Check_qa

Εφαρμογή ελέγχου των σελίδων περιγραφής μαθημάτων, όπως αυτά παρουσιάζονται στον ιστότοπο της Μονάδας Διασφάλισης Ποιότητας του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.

Ο σκοπός

- Το παρόν project έγινε στο πλαίσιο του μαθήματος "Τεχνολογία Λογισμικού" (5ο εξάμηνο, ακαδ. έτος 2019-2020) του προπτυχιακού προγράμματος σπουδών του τμήματος Πληροφορικής Α.Π.Θ.
- Ο σκοπός της εφαρμογής είναι να υποδεικνύει στο προσωπικό του τμήματος ποια μαθήματα έχουν ελλείψεις στις σελίδες περιγραφής τους, ώστε να ενημερώνονται οι υπεύθυνοι των μαθημάτων και να τις συμπληρώνουν με τις απαραίτητες πληροφορίες.
- Οι περιγραφές των μαθημάτων (π.χ. προαπαιτήσεις, περιεχόμενο μαθήματος) ανά εξάμηνο σπουδών είναι ιδιαίτερα σημαντικές για τους φοιτητές, ειδικά στα εξάμηνα 6,7 και 8 κατά τα οποία καλούνται να επιλέξουν τα μαθήματα που θα παρακολουθήσουν .

Το εργαλείο DeiXTo

- Για την λήψη των δεδομένων των ιστοσελίδων χρησιμοποιήθηκε το εργαλείο εξαγωγής δεδομένων DeiXTo (web data extraction tool), ένα λογισμικό ανοικτού κώδικα που χρησιμοποιείται για την δημιουργία κανόνων εξαγωγής δεδομένων από ιστοτόπους.
- Με το DeiXTo ο κώδικας HTML αναλύεται και δημιουργείται μια δενδροειδής αναπαράσταση με κόμβους τις περιοχές τις ιστοσελίδας (με βάση τις HTML ετικέτες) ακολουθώντας το Μοντέλο Αντικειμένου Εγγράφου (Document Object Model).
- Το μοντέλο DOM είναι ένα αντικειμενοστρεφές μοντέλο περιγραφής εγγράφων Ιστού που δημοσιεύθηκε το 1998 από την Κοινοπραξία του Παγκοσμίου Ιστού (W3C).

Το εργαλείο DeiXTo

- Το DeiXTo μας επιτρέπει να δημιουργήσουμε κανόνα εξαγωγής, με βάση μια ιστοσελίδα. Έτσι, για την δημιουργία του κανόνα, αρχικά σχηματίζεται η DOM αναπαράσταση της ιστοσελίδας που επιλέγουμε : στην περίπτωσή μας της σελίδας ενός μαθήματος που έχει πλήρη περιγραφή – δηλαδή όλα τα πεδία πληροφοριών.
- Στη συνέχεια επεξεργαζόμαστε τους κόμβους του δένδρου. Το πρόγραμμα μας δίνει τη δυνατότητα να ορίσουμε για κάθε κόμβο εάν θέλουμε να είναι υποχρεωτικός ή προαιρετικός στα στιγμιότυπα που θα εξετάσει ο κανόνας και εάν θα κρατείται το περιεχόμενό του.
- Μπορούμε να επιλέξουμε ανάμεσα σε 5 δυνατές καταστάσεις για την ρύθμιση των κόμβων.

