγραφα

[illegible]

Σ. ΜΑΡΚΕΖΙΝΗΣ (Τπ. Συντον.). Τὸ λέγει ἀπὸ
τῆς παντός. Χάρω ἐπὶ δὲν γίνονται διακρίσεις καὶ τοῦτε
δοῖσι εἰ ἀνθρώποις μοι εἶναι μορὰ καὶ θά ἔχοντες μακρο-
τάτην ἀσπίδα, πρῶτος τὸ ἀπὸ δὲν ἐλπίδω. Ἄν εἰ ἐρω-
τῆς σας, εἶναι ἐπὶ τοῦ θέματος καὶ συγκεκριμένην τὴν κί-
νῃ.

Σ. ΜΑΡΚΕΖΙΝΗΣ (Ἰπ. Συντον.). Ἀκόμη εἴμεθα
ἐν τὴν ἀρχήν. Θὰ δῶτε καὶ ἄλλα ἀκόμη.

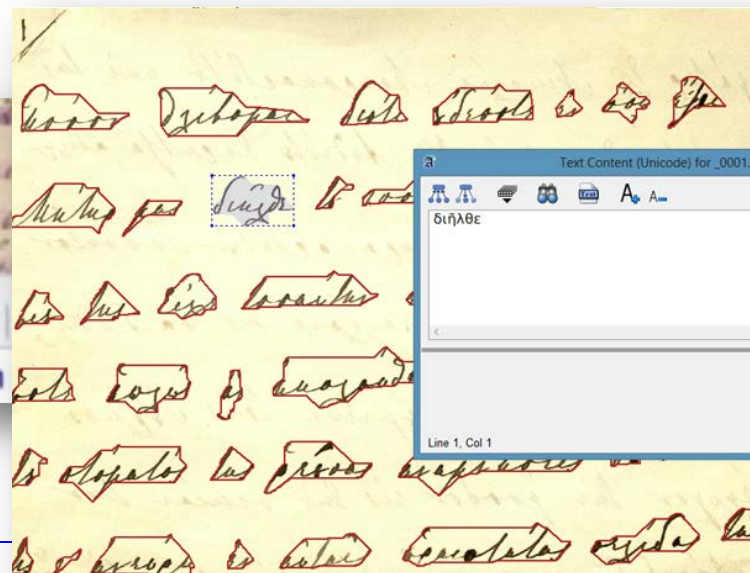
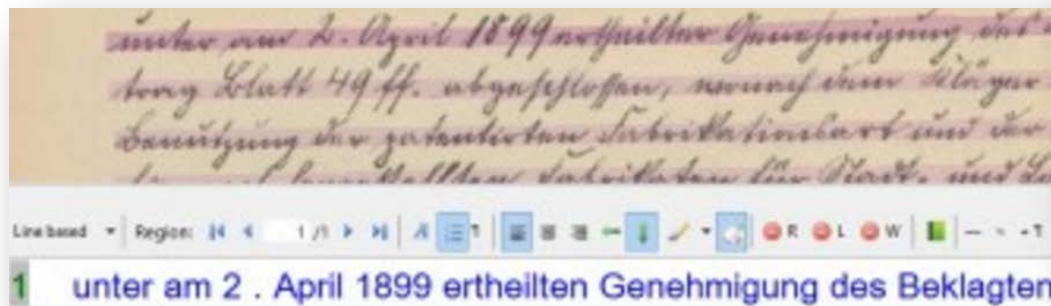
ΣΤΗ ΜΑΡΚΕΖΙΝΗΣ (Τη Συντονιστ.). Έχετε κάνει
 τη συνάντηση με τους δικούς σας εκπαιδευτικούς και τους λό-
 γιστές; Δι' όρους του Θεού, όχι εξακολουθώ να τους
 έχω συναντήσει. Είναι όπως συνέβη το 1983 (για τους
 δικούς τους).

Έχετε λοιπόν, κύριε προέδρε, να τηρήσει πιστότα-
 τα αυτή τη θέση, αλλά πώς να την αποδέχεται το Τμήμα;
 Όργανωσαν διάσκεψη σύντομη, ακόμη ελκυστική, με
 σκοπό, όχι να καθορίσει αυτή η διάσκεψη, οι αντιπροσώποι
 της Γενικής και να περάσει την. Όχι. Ζούσαμε, περνούσαμε
 τις Γενικές διά κάποια συνέδρια, όχι. Μαντζινάριος με
 κ. Κονταρίδη και τον Σαββατοκύριακο έκαναν διά να πάνε

Β. Γάτος, Σεμινάριο «Ψηφιακές Εκδόσεις και Νεοελληνικές Σπουδές», Ανάλυση και Αναγνώριση Ιστορικών Εγγράφων

