ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ:
«ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΩΝ ΣΕ ΒΙΒΛΙΟΘΗΚΕΣ, ΑΡΧΕΙΑ, ΜΟΥΣΕΙΑ»

ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ, ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΗΣΗΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΤΙΚΩΝ, ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΚΟΙΝΩΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

DEPARTMENT OF ARCHIVAL, LIBRARY AND INFORMATION STUDIES
SCHOOL OF MANAGEMENT, ECONOMICS AND SOCIAL SCIENCES

Διπλωματική Εργασία

Συγκέντρωση και καταγραφή ελληνικών γλωσσικών πόρων του ΠαΔΑ

Συγγραφέας

Αγγελική Μπαμνιώτη (ΑΜ: mslam206682018)

Επιβλέπων: Σαράντος Καπιδάκης

Αθήνα, Δεκέμβριος 2022

ΕΠΙΤΡΟΠΗ ΕΞΕΤΑΣΗΣ

1. Σαράντος Καπιδάκης

Sarantos Kapidakis
2023.01.25
23:17:16 Z

2. Ιωάννης Τριανταφύλλου

Ioannis Triantafyllou

Digitally signed by Ioannis Triantafyllou
Date: 2023.01.25 16:53:51 +02'00'

3. Γεώργιος Γιαννακόπουλος

Georgios Giannakopoulos

Digitally signed by Georgios Giannakopoulos
Date: 2023.01.25 22:51:21 +02'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Η κάτωθι υπογεγραμμένη Αγγελική Μπαμνιώτη του Κωνσταντίνου με αριθμό μητρώου 206682018, φοιτήτρια του Προγράμματος Μεταπτυχιακών Σπουδών «Διαχείριση Πληροφοριών σε Βιβλιοθήκες, Αρχεία, Μουσεία του Τμήματος Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης της Σχολής Διοικητικών, Οικονομικών και Κοινωνικών επιστημών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι: «Είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Αθήνα, 31/12/2022

Αγγελική Μπαμνιώτη

Ευχαριστίες – Αφιερώσεις

Ευχαριστώ ιδιαίτερα τον επιβλέποντα καθηγητή μου, κύριο Σαράντο Καπιδάκη, για τη στήριξη και βοήθεια του, στην εκπόνηση της παρούσας διπλωματικής εργασίας. Επίσης ευχαριστώ πολύ τους καθηγητές του τμήματος «Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης» που με βοήθησαν στη δύσκολη και επίπονη προσπάθεια μου να συλλέξω το απαραίτητο υλικό, όπως και στους υπόλοιπους καθηγητές του Πανεπιστημίου Δυτικής Αττικής που με συνέδραμαν εξίσου. Ευχαριστώ την ομάδα του Clarin:el που ήταν πάντα πρόθυμοι να με βοηθήσουν σε όποια απορία και η συνεργασία μαζί τους ήταν άριστη.

Η παρούσα διπλωματική εργασία αφιερώνεται στην οικογένεια μου, και ιδιαίτερα στη μητέρα μου Σοφία, που πάντα ήθελε να μας προσφέρει ό,τι καλύτερο, ακόμα και με προσωπικές θυσίες.

Δεκέμβριος 2022

Αγγελική Μπαμνιώτη

Περίληψη στα ελληνικά

Ως γλωσσικός πόρος νοείται οποιοδήποτε σύνολο δεδομένων σε κάθε μορφή, σχετιζόμενο με τη γλώσσα, σε δομημένη ή αδόμητη μορφή. Το περιεχόμενο τους μπορεί να είναι πρωτογενές, επεξεργασμένο, πόροι οργανωμένης γνώσης ή ακόμα να εμπίπτει στην κατηγορία των γλωσσικών τεχνολογιών. Η συλλογή και καταγραφή γλωσσικών πόρων, εκτός από τη διάχυση της γνώσης και την προβολή του έργου των δημιουργών τους, συμβάλει σημαντικά στην εξέλιξη των γλωσσικών τεχνολογιών, οι οποίες αναπτύσσουν διάφορα εργαλεία και εφαρμογές γλωσσικής ανάλυσης και επεξεργασίας. Στη συγκεκριμένη εργασία συγκεντρώθηκαν και καταγράφηκαν ελληνικοί γλωσσικοί πόροι οι οποίοι έχουν παραχθεί στα πλαίσια του διδακτικού και ερευνητικού έργου του Πανεπιστημίου Δυτικής Αττικής, από τους διδάσκοντες, ερευνητές ή φοιτητές του. Στη συνέχεια, οι συγκεκριμένοι γλωσσικοί πόροι οργανώθηκαν, περιεγραφήκαν και τεκμηριώθηκαν στην ελληνική εκδοχή της ευρωπαϊκής διαδικτυακής υποδομής του Clarin. Το Clarin συσσωρεύει γλωσσικούς πόρους, τεχνολογίες και υπηρεσίες, σε διάφορες γλώσσες, με στόχο τη διάθεση τους προς την ερευνητική κοινότητα και τον απλό ιδιώτη. Το υλικό μπορεί να καταστεί επεξεργάσιμο μέσω διαφόρων γλωσσικών τεχνολογιών. Σημαντικός αριθμός πανεπιστήμιων και ερευνητικών κέντρων της Ελλάδας διαθέτουν ήδη ψηφιακό αποθετήριο στο Clarin, το οποίο φιλοξενεί τους παραγόμενους γλωσσικούς τους πόρους. Μέσω της εκπόνησης της συγκεκριμένης μεταπτυχιακής εργασίας, επιχειρήθηκε η δημιουργία ανάλογου ψηφιακού αποθετηρίου, μέσα στην υποδομή του Clarin, για το Πανεπιστήμιο Δυτικής Αττικής. Το υλικό συλλέχθηκε, επεξεργάστηκε, περιγράφηκε, τεκμηριώθηκε και έγινε προσβάσιμο προς την επιστημονική κοινότητα. Η συλλογή των πόρων πραγματοποιήθηκε έπειτα από επικοινωνία με τους δημιουργούς τους, οι οποίοι είναι και κάτοχοι των δικαιωμάτων διάθεσης τους. Επίσης συμπεριλήφθηκε υλικό που διατίθεται ήδη με ελεύθερες μορφές πνευματικών δικαιωμάτων και έχει παραχθεί στα πλαίσια του εκπαιδευτικού έργου του πανεπιστημίου. Η διάθεση του υλικού στο Clarin γίνεται με άδειες ανοιχτής πρόσβασης Creative Commons, σεβόμενοι την επιθυμία των δημιουργών του. Έπειτα από μια χρονοβόρα και επίπονη διαδικασία συλλέχθηκε υλικό από μεγάλο αριθμό δημιουργών, οι οποίοι εκπροσωπούν την πλειοψηφία των σχολών και τμημάτων των δύο ΤΕΙ, Αθηνών και Πειραιά, με τη συγχώνευση των οποίων δημιουργήθηκε το Πανεπιστήμιο Δυτικής Αττικής, καθώς και από το ίδιο το ΠαΔΑ. Συνολικά συλλέχθηκαν 193 γλωσσικοί πόροι. Αναφορικά με το ΤΕΙ Αθηνών, ενσωματώθηκε στην υποδομή του Clarin:el υλικό από 5 σχολές και 19 τμήματα, ενώ από το ΤΕΙ Πειραιά, υλικό από 2 σχολές και 8 τμήματα. Το υλικό από το ΠαΔΑ καλύπτει 4 σχολές και 5 τμήματα. Το

συγκεκριμένο υλικό, που βρίσκεται εξολοκλήρου σε μορφή κειμένου (text), αποτελείται από 188 σώματα κειμένου και 5 λεξικό /εννοιολογικούς πόρους. Οι 189 γλωσσικοί πόροι είναι μονόγλωσσοι και οι 4 δίγλωσσοι, ενώ οι εκπροσωπούμενες γλώσσες είναι τα ελληνικά, τα αγγλικά και η τοπική διάλεκτο της Μεσσηνίας.

Clarin	Μεταδεδομένα
Γλωσσικοί Πόροι	Τεκμηρίωση
Γλωσσικές Τεχνολογίες	Πανεπιστήμιο Δυτικής Αττικής
Γλωσσική Επεξεργασία	Ψηφιακές Ανθρωπιστικές Επιστήμες

Περίληψη στα αγγλικά

A language resource is any set of data in any form, language-related, structured, or unstructured. Their content can be raw, processed, organized knowledge resources, or even fall into the category of language technologies. The collection and recording of language resources, in addition to the dissemination of knowledge and the promotion of the work of their creators, contributes significantly to the evolution of language technologies, which develop various tools and applications of language analysis and processing. In this specific work, Greek language resources which have been produced in the context of the teaching and research work of the University of West Attica, by its teachers, researchers or students were collected and recorded. The language resources were then organized, described, and documented in the Greek version of Clarin's European web infrastructure. Clarin accumulates language resources, technologies and services, in various languages, with the aim of making them available to the research community and the ordinary individual. The material can be made editable through various language technologies. A significant number of universities and research centers in Greece already have a digital repository in Clarin, which hosts their produced language resources. Through the elaboration of this postgraduate thesis, the creation of a similar digital repository, within the infrastructure of Clarin, was attempted for the University of West Attica. The material was collected, processed, described, documented, and made accessible to the scientific community. The collection of the resources was carried out after contacting their creators, who are also the owners of their distribution rights. Also included was material that is already available in free forms of copyright and has been produced in the context of the university's educational work. The material on Clarin is made available under Creative Commons open access licenses, respecting the wishes of its creators. After a time-consuming and laborious process, material was collected from many authors, who represent the majority of schools and departments of the two TEIs, Athens and Piraeus, with the merger of which the University of West Attica was created, as well as from the PADA itself. A total of 193 language resources were collected. Regarding the TEI of Athens, material from 5 schools and 19 departments was integrated with the Clarin:el infrastructure, while from TEI Piraeus, material from 2 schools and 8 departments. The material from the PADA covers 4 schools and 5 departments. The material, which is entirely in text format, consists of 188 text corpuses and 5 lexical /conceptual resources. The 189 language resources are monolingual and 4 are bilingual, while the languages represented are Greek, English and the local dialect of Messinia.

Keywords:

Clarin
Language Resources
Language Technologies
Language Processing
Metadata
Documentation
University of West Attica
Digital Humanities

Πίνακας περιεχομένων

Κεφάλαιο 1. Εισαγωγή

Οι γλωσσικοί πόροι είναι σύνολα δεδομένων που χαρακτηρίζονται από ποικίλες μορφές. Μπορεί να είναι δομημένοι ή αδόμητοι, πρωτογενείς, επεξεργασμένοι, σε οργανωμένη μορφή ή σχετιζόμενοι με εργαλεία γλωσσικών τεχνολογιών. Ορισμένα χαρακτηριστικά παραδείγματα είναι τα σώματα κειμένου, τα γλωσσάρια, τα μονόγλωσσα ή πολύγλωσσα λεξικά, οι θησαυροί κτλ. Αποτελούν σημαντική παρακαταθήκη της γλώσσας μας αλλά και εργαλεία για τη διαμόρφωση και εκπαίδευση έξυπνων διαδικασιών καθώς και την ανάπτυξη τεχνολογιών γλωσσικής επεξεργασίας, όπως είναι για παράδειγμα η γραμματική ή συντακτική ανάλυση, η αυτόματη μετάφραση, η κατανόηση κειμένου, η εξαγωγή περιλήψεων, κλπ. Η περιγραφή τους γίνεται με ιδιαίτερο τρόπο που τους επιτρέπει να είναι αναζητήσιμοι και εκμεταλλεύσιμοι σε υπολογιστικό περιβάλλον.

Στόχος της παρούσας μεταπτυχιακής εργασίας είναι να διερευνήσει ποιοι είναι οι γλωσσικοί πόροι που έχουν παραχθεί μέσα στα πλαίσια της λειτουργίας και εκπαιδευτικού και ερευνητικού έργου του Πανεπιστημίου Δυτικής Αττικής, από τους διδάσκοντες ή ενδεχομένως και τους φοιτητές του, ανά τμήμα και σχολή. Θα διερευνηθεί ποια είναι η μορφή και το περιεχόμενο των σχετικών πόρων, καθώς και με ποιους όρους είναι διαθέσιμοι στο κοινό από τους δημιουργούς τους. Στη συνέχεια, θα καταγραφούν με κατάλληλο τρόπο και αναφορτωθούν (ή συνδεθούν) με το αντίστοιχο ιδρυματικό αποθετήριο του Πανεπιστημίου Δυτικής Αττικής στην ελληνική εκδοχή του Clarin. Το Clarin είναι μια ευρωπαϊκή διαδικτυακή υποδομή που συσσωρεύει γλωσσικούς πόρους, τεχνολογίες και υπηρεσίες, σε διάφορες γλώσσες, ώστε να τους διαθέσει προς την ερευνητική κοινότητα γενικότερα και της γλωσσολογίας ειδικότερα αλλά και τον απλό ιδιώτη, προς προώθηση της γνώσης και επεξεργασία του υλικού μέσω διαφόρων γλωσσικών τεχνολογιών (https://www.clarin.eu/).

Σημαντικός αριθμός πανεπιστήμιων και ερευνητικών κέντρων της Ελλάδας διαθέτουν ήδη ψηφιακό αποθετήριο στο Clarin:el, το οποίο φιλοξενεί τους παραγόμενους γλωσσικούς τους πόρους. Μέσω της εκπόνησης της παρούσας μεταπτυχιακής εργασίας, επιχειρείται η δημιουργία ανάλογου ψηφιακού αποθετηρίου, μέσα στην υποδομή του Clarin:el, για το Πανεπιστήμιο Δυτικής Αττικής (https://www.clarin.eu/). Η εργασία περιλαμβάνει και περιγράφει αναλυτικά όλα τα στάδια της διερεύνησης, καταγραφής και συγκέντρωσης του υλικού, καθώς και την επικοινωνία με τους σχετικούς δημιουργούς των γλωσσικών πόρων, προς διευκρίνιση, μεταξύ άλλων, των όρων διάθεσης τους και μεταφόρτωσης τους στη

γλωσσική υποδομή του Clarin:el. Επίσης παραθέτει τη μορφή των συλλεγόμενων πόρων και των μεταδεδομένων τους.

1.1 Κίνητρο της έρευνας

Μεγάλος αριθμός πανεπιστημιακών ιδρυμάτων και ερευνητικών κέντρων της Ελλάδας διαθέτουν ιδρυματικό ψηφιακό αποθετήριο στη διαδικτυακή υποδομή του Clarin:el. Η συγκεκριμένη διασύνδεση τους επιτρέπει να προβάλλουν το έργο τους, να οργανώνουν τους γλωσσικούς τους πόρους σε ένα ψηφιακό, υπολογιστικό περιβάλλον, να διασυνδέονται με την επιστημονική κοινότητα ή να γίνονται μέλη μιας ευρύτερης κοινωνίας που επικοινωνεί και που υφίσταται σε διεθνές επίπεδο, με πολλά οφέλη και προνόμια στην προώθηση της γνώσης. Μέχρι πριν την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας, δεν είχαν διατεθεί γλωσσικοί πόροι κατά τη διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με τη διεθνή γλωσσική υποδομή του Clarin, μέσω του ελληνικού της παραρτήματος, που είναι το Clarin:el. Θα μπορούσαμε να συμπεράνουμε οπότε ότι το βαθύτερο κίνητρο της συγκεκριμένης ερευνητικής εργασίας είναι η διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με τη συγκεκριμένη εφαρμογή, και η λεπτομερής καταγραφή και περιγραφή της διεργασίας που πραγματοποιήθηκε.

1.2 Πλαίσιο, σκοπός και στόχοι της διπλωματικής εργασίας

Σκοπός της παρούσας διπλωματικής εργασίας είναι η συγκέντρωση και καταγραφή των ελληνικών γλωσσικών πόρων, οι οποίοι έχουν παραχθεί στα πλαίσια του ερευνητικού και εκπαιδευτικού έργου του Πανεπιστημίου Δυτικής Αττικής, καθώς και η διασύνδεση τους με τη διαδικτυακή υποδομή του Clarin, ώστε το υλικό να οργανωθεί, να τεκμηριωθεί και να περιγραφεί με τρόπο που θα το καταστήσει αναζητήσιμο και αξιοποιήσιμο από την ερευνητική κοινότητα ή το απλό κοινό, με στόχο την προώθηση της γλωσσικής επεξεργασίας.

Αναλυτικότερα, οι κύριοι στόχοι της έρευνας έχουν ως εξής:

- Συγκέντρωση των γλωσσικών πόρων που έχουν παραχθεί από το Πανεπιστήμιο Δυτικής Αττικής
- Καταγραφή των συγκεκριμένων πόρων

- Διασαφήνιση των όρων με τους οποίους παρέχονται και χρησιμοποιούνται οι πόροι από τους χρήστες
- Οργάνωση και ταξινόμηση των πόρων ώστε να είναι ανακτήσιμοι από την επιστημονική κοινότητα και το ενδιαφερόμενο κοινό, με τα ανάλογα δικαιώματα μέσω της καταγραφής
- Διασύνδεση και τεκμηρίωση των πόρων μέσω του Clarin, το οποίο είναι μία πανευρωπαϊκής εμβέλειας προσπάθεια να συγκεντρωθούν, να συντονιστούν και να διατεθούν στην ερευνητική κοινότητα γλωσσικοί πόροι, τεχνολογίες και υπηρεσίες σε όλες τις γλώσσες, μέσω μιας διαδικτυακής Ερευνητικής Υποδομής που θα περιλαμβάνει και εργαλεία γλωσσικής επεξεργασίας
- Εμπλουτισμός και οργάνωση του ψηφιακού αποθετηρίου του Πανεπιστημίου Δυτικής Αττικής στην υποδομή του Clarin:el.

1.3 Μεθοδολογία

Η παρούσα πτυχιακή εργασία συνίσταται στη συγκέντρωση των διαφόρων πόρων που παράχθηκαν ανά τμήμα από το Πανεπιστήμιο Δυτικής Αττικής, την καταγραφή και τεκμηρίωση τους, σύμφωνα με το σχήμα του Clarin για γλωσσικούς πόρους. Η ακολουθούμενη μεθοδολογία αποτελεί έρευνα πεδίου για τη συγκέντρωση του υλικού. Οι πιθανοί παραγωγοί γλωσσικών πόρων και κάποιο από το ζητούμενο υλικό εντοπίστηκαν μέσω της ιστοσελίδας των τμημάτων του Πανεπιστημίου Δυτικής Αττικής. Επικοινωνήσαμε με τα μέλη του ΠΑΔΑ που είναι οι παραγωγοί των πόρων, μέσω των αντίστοιχων γραμματειών των τμημάτων, των ιστοσελίδων των σχολών και τμημάτων, με μήνυμα ηλεκτρονικού ταχυδρομείου, μέσω συνάντησης στο teams, τηλεφωνικά ή δια ζώσης. Επίσης, απευθείας, μέσω του ψηφιακού αποθετηρίου Πολυνόη του πανεπιστημίου ή του διαδικτύου εντοπίσαμε μεγάλο αριθμό γλωσσικών πόρων που παράχθηκαν από το ΠΑΔΑ.

Κάποιο γράμμα – κοινή φόρμα αποστάλθηκε με ηλεκτρονικό ταχυδρομείο, με στόχο να ενημερώσει τους παραγωγούς των αναζητούμενων πόρων για τη έρευνα μας και τους πιθανούς της στόχους. Προβήκαμε σε ανταλλαγή μηνυμάτων ηλεκτρονικού ταχυδρομείου. Πραγματοποιήθηκε κάποιος αριθμός ενημερωτικών συνεντεύξεων με τους πιθανούς παραγωγούς των πόρων με στόχο να διερευνηθεί αν όντως διαθέτουν κατάλληλο υλικό, να εξηγήσουν τις επιλογές τους και να συμβάλλουν στη βέλτιστη τεκμηρίωση των πόρων. Στη συνέχεια συλλέχθηκαν, τεκμηριώθηκαν οι πόροι και απέκτησαν πρόσβαση και διασύνδεση με το Clarin:el.

Επίσης συντάχθηκε ένα κείμενο με συχνές ερωταποκρίσεις. Το συγκεκριμένο κείμενο παρατίθεται στο παράρτημα της παρούσας διπλωματικής εργασίας και περιλαμβάνει είκοσι ερωτήσεις. Η σύνταξη του συγκεκριμένου εγγράφου ήταν το πρώτο πράγμα που έγινε, πριν αρχίσουμε τη συγγραφή της διπλωματικής εργασίας και τη συλλογή και διασύνδεση του υλικού με την υποδομή. Στη συνέχεια εμπλουτίστηκε με επιπλέον ερωτήσεις που προέκυψαν από την επικοινωνία με τους παραγωγούς του υλικού και τις απορίες τους. Μας βοήθησε τόσο στην αρχική κατανόηση του θέματος, όσο και στην προσέγγιση και ενημέρωση των πιθανών παραγωγών των γλωσσικών πόρων σχετικά με τους στόχους της έρευνας, την εφαρμογή του Clarin και του Clarin:el, το όφελος που θα υπάρχει με τη διάθεση των πόρων τους, κάποια απαραίτητη ορολογία, καθώς και τη μορφή υλικού που συλλέγουμε. Οι απορίες των χρηστών αφορούσαν κυρίως την υποδομή και τους στόχους της, όπως και τη μορφή και περιεχόμενο του υλικού που αναζητούμε.

1.4 Περιορισμοί

Το έργο της συλλογής και τεκμηρίωσης των γλωσσικών πόρων προς διασύνδεση με τη γλωσσική υποδομή του Clarin:el χαρακτηρίζεται από πολυπλοκότητα λόγω ποικίλων παραγόντων. Εμείς καλούμαστε να ανταπεξέλθουμε στις προκλήσεις και στα εμπόδια και να καταφέρουμε να επιτύχουμε το έργο μας. Κάποιοι από τους περιορισμούς και δυσκολίες που κληθήκαμε να αντιμετωπίσουμε είναι οι παρακάτω:

- Καχυποψία των παραγωγών των γλωσσικών πόρων ως προς τους στόχους της έρευνας
- Άγνοια σχετικά με την υποδομή του Clarin:el και το έργο του
- Φόβος για πιθανή μη ηθελημένη ευρεία διάθεση του υλικού και μη σεβασμό των πνευματικών δικαιωμάτων
- Υλικό στο οποίο οι δημιουργοί των γλωσσικών πόρων δεν έχουν πλήρως τα δικαιώματα διάθεσης τους
- Χρονικός περιορισμός
- Φόβος ότι μετά το πέρας της παρούσας διπλωματικής εργασίας, η διασύνδεση του ΠαΔΑ με το Clarin:el, μέσω του εμπλουτισμού της υποδομής με νέο υλικό, δεν θα συνεχιστεί
- Περιορισμοί ως προς τη μορφή του υλικού και τους μορφότυπους του που σε κάποιες περιπτώσεις δεν είναι υποστηριζόμενοι από την υποδομή

1.5 Οργάνωση Κεφαλαίων / Διάρθρωση της Εργασίας

Η διάρθρωση της παρούσας διπλωματικής εργασίας οργανώνεται σε πέντε κεφάλαια. Το πρώτο είναι ένα εισαγωγικό κεφάλαιο που περιγράφει τους στόχους, τους σκοπούς και τα κίνητρα της έρευνας. Στη συνέχεια, στο δεύτερο κεφάλαιο, δίδονται κάποιοι χρήσιμοι ορισμοί της χρησιμοποιούμενης ορολογίας, οι οποίοι είναι αναγκαίοι στην εμβάθυνση του θέματος. Το θεωρητικό κομμάτι της εργασίας παρατίθεται, με τις προεκτάσεις του. Το τρίτο κεφάλαιο αφιερώνεται στη μεθοδολογία της έρευνας, η οποία αναλύεται διεξοδικά. Επίσης παρουσιάζεται και αναλύεται ο τρόπος περιγραφής και τεκμηρίωσης του υλικού στην υποδομή του Clarin:el και δίδονται ενδεικτικά δύο παραδείγματα διαφορετικών τύπων υλικού που έχουν διασυνδεθεί με αυτή. Στο τέταρτο κεφάλαιο περιγράφονται αναλυτικά τα αποτελέσματα και τα ευρήματα της έρευνας. Ενώ, τέλος, στο πέμπτο και τελευταίο κεφάλαιο, αναφέρονται τα συμπεράσματα που απορρέουν από την έρευνα, γίνεται μια ανακεφαλαίωση και δίδονται κάποιες μελλοντικές προεκτάσεις οι οποίες μπορούν να ακολουθηθούν μετά το πέρας της συγκεκριμένης έρευνας, που έγινε στα πλαίσια μιας διπλωματικής εργασίας.

Κεφάλαιο 2. Θεωρητικό μέρος –Βιβλιογραφική έρευνα

Στο συγκεκριμένο κεφάλαιο θα επικεντρωθούμε στην αναλυτική παρουσίαση της υπάρχουσας βιβλιογραφίας σχετικά με θέματα που πραγματεύεται η παρούσα διπλωματική εργασία, αφού πρώτα δώσουμε κάποιους αναγκαίους ορισμούς όρων που θα συναντήσουμε στην έρευνα μας.

2.1 Ορισμοί

Αρχικά θα παραθέσουμε την επεξήγηση κάποιων συγκεκριμένων ορισμών που είναι απαραίτητη για την κατανόηση και περαιτέρω εμβάθυνση του θέματος της διπλωματικής εργασίας, οι οποίοι περιγράφονται παρακάτω.

2.1.1 Clarin

Το Clarin είναι μια ευρωπαϊκή διαδικτυακή υποδομή η οποία συγκεντρώνει γλωσσικούς πόρους, τεχνολογίες και υπηρεσίες, σε διάφορες γλώσσες, με σκοπό να τους διαθέσει προς την ερευνητική κοινότητα ή και τον απλό ιδιώτη. Είναι οργανωμένο σε κατά τόπους και γλώσσα υποδομές του Clarin. Το υλικό που συσσωρεύει τεκμηριώνεται κατάλληλα και μπορεί να είναι επεξεργάσιμο μέσω γλωσσικών τεχνολογιών. Παρέχει επίσης εκπαίδευση αναφορικά με τις γλωσσικές τεχνολογίες και οργανώνει παγκοσμίως διάφορες δράσεις. Το υλικό από τα κατά τόπου παραρτήματα του, συλλέγεται και διασυνδέεται με την κεντρική υποδομή του Clarin (https://www.clarin.eu/).

2.1.2 Clarin : el

Το CLARIN:EL είναι το ελληνικό παράρτημα της ευρωπαϊκής υποδομής Clarin. Περιέχει κεντρικό κατάλογο, όπου εμφανίζονται προς το κοινό όλοι οι γλωσσικοί πόροι που φιλοξενεί. Οι πόροι είναι οργανωμένοι ανά ψηφιακό αποθετήριο. Συνολικά περιέχει μέχρι στιγμής δεκατρία ψηφιακά αποθετήρια, οργανωμένα ανά τον ανάλογο φορέα που εκπροσωπούν. Περιλαμβάνει οκτώ ακαδημαϊκά αποθετήρια (το όγδοο είναι αυτό που εμπλουτίσαμε για το Πανεπιστήμιο Δυτικής Αττικής, στα πλαίσια της παρούσας διπλωματικής εργασίας), τέσσερα

αποθετήρια ερευνητικών κέντρων και ένα αποθετήριο φιλοξενούμενων πόρων, στο οποίο εντάσσονται μεμονωμένοι πόροι που δεν συσχετίζονται με κάποιο συγκεκριμένο ψηφιακό αποθετήριο από τα προϋπάρχοντα. Πιο συγκεκριμένα, τα αποθετήρια που φιλοξενούνται από το Clarin:el ανήκουν στα παρακάτω ιδρύματα και σε παρένθεση αναφέρεται ο αριθμός των γλωσσικών πόρων που περιέχουν (https://www.clarin.gr/):

- Ερευνητικό Κέντρο Αθηνά (340 γλωσσικοί πόροι)
- Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (52 γλωσσικοί πόροι)
- Πανεπιστήμιο Αιγαίου (45 γλωσσικοί πόροι)
- Κέντρο Ελληνικής Γλώσσας (39 γλωσσικοί πόροι)
- Πανεπιστήμιο Κρήτης (31 γλωσσικοί πόροι)
- Ιόνιο Πανεπιστήμιο (26 γλωσσικοί πόροι)
- Πανεπιστήμιο Δυτικής Αττικής (193 γλωσσικοί πόροι)
- Αποθετήριο Φιλοξενούμενων Πόρων (18 γλωσσικοί πόροι)
- Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (17 γλωσσικοί πόροι)
- Εθνικό Κέντρο Κοινωνικών Ερευνών (13 γλωσσικοί πόροι)
- Οικονομικό Πανεπιστήμιο Αθηνών (12 γλωσσικοί πόροι)
- ΕΚΕΦΕ Δημόκριτος (8 γλωσσικοί πόροι)
- Πάντειο Πανεπιστήμιο (5 γλωσσικοί πόροι)

Επίσης το Πανεπιστήμιο Πατρών έχει διασυνδεθεί με το Clarin:el, χωρίς ωστόσο, μέχρι τώρα να έχει εμπλουτίσει το ψηφιακό του αποθετήριο ώστε να φιλοξενεί γλωσσικούς πόρους παραγόμενους από αυτό.

Η ελληνική υποδομή του Clarin έχει περισσότερους από 1.350 εγγεγραμμένους χρήστες και φιλοξενεί συνολικά 762 γλωσσικούς πόρους (https://www.clarin.gr/).

2.1.3 Άλλες παρεμφερείς υποδομές

Εκτός από το CLARIN, στη διεθνή εκδοχή του ή τα κατά τόπους παραρτήματα του, υπάρχουν και άλλες παρεμφερείς υποδομές που συλλέγουν γλωσσικούς πόρους (ανάμεσα σε άλλο υλικό του ενδιαφέροντός τους), ψηφιοποιημένο ή ψηφιακό υλικό αναφορικά με την παγκόσμια πολιτιστική κληρονομιά και προάγουν τις ανθρωπιστικές επιστήμες. Ενδεικτικά θα αναφέρουμε παρακάτω κάποια παραδείγματα από τις συγκεκριμένες υποδομές:

- *Dariah.eu*: Πρόκειται για μια ευρωπαϊκή ερευνητική υποδομή που προάγει τις τέχνες και τις ανθρωπιστικές επιστήμες, ενισχύει την έρευνα που σχετίζεται με την ψηφιακή τεχνολογία και τη διδασκαλία. Δραστηριοποιείται στη δημιουργία, ανάλυση και ερμηνεία ψηφιακών πόρων (https://www.dariah.eu/). Η Ελλάδα αποτελεί ένα από τα είκοσι μέλη της υποδομής με συνιστώσα την ελληνική εθνική υποδομή Dariah-gr/ ΔΥΑΣ που δρα σε εθνικό επίπεδο (https://dyas-net.gr/).

- *Europeana.eu*: Η Europeana αποτελεί μια ψηφιακή υποδομή που χρηματοδοτείται από την Ευρωπαϊκή Ένωση. Συλλέγει και φιλοξενεί την ψηφιακή πολιτιστική κληρονομιά της Ευρώπης (https://www.europeana.eu/). Ξεκίνησε τη λειτουργία της το Νοέμβριο του 2008. Περιλαμβάνει υλικό που περιέχουν ψηφιακές βιβλιοθήκες από φορείς όπως είναι οι βιβλιοθήκες, τα μουσεία, τα αρχεία, οι ταινιοθήκες ή άλλοι φορείς πολιτισμού στην Ευρώπη. Προάγει την ανοιχτή πρόσβαση (Καπιδάκης, 2014).

- *Απολλωνίς*: Η Απολλωνίς δημιουργήθηκε από τη σύμπραξη του Εθνικού Δικτύου Γλωσσικής Τεχνολογίας Clarin:el και του Εθνικού Δικτύου Ψηφιακών Υποδομών για τις Ανθρωπιστικές Επιστήμες Dariah-GR/ΔΥΑΣ. Στόχος της είναι η προώθηση των ψηφιακών ανθρωπιστικών επιστημών, τεχνών, γλωσσικών τεχνολογιών και καινοτομίας στην Ελλάδα. Διαθέτει συγχρηματοδότηση από την Ευρωπαϊκή Ένωση, μέσω προγράμματος ΕΣΠΑ και από εθνικούς πόρους (https://apollonis-infrastructure.gr/).

2.1.4 Γλωσσικοί πόροι

Γλωσσικός / ορολογικός πόρος είναι οποιοδήποτε σύνολο δεδομένων, σε κάθε μορφή, δομημένο ή αδόμητο, που συνδέεται με τη γλώσσα. Μπορεί να είναι πόροι με **πρωτογενές περιεχόμενο** (λόγος σε ψηφιακή ή ψηφιοποιημένη μορφή, βιβλία, κείμενα, διάφορες σημειώσεις, σημειώσεις από τα μαθήματα, σώματα κειμένων, κείμενα που προέρχονται από το διαδίκτυο, εφημερίδες, συνεντεύξεις, εκπομπές, βιντεοσκοπημένο υλικό, περιγραφές μαθημάτων κτλ.), **επεξεργασμένοι πόροι** (υποσημειώσεις που έχουν δημιουργηθεί αυτόματα ή από το χρήστη, μεταγραφές ηχογραφημένων ή βιντεοσκοπημένων αρχείων, ηλεκτρονικά εγχειρίδια κτλ.), **πόροι οργανωμένης μορφής της γνώσης** (μονόγλωσσα ή πολύγλωσσα λεξικά, γλωσσάρια, λίστες λέξεων, θησαυροί κτλ.) ή **διάφορες εφαρμογές και εργαλεία γλωσσικής τεχνολογίας** (εργαλεία λογισμικού σε κείμενα, εργαλεία εξόρυξης γνώσης, λημματοποίησης, παρουσίασης δεδομένων κτλ.) (https://www.clarin.gr/).

2.1.5 Γλωσσικές τεχνολογίες

Οι γλωσσικές τεχνολογίες είναι διάφορα υπολογιστικά εργαλεία γλωσσικής ανάλυσης μέσω των οποίων μπορούν να πραγματοποιηθούν ενέργειες όπως η ανάλυση, επισημείωση, επεξεργασία και τροποποίηση των διαφόρων γλωσσικών / ορολογικών δεδομένων (https://www.clarin.gr/).

2.1.6 Υπηρεσίες γλωσσικής επεξεργασίας

Οι υπηρεσίες γλωσσικής επεξεργασίας επιτρέπουν τη χρήση των γλωσσικών πόρων και τεχνολογιών, όπως επίσης και των εφαρμογών αυτών στο διαδίκτυο (https://www.clarin.gr/).

2.2 Βιβλιογραφική έρευνα

Έχοντας επεξηγήσει ήδη κάποιους όρους, οι οποίοι είναι απαραίτητοι στην κατανόηση της θεματικής της παρούσας διπλωματικής εργασίας, στη συνέχεια θα προβούμε σε μια βιβλιογραφική έρευνα σχετικά με το θεωρητικό υπόβαθρο της έρευνας μας.

2.2.1 Ψηφιακά αποθετήρια

Η εξέλιξη της τεχνολογίας και η έλευση της ψηφιακής εποχής, οδήγησε στη δημιουργία των λεγόμενων ψηφιακών βιβλιοθηκών. Οι ψηφιακές βιβλιοθήκες είναι οντότητες που καθιστούν διαθέσιμους στο κοινό πόρους και ψηφιακές ή ψηφιοποιημένες συλλογές τεκμηρίων, οι οποίες έχουν επιλεγεί, οργανωθεί, τεκμηριωθεί, διανεμηθεί και συντηρηθεί από εξειδικευμένο προσωπικό. Στόχος των ψηφιακών βιβλιοθηκών είναι η παροχή πληροφοριακών υπηρεσιών σε ψηφιακό περιβάλλον, οι οποίες είναι εφάμιλλες με αυτές των συμβατικών βιβλιοθηκών (Κυριάκη – Μάνεση και Κουλούρης, 2015). Η σημαντικότητα των ψηφιακών βιβλιοθηκών έγκειται στο ότι οργανώνουν το ψηφιακό τους περιεχόμενο με τρόπο ανεξάρτητο από την προσβασιμότητα του (Καπιδάκης, 2014). Πολλοί χρήστες ταυτόχρονα, σε διαφορετικά μέρη είναι δυνατόν να έχουν πρόσβαση στο ίδιο υλικό που παρέχει μια ψηφιακή βιβλιοθήκη.

Επέκταση ή γενικά εξέλιξη των ψηφιακών βιβλιοθηκών αποτελούν και τα ψηφιακά αποθετήρια. Στα αποθετήρια κατατίθεται ψηφιακό υλικό και δεδομένα για φύλαξη και διάχυση στο διαδίκτυο. Τα ακαδημαϊκά ιδρύματα της χώρας μας διαθέτουν ψηφιακά

αποθετήρια, όπου οι ερευνητές, τα μέλη του διδακτικού προσωπικού, οι φοιτητές και οι λοιποί εμπλεκόμενοι με το έκαστο πανεπιστήμιο, μπορούν ή οφείλουν να καταθέσουν τη συγγραφική και ερευνητική τους παραγωγή για φύλαξη, κρίση του έργου, τεκμηρίωση, παραγωγή μεταδεδομένων, διάχυση στο διαδίκτυο, προβολή, προαγωγή της έρευνας και χρήση. Τα χαρακτηριστικά στοιχεία ενός ιδρυματικού αποθετηρίου είναι τα εξής (Shirley, 2005):

- Έχει δημιουργηθεί από κάποιο πανεπιστήμιο ή ακαδημαϊκό ή άλλου τύπου ίδρυμα
- Το περιεχόμενο του είναι επιστημονικό
- Ο χαρακτήρας που το διέπει είναι συγκεντρωτικός. Αυτό σημαίνει ότι σκοπός και στόχος του είναι να συλλέγει το σύνολο της παραγωγής των ψηφιακών τεκμηρίων του οργανισμού μέσα στα πλαίσια του οποίου λειτουργεί
- Έχει άμεση σύνδεση με την ανοιχτή πρόσβαση αναφορικά με το περιεχόμενο του

Τα ψηφιακά αποθετήρια τα οποία έχουν δημιουργηθεί στο Clarin:el, βρίσκονται σε αναφορά και σύνδεση με το εκάστοτε ακαδημαϊκό ίδρυμα ή ερευνητικό κέντρο που διαμοιράζει το υλικό του και διαθέτουν όλα τα παραπάνω χαρακτηριστικά ενός ιδρυματικού αποθετηρίου.

2.2.2 Ψηφιακές ανθρωπιστικές επιστήμες

Με την εξέλιξη της τεχνολογίας οι ανθρωπιστικές επιστήμες μεταλλάσσονται και γίνονται προσβάσιμες στο κοινό με ποικίλους τρόπους πλέον, οι οποίοι δεν είναι μόνον οι κλασσικοί συμβατικοί. Μια σημαντική εξέλιξη είναι οι ψηφιακές ανθρωπιστικές επιστήμες οι οποίες αξιοποιούν ψηφιακούς πόρους και εργαλεία προάγοντας την έρευνα στον τομέα των ανθρωπιστικών σπουδών (Γούτσος και Φραγκάκη, 2015).

Ήδη από τη δεκαετία του ’50 άρχισαν οι πρώτες προσπάθειες αξιοποίησης της τότε τεχνολογίας στις ανθρωπιστικές επιστήμες. Ο καθολικός ιερέας Roberto Busa δημιούργησε, σε συνεργασία με την IBM, τον πρώτο ηλεκτρονικό συμφραστικό πίνακα του Θωμά Ακινάτη. Ως συμφραστικός πίνακας νοείται ο κατάλογος όλων των λέξεων ενός κειμένου. Επίσης υπήρξαν προσπάθειες μελέτης και απόδοσης πατρότητας των έργων του Shakespeare και άλλων συγγραφέων, εγχειρήματα αυτόματης μετάφρασης και εφαρμογές λεξικών, ορολογίας ή γραμματικών (Δημητρούλια και Τικτοπούλου, 2015).

Οι ψηφιακές ανθρωπιστικές επιστήμες αφορούν το σύνολο των ανθρωπιστικών και κοινωνικών επιστημών, των επιστημών της τέχνης και της φιλολογίας. Το βιβλίο *A Companion to Digital Humanities* θεωρείται σημαντικό αφού περιλαμβάνει άρθρα επιστημόνων και ερευνητών, θέτοντας την περιγραφή, τις βάσεις και το θεωρητικό και πρακτικό πλαίσιο των ψηφιακών ανθρωπιστικών επιστημών που ονοματίζονται με αυτό τον όρο για πρώτη φορά (Schreibman, Siemens and Unsworth, 2004).

Σημαντικό είναι επίσης να επισημάνουμε ότι μεγάλο μέρος της κοινότητας των επιστημόνων αναζητά γενικά την ελεύθερη πρόσβαση σε δεδομένα και μεταδεδομένα που αφορούν πόρους και υλικό, στα πλαίσια των ψηφιακών ανθρωπιστικών επιστημών, με σκοπό να προαχθεί η γνώση και επιστήμη απρόσκοπτα (Dacos, 2011).

2.2.3 Υπολογιστική γλωσσολογία

Εφαρμογές της υπολογιστικής γλωσσολογίας συναντάμε κάθε μέρα, χωρίς να έχουμε αναγνωρίσει ότι πρόκειται πράγματι για το συγκεκριμένο πεδίο. Κάποια παραδείγματα της καθημερινότητας που εμπίπτουν σε αυτή, είναι η αναζήτηση με λέξεις κλειδιά στο Google ή άλλες μηχανές αναζήτησης, οι φωνητικές εντολές που έχουμε στο κινητό τηλέφωνο, οι σελίδες στο διαδίκτυο με ομιλία ή οι οδηγίες πλοήγησης με συνθετική φωνή που χρησιμοποιούν τα αυτοκίνητα. Επίσης, υπολογιστική γλωσσολογία είναι ο ορθογραφικός έλεγχος στον επεξεργαστή κειμένου, οι αυτόματες περιλήψεις κειμένων που πραγματοποιούν οι υπολογιστές, η μηχανική μετάφραση, η επίλυση λεξικών αμφισημιών, η επίλυση συντακτικών αμφισημιών ή η επίλυση αναφορών στα κείμενα (Τάντος κ.ά., 2015).

Η υπολογιστική γλωσσολογία αποτελεί ένα διεπιστημονικό πεδίο της πληροφορικής και της γλωσσολογίας που ασχολείται με την επεξεργασία της φυσικής γλώσσας. Διάφορα επίπεδα της, όπως είναι η μορφολογία, η σύνταξη, η φωνολογία κτλ, μπορούν να υποστούν επεξεργασία η οποία έχει ως σκοπό τη δημιουργία διαφόρων υπολογιστικών εφαρμογών μέσω των οποίων οι ηλεκτρονικοί υπολογιστές θα είναι σε θέση να αναγνωρίσουν, επεξεργαστούν και παράγουν τη φυσική γλώσσα. Στο θεωρητικό επίπεδο, η υπολογιστική γλωσσολογία αναπτύσσει υπολογιστικά μοντέλα αναφορικά με τις δομές τις γλώσσας, τα οποία εφαρμόζει σε πρακτικό επίπεδο, για την ανάπτυξη λογισμικών που αναλύουν την εισαγόμενη φυσική γλώσσα, αναγνωρίζοντας τη φωνή, γραμματικές, συντακτικές ή σημασιολογικές δομές, πραγματοποιώντας ανάλυση τους και παράγοντας κάποιο γλωσσικό εξαγόμενο αποτέλεσμα, που μπορεί να είναι η παραγωγή συνθετικής φωνής, η επεξεργασία κειμένου, η μηχανική μετάφραση κτλ. (Γούτσος και Φραγκάκη, 2015). Γραμματικοί

φορμαλισμοί που αντιστοιχούν στα διάφορα επίπεδα της δομής της γλώσσας αναπτύσσονται και χρησιμοποιούνται για τη δημιουργία υπολογιστικών εφαρμογών (Μαρκόπουλος, 2006).

Η υπολογιστική γλωσσολογία, αρχικά ασχολήθηκε με την ανάλυση υλικού κειμένων με στατιστικές μεθόδους. Έπειτα προσανατολίστηκε στο θεωρητικό πλαίσιο και σύνταξη κανόνων. Ενώ τελικά επέστρεψε στην ανάλυση οργανωμένων, δομημένων σωμάτων κειμένου και τις στατιστικές προσεγγίσεις, χωρίς απαραίτητα αυτά να υπακούουν σε φορμαλιστικούς κανόνες (Dipper, 2008).

Τα κύρια χαρακτηριστικά της υπολογιστικής γλωσσολογία είναι τα εξής (Τάντος κ.ά., 2015):

- Χρησιμοποιεί αναπαραστάσεις της γλώσσας και αλγορίθμους για τη διαχείριση τους
- Πρόκειται για διεπιστημονικό πεδίο που συνδυάζει την πληροφορική, τα μαθηματικά και τη γλωσσολογία
- Ως επιστήμη, υποστηρίζει πληθώρα σημαντικών εφαρμογών και έχει πολύ μέλλον

Στην περίπτωση των γλωσσικών πόρων που διαμοιράζονται μέσω του Clarin:el και της γλωσσικής επεξεργασίας που δύνανται να υποστούν, με στόχο την επίτευξη των ήδη περιγραφόμενων αποτελεσμάτων, μπορούμε να διακρίνουμε καθαρά εφαρμογές που εμπίπτουν στον τομέα της υπολογιστικής γλωσσολογίας.

2.2.4 Τεχνητή νοημοσύνη

Η τεχνητή νοημοσύνη αναφέρεται στην ικανότητα των μηχανών, οι οποίες μέσω της τεχνολογίας, είναι δυνατόν να αναπαράγουν διάφορες λειτουργίες που προσομοιάζουν αυτές που παράγονται από το ανθρώπινο νου. Οι μηχανές, μέσω ηλεκτρονικών υπολογιστών και εξειδικευμένων αλγορίθμων δύνανται να κατανοήσουν κάποιο πρόβλημα ή κατάσταση, να προσφέρουν λύση, να επεξεργαστούν στοιχεία και να δώσουν τα επιδιωκόμενα αποτελέσματα. Η τεχνητή νοημοσύνη μπορεί να έχει πρακτικές εφαρμογές σε ποικίλους τομείς όπως είναι το διαδίκτυο, τα έξυπνα κινητά τηλέφωνα, οι αυτόματες μεταφράσεις, τα συστήματα πλοήγησης στα αυτοκίνητα, η κυβερνασφάλεια, η υγεία, οι μεταφορές, οι δημόσιες υπηρεσίες κτλ.

Η τεχνητή νοημοσύνη, μεταξύ άλλων εφαρμογών, μπορεί να έχει πρακτική εφαρμογή και με τις γλωσσικές υπηρεσίες. Στόχος της είναι να μιμηθεί τη φυσική γλώσσα, αναπαράγοντας την κωδικοποίηση και επικοινωνία. Με την ολοένα αυξανόμενη τεχνολογική εξέλιξη και τον παγκόσμιο ιστό, κατέστη δυνατή η παραγωγή και διάθεση πληθώρας ψηφιακών γλωσσικών δεδομένων, τα οποία χρησιμοποιήθηκαν στο να εξελιχθούν οι μηχανές γλωσσικής επεξεργασίας μέσω της αξιοποίησης τους. Η ταχύτατη επεξεργασία μεγάλου όγκου δεδομένων και η μεγάλη τεχνολογική εξέλιξη οδήγησε στην ανάπτυξη των μοντέλων μηχανικής μάθησης (machine learning) και βαθιάς μάθησης (deep learning), τα οποία υποστηρίζουν πράξεις όπως η αυτόματη μετάφραση, η αναγνώριση φωνής, η σύνθεση φωνής, τα διαλογικά συστήματα ή την παραγωγή γλώσσας. Ο τρόπος λειτουργίας τους είναι ουσιαστικά η εκπαίδευση των κατάλληλων, εξειδικευμένων αλγορίθμων (Γαβριηλίδου, Πιπερίδης, 2021).

Το Clarin:el συλλέγει τους ανάλογους γλωσσικούς πόρους, οι οποίοι, μέσω γλωσσικών εργαλείων και επεξεργασίας συμβάλουν στην ανάπτυξη εφαρμογών που βασίζονται στην τεχνητή νοημοσύνη και τις εξελίξεις της.

2.3 Ανακεφαλαίωση - Συμπεράσματα

Στο κεφάλαιο αυτό αναλύσαμε κάποια βασική ορολογία που σχετίζεται με το προς διερεύνηση θέμα της παρούσας διπλωματικής εργασίας. Επίσης εξετάσαμε το θεωρητικό υπόβαθρο και τα επιστημονικά πεδία στα οποία εντάσσεται η συλλογή γλωσσικού υλικού και η τεκμηρίωση του σε κάποια συγκεκριμένη υποδομή που φιλοξενεί γλωσσικούς πόρους. Μένει τώρα να αναλυθεί η πρακτική πλευρά της εφαρμογής και η μεθοδολογία υλοποίησης της. Η ψηφιακή υποδομή όπου επιχειρείται να συλλεχθεί και τεκμηριωθεί το υλικό που σχετίζεται με το Πανεπιστήμιο Δυτικής Αττικής είναι το Clarin:el.

Κεφάλαιο 3. Μεθοδολογία

Έχοντας εξετάσει το θεωρητικό πλαίσιο που διέπει τη διαχείριση και επεξεργασία των δεδομένων και τη χρησιμότητα της συλλογής γλωσσικών πόρων, στη συνέχεια θα αναλύσουμε τη μεθοδολογία που ακολουθήθηκε στην παρούσα εργασία αναφορικά με την αναζήτηση, τεκμηρίωση και διασύνδεση των γλωσσικών πόρων με την ελληνική εκδοχή της πλατφόρμας του Clarin. Θα περιγραφεί η εμπειρία της υλοποίησης του έργου, θα δοθούν οι οδηγίες για το ανέβασμα των πόρων και τέλος, θα αναλυθούν κάποιες επιλογές και περιορισμοί που λάβαμε υπόψη.

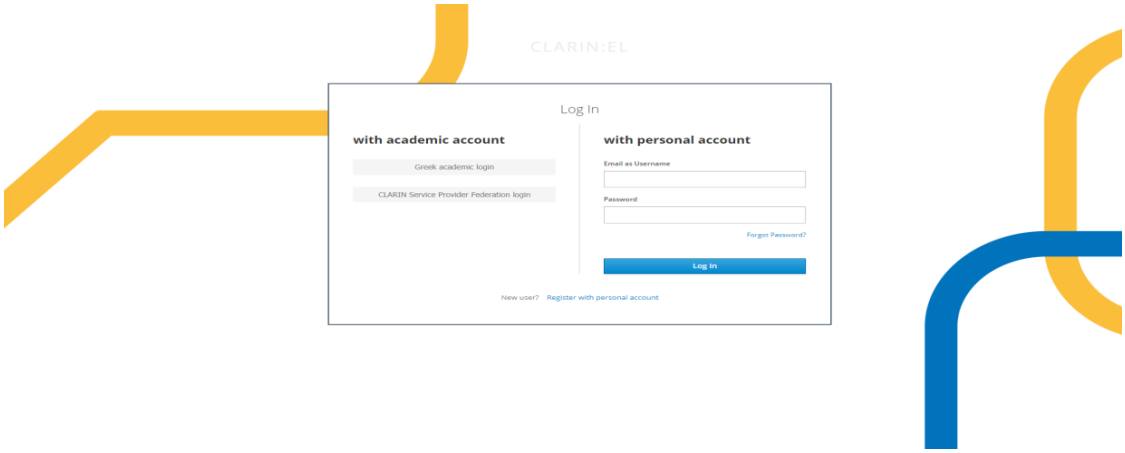
3.1 Κατανόηση θέματος και σχέδιο εργασιών

Το θέμα της παρούσας διπλωματικής εργασίας περιλαμβάνει και θεωρητικό αλλά και πρακτικό κομμάτι. Το θεωρητικό κομμάτι, εκτός από κάποια θεωρία σχετική με τη θεματολογία της διπλωματικής, περιλαμβάνει τη συγκέντρωση και διερεύνηση του υλικού, την οργάνωση της επικοινωνίας με τους δημιουργούς και την τεκμηρίωση και διάθεση μέσω συστήματος. Το πρακτικό κομμάτι είναι η εργασία που πραγματοποιήθηκε απευθείας στην εφαρμογή. Πως δηλαδή περαστήκαν τα στοιχεία της περιγραφής και τεκμηρίωσης ενός γλωσσικού πόρου.

Λόγω της πολυπλοκότητας του, ο προγραμματισμός των εργασιών που έπρεπε να πραγματοποιηθούν και οι ανάλογες προτεραιότητες και βήματα που ακολουθήθηκαν, ήταν αναγκαίο να σχεδιαστούν προσεκτικά και μεθοδικά. Πρώτα απ’ όλα, έπρεπε να υλοποιηθεί το πρακτικό κομμάτι, το οποίο ήταν ο εμπλουτισμός ενός ψηφιακού αποθετηρίου που θα φιλοξενούσε τους γλωσσικούς πόρους του Πανεπιστημίου Δυτικής Αττικής, μέσα στην ελληνική υποδομή του Clarin. Στη συνέχεια, πραγματοποιήθηκε η ενημέρωση των διδασκόντων του Πανεπιστημίου Δυτικής Αττικής σχετικά με την προσπάθεια συλλογής γλωσσικών πόρων, παραγόμενων στα πλαίσια της λειτουργίας του ΠαΔΑ, με στόχο τη διασύνδεση τους με την υποδομή του Clarin:el. Το συλλεγόμενο υλικό τεκμηριώθηκε, περιγράφηκε και μεταφορτώθηκε στη γλωσσική υποδομή. Ενώ, τέλος, η περιγραφή, οι στόχοι και τα αποτελέσματα της συγκεκριμένης δράσης, καταγράφονται στην παρούσα διπλωματική εργασία.

3.2 Σύνδεση και δημιουργία ψηφιακού αποθετηρίου στο Clarin:el

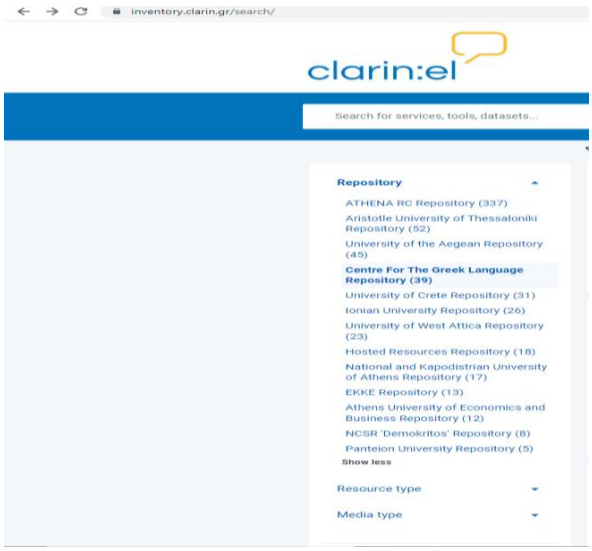
Στη συνέχεια θα εξετάσουμε τις διαδικασίες της διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με το Clarin:el, από το αρχικό στάδιο. Πρώτα όλα, πραγματοποιήθηκε σύνδεση στο Clarin:el, με δημιουργία λογαριασμού. Η υποδομή προσφέρει τη δυνατότητα σύνδεσης είτε με προσωπικό λογαριασμό, είτε με λογαριασμό ελληνικού ακαδημαϊκού ιδρύματος. Δεδομένου ότι όλοι οι εμπλεκόμενοι είναι μέλη ακαδημαϊκού ιδρύματος αλλά και ο στόχος της παρούσας διπλωματικής εργασίας ήταν η διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με το Clarin:el, επιλέχτηκε η σύνδεση με τη διεύθυνση ηλεκτρονικού ταχυδρομείου που παρέχει το ΠαΔΑ στους φοιτητές και διδάσκοντες του. Οπότε πραγματοποιήθηκε ακαδημαϊκή σύνδεση. Γενικά εφόσον ο τεκμηριωτής ή ο χρήστης είναι μέλος κάποιου ακαδημαϊκού ιδρύματος, ο ορθός τρόπος σύνδεσης του είναι με τον ιδρυματικό του λογαριασμό και όχι με κάποιον ιδιωτικό που πιθανόν έχει δηλώσει στην υποδομή.



Εικόνα 1. Σύνδεση με λογαριασμό στο Clarin:el

Έπειτα από επικοινωνία με την ομάδα υπευθύνων του Clarin:el, το ψηφιακό αποθετήριο του ΠαΔΑ, το οποίο υπήρχε ήδη αλλά δεν εμφανιζόταν στην αναζήτηση, αφού δεν περιείχε καθόλου υλικό, έγινε ορατό και αναζητήσιμο. Και επίσης προστέθηκε το λογότυπο του πανεπιστημίου (πριν το αποθετήριο είχε το λογότυπο του ιδρύματος που υποστηρίζει το Clarin:el, δηλαδή το ινστιτούτο για την έρευνα και καινοτομία Αθηνά (https://www.athenarc.gr/). Η υποδομή περιλάμβανε ήδη ψηφιακά αποθετήρια από επτά ακόμα ελληνικά πανεπιστήμια, τέσσερα ερευνητικά κέντρα καθώς και ένα αποθετήριο

φιλοξενούμενων πόρων, για μεμονωμένες εισαγωγές υλικού. Το αποθετήριο, στη συνέχεια, εμπλουτίστηκε σιγά – σιγά με γλωσσικό υλικό.



Εικόνα 2. Ιδρυματικά αποθετήρια στο Clarin:el

3.3 Αναζήτηση γλωσσικών πόρων

Έπειτα από τη δημιουργία του λογαριασμού στο Clarin:el και τις διαδικασίες με το συντονισμό του ψηφιακού αποθετηρίου του Πανεπιστημίου Δυτικής Αττικής, στη συνέχεια ασχοληθήκαμε με την αναζήτηση υλικού που θα διασυνδεόταν με το αντίστοιχο αποθετήριο.

Λόγω μεγαλύτερης οικειότητας και ως αρχικό μικρό δείγμα και επίσης για να καταλάβουμε τη σχετική ενημέρωση, την απήχησή της και τους ενδοιασμούς των χρηστών, πριν τη μαζική επικοινωνία μαζί τους, αρχικά στάλθηκε μήνυμα ηλεκτρονικού ταχυδρομείου στους καθηγητές του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*. Στο συγκεκριμένο μήνυμα οι διδάσκοντες ενημερώνονταν για την αναζήτηση γλωσσικών πόρων, στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας, και την πιθανή τους διασύνδεση με το Clarin:el. Επίσης, το μήνυμα περιείχε περιγραφή του υλικού που αναζητούσαμε, του Clarin:el και των στόχων του, καθώς και των οφελών της διάθεσης γλωσσικών πόρων, στους οποίους είχαν τα πνευματικά δικαιώματα οι δημιουργοί τους. Η διάθεση των πόρων, πέραν του αυτονόητου οφέλους της προαγωγής της γνώσης και των τεχνολογιών γλωσσικής επεξεργασίας, παρέχει προβολή στην επιστημονική κοινότητα του έργου, τόσο του Πανεπιστημίου Δυτικής Αττικής, όσο και των δημιουργών του. Στο τέλος του ηλεκτρονικού μηνύματος, ζητήθηκε επικοινωνία μέσω της πλατφόρμας Teams, ώστε οι

διδάσκοντες του Τμήματος να ενημερωθούν καλύτερα και να επιλυθεί όποια απορία τους σχετικά με το ζητούμενο υλικό και τους στόχους της παρούσας διπλωματικής εργασίας ή γενικά της υποδομής του Clarin. Στο συγκεκριμένο μήνυμα ηλεκτρονικού ταχυδρομείου, η ανταπόκριση ήταν μικρή. Κάποιοι καθηγητές απάντησαν ότι δεν είχαν το αντίστοιχο υλικό και οπότε δεν θα ήταν σε θέση να συνδράμουν στον εμπλουτισμό του αποθετηρίου του ΠαΔΑ, κάποιοι, ενώ δεν είχαν γλωσσικούς πόρους προς διάθεση ή σε αυτούς που είχαν, δεν ήταν κάτοχοι των πνευματικών τους δικαιωμάτων και δεν μπορούσαν να τους διαθέσουν ελεύθερα, απάντησαν ότι θα μπορούσε να πραγματοποιηθεί μια ενημερωτική επικοινωνία μαζί τους, μέσω Teams, προς πληροφόρηση και ενδιαφέρον για το εγχείρημα. Τρεις καθηγητές ανταποκρίθηκαν και έπειτα από συνάντηση μέσω ηλεκτρονικής πλατφόρμας, απέστειλαν σημαντικό αριθμό γλωσσικών πόρων για διασύνδεση με το Clarin. Η πλειοψηφία των διδασκόντων είναι αρκετά απασχολημένοι και συχνά δεν διαθέτουν χρόνο για να τον αφιερώσουν σε δραστηριότητες που δεν άπτονται των διδακτικών ή επιστημονικών καθηκόντων τους και έπρεπε να τους ενημερώσουμε πρώτα κατάλληλα για να καταλάβουν τη σημασία του εγχειρήματος. Το συγκεκριμένο μήνυμα αποστάλθηκε άλλες δύο φορές, δηλαδή συνολικά τρεις φορές.

Διαπιστώνοντας τη χαμηλή ανταπόκριση στο μήνυμα ηλεκτρονικού ταχυδρομείου που στάλθηκε στους διδάσκοντες του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*, αποφασίστηκε η προσπάθεια αυθόρμητης δια ζώσης ενημερωτικής συνάντησης μαζί τους. Πράγματι, συναντηθήκαμε με τους περισσότερους διδάσκοντες, στους χώρους τις σχολής, τις αίθουσες μαθημάτων, τους διαδρόμους ή το γραφείο τους. Η προσπάθεια ενημέρωσης δεν κατέληξε σε άμεση συγκομιδή υλικού. Παρόλα αυτά, παρασχέθηκε καλύτερη και άμεση ενημέρωση για τους στόχους της διπλωματικής και το είδος και μορφή του αναζητούμενου υλικού και επίσης, δόθηκαν χρήσιμες συμβουλές, από την πλευρά των διδασκόντων, για το που μπορεί να αναζητηθεί άλλο υλικό, ελεύθερο δικαιωμάτων, για διασύνδεση με το Clarin, το οποίο, κατά κανόνα, έχει παραχθεί στα πλαίσια του διδακτικού έργου του ΠαΔΑ.

Έχοντας έρθει σε προσωπική (φυσική ή εξ αποστάσεων) επικοινωνία με το πιο οικείο προσωπικό του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* και καταφέρει να μαζέψουμε ό,τι γλωσσικούς πόρους ήταν διαθέσιμοι, στη συνέχεια απευθυνθήκαμε στους διδάσκοντες των υπόλοιπων σχολών και τμημάτων του Πανεπιστημίου Δυτικής Αττικής, στέλνοντας τους κάποιο παρόμοιο αλλά περισσότερο κατατοπιστικό και εξειδικευμένο (σύμφωνα με την εμπειρία μας από την πρώτη προσπάθεια συγκέντρωσης υλικού) ενημερωτικό μήνυμα ηλεκτρονικού ταχυδρομείου. Και στη

συγκεκριμένη περίπτωση, η ανταπόκριση δεν ήταν πολύ μεγάλη. Κάποιοι μεμονωμένοι διδάσκοντες από διάφορα τμήματα, ενδιαφέρθηκαν και έστειλαν το ανάλογο υλικό γλωσσικών πόρων.

Έπειτα, και έχοντας εξαντλήσει την πιθανότητα να μαζέψουμε υλικό μετά από απευθείας επικοινωνία με τους διδάσκοντες του ΠαΔΑ, αποφασίσαμε να ενσωματώσουμε γλωσσικούς πόρους οι οποίοι έχουν ανοιχτή πρόσβαση και άδειες διάθεσης Creative Commons και αφορούν παραγωγή μέσω του διδακτικού και ερευνητικού έργου του ακαδημαϊκού ιδρύματος. Στην περίπτωση αυτή εμπίπτουν τα Ανοιχτά Ακαδημαϊκά Μαθήματα στο ΤΕΙ Αθήνας (https://ocp.teiath.gr) και στο ΤΕΙ Πειραιά (https://opencourses.gr/results.xhtml?ln=el&uni=ΤΕΙ+Πειραιά).

Το Τεχνολογικό Εκπαιδευτικό Ίδρυμα Αθηνών μετασχηματίστηκε σε Πανεπιστήμιο Δυτικής Αττικής (https://www.uniwa.gr) , το Μάρτιο του 2018, έπειτα από τη συγχώνευση του με το Τεχνολογικό Εκπαιδευτικό Ίδρυμα Πειραιά, σύμφωνα με το Νόμο 4521 (Ν. 4521/2018). Το 2019 εντάχθηκε στο ΠαΔΑ και η Εθνική Σχολή Δημόσιας Υγείας. Τα Ανοιχτά Ακαδημαϊκά Μαθήματα αποτελούν ένα έργο που υλοποιήθηκε στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση», με συγχρηματοδότηση από το Ευρωπαϊκό Κοινωνικό Ταμείο της Ευρωπαϊκής Ένωσης και εθνικούς πόρους. Αποτελεί την ανοιχτή και δωρεάν ψηφιακή δημοσίευση εκπαιδευτικού υλικού από μαθήματα του ΤΕΙ Αθηνών και του ΤΕΙ Πειραιά. Στόχος δημιουργίας τους αποτελεί η προώθηση της ασύγχρονης τηλεκπαίδευσης, χωρίς περιορισμούς όσον αφορά τον τόπο, χρόνο ή κοινό που θα τα παρακολουθήσει (https://ocp.teiath.gr). Το υλικό από τα μαθήματα αυτά συλλέχτηκε, υπέστη την ανάλογη επεξεργασία, και διασυνδέθηκε με το Clarin.

Μια άλλη κατηγορία υλικού που συλλέχτηκε και διασυνδέθηκε με το Clarin αποτελούν τα συγγράμματα διδασκόντων στο Τμήμα *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* στο Πανεπιστήμιο Δυτικής Αττικής, τα οποία έχουν δημοσιευθεί στον Κάλλιππο, με άδειες ανοιχτής πρόσβασης Creative Commons. Ο Κάλλιππος είναι μια δράση που δημοσιεύει σε ψηφιακή μορφή ελληνικά ακαδημαϊκά ηλεκτρονικά συγγράμματα (https://www.kallipos.gr) και τα καθιστά διαθέσιμα προς το ενδιαφερόμενο κοινό και την ακαδημαϊκή κοινότητα.

Τέλος, στο υλικό που συλλέχθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας περιλαμβάνονται, ενδεικτικά, κάποιες διπλωματικές εργασίες μεταπτυχιακών φοιτητών στο Μεταπτυχιακό Πρόγραμμα *Σπουδών «Διαχείριση Πληροφορίας σε Βιβλιοθήκες, Αρχεία και Μουσεία»*, καθώς και η πτυχιακή εργασία της συγγραφέως της παρούσας διπλωματικής εργασίας στο τμήμα *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*.

3.4 Περιγραφή Υλοποίησης – Εφαρμογής

Το γλωσσικό υλικό που μπορεί να μεταφορτωθεί στο Clarin κατατάσσεται σε τέσσερις κατηγορίες, οι οποίες είναι οι εξής:

- Εργαλεία ή υπηρεσίες γλωσσικής επεξεργασίας
- Σώματα κειμένου
- Γλωσσικές περιγραφές και υπολογιστικά μοντέλα
- Λεξικά, γλωσσάρια, θησαυροί, οντολογίες ή λίστες λέξεων, φράσεων κτλ., που αποτελούν τους εννοιολογικούς πόρους

Στα πλαίσια της παρούσας έρευνας καταφέραμε να συλλέξουμε συνολικά 193 γλωσσικούς πόρους, οι οποίοι εμπίπτουν σε δύο από τις κατηγορίες του προαναφερόμενου υλικού. Πρόκειται για τα σώματα κειμένου και τα λεξικά /γλωσσάρια/ εννοιολογικούς πόρους. Τα σώματα κειμένου αποτελούν τη μεγάλη πλειοψηφία και αριθμούν στα 188, ενώ τα λεξικά /γλωσσάρια είναι 5. Οι συγκεκριμένοι πόροι είναι σε αναλογία με τη μορφή του υλικού που σωρεύεται γενικά, σε όλα τα αποθετήρια των ακαδημαϊκών ιδρυμάτων που φιλοξενούνται στο Clarin, και το οποίο αντιστοιχεί τη στιγμή της συγγραφής της παρούσας διπλωματικής σε 657 σώματα κειμένου, 92 λεξικά / γλωσσάρια, 48 εργαλεία / υπηρεσίες και 2 γλωσσικές περιγραφές. Αξίζει να αναφερθεί ότι επιχειρήσαμε να μεταφορτώσουμε μεγαλύτερο αριθμό γλωσσικών πόρων, 238 συνολικά. Αλλά κάποιους από αυτούς τους αποσύραμε, κάνοντας τους unpublish αλλά κρατώντας τα μεταδεδομένα τους στη βάση του Clarin. Ο λόγος ήταν ότι αποτελούσαν τμηματικό μέρος υλικού, το οποίο στην οντότητα τους είχε συμπεριληφθεί ήδη στο Clarin. Πρόκειται κυρίως για κεφάλαια συγγραμμάτων του Κάλλιπου. Τα συγκεκριμένα δεν πρέπει να καταγράφονται χωριστά.

3.4.1 Προετοιμασία του υλικού

Όπως αναφέρθηκε, οι γλωσσικοί πόροι που συνδέονται με το Πανεπιστήμιο Δυτικής Αττικής και έχουμε στη διάθεση μας, ανταποκρίνονται σε δύο κατηγορίες υλικού από αυτό που συλλέγει το Clarin. Πρόκειται για τα σώματα κειμένου και τα λεξικά / γλωσσάρια.

Ως σώματα κειμένου νοούνται οι συλλογές από πρωτογενή δεδομένα σε διάφορα μέσα, όπως είναι τα ψηφιακά και ψηφιοποιημένα κείμενα γραπτού λόγου, ο ηχογραφημένος προφορικός λόγος, οι βιντεοσκοπήσεις ή οι εικόνες. Επίσης είναι επεξεργασμένα δεδομένα

όπως οι επισημειώσεις κειμένων, αυτόματα ή όχι δημιουργημένα κείμενα, ήχος, βίντεο, μεταγραφές προφορικών δεδομένων κτλ. Στην περίπτωση της παρούσας διπλωματικής εργασίας, το υλικό που συλλέχτηκε αποτελείται σχεδόν αποκλειστικά από ψηφιακά και ψηφιοποιημένα κείμενα γραπτού λόγου, σε διάφορες μορφές: άρθρα, εργασίες, υλικό από μαθήματα του ΠαΔΑ, περιγραφές μαθημάτων, ακαδημαϊκά συγγράμματα, περιγραφές συνεδρίων, εισηγήσεις κτλ. (Clarin:el, 2020).

Τα λεξικά ή γλωσσάρια όρων και ορολογίας είναι δομημένα γλωσσικά δεδομένα, όπως οι λίστες και κατάλογοι λέξεων, οι θησαυροί, ή άλλοι εννοιολογικοί πόροι που χρησιμοποιούνται για την επεξεργασία των πρωτογενών και επεξεργασμένων δεδομένων (Clarin:el, 2020).

Το Clarin:el συλλέγει κυρίως υλικό στην ελληνική γλώσσα. Οι γλωσσικοί μας πόροι είναι όντως στα ελληνικά, εκτός από τρεις εξαιρέσεις που είναι μόνο στην αγγλική γλώσσα και που περιελάβαμε αφού αποτελούν παραγωγή του Πανεπιστημίου Δυτικής Αττικής και αποστάλθηκαν ευγενικά από τους διδάσκοντες του. Επίσης, τα περισσότερα γλωσσάρια είναι δίγλωσσα, στα ελληνικά και αγγλικά και περιλαμβάνουν δίγλωσση ορολογία. Ακόμα, καταφέραμε να συλλέξουμε ένα γλωσσάρι με λέξεις στη μεσσηνιακή τοπική διάλεκτο και την αντιστοιχία τους στα επίσημα νέα ελληνικά.

Στη συνέχεια προχωρήσαμε με την προετοιμασία και διασύνδεση του υλικού με την ελληνική υποδομή του Clarin. Το Clarin θέτει κάποιους περιορισμούς ως προς τη μορφή του υλικού που δέχεται. Όσον αφορά τις κατηγορίες του υλικού που συλλέξαμε, κανονικά οι πόροι θα πρέπει να είναι σε μορφή XML ή TXT, ενώ, η υποδομή σωρεύει και κείμενα σε PDF και MS-Word, αφού βρίσκεται σε φάση δρομολόγησης μιας υπηρεσίας που μετατρέπει τα αρχεία σε TXT (Clarin:el, 2020). Οπότε, στην παρούσα φάση, μόνο το υλικό στους συγκεκριμένους μορφότυπους είναι συμβατό με το Clarin για τις δύο κατηγορίες πόρων που συλλέξαμε. Για το λόγο αυτό, αποφύγαμε να μεταφορτώσουμε πόρους αμιγώς σε άλλους μορφότυπους (πχ PPT, TIFF) που βρήκαμε και που θα μπορούσαν πιθανόν να μεταφορτωθούν σε κάποια μεταγενέστερη φάση, αν το υλικό στους συγκεκριμένους μορφότυπους καταστεί επεξεργάσιμο και το συγκεκριμένο υλικό είναι επιθυμητό από την ομάδα του Clarin. Αντίθετα, κάποιοι πόροι σε PPT μεταφορτώθηκαν όταν ήταν μέρος μικτού υλικού που περιλάμβανε και επεξεργάσιμο υλικό σε TXT, Word ή PDF. Ακόμα, σε κάποιες περιπτώσεις που δεν ήταν πολύ δύσκολο και που δεν υποβαθμιζόταν η ποιότητα του υλικού, πραγματοποιήσαμε μετατροπή του κειμένου σε TXT. Στην παρακάτω εικόνα, διακρίνουμε τους συνιστώμενους μορφότυπους (formats) αρχείων που δέχεται το Clarin.

	Recommended	Acceptable
CLARIN:EL processable data	**Monolingual textual data**: plain text **Monolingual encoded data**: XCES-ILSP variant (XML based format compliant with the XCES model for corpora) **Bi-/Multilingual encoded data**: TMX (XML based format for aligned data), MOSES (text-based format for parallel data)	
Textual Data	**File Formats**: plain text **Formatted/Encoded**: ODT, DOCX, PDF/A, HTML, Latex, TeX, MOSES	PDF, SGML, Rich Text Format (.rtf), Microsoft Word (.doc, .docx), PostScript
Text Annotation	**File Formats**: XML, XMI, CSV, TSV, RDF (all serialisation formats RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), JSON	SGML, Plain Text, Microsoft Excel (.xlsx, .xls), ELLOGON
	Models: XCES for corpora and structural annotation, TEI for structural and linguistic annotation, GrAF linguistic annotation, TMX for aligned, GATE linguistic annotation, CoNLL family (CoNLL-U, CoNLL-2000, CoNLL-2002, CoNLL-2003, CoNLL-2006, CoNLL-2008, CoNLL-2009, CoNLL-2012) for linguistic annotation, NIF linguistic annotation for RDF data, WARC for web crawled data	
Language Description	**ML Model**: H5, ProtoBuf, ONNX, PMML, Pickle, MLeap, YAML, JSON **N-gram model**: ARPA	
Lexical/Conceptual Resource	**File Formats**: XML, CSV, TSV, RDF (RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), OWL	Microsoft Excel (.xlsx, .xls), Plain Text, SQL
	Models: LMF for lexica, OWL for ontologies, SKOS for thesauri, OntoLex-Lemon for lexica, TBX for terminological data	
Image data	**All images**: TIFF, SVG, JPEG 2000, PNG, GIF **Scanned images**: PDF/A	JPEG, BMP, Photoshop, NifTi, FlashPix, PDF
Audio data	WAV, AIFF, FLAC	MP3, MPEG, Windows Media Audio
Video data	AVI	MPEG-4, RealNetworks 'Real Video', Windows Media Video, Flash Video, QuickTime Video

Εικόνα 3. Συνιστώμενοι μορφότυποι. Πηγή: https://www.clarin.gr/

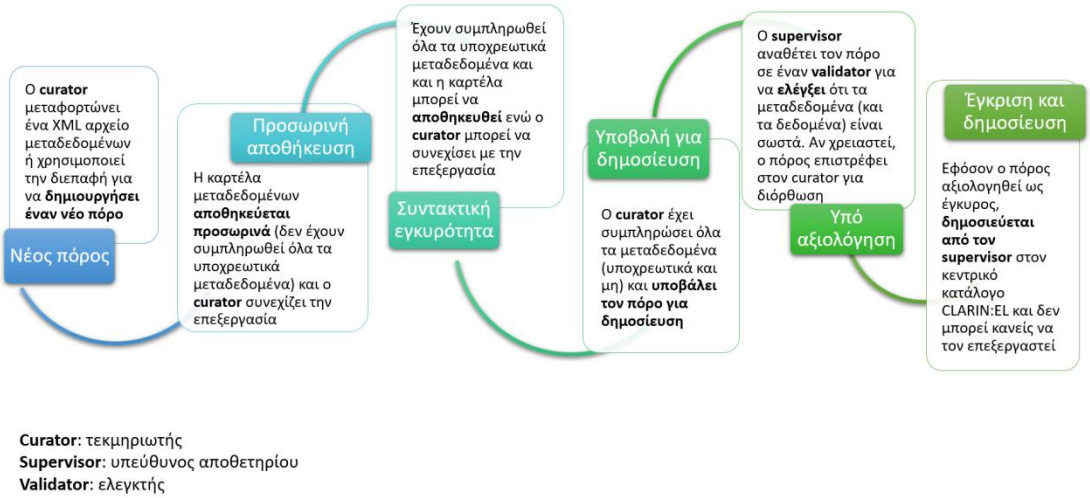
Το υλικό στους μορφότυπους που αναφέραμε, δεν είναι δυνατόν να μεταφορτωθεί στη μορφή που βρίσκεται. Τα αρχεία, ένα ή περισσότερα, είναι αναγκαίο να βρίσκονται σε έναν συμπιεσμένο φάκελο σε έναν από τους ακόλουθους μορφότυπους .zip, .tgz, .gz, .tar. Εμείς επιλέξαμε να τους συμπιέσουμε σε μορφότυπο .zip. Η ονομασία του φάκελου πρέπει να είναι σε λατινικό αλφάβητο, χωρίς κενά.

Η παρακάτω εικόνα περιγράφει τη διαδικασία της προετοιμασίας του υλικού για τη διασύνδεση του με το Clarin.

Εικόνα 4. Προετοιμασία του υλικού. Πηγή: https://www.clarin.gr/

3.4.2 Μεταφόρτωση και μεταδεδόμενα του πόρου

Στη συνέχεια, θα παραθέσουμε πως γίνεται η μεταφόρτωση και τεκμηρίωση στο Clarin των δύο προαναφερόμενων μορφών υλικού, δηλαδή των σωμάτων κειμένου και των γλωσσάριων που συλλέξαμε. Η συγκεκριμένη διεργασία ανταποκρίνεται στην παρακάτω εικόνα, την οποία θα αναλύσουμε διεξοδικά. Παρατηρώντας τη συγκεκριμένη εικόνα, διαπιστώνουμε ότι υπάρχουν κάποιοι ρόλοι, που καταλαμβάνουν διαφορετικά άτομα στην προετοιμασία, υποβολή και δημοσίευση ενός γλωσσικού πόρου. Πρόκειται για τον curator (τεκμηριωτή), metadata validator (ελεγκτή μεταδομένων), legal validator (νομικό ελεγκτή) και το supervisor (υπεύθυνο αποθετηρίου), οι οποίοι ελέγχουν τα διαφορετικά στάδια της πορείας του πόρου προς τη δημοσίευση του στο αποθετήριο.

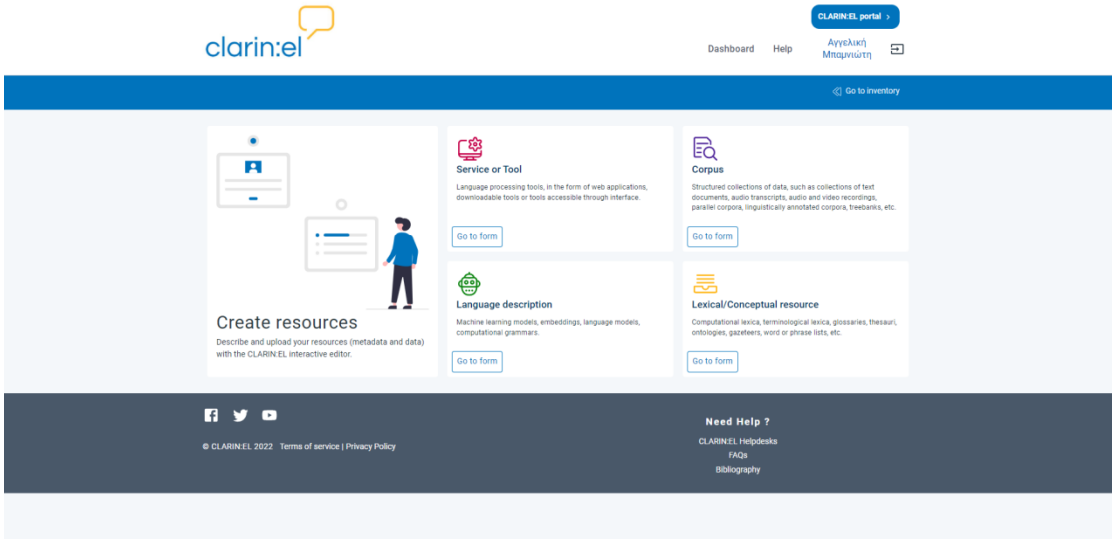


Εικόνα 5. Πορεία δημοσίευσης των πόρων. Πηγή: https://www.clarin.gr/

Ο curator (τεκμηριωτής) είναι αυτός που, αφού έχει ήδη προετοιμάσει το υλικό όπως περιγράφηκε στο προηγούμενο υποκεφάλαιο, αναλαμβάνει να το μεταφορτώσει στην υποδομή, παραθέτοντας την περιγραφή του με τα κατάλληλα τεκμηριωμένα μεταδεδομένα.

Πρώτα απ' όλα, συνδεθήκαμε με τον ιδρυματικό λογαριασμό μας στο Clarin. Στη συνέχεια, κατευθυνθήκαμε στον κεντρικό κατάλογο της υποδομής (central inventory). Επιλέξαμε τον πίνακα dashboard και μετά, δημιουργήσαμε τον πόρο μας (create resources). Στην παρακάτω εικόνα διακρίνουμε τις τέσσερις κατηγορίες ταξινόμησης του υλικού. Εμείς, λόγω της μορφής του υλικού που διαθέτουμε, επιλέξαμε είτε Corpus, είτε Lexical Conceptual resource. Στο σημείο αυτό πρέπει να επισημάνουμε ότι, αν και το Clarin έχει διεπαφή και στα

ελληνικά, στις σελίδες της μεταφόρτωσης και της τεκμηρίωσης, τα πεδία είναι μόνο στα αγγλικά.



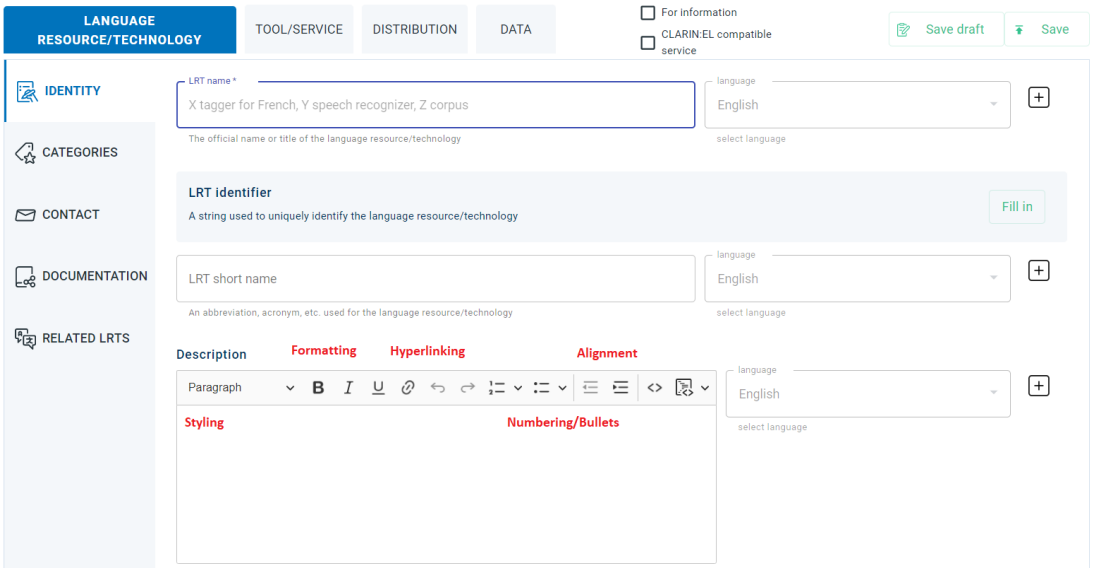
Εικόνα 6. Δημιουργία πόρων. Πηγή: https://www.clarin.gr/

Έχοντας επιλέξει την κατηγορία του πόρου, στη συνέχεια περνάμε τα μεταδεδομένα. Ως μεταδεδομένα εννοούνται τα δεδομένα που αφορούν άλλα δεδομένα (Κυριάκη – Μάνεση και Κουλούρης, 2015). Τα μεταδεδομένα έχουν ως στόχο να διευκολύνουν τη χρήση, ανεύρεση και διαχείριση των πόρων, ψηφιακών ή συμβατικών (Καπιδάκης, Λαζαρίνης και Τοράκη, 2015).

Γενικά η υποδομή και το σχήμα μεταδεδομένων που υποστηρίζει ανταποκρίνονται στις αρχές FAIR: Findability (ευρεσιμότητα), Accessibility (προσβασιμότητα), Interoperability (διαλειτουργικότητα) και Reuse (επαναχρησιμοποίηση) (Clarin:el, 2020). Υπάρχουν κάποια υποχρεωτικά μεταδεδομένα και άλλα προαιρετικά, ανάλογα με την κατηγορία του υλικού, με σκοπό την καλύτερη περιγραφή του. Τα κοινά υποχρεωτικά μεταδεδομένα που πρέπει να έχουν στην περιγραφή τους όλοι οι πόροι, όποια και αν είναι η μορφή τους, είναι τα εξής (Clarin:el, 2020):

- **Όνομα του πόρου** (LRT name): δηλαδή πως ονομάζεται ο γλωσσικός πόρος που επιχειρούμε να διασυνδέσουμε με την υποδομή
- **Περιγραφή** (Description): περιλαμβάνει στοιχεία για το περιεχόμενο του γλωσσικού πόρου

- **Αναγνωριστικό γλωσσικού πόρου** (LRT identifier): μπαίνει αυτόματα και είναι σε μορφή handle
- **Αριθμός εκδοχής του πόρου** (version): μπαίνει αυτόματα αλλά μπορούμε να τον δώσουμε και εμείς
- **Λέξεις κλειδιά** (keywords): δηλαδή λέξεις που χαρακτηρίζουν ένα γλωσσικό πόρο και βοηθούν στην ευρετηρίαση, τεκμηρίωση και αναζήτηση του
- **Επαφή** (contact): Email ή ιστοσελίδα όπου κάποιος μπορεί να λάβει περισσότερες πληροφορίες σχετικά με τον πόρο
- **Τρόπος διανομής του πόρου** (distribution): δηλαδή τη μορφή του υλικού
- **Όροι χρήσης του πόρου** (licence terms):αναφέρεται στο πως μπορεί να έχει κάποιος πρόσβαση στο γλωσσικό πόρο και με ποιους όρους και συνθήκες μπορεί να τον χρησιμοποιήσει.



Εικόνα 7. Κοινά μεταδεδομένα σε κάθε μορφή γλωσσικών πόρων. Πηγή: https://www.clarin.gr/

Να σημειώσουμε ότι στην τεκμηρίωση των πόρων μας στο Clarin, εισαγάγαμε τον τίτλο, την περιγραφή και τις λέξεις – κλειδιά και στα αγγλικά, που ήταν υποχρεωτικό, αφού πρόκειται για το ελληνικό παράρτημα της διεθνούς υποδομής, στην οποία καταλήγει το υλικό από όλα τα παραρτήματα, στις διαφορετικές γλώσσες και χώρες που υπάρχουν, άλλα και στα ελληνικά, για καλύτερη πρόσβαση και αναζήτηση.

Στην περίπτωση όλων των πόρων που μεταφορτώσαμε στο Clarin, είτε είναι σώμα κειμένου ή λεξικό / εννοιολογικός πόρος, επιλέξαμε επίσης να αναφέρουμε, σε κάθε έναν, το όνομα του δημιουργού τους, δηλαδή του ατόμου που έγραψε / δημιούργησε το

συγκεκριμένο γλωσσικό πόρο. Πρόκειται για προαιρετικό πεδίο σύμφωνα με την εφαρμογή, αλλά κρίθηκε απαραίτητο να αναφερθεί το όνομα του δημιουργού του υλικού, σε ελληνικό και λατινικό αλφάβητο, εφόσον γνωρίζουμε ποιος είναι κάθε φορά. Είναι ιδιαίτερα σημαντικό να ξέρει κάποιος ποιος είναι ο δημιουργός ενός πόρου για πολλούς λόγους όπως πχ ότι καταλαβαίνουμε καλύτερα την αξία του περιεχομένου του, μπορούμε να τον αναζητήσουμε με βάση το όνομα του δημιουργού του, προσφέρεται αναγνωρισιμότητα, πιθανόν να θέλουμε να τιμήσουμε κάποιον για τη συνεισφορά του κτλ. Γενικά, η πολιτική που ακολουθήθηκε με τα προαιρετικά πεδία της εφαρμογής είναι να συμπληρώνονται όταν τα γνωρίζουμε, δεν παρουσιάζεται ιδιαίτερη δυσκολία που μας αποτρέπει ή θεωρούμε ότι η περιγραφή μας θα είναι πιο πλήρης.

Στην περίπτωση που εκτός από διδάσκοντες ή άλλα άτομα σχετιζόμενα με το Πανεπιστήμιο Δυτικής Αττικής ανάμεσα στους δημιουργούς του πόρου περιλαμβάνονται και άλλα, τρίτα άτομα, πλειοψηφικά επιλέχτηκε να δοθεί το όνομα του συσχετιζόμενου ατόμου με το ΠαΔΑ στο πεδίο του δημιουργού, αλλά στην περιγραφή να αναφερθούν τα ονόματα όλων των δημιουργών. Η εφαρμογή μπορεί να περιλάβει περισσότερους από έναν δημιουργό.

Επίσης, εισαγάγαμε σε όλους τους διαθέσιμους γλωσσικούς πόρους τον επιστημονικό τομέα στον οποίο κατατάσσεται ο πόρος, στα ελληνικά και αγγλικά, ο οποίος είναι προαιρετικό πεδίο. Η περιγραφή είναι πληρέστερη όταν ξέρουμε σε ποιο επιστημονικό πεδίο ανήκει κάποιος γλωσσικός πόρος, ενώ και η προσβασιμότητα σε αυτόν είναι πιο εύκολη.

Τέλος, σε όλους τους σχετιζόμενους γλωσσικούς πόρους με το Πανεπιστήμιο Δυτικής Αττικής προσθέσαμε το λογότυπο του πανεπιστημίου. Ο λόγος που έγινε αυτό είναι γιατί αρχικά, και όπως προαναφέρθηκε, το αποθετήριο του Πανεπιστημίου Δυτικής Αττικής, είχε λανθασμένο λογότυπο άλλου ερευνητικού κέντρου, το οποίο πήρε κάποιο καιρό, λόγω κάποιας πολυπλοκότητας με το σύστημα, στους εκπροσώπους της εφαρμογής για να το διορθώσουν δίνοντας το σωστό λογότυπο του ΠαΔΑ. Εν τω μεταξύ είχε αρχίσει η μαζική μεταφόρτωση των πόρων του ΠαΔΑ στην εφαρμογή και για πρακτικούς λόγους ξεκινήσαμε να προσθέτουμε το λογότυπο σε κάθε πόρο ξεχωριστά, για να φαίνεται ευκρινώς σε ποιο αποθετήριο ανήκουν. Για λόγους ομοιομορφίας, συνεχίστηκε να προστίθεται το λογότυπο και στους νέους πόρους που ανεβήκαν με τη διόρθωση του προβλήματος.

3.4.2.1 Μεταφόρτωση γλωσσικών πόρων σε μορφή σωμάτων κειμένου

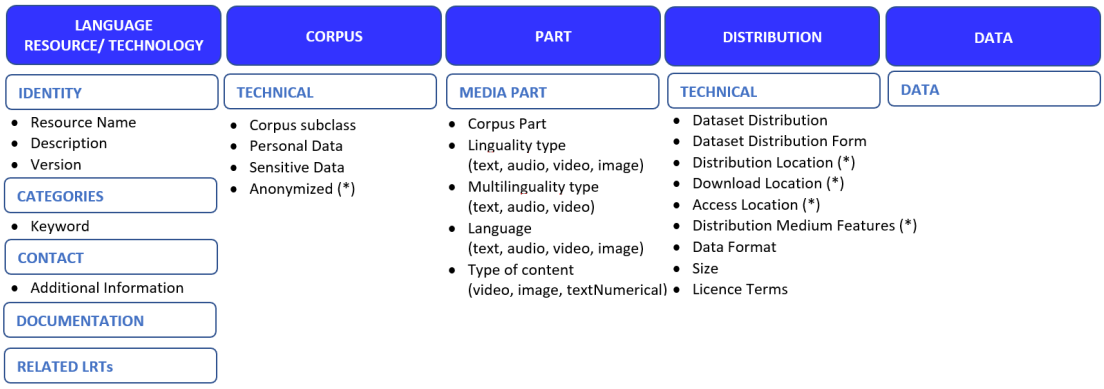
Στη συνέχεια θα εξετάσουμε τα υπόλοιπα υποχρεωτικά μεταδεδομένα που είναι διαφορετικά στις δύο κατηγορίες υλικού που ενσωματώσαμε στο Clarin. Στην περίπτωση των σωμάτων κειμένου, πρόκειται για τα εξής (Clarin:el, 2020):

- **Υποκατηγορία (subclass)**, δηλαδή αν πρόκειται για πρωτογενή (raw) ή επισημειωμένο (annotated) υλικό
- Δήλωση αν περιλαμβάνουν **προσωπικά ή ευαίσθητα προσωπικά δεδομένα** και πρέπει να **ανωνυμοποιηθούν**
- Το **είδος μέσου** του σώματος κειμένου (πχ κείμενο, βίντεο, ήχος, εικόνα, κείμενο με αριθμητικά δεδομένα)
- Το **μέγεθος** του σώματος κειμένου, με τιμές από μια μακρά αναπτυσσόμενη προεπιλεγμένη λίστα (dropdown list) απ' όπου μπορούμε να επιλέξουμε τις τιμές που χρειαζόμαστε ανάλογα με την περίπτωση. Αρχικά επιλέγουμε το μέγεθος μέτρησης του πόρου αριθμητικά (1,2,3,4….). Και έπειτα επιλέγουμε την τιμή που περιγράφει καλύτερα το μέγεθος του πόρου από την ανάλογη μακρά λίστα. Ενδεικτικά κάποιες από τις τιμές που μπορούν να επιλεγούν είναι: article, byte, class, concept, element, entry, expression, file, frame, hour, idiomatic expression, image, ingested record, internal record, item, kilobyte, keyword, lexical type, minute, multiword unit, phoneme, phonetic unit, neologism, phrase, predicate, published record, question, rule, second, segment, semantic unit, sentence, shot, syllable, sentence, syntactic unit, terabyte, term, text, token, translation unit, trigram, turn, unigram, unit, utterance, word κτλ.
- Ο **μορφότυπος** του σώματος κειμένου, με τιμές από μια μακρά αναπτυσσόμενη προεπιλεγμένη λίστα (dropdown list) απ' όπου μπορούμε να επιλέξουμε τις τιμές που εκφράζουν την ψηφιακή αναπαράσταση του πόρου ανάλογα με την περίπτωση. Κάποιες από τις τιμές ανάμεσα στις οποίες μπορούμε να επιλέξουμε είναι οι εξής: ACL Anthology Corpus format, AIMED corpus format, anafora, annotation format, audio format, Avro, blast, binary CAS, binary format, BioNLP, BioNLP format, bliki Wikipedia, BNC format, BRAT, CBOR, CHAT, Cochrane, CoNLL format, corpus format, CSV, database format, datasift/JSON, DIAML, document format, EMMA, Fast infoset, FoLIA, GATE format, GrAF, HTML, image format, JSON, KAF, linked data format, media wiki markup, MS-Access database, MS-Excel, MS-PowerPoint, MS-Word, NAF, NIF,

Oasis text, OBO, open format, OWL, OWL/XML, PDF, PLS, PML, postscript, PTB, PubMed, raw audio format, RDF format, RDF/XML, RTF, SGML, tabular format, tbx, TCF, TEI, TEX, text/plain, tika, TMX, TSV, tuepp, turtle, VideoFormat, Web ARChive format, wiki format, web annotation format, Wikipedia article, Wikipedia format, Wikipedia link, XHTML, XMI, XML, AVI, GIF, JPEG, audio mp3, PNG, SVG, TIFF κτλ.

- Η **γλώσσα** του πόρου, την οποία μπορούμε να επιλέξουμε από την υπάρχουσα αναπτυσσόμενη προεπιλεγμένη λίστα (dropdown list).

Στην παρακάτω εικόνα διακρίνουμε τα υποχρεωτικά μεταδεδομένα που χρειάζονται στην τεκμηρίωση και την περιγραφή στα σώματα κειμένου.



Εικόνα 8.Υποχρεωτικά μεταδεδομένα σε σώματα κειμένου. Πηγή: https://www.clarin.gr/

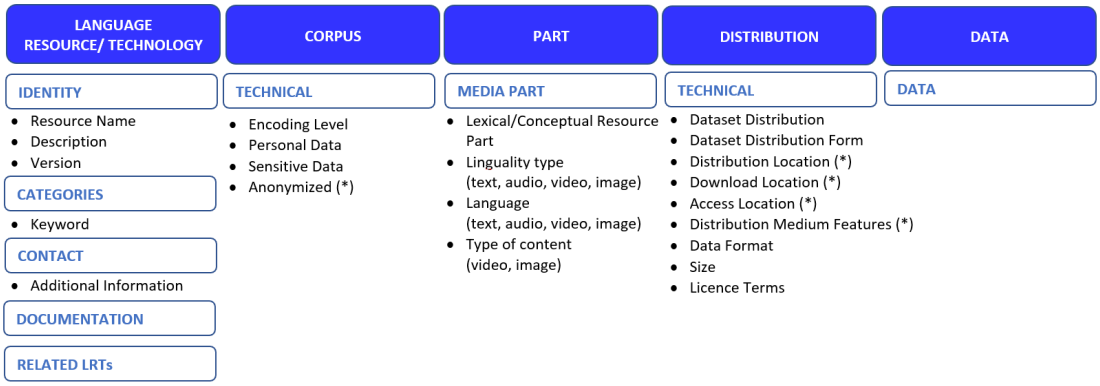
3.4.2.2 Μεταφόρτωση γλωσσικών πόρων σε μορφή λεξικό-εννοιολογικών πόρων

Τα υπόλοιπα υποχρεωτικά δεδομένα τα οποία είναι απαραίτητα για την εισαγωγή υλικού σε μορφή λεξικο-εννοιολογικών πόρων στο Clarin, είναι τα εξής (Clarin:el, 2020):

- Πληροφορίες σχετικά με το **επίπεδο κωδικοποίησης** του πόρου, δηλαδή αν πρόκειται για μορφολογία, φωνολογία, σημασιολογία κτλ. στο επίπεδο της γλωσσικής του ανάλυσης
- Δήλωση αν περιλαμβάνουν **προσωπικά ή ευαίσθητα προσωπικά δεδομένα** που πρέπει να **ανωνυμοποιηθούν**
- Το **είδος μέσου** του λεξικό / εννοιολογικού πόρου (πχ κείμενο, βίντεο, ήχος, εικόνα)

- Το **μέγεθος** του πόρου. Τα παραδείγματα από τις υπάρχουσες τιμές από τις οποίες μπορεί να γίνει η επιλογή, ανάλογα με την περίπτωση στην οποία ανταποκρίνεται ο γλωσσικός πόρος, είναι τα ίδια με αυτά που αναφέρονται στο κεφάλαιο 3.4.2.1.
- Ο **μορφότυπος** του πόρου. Τα παραδείγματα από τις υπάρχουσες τιμές από τις οποίες μπορεί να γίνει η επιλογή, ανάλογα με την περίπτωση στην οποία ανταποκρίνεται ο γλωσσικός πόρος, είναι τα ίδια με αυτά που αναφέρονται στο κεφάλαιο 3.4.2.1.
- Η **γλώσσα** του πόρου, την οποία μπορούμε να επιλέξουμε από την υπάρχουσα αναπτυσσόμενη προεπιλεγμένη λίστα (dropdown list).

Στην παρακάτω εικόνα διακρίνουμε τα υποχρεωτικά μεταδεδομένα που χρειάζονται στην τεκμηρίωση των λεξικό /εννοιολογικών πόρων.



Εικόνα 9.Υποχρεωτικά μεταδεδομένα σε λεξικό-εννοιολογικό πόρο. Πηγή: https://www.clarin.gr/

3.4.2.3 Τα προαιρετικά πεδία γλωσσικών πόρων σε μορφή σωμάτων κειμένου

Στη συνέχεια θα αναφερθούμε στα προαιρετικά πεδία τεκμηρίωσης και περιγραφής των γλωσσικών πόρων. Κάποια από αυτά χρησιμοποιήθηκαν στην τεκμηρίωση των γλωσσικών πόρων του Πανεπιστημίου Δυτικής Αττικής στο Clarin. Τα προαιρετικά πεδία δίδονται στα αγγλικά, που είναι και η υποχρεωτική γλώσσα τεκμηρίωσης και η μόνη στην οποία παρέχει το Clarin τα πεδία περιγραφής και τεκμηρίωσης, υποχρεωτικά ή προαιρετικά. Η συμπλήρωση τους από το χρήστη γίνεται στα αγγλικά επίσης, εφόσον η διεθνής υποδομή του Clarin θα μπορεί να κάνει τη συγκομιδή τους. Πρόκειται για τα παρακάτω πεδία:

- **Σύντομο όνομα γλωσσικού πόρου** (LRT short name): πρόκειται για κάποια αρχικά, συντόμευση, ακρωνύμιο κτλ. το οποίο χρησιμοποιείται για τον περιγραφόμενο γλωσσικό πόρο. Το συγκεκριμένο πεδίο δεν βρήκε πρακτική εφαρμογή στους γλωσσικούς πόρους του ΠαΔΑ που μεταφορτώσαμε στο Clarin.
- **Πάροχος γλωσσικού πόρου** (LRT provider): Πρόκειται για το φορέα ή άτομο που έχει την ευθύνη για την παροχή, επιμέλεια, διατήρηση και δημοσίευση του γλωσσικού πόρου. Αναφορικά με τους πόρους του ΠαΔΑ, δεν χρησιμοποιήθηκε το συγκεκριμένο προαιρετικό πεδίο γιατί τα στοιχεία των δύο υπευθύνων ατόμων για τις συγκεκριμένες δράσεις στο ΠαΔΑ, τα οποία είναι η συγγραφέας της παρούσας διπλωματικής εργασίας και ο επιβλέποντας καθηγητής της, έχουν ήδη δοθεί στο υποχρεωτικό πεδίο της επαφής (contact). Οπότε κρίθηκε μάλλον ως πλεονασμός να ξαναναφερθούν τα στοιχεία τους.
- **Δημιουργός γλωσσικού πόρου** (LRT creator): Αναφορικά με το δημιουργό του γλωσσικού πόρου, αντίθετα, επιλέχθηκε να αναφέρεται πάντα το όνομα του. Ένα έργο αποτελεί μια οντότητα στην οποία θα πρέπει να φαίνεται ποιος είναι υπεύθυνος για τη δημιουργία του και του έχει προσδώσει τα ιδιαίτερα χαρακτηριστικά και την ποιότητα του.
- **Χρηματοδότηση έργου** (Funding project): Στο συγκεκριμένο προαιρετικό πεδίο αναφέρεται αν η δημιουργία, εμπλουτισμός ή επέκταση του γλωσσικού πόρου έχει χρηματοδοτηθεί από κάποιο έργο (project). Αν ναι, στη συνέχεια μπορεί να δοθεί και το **όνομα του έργου**. Τα πεδία αυτά δεν έχουν πρακτική εφαρμογή στους παρεχόμενους γλωσσικούς πόρους του ΠαΔΑ. Οπότε δεν έχουν συμπληρωθεί.
- **Μεταφόρτωση λογότυπου** (Upload a logo): Σε όλους τους γλωσσικούς πόρους του Πανεπιστημίου Δυτικής Αττικής έχουμε μεταφορτώσει το λογότυπο του ακαδημαϊκού ιδρύματος, για λόγους άμεσης αναγνωρισιμότητας της προέλευσης τους για το χρήστη.
- **Ανωνυμοποίηση** (Anomymized): Το πεδίο στο οποίο επισημαίνουμε αν στοιχεία του γλωσσικού πόρου έχουν ανωνυμοποιηθεί, λόγω του ότι μπορεί να είναι ευαίσθητα ή προσωπικά δεδομένα, είναι προαιρετικό. Αντίθετα η δήλωση του αν ο πόρος περιλαμβάνει προσωπικά ή ευαίσθητα στοιχεία είναι υποχρεωτικά πεδία. Εμείς επιλέξαμε να απαντήσουμε, αναφορικά με όλο το υλικό του ΠαΔΑ ότι δεν έχει γίνει κάποια ανωνυμοποίηση.

- **Τοποθεσία δειγμάτων του υλικού** (Samples location): Στο συγκεκριμένο πεδίο συμπληρώνεται κάποιο URL σχετικό με δείγματα από τη διανομή / διάθεση ενός σώματος κειμένου. Δεν έχει χρησιμοποιηθεί.
- **Κόστος** (Cost): Αφορά το χρηματικό κόστος για τη χρήση του υλικού. Το υλικό διατίθεται δωρεάν. Οπότε δεν έχει συμπληρωθεί το ανάλογο πεδίο.
- **Δικαιώματα πρόσβασης** (Access rights): Πρόκειται για προαιρετικό πεδίο αναφορικά με τα δικαιώματα πρόσβασης στο υλικό, ανά κατηγορίες. Το πεδίο δεν έχει συμπληρωθεί γιατί δεν έχουμε διαχωρίσει τις κατηγορίες του κοινού το οποίο θα μπορεί να έχει πρόσβαση στο υλικό που παράχθηκε από το ΠαΔΑ. Κοινώς όποιος επιθυμεί μπορεί να έχει πρόσβαση στους πόρους.
- **Προβλεπόμενη εφαρμογή** (Intended application): Αφορά την εφαρμογή στην οποία θα μπορούσε να χρησιμοποιηθεί ο γλωσσικός πόρος. Το συγκεκριμένο πεδίο περιλαμβάνει μια μακρά αναπτυσσόμενη προεπιλεγμένη λίστα τιμών (dropdown list). Οι πόροι δεν δημιουργήθηκαν για συγκεκριμένο στόχο, και μπορούν να έχουν μια γενική, μη προβλεπόμενη αρχικά, εφαρμογή.
- **Τομέας** (Domain): Αναφέρεται στο πεδίο γνώσης ή δράσης στα οποία κατατάσσεται ο γλωσσικός πόρος. Επιλέξαμε να συμπληρώσουμε το συγκεκριμένο πεδίο σε όλους τους πόρους του ΠαΔΑ που μεταφορτώσαμε για λόγους καλύτερης τεκμηρίωσης και ευκολότερης ανεύρεσης του πόρου από το χρήστη.
- **Γεωγραφική κάλυψη** (Geographic coverage): Αφορά τη γεωγραφική περιοχή που καλύπτουν τα περιεχόμενα ενός γλωσσικού πόρου. Θα μπορούσε να συμπληρωθεί στην περίπτωση που κάποιος πόρος αναφέρεται καθαρά σε κάποια συγκεκριμένη περιοχή. Στην περίπτωση των γλωσσικών πόρων που παράχθηκαν από το ΠαΔΑ, είχαμε κάποιες περιπτώσεις όπου συμπληρώθηκε το συγκεκριμένο πεδίο, όπως πχ σε ένα γλωσσάρι που περιείχε λεξιλόγιο από την τοπική διάλεκτο της περιοχής της Μεσσηνίας.
- **Χρονική κάλυψη** (Time coverage): Συμπληρώνεται όταν το περιεχόμενου του υλικού αναφέρεται σε κάποια συγκεκριμένη χρονική περίοδο. Στην περίπτωση των γλωσσικών πόρων που παράχθηκαν από το ΠαΔΑ, είχαμε κάποιες περιπτώσεις όπου συμπληρώθηκε το συγκεκριμένο πεδίο.
- **Τεκμηριωμένο στο…** (Documented in): Πρόκειται για πεδίο που δεν έχουμε συμπληρώσει λόγω πρακτικής δυσκολίας. Περιλαμβάνει στοιχεία τεκμηρίων που είναι σχετικά με το γλωσσικό πόρο, όπως κάποιο ερευνητικό άρθρο κτλ. Μπορεί να συμπληρωθεί ακόμα και η βιβλιογραφική αναφορά του άρθρου ή άλλου σχετικού

τεκμηρίου. Πιθανόν, κάποιο από το υλικό που έχουμε στη διάθεση μας να εμπίπτει σε αυτή την κατηγορία. Χρειάζεται μεγαλύτερη έρευνα ή περεταίρω επικοινωνία με τους δημιουργούς των γλωσσικών πόρων για να διαπιστωθεί αυτό και λόγω της δυσκολίας του χρονικού περιορισμού, δεν κατέστη δυνατόν να διερευνηθεί περισσότερο.

- **Προηγούμενες εκδοχές** (Previous versions): Αναφέρεται σε παλαιότερες εκδοχές κάποιου γλωσσικού πόρου ο οποίος έχει αντικατασταθεί με το συγκεκριμένο πόρο. Το συγκεκριμένο πεδίο δεν έχει πρακτική εφαρμογή στους πόρους του ΠαΔΑ που έχουμε μεταφορτώσει στο Clarin αφού δεν υπάρχουν παλαιότερες εκδοχές τους που να έχουν κατατεθεί στην εφαρμογή.
- **Εκδοχή του…** (Version of): Περιλαμβάνει έναν γλωσσικό πόρο που θεωρείται διαφορετική εκδοχή (διορθωμένη, σχολιασμένη, εμπλουτισμένη, επεξεργασμένη κτλ.) του περιγραφόμενου γλωσσικού πόρου. Δεν έχει συμπληρωθεί για τους πόρους του ΠαΔΑ γιατί δεν έχει πρακτική εφαρμογή.
- **Μέρος του…** (Part of): Στο συγκεκριμένο πεδίο συμπληρώνεται κάποιος άλλος γλωσσικός πόρος ο οποίος περιλαμβάνει το γλωσσικό πόρο που περιγράφουμε. Δεν έχει συμπληρωθεί για τους πόρους του ΠαΔΑ γιατί δεν έχει πρακτική εφαρμογή.
- **Παρόμοιο με…** (Similar to): Πρόκειται για πεδίο όπου συμπληρώνεται με κάποιο γλωσσικό πόρο ο οποίος είναι παρόμοιος με το γλωσσικό πόρο που περιγράφουμε. Δεν έχει συμπληρωθεί για τους πόρους του ΠαΔΑ γιατί δεν έχει πρακτική εφαρμογή.
- **Σχέση** (Relation): Στο συγκεκριμένο πεδίο συμπληρώνονται πληροφορίες σχετικά με τη σχέση των δύο πόρων σε ελεύθερο κείμενο, στην περίπτωση που δεν καλύπτεται η σχέση τους με τις περιγραφές των τεσσάρων προηγουμένων περιπτώσεων.

3.4.2.4 Τα προαιρετικά πεδία γλωσσικών πόρων σε μορφή λεξικό-εννοιολογικών πόρων

Αντίστοιχα τα προαιρετικά πεδία της τεκμηρίωσης και περιγραφής των γλωσσικών πόρων σε μορφή λεξικο-εννοιολογικών πόρων δίδονται παρακάτω. Θα παρατηρήσουμε ότι ορισμένα από αυτά είναι κοινά με τους γλωσσικούς πόρους σε

σώμα κειμένου. Τα κοινά προαιρετικά πεδία, για τα οποία ισχύει η περιγραφή που δώσαμε στο προηγούμενο κεφάλαιο, είναι τα εξής:

- **Σύντομο όνομα γλωσσικού πόρου** (LRT short name)
- **Πάροχος γλωσσικού πόρου** (LRT provider)
- **Δημιουργός γλωσσικού πόρου** (LRT creator)
- **Χρηματοδότηση έργου** (Funding project)
- **Μεταφόρτωση λογότυπου** (Upload a logo)
- **Ανωνυμοποίηση** (Anomymized)
- **Τοποθεσία δειγμάτων του υλικού** (Samples location)
- **Κόστος** (Cost)
- **Δικαιώματα πρόσβασης** (Access rights)
- **Προβλεπόμενη εφαρμογή** (Intended application)
- **Τομέας** (Domain)
- **Γεωγραφική κάλυψη** (Geographic coverage)
- **Χρονική κάλυψη** (Time coverage)
- **Τεκμηριωμένο στο…** (Documented in)
- **Προηγούμενες εκδοχές** (Previous versions)
- **Εκδοχή του…** (Version of)
- **Μέρος του…** (Part of)
- **Παρόμοιο με…** (Similar to)
- **Σχέση** (Relation)

Αντίστοιχα τα προαιρετικά πεδία, τα οποία έχουν εφαρμογή στην περίπτωση των γλωσσικών πόρων σε μορφή λεξικο-εννοιολογικών πόρων, είναι τα παρακάτω:

- **Υποδιαίρεση λεξικο-εννοιολογικού πόρου** (LCR subclass): Περιλαμβάνει υποδιαιρέσεις στην περιγραφή της μορφής του πόρου. Οι λεξικο-εννοιολογικοί πόροι που είχαμε στη διάθεση μας ήταν ελάχιστοι και μικρής σχετικά έκτασης. Οπότε δεν ήταν πάντα δυνατή η χρήση των τριών προαιρετικών πεδίων στα οποία αναφερόμαστε.
- **Τύπος περιεχομένου** (Content type): Στο συγκεκριμένο πεδίο επιλέγεται από μακρά λίστα ο αναλυτικός τύπος της γλωσσικής πληροφορίας που περιέχεται

στον πόρο. Σε ορισμένους από τους γλωσσικούς πόρους του ΠαΔΑ έχει χρησιμοποιηθεί αυτό το προαιρετικό πεδίο. Σε πόρους με πρακτική δυσκολία ως προς τη συμπλήρωση του πεδίου αφού δεν ήταν εύκολη η επιλογή του σωστού τύπου περιεχομένου, αποφεύχθηκε η συμπλήρωση.

- **Συμβατότητα / συμμόρφωση** (Compliance): Πρόκειται για εξειδικευμένο πεδίο όπου επιλέγεται το λεξιλόγιο, πρότυπο κτλ. με το οποίο συμμορφώνεται ο γλωσσικός πόρος.

3.4.2.5 Υποχρεωτικά και προαιρετικά πεδία που χρησιμοποιήθηκαν

Στο κεφάλαιο αυτό παραθέτουμε ένα συνοπτικό πίνακα που περιλαμβάνει τα υποχρεωτικά και τα προαιρετικά πεδία που χρησιμοποιήσαμε στην τεκμηρίωση των γλωσσικών πόρων του Πανεπιστημίου Δυτικής Αττικής στο Clarin:el.

Υποχρεωτικά πεδία	Προαιρετικά πεδία
Κοινά πεδία σε σώματα κειμένου και λεξικό/εννοιολογικούς πόρους	
Όνομα του πόρου (LRT name)	Δημιουργός γλωσσικού πόρου (LRT creator)
Αναγνωριστικό γλωσσικού πόρου (LRT identifier	Μεταφόρτωση λογότυπου (Upload a logo)
Περιγραφή (Description)	Ανωνυμοποίηση (Anomymized)
Αριθμός εκδοχής του πόρου (version)	Τομέας (Domain)
Λέξεις κλειδιά (keywords)	Τύπος κειμένου (text type)
Επαφή (contact)	Τύπος περιεχομένου (Content type)
Όροι χρήσης του πόρου (licence terms)	Γεωγραφική κάλυψη (Geographic coverage)
Τρόπος διανομής του πόρου (distribution)	Χρονική κάλυψη (Time coverage)
Πεδία για σώματα κειμένου	
Υποκατηγορία (subclass)	
Περιλαμβάνονται προσωπικά ή ευαίσθητα προσωπικά δεδομένα (personal/sensitive data included)	
Είδος μέσου του σώματος κειμένου (corpus media type)	
Μέγεθος του σώματος κειμένου (amount/size unit)	
Μορφότυπος του σώματος κειμένου (data format)	
Γλώσσα του πόρου (language)	
Πεδία για λεξικο-εννοιολογικούς πόρους	
Επίπεδο κωδικοποίησης του πόρου (encoding level)	
Περιλαμβάνονται προσωπικά ή ευαίσθητα προσωπικά δεδομένα (personal/sensitive data included)	
Είδος μέσου του λεξικό/εννοιολογικού πόρου (media type)	
Μέγεθος του λεξικό/εννοιολογικού πόρου (amount/size unit)	
Μορφότυπος του λεξικό/εννοιολογικού πόρου (data format)	
Γλώσσα του πόρου (language)	

Πίνακας 1.Υποχρεωτικά μεταδεδομένα και προαιρετικά πεδία. Πηγή: https://www.clarin.gr/

3.5 Η επεξεργασία και δημοσίευση του υλικού

Ο curator (τεκμηριωτής), έχοντας εισαγάγει τα ανάλογα μεταδεδομένα σε όλους τους γλωσσικούς πόρους σε μορφή σώματος κειμένου ή λεξικό / εννοιολογικού πόρου που σχετίζονται με το ΠαΔΑ, στη συνέχεια τους προωθεί προς δημοσίευση. Ο metadata validator

(ελεγκτής μεταδομένων) ελέγχει την εγκυρότητα των μεταδεδομένων. Αν είναι εντάξει επικυρώνει τη δημοσίευση του πόρου ως συντακτικά ορθό. Αν όχι, τον επιστρέφει στον τεκμηριωτή για διορθώσεις. Στη συνέχεια, ο legal validator (νομικός ελεγκτής) ελέγχει αν ο πόρος είναι νομικά έγκυρος. Αν όλα είναι σωστά, θεωρεί τον πόρο ως νομικά έγκυρο και εγκρίνει τη δημοσίευση του. Αν όχι, ο τεκμηριωτής προχωράει στις ανάλογες διορθώσεις. Τέλος, και εφόσον όλα είναι ορθά, ο supervisor (υπεύθυνος αποθετηρίου) εγκρίνει τη δημοσίευση του γλωσσικού πόρου, ο οποίος εμφανίζεται στον κεντρικό κατάλογο του Clarin πλέον. Μπορούμε να κάνουμε εξαγωγή (export) της περιγραφής του γλωσσικού πόρου σε μορφότυπο XML.

3.6 Άδειες χρήσης και πνευματικά δικαιώματα

Ο δημόσιος διαμοιρασμός πληροφοριών και υλικού, το οποίο έχει παραχθεί από κάποιο δημιουργό προστατεύεται από κάποιους κανόνες δικαίου της πνευματικής ιδιοκτησίας. Ο δημιουργός έχει το δικαίωμα να επιτρέψει ή όχι την εκμετάλλευση του έργου του (Κανελλοπούλου-Μπότη, 2004).

Η ανοιχτή πρόσβαση στα δεδομένα, με στόχο την προώθηση της επιστήμης και της γνώσης, κερδίζει ολοένα έδαφος στον ακαδημαϊκό κόσμο και την επιστημονική κοινότητα. Ένα σημαντικό θέμα που τίθεται είναι αυτό των πνευματικών δικαιωμάτων και πως αυτά διασφαλίζονται, στην εποχή της ψηφιοποίησης και του διαδικτύου. Μια από τις άδειες που χρησιμοποιούνται πιο συχνά στο περιεχόμενο ανοιχτής πρόσβασης είναι οι άδειες χρήσης Creative Commons. Τα κύρια χαρακτηριστικά των άδειών αυτών είναι τα εξής (Αρτέμη κ.ά., 2010):

- Το υλικό διατίθεται δωρεάν μέσω του διαδικτύου
- Οι άδειες δεν χαρακτηρίζονται από αποκλειστικότητα
- Οι άδειες διατηρούν το εμπορικό ή ηθικό δικαίωμα πνευματικής ιδιοκτησίας του δημιουργού του έργου

Όπως προαναφέρθηκε, η υποδομή Clarin:el προωθεί την ανοιχτότητα και το διαμοιρασμό των δεδομένων σύμφωνα με τις αρχές FAIR: Findability (ευρεσιμότητα), Accessibility (προσβασιμότητα), Interoperability (διαλειτουργικότητα) και Reuse (επαναχρησιμοποίηση). Οι διάφοροι γλωσσικοί πόροι διατίθενται για ερευνητικούς σκοπούς σύμφωνα με συγκεκριμένες ανοιχτές άδειες χρήσης Creative Commons, έκδοσης 4.0 ή

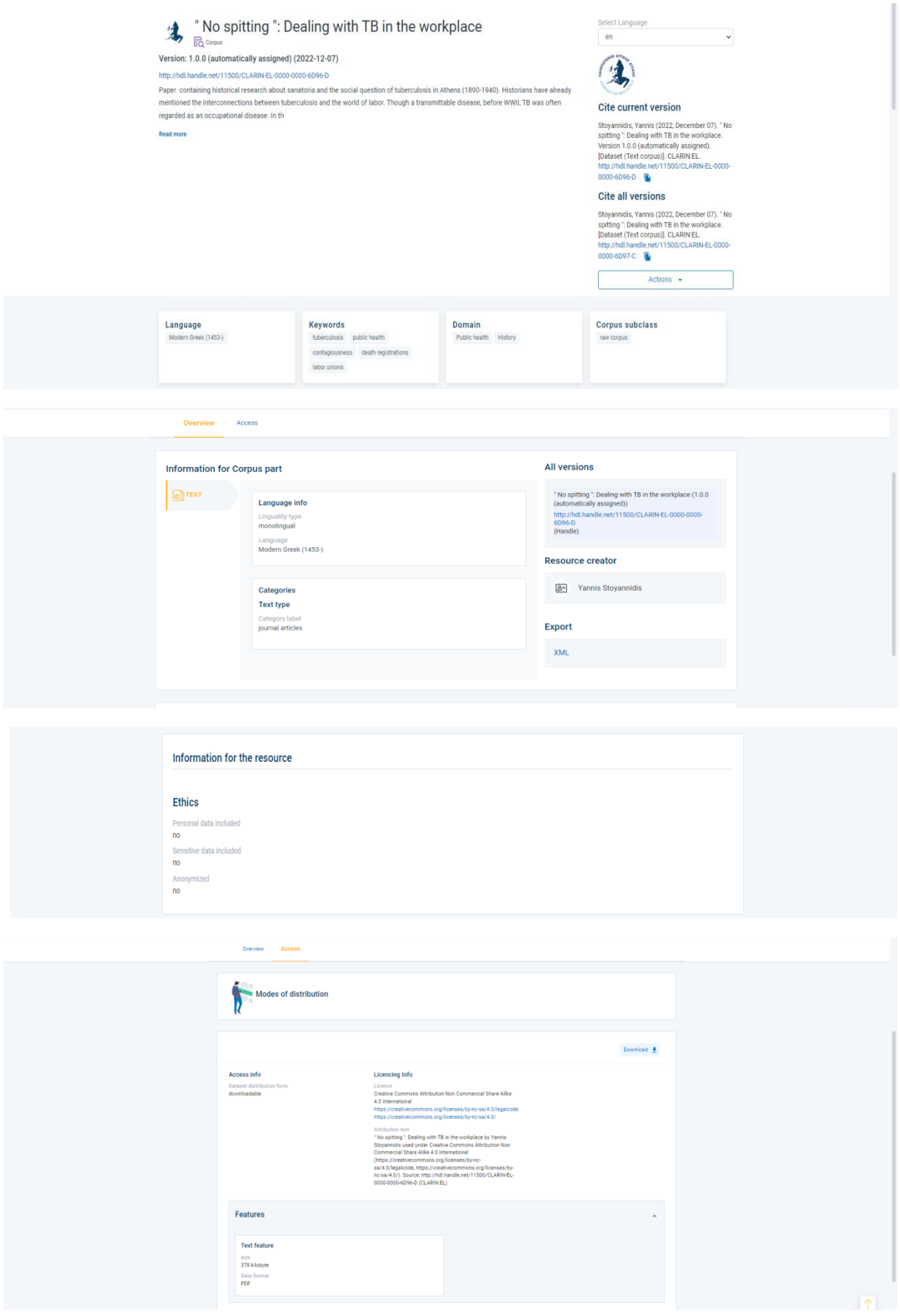
ανώτερη (Clarin:el, 2020). Στο διαμοιρασμό των πόρων του ΠαΔΑ στο Clarin:el ακολουθήσαμε την αρχική άδεια που είχε ο πόρος όταν πρωτοδημοσιεύτηκε και παρατέθηκε στην περιγραφή των μεταδεδομένων του. Πρόκειται για δύο άδειες Creative Commons, την CC-BY-NC-SA και την CC-BY-NC-ND.

Η άδεια CC-BY-NC-SA επιτρέπει μη εμπορική χρήση (non commercial) του πόρου που μπορεί να χρησιμοποιηθεί και διαμοιραστεί με παρόμοιο τρόπο (share alike) και κάτω από την ίδια άδεια με το πρωτότυπο. Η άδεια CC-BY-NC-ND επιτρέπει μη εμπορική χρήση (non commercial) του πόρου, άλλα χωρίς να μπορούν να διανεμηθούν τα παράγωγα του (non derivatives) (https://creativecommons.org/).

3.7 Τεκμηρίωση των γλωσσικών πόρων και πρακτική εφαρμογή

Η τεκμηρίωση των γλωσσικών πόρων είναι μια χρονοβόρα αλλά και ταυτόχρονα επίπονη διαδικασία, όπου συχνά καλούμαστε να λάβουμε προσωπικές αποφάσεις σχετικά με τις τιμές και τα πεδία που θα συμπληρώσουμε. Στο τρέχον κεφάλαιο θα προσπαθήσουμε να εξετάσουμε κάποια έμπρακτα παραδείγματα τεκμηρίωσης γλωσσικών πόρων που διασυνδέσαμε με το Clarin:el στα πλαίσια της παρούσας διπλωματικής εργασίας. Θα εξετάσουμε περιπτώσεις που ανταποκρίνονται σε διαφορετικές κατηγορίες υλικού.

Έχοντας παραθέσει σε παραπάνω κεφάλαιο συγκεντρωτικό πίνακα με τα υποχρεωτικά και τα προαιρετικά πεδία που χρησιμοποιήθηκαν στην περιγραφή και τεκμηρίωση των δύο κατηγοριών γλωσσικών πόρων που παράχθηκαν από το ΠαΔΑ στα πλαίσια του διδακτικού του έργου, στη συνέχεια θα προχωρήσουμε στην απτή εφαρμογή τους, δίδοντας δύο παραδείγματα, ένα ανά κατηγορία, από το διαθέσιμο υλικό μας.



Εικόνα 10. Τεκμηρίωση γλωσσικού πόρου (σώμα κειμένου). Πηγή: https://www.clarin.gr/

Στο παραπάνω στιγμιότυπο οθόνης διακρίνουμε ένα γλωσσικό πόρο που έχει διασυνδεθεί με την υποδομή. Πρόκειται για ένα άρθρο καθηγητή του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* το οποίο περιέχει ιστορικές έρευνες για τα σανατόρια και το κοινωνικό ζήτημα της φυματίωσης στην Αθήνα (1890-1940). Ο τίτλος δίδεται υποχρεωτικά στα αγγλικά, καθώς και στα ελληνικά που είναι η γλώσσα στην οποία έχει γραφτεί. Όλα τα δεδομένα δίδονται υποχρεωτικά στα αγγλικά και επίσης στα ελληνικά. Έχουμε προσθέσει το λογότυπο του ΠαΔΑ. Στη συνέχεια προχωρούμε στην περιγραφή του πόρου. Η συγκεκριμένη περιγραφή περιέχει βασικά στοιχεία του περιεχομένου του πόρου. Θα μπορούσαμε όμως να την εμπλουτίσουμε περισσότερο και να περιλαμβάνει και άλλα περιγραφικά στοιχεία, αν το κρίναμε σκόπιμο.

Στη συνέχεια περνάμε τα μεταδεδομένα. Προσθέτουμε τη γλώσσα στην οποία είναι γραμμένο το κείμενο, τις λέξεις – κλειδιά οι οποίες θα βοηθήσουν στην ευκολότερη εύρεση του πόρου μέσα στην πληθώρα των πόρων που ευρετηριάζει το Clarin:el και τη μορφή του που είναι «σώμα κειμένου» (raw corpus). Προσθέσαμε και τον τομέα στον οποίο αναφέρεται το κείμενο, αν και δεν πρόκειται για υποχρεωτικό μεταδεδομένο περιγραφής. Στο συγκεκριμένο παράδειγμα οι τομείς που εντάσσεται θεματικά το άρθρο είναι η δημόσια υγεία και η ιστορία.

Στη συνέχεια περνάμε κάποια στοιχεία αναφορικά με τη μορφή του σώματος κειμένου. Η γλώσσα γραφής αναφέρεται, καθώς και αν είναι μονόγλωσσο ή περιλαμβάνει περισσότερες από μια γλώσσες. Στις κατηγορίες αναφορικά με τον πόρο αναφέρεται η μορφή του (κειμενικός τύπος) και ο τρόπος που παρουσιάζεται. Προσθέσαμε τον τύπο κειμένου (Text type) που είναι ακόμα ένα προαιρετικό πεδίο και ανταποκρίνεται σε άρθρα περιοδικού (journal articles).

Το όνομα του δημιουργού του πόρου πάντα καταγράφεται. Δεν πρόκειται για υποχρεωτικό πεδίο αλλά κρίθηκε απαραίτητο να αναφερθεί καθαρά το όνομα του δημιουργού, για λόγους που έχουν δοθεί σε προηγούμενα κεφάλαια.

Στο πεδίο των πληροφοριών αναφορικά με τον πόρο συμπληρώνουμε αν το υλικό περιλαμβάνει προσωπικά δεδομένα, ευαίσθητα στοιχεία ή πρέπει να γίνει ανωνυμοποίηση σε κάποια δεδομένα. Στη συγκεκριμένη περίπτωση το κείμενο δεν περιλαμβάνει ούτε προσωπικά δεδομένα, ούτε ευαίσθητα στοιχεία και ως εκ τούτου, η ανωνυμοποίηση δεν κρίνεται απαραίτητη.

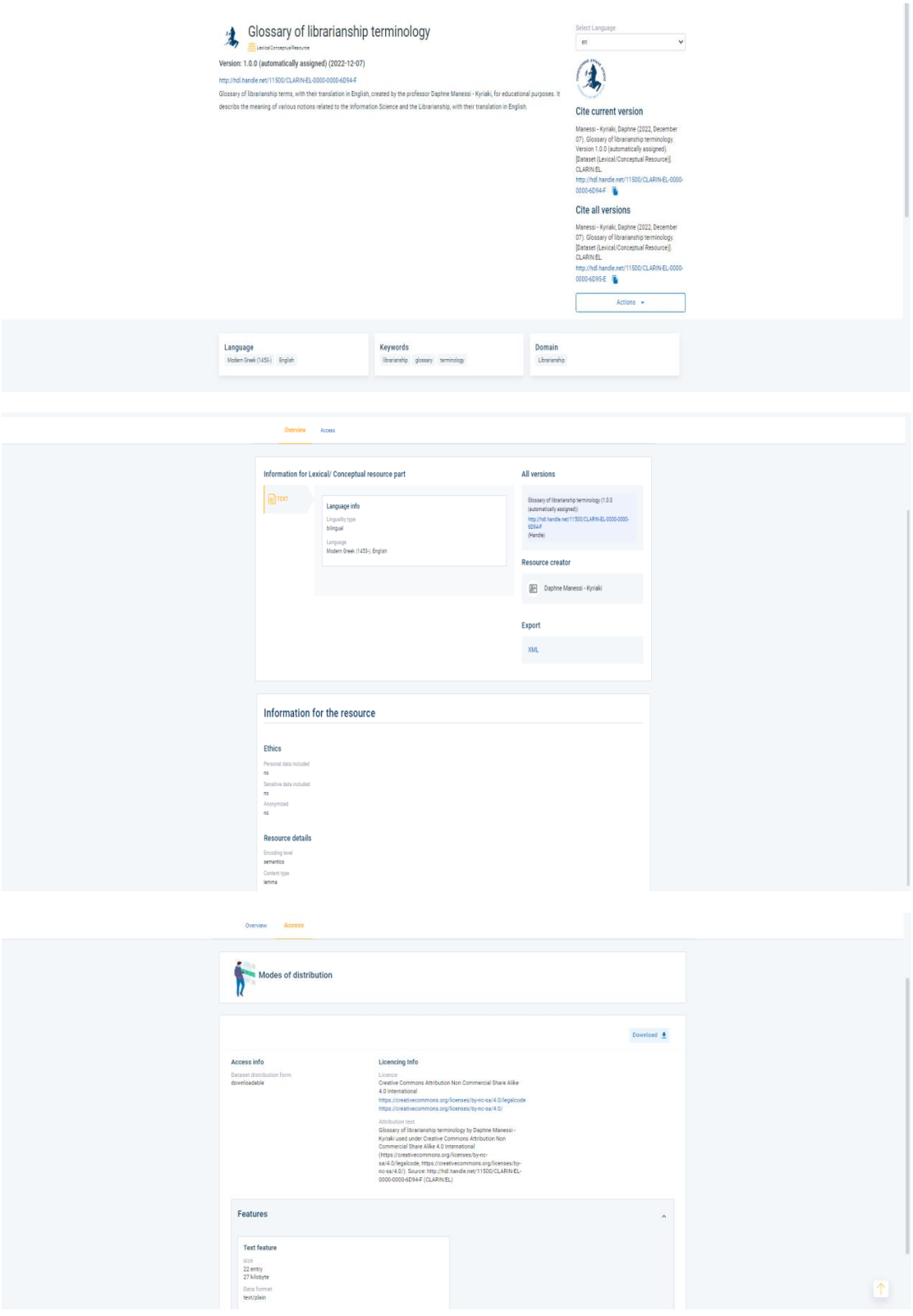
Το επόμενο πεδίο που συμπληρώσαμε αφορά τους τρόπους διάθεσης του γλωσσικού πόρου. Στις πληροφορίες πρόσβασης των στοιχείων του πόρου (dataset) επιλέξαμε να είναι αυτά μεταφορτώσιμα αλλά όχι επεξεργάσιμα. Πρόκειται για προσωπική επιλογή η οποία

είναι δυνατόν να αλλάξει αν κριθεί σκόπιμο. Στο συγκεκριμένο πεδίο συμπληρώνουμε την άδεια χρήσης με την οποία επιθυμούμε να διατεθεί ο πόρος. Πρόκειται για άδεια Creative Commons, Non Commercial, Share Alike 4.0 International.

Επιπλέον η μεταφόρτωση (download) του, για όραση και ανάλογη χρήση του υλικού, γίνεται από το συγκεκριμένο πεδίο. Έπειτα συμπληρώνουμε τα στοιχεία του πόρου. Στη συγκεκριμένη περίπτωση αναφέραμε ότι το μέγεθος του είναι 378 kilobytes. Το μέγεθος θα ήταν δυνατόν να δοθεί και με πληθώρα άλλων μετρήσιμων μορφών όπως τα megabytes, οι λέξεις, οι φάκελοι, το τεκμήριο, το αρχείο κτλ. Επίσης συμπληρώσαμε τη μορφή που βρίσκονται τα δεδομένα που είναι PDF στο συγκεκριμένο γλωσσικό πόρο.

Αυτόματα προστίθεται η εκδοχή (version) την καταχώρισης. Αντίστοιχα, μπορούμε να προσθέσουμε κάποια νέα version με μη αυτόματο τρόπο. Επίσης φαίνεται η ημερομηνία δημιουργίας και κατάθεσης στο Clarin:el, προς δημοσίευση του πόρου. Αυτόματα δημιουργείται και ο τρόπος βιβλιογραφικής παραπομπής του πόρου (cite current version) και δίδεται ένα μοναδικό αναγνωριστικό (handle). Από τη συγκεκριμένη version μπορούμε να δημιουργήσουμε μια νέα version, να αντιγράψουμε το αρχείο ή να ζητήσουμε να κατέβει η δημοσίευση του. Τα μεταδεδομένα του γλωσσικού πόρου μπορούν να εξαχθούν σε μορφή XML.

Στη συνέχεια θα εξετάσουμε έναν γλωσσικό πόρο σε λεξικο-εννοιολογική μορφή ο οποίος έχει παραχωρηθεί από καθηγήτρια του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης.* Πρόκειται για τον πόρο του οποίου στιγμιότυπο οθόνης διακρίνουμε παρακάτω. Πρόκειται για ένα γλωσσάρι βιβλιοθηκονομικών όρων.



Εικόνα 11. Τεκμηρίωση γλωσσικού πόρου (λεξικο-εννοιολογικός πόρος).

Πηγή: https://www.clarin.gr/

Αναφορικά με τη μορφή του υλικού, επιλέγουμε την τιμή του λεξικο-εννοιολογικού πόρου. Και προσθέτουμε την ανάλογη περιγραφή, η οποία είναι σύντομη αλλά θα μπορούσε να εμπλουτιστεί παραπάνω μελλοντικά.

Στο πεδίο των μεταδεδομένων, συμπληρώνουμε ότι η γλώσσα του πόρου είναι τα ελληνικά άλλα και τα αγγλικά, αφού περιλαμβάνεται μετάφραση της ορολογίας στην αγγλική γλώσσα. Οι λέξεις κλειδιά είναι βιβλιοθηκονομία, γλωσσάρι και ορολογία. Ο τομέας είναι βιβλιοθηκονομία.

Αναφορικά με τις πληροφορίες της γλώσσας, πρόκειται για έναν δίγλωσσο πόρο, σε ελληνικά και αγγλικά. Το όνομα του δημιουργού του που μας τον παραχώρησε αναφέρεται επίσης.

Ο συγκεκριμένος πόρος δεν περιέχει προσωπικά δεδομένα, ευαίσθητα στοιχεία και δεν χρειάζεται ανωνυμοποίηση. Ενώ η κωδικοποίησή του είναι σημειολογική (semantics). Ο τύπος κειμένου (Text type) είναι λήμμα (lemma). Το υλικό διατίθεται για μεταφόρτωση, σε άδεια Creative Commons, Non Commercial, Share Alike 4.0 International.

Όσον αφορά τα μεταδεδομένα των χαρακτηριστικών του πόρου, επιλέξαμε ως μέγεθος ότι πρόκειται για ένα 22 «λήμματα» (entries) και 27 kilobytes. Οπότε πρόκειται προφανώς για ένα μικρό σε μέγεθος πόρο. Ενώ η μορφή των δεδομένων είναι text. Να επισημάνουμε ότι εμείς μετατρέψαμε το αρχείο σε text, ώστε να έχει επεξεργάσιμη μορφή από το Clarin:el. Δεν ήταν πολύ δύσκολη διαδικασία γιατί ο πόρος ήταν μικρού μεγέθους.

Όπως παρατηρήσαμε η περιγραφή και τα μεταδεδομένα ενός γλωσσικού πόρου δίδονται και συμπληρώνονται ώστε να γίνει η σωστότερη και πληρέστερη δυνατόν τεκμηρίωση. Παρόλα αυτά, τα όρια ανάμεσα στις διαθέσιμες τιμές είναι συχνά τόσο κοντά που είναι εύκολο να γίνει κάποια λανθασμένη ή άστοχη επιλογή μεταδεδομένων, ιδίως όταν κάποιος τεκμηριωτής δεν είναι απόλυτα εξοικειωμένος με την υποδομή.

3.8 Κάποιες οδηγίες για την τεκμηρίωση των γλωσσικών πόρων

Στο κεφάλαιο αυτό θα προσπαθήσουμε να δώσουμε κάποιες οδηγίες σχετικά με την τεκμηρίωση και περιγραφή των γλωσσικών πόρων, οι οποίες θα μπορούσαν να ακολουθηθούν από κάποιον που επιθυμεί να τεκμηριώσει γλωσσικούς πόρους για κάποια συγκεκριμένο ιδρυματικό αποθετήριο που φιλοξενείται στην υποδομή του Clarin:el.

- **Δημιουργία γλωσσικού πόρου**: για τη δημιουργία μιας εγγραφής γλωσσικού πόρου στο Clarin:el επιλέγουμε πρώτα απ' όλα τη μορφή του. Έχουμε στη διάθεση μας τέσσερις επιλογές μορφών γλωσσικών πόρων. Τα σώματα κειμένου (corpus), τους λεξικο-εννοιολογικούς πόρους (lexical/conceptual resource), τις υπηρεσίες ή εργαλεία (service or tool) και τις γλωσσικές περιγραφές (language description). Η πιο διαδεδομένη μορφή είναι τα σώματα κειμένου. Έπειτα έχουμε τους λεξικο-εννοιολογικούς πόρους. Ενώ οι δύο τελευταίοι γλωσσικοί πόροι είναι πιο σπάνιοι.
- **Όνομα του πόρου (LRT name):** αναγράφουμε το όνομα του γλωσσικού πόρου. Το όνομα του στα αγγλικά είναι υποχρεωτικό, αφού είναι υποχρεωτική γλώσσα περιγραφής και τεκμηρίωσης στο Clarin:el (γενικά όλα τα στοιχεία πρέπει να μπουν υποχρεωτικά στα αγγλικά). Έπειτα μπορούμε