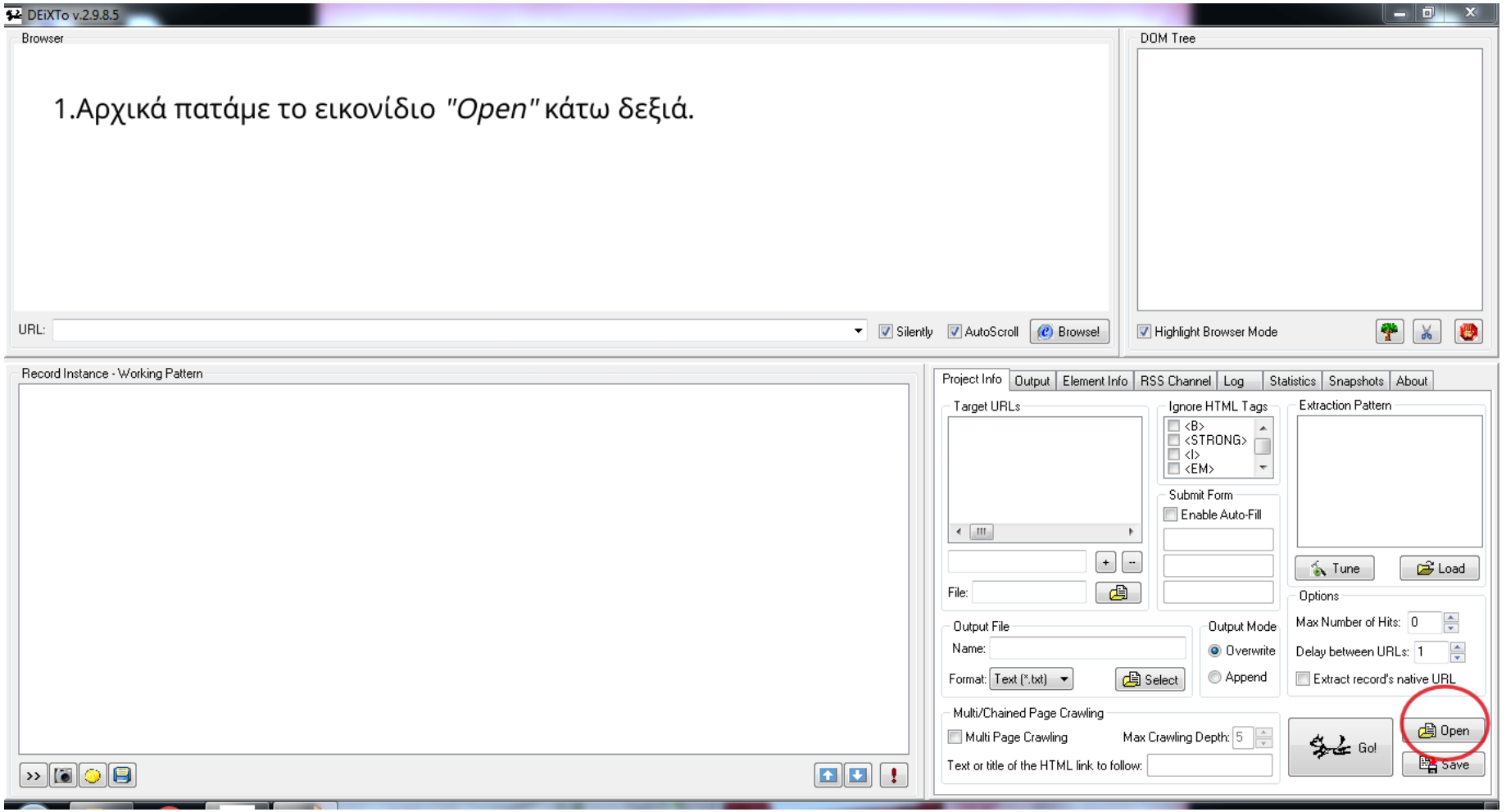
Το εργαλείο DeiXTo

- Αφού εργαστούμε πάνω στο δένδρο της σελίδας, έχουμε ρυθμίσει δηλαδή τον κανόνα, μπορούμε να τον αποθηκεύσουμε (το αρχείο θα έχει την κατάληξη .wpf).
- Στη συνέχεια για να τον χρησιμοποιήσουμε τον φορτώνουμε και εκτελούμε την λειτουργία του ταιριάσματος για έναν αριθμό από ιστοσελίδες της ίδιας μορφής με εκείνη που χρησιμοποιήθηκε για την δημιουργία του κανόνα.
- Μπορούμε να φανταστούμε την διαδικασία σαν “κοσκίνισμα”: το “κόσκινο” αποτελεί ο κανόνας και μέσα από αυτό περνούν από έλεγχο οι ιστοσελίδες που επιθυμούμε.
- Αυτό που θα αποθηκευθεί από τις ιστοσελίδες που περνάνε επιτυχώς τον έλεγχο ταιριάσματος, είναι τα δεδομένα των κόμβων εκείνων που έχουμε ορίσει ότι θέλουμε να κρατήσουμε το περιεχόμενό τους.

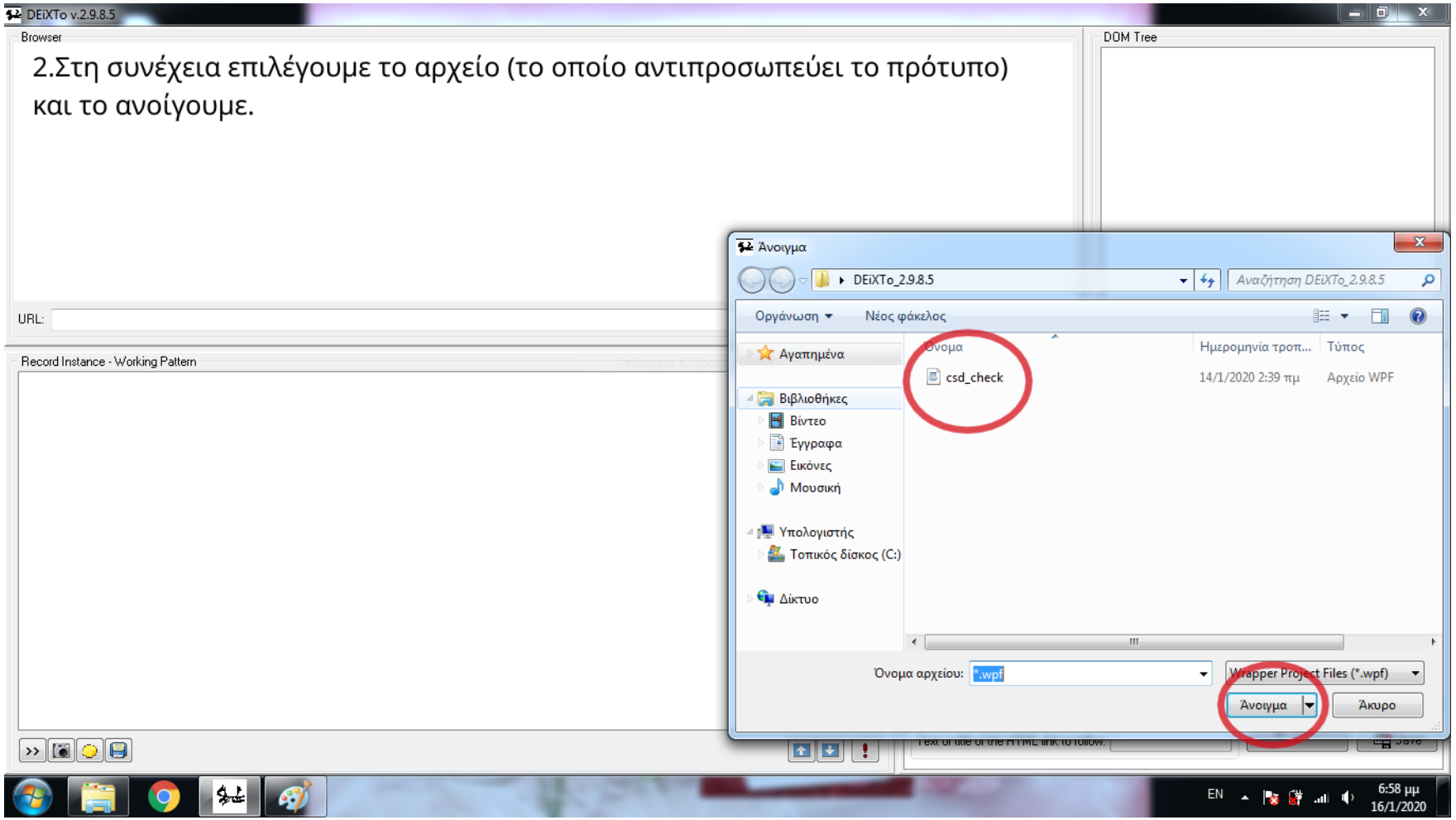
Το εργαλείο DeiXTo

- Στην περίπτωσή μας θέλουμε να επιστρέφεται το περιεχόμενο των κόμβων που αναπαριστούν τα πεδία περιγραφής (π.χ. μέθοδοι αξιολόγησης φοιτητών).
- Τα δεδομένα μπορούν να αποθηκευθούν σε μορφή απλού κειμένου, XML ή RSS.
- Για την εργασία, έχει δημιουργηθεί ήδη ο κανόνας οπότε αρκεί η εκτέλεση του προγράμματος δίνοντας ως είσοδο τις διευθύνσεις των ιστοσελίδων προς έλεγχο.
- Ακολουθεί η διαδικασία χρήσης του κανόνα στο DeiXTo για την εξαγωγή των δεδομένων από τις σελίδες περιγραφής μαθημάτων.

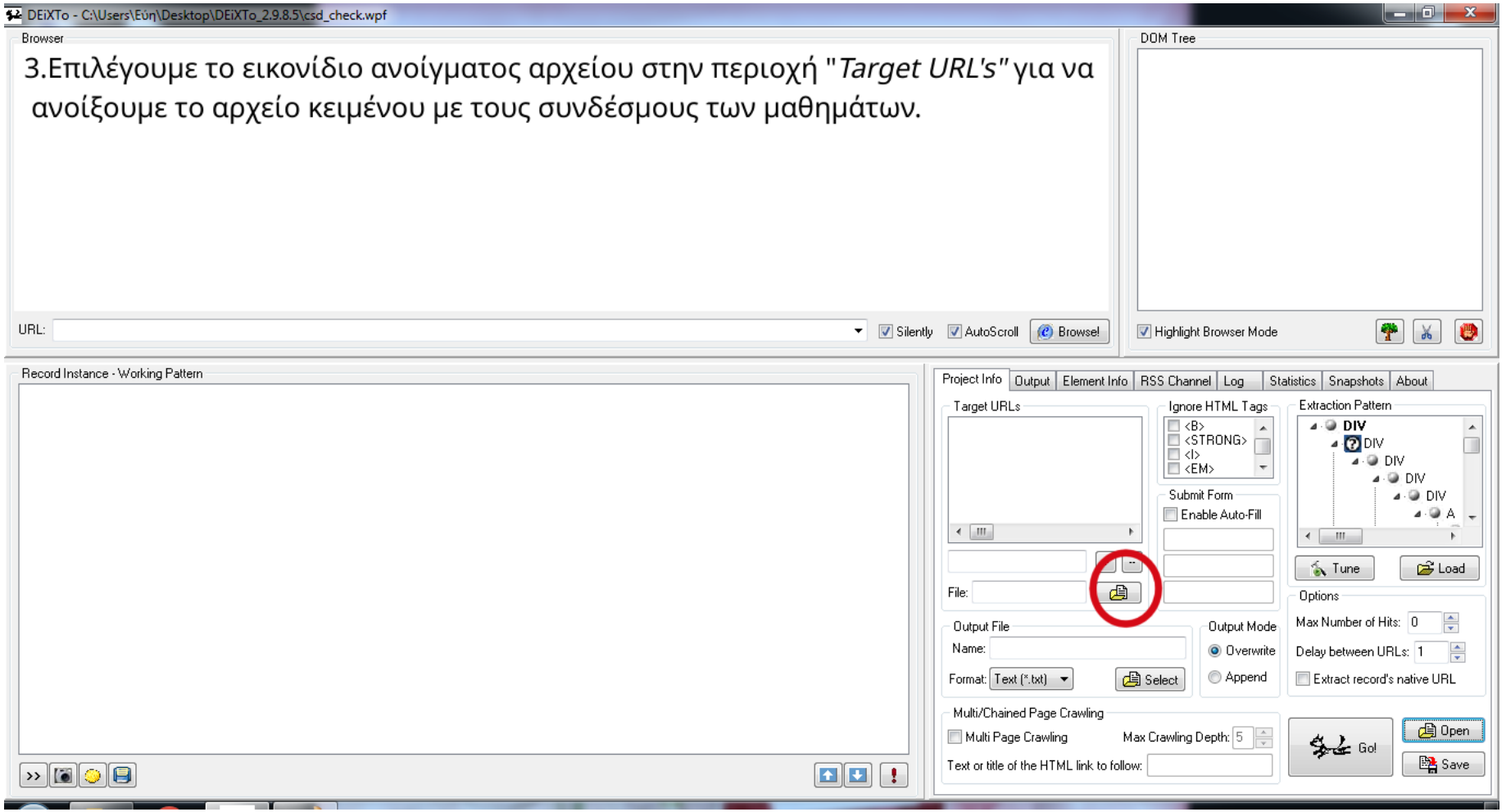
Εκτέλεση του κανόνα εξαγωγής (1)



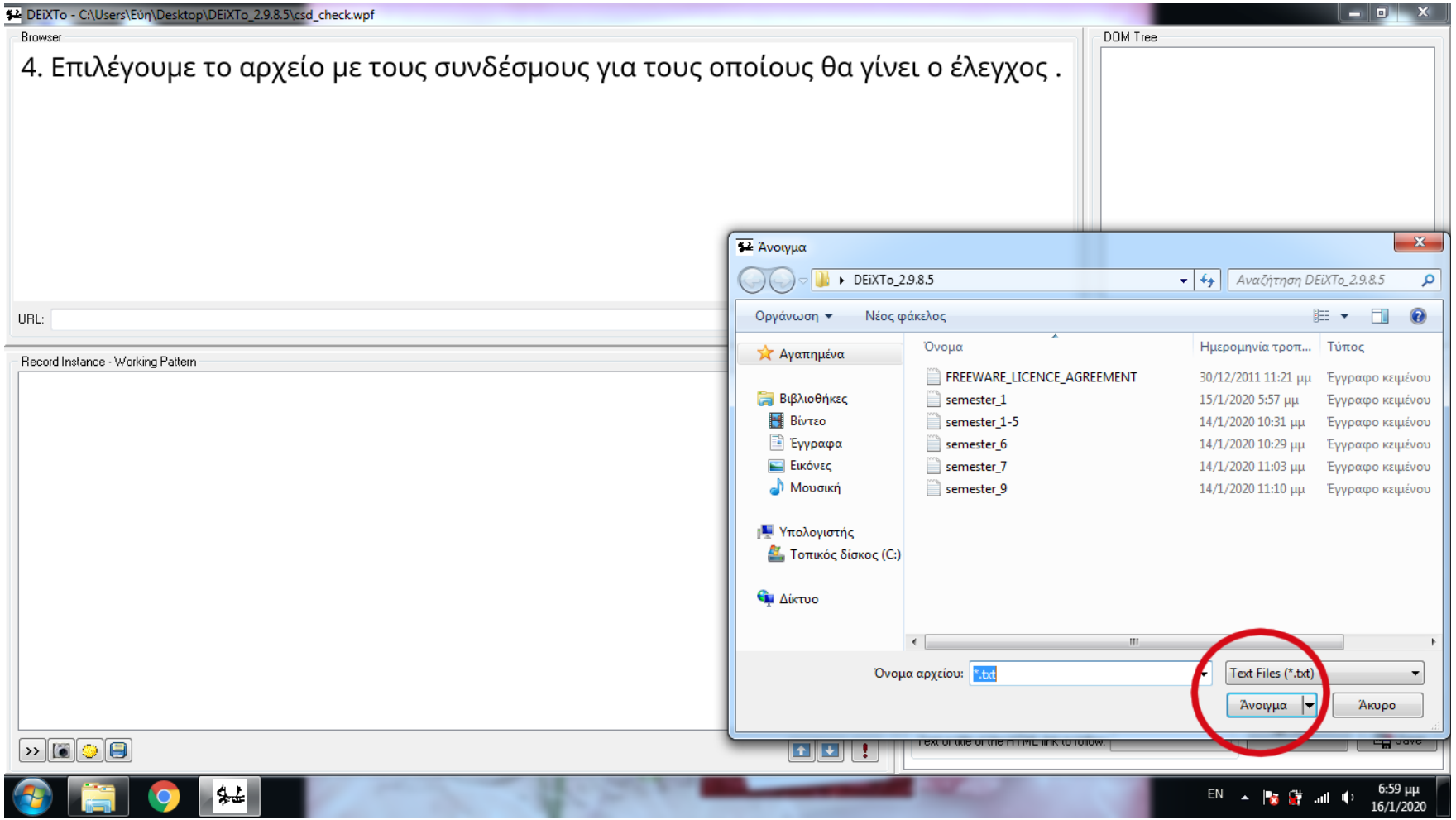
Εκτέλεση του κανόνα εξαγωγής (2)



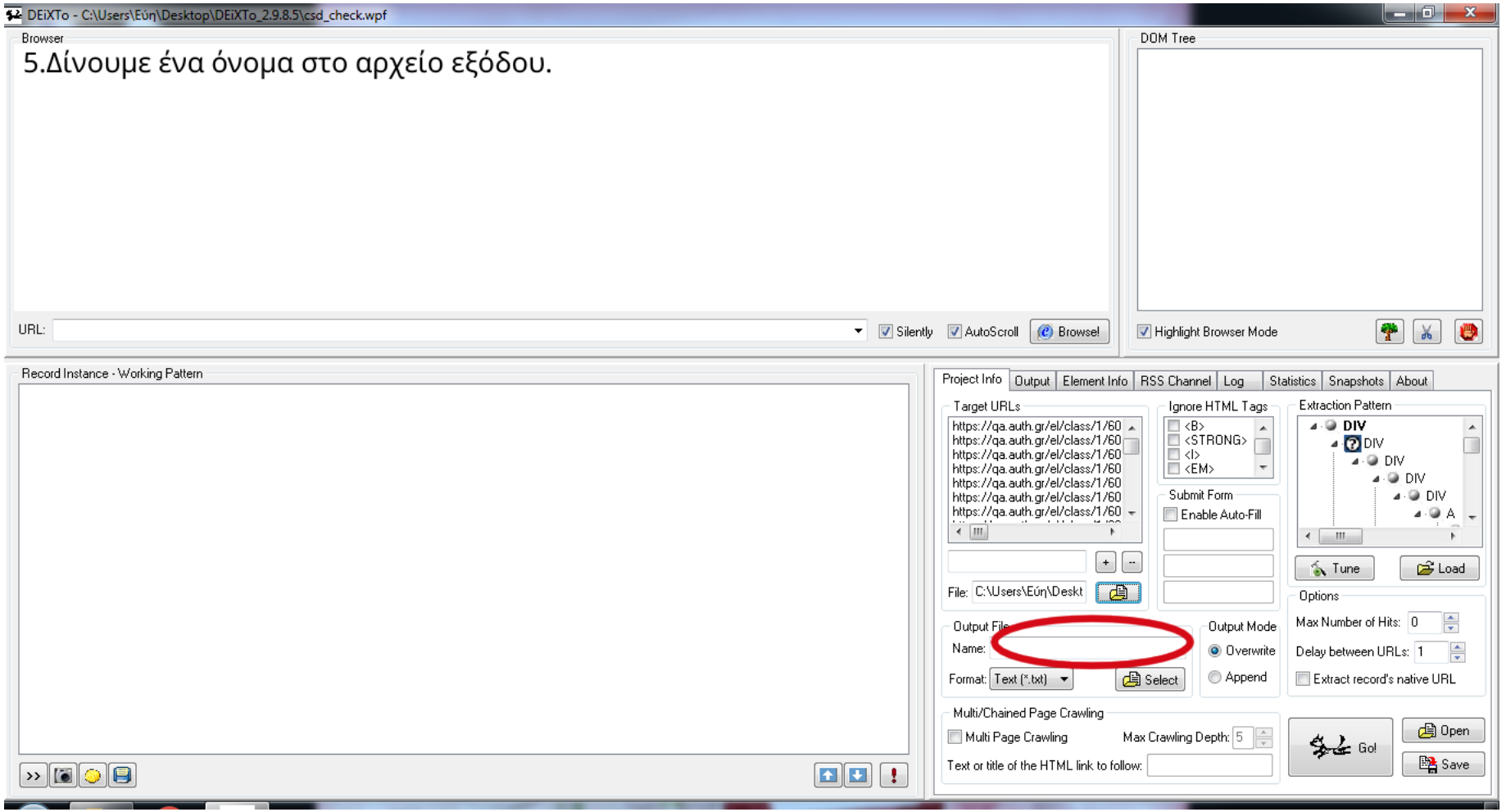
Εκτέλεση του κανόνα εξαγωγής (3)



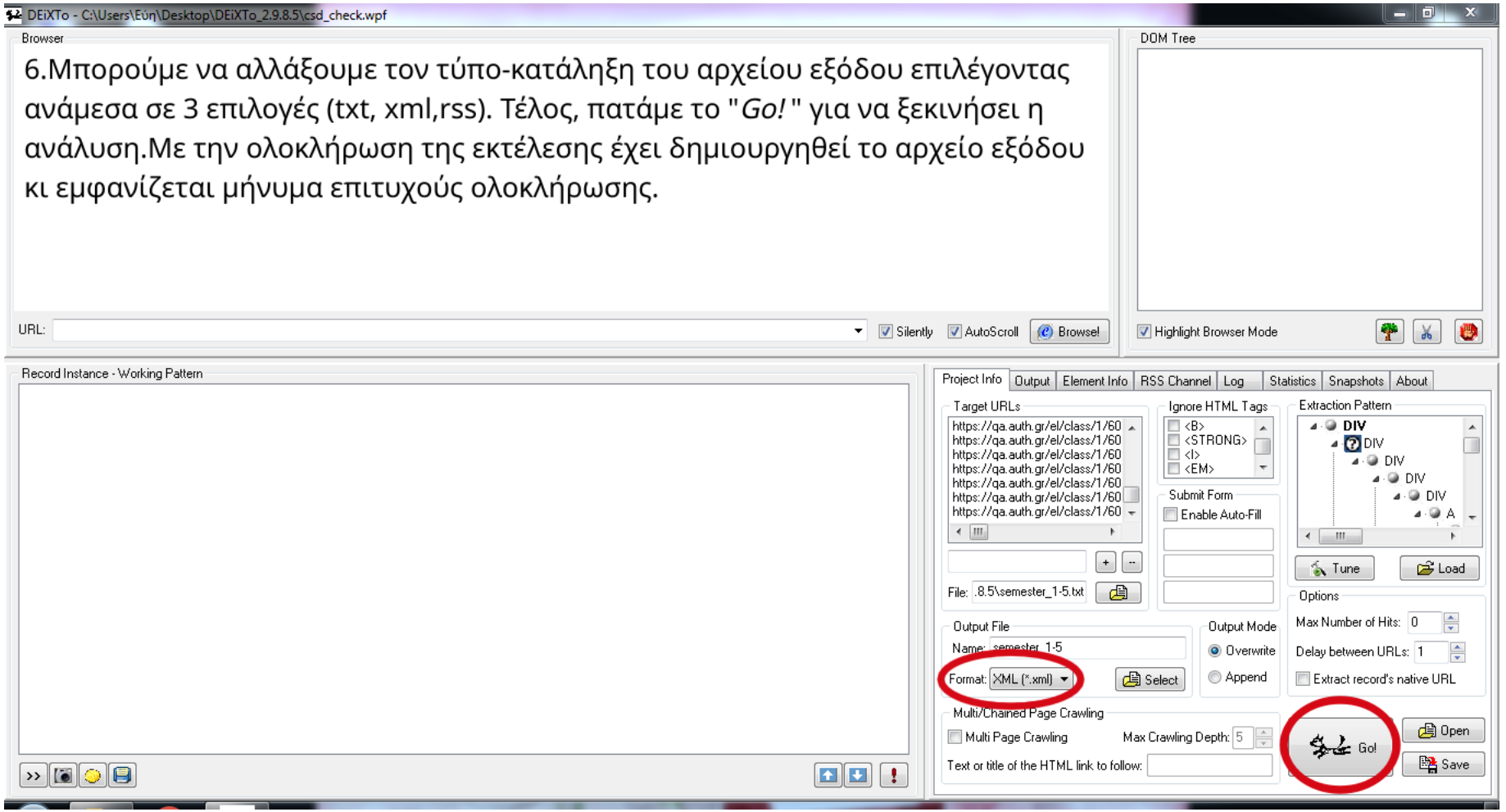
Εκτέλεση του κανόνα εξαγωγής (4)



Εκτέλεση του κανόνα εξαγωγής (5)



Εκτέλεση του κανόνα εξαγωγής (6)



Το πρόγραμμα ελέγχου

- Επειδή δεν μπορούσαν να περάσουν από το ταίριασμα οι ιστοσελίδες στην περίπτωση που θέλαμε να γίνεται η εξαγωγή ανά κόμβο με βάση το πεδίο περιγραφής , η ρύθμιση του κανόνα έγινε έτσι ώστε να επιστρέφεται σε μια ετικέτα το όνομα του μαθήματος και σε μια άλλη όλο το περιεχόμενο της περιγραφής του.
- Οπότε δημιουργήθηκε κώδικας σε Java ώστε το πρόγραμμα ελέγχου να δέχεται το xml αρχείο με τα αποτελέσματα του ελέγχου του DeiXTo και ένα αρχείο κειμένου με τα πεδία για το οποία θέλουμε να γίνει ο έλεγχος για το αν υπάρχουν στην σελίδα περιγραφής.
- Η μορφή του προγράμματος είναι jar αρχείο που εκτελείται από το τερματικό και τα ονόματα των δύο αρχείων δίνονται ως παράμετροι κατά την εκτέλεση.
- Το αποτέλεσμα είναι η δημιουργία αρχείου κειμένου, στο οποίο αναγράφεται για κάθε μάθημα τα πεδία που λείπουν από την περιγραφή.

Το πρόγραμμα ελέγχου

Αποτελείται από 6 κλάσεις:

- *Course*: για την αναπαράσταση των μαθημάτων.
- *XMLParserSAX* και *MyHandler :* Για τη διαχείριση του αρχείου xml και την εξαγωγή των δεδομένων από αυτό.
- *Controller*:όπου γίνεται ο έλεγχος για την πληρότητα των πληροφοριών στην σελίδα περιγραφής του μαθήματος.
- *File* :για το άνοιγμα των αρχείων εισόδου και την δημιουργία του αρχείου εξόδου.
- *Start*: για την λήψη των παραμέτρων από το τερματικό και την εκκίνηση του προγράμματος.

Συμπεράσματα και οφέλη

- Γνωριμία με ένα λογισμικό ανοικτού κώδικα, ειδικού σκοπού και την ιστορία του (επικοινωνία μέσω email με έναν από τους δύο υπεύθυνους του DeiXTo , τον κ. Φώτη Κόκκορα).
- Γνωριμία με μια σημαντική γνωστική περιοχή , αυτή της εξαγωγής δεδομένων ιστού (Web Mining) .
- Γνωριμία με τα αρχεία xml και τη διαχείρισή τους στη Java.
- Δυνατότητα διερεύνησης μεθόδων βελτίωσης του τρόπου που διαχειριζόμαστε τα δεδομένα στο πανεπιστήμιο και τρόπων αξιοποιήσής τους.

Βιβλιογραφία κι επίλογος

- Πληροφορίες από τη Βικιπαίδεια για την περιγραφή του DOM.
- Πληροφορίες από το εγχειρίδιο του DeiXTo.
- Το αποθετήριο του προγράμματος ελέγχου είναι δημόσιο: https://github.com/terilias/check_qa