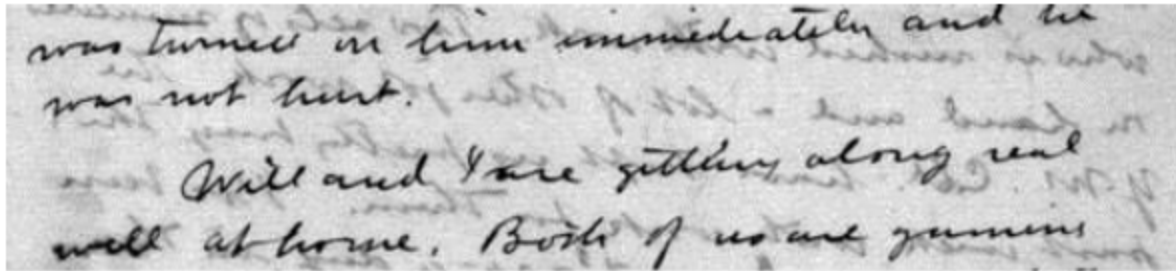
Εισαγωγή στην Ψηφιακή Επεξεργασία Εγγράφων

- Ground-truth
- Κύκλοι εκπαίδευσης:
 - Έστω το σύνολο εκπαίδευσης αποτελείται από 50 εικόνες.
 - Ένα δεύτερο σύνολο από 50 σελίδες αναγνωρίζεται με σφάλμα σε επίπεδο λέξης 30% (3 στις 10 λέξεις δεν έχουν αναγνωρισθεί σωστά).
 - Το αποτέλεσμα από το δεύτερο σύνολο διορθώνεται από τον χρήστη.
 - Ένα τρίτο σύνολο 50 σελίδων αναγνωρίζεται με σφάλμα 25%.
 - ...



Στάδια Ψηφιακής Επεξεργασίας Εγγράφων

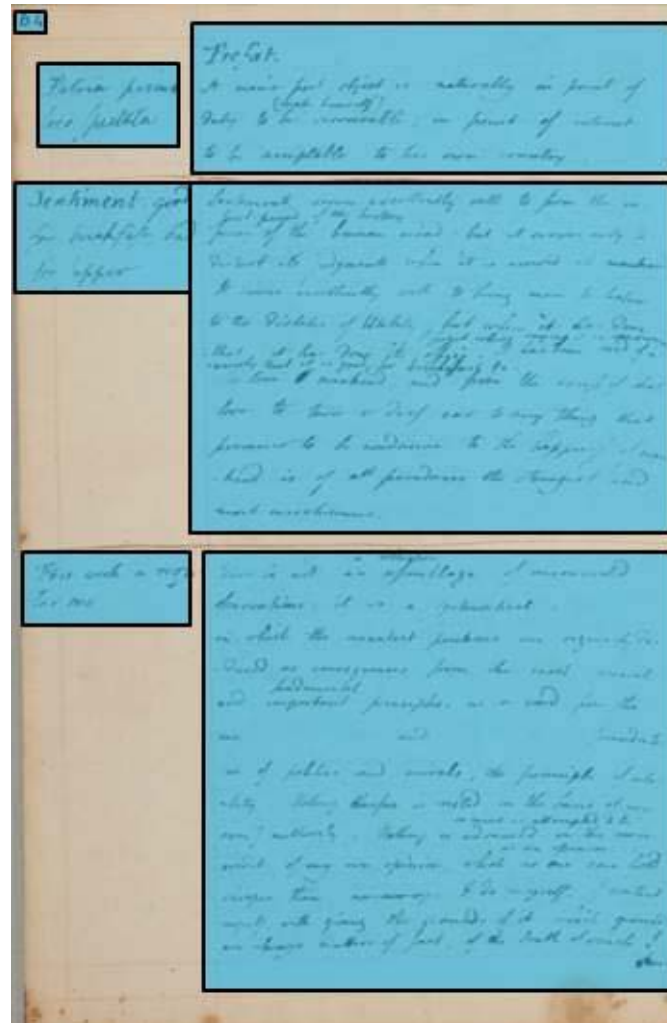
- Βελτίωση της ποιότητας της εικόνας



was turned in him immediately and he
was not hurt.
Will and I are getting along real
well at home. Both of us are getting

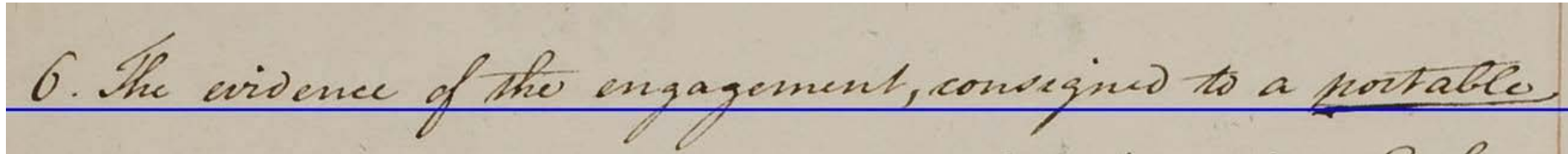
Στάδια Ψηφιακής Επεξεργασίας Εγγράφων

- Κατάτμηση, ανάλυση της δομής (layout analysis)



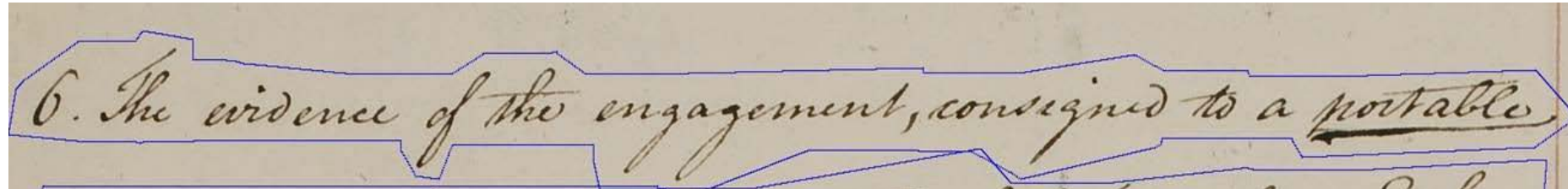
Στάδια Ψηφιακής Επεξεργασίας Εγγράφων

- Εντοπισμός γραμμών κειμένου



6. The evidence of the engagement, consigned to a portable

This image shows a horizontal strip of a handwritten document. The text is written in a cursive script in dark ink on aged, slightly textured paper. A solid blue horizontal line is drawn across the entire width of the strip, positioned just below the baseline of the handwriting.



6. The evidence of the engagement, consigned to a portable

This image shows the same horizontal strip of handwritten text as above. However, a blue, irregular, jagged line has been drawn over the text, following its contours. This line represents a segmentation or mask applied to the text for digital processing.

Στάδια Ψηφιακής Επεξεργασίας Εγγράφων

- Διόρθωση κλίσης γραμμών κειμένου και γραμμάτων

upon the above mentioned terms he would engage
upon the above mentioned terms he would engage

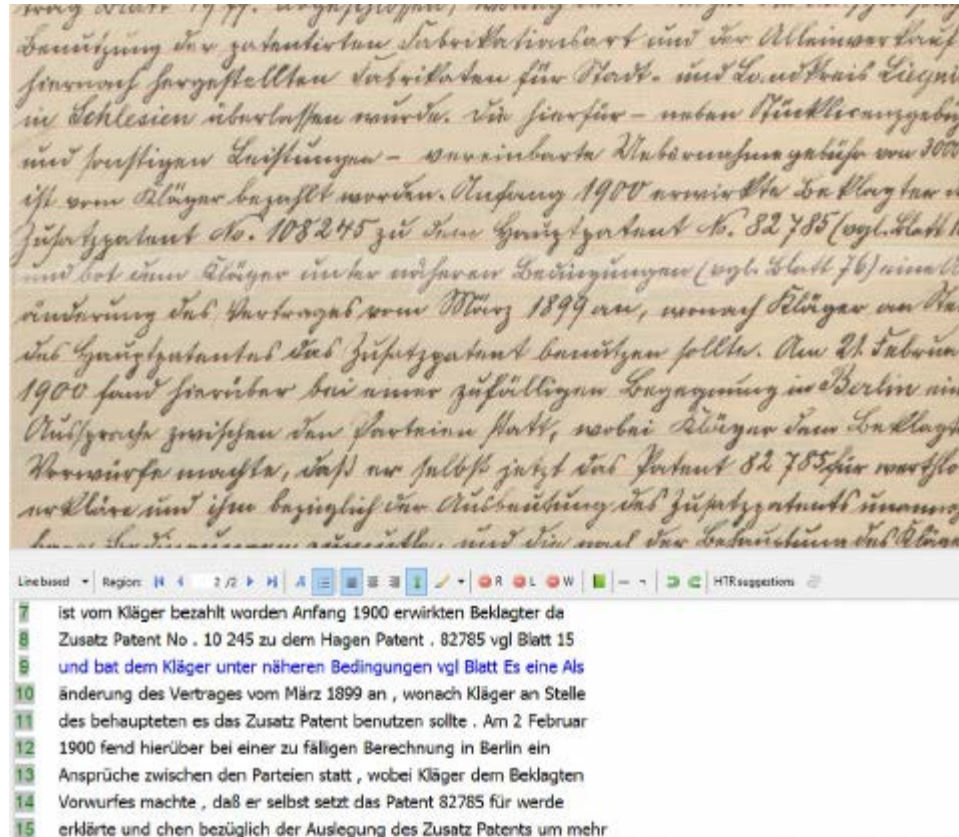
(a) *The distinction is material: for as fixed income comes*

(a) *The distinction is material: for as fixed income comes*

of Justifications. power
↓ ↓
of Justifications. power

Στάδια Ψηφιακής Επεξεργασίας Εγγράφων

- Αναγνώριση του κειμένου (Handwritten Text Recognition - HTR)



Στάδια Ψηφιακής Επεξεργασίας Εγγράφων

■ Εντοπισμό λέξεων (Keyword Spotting - KWS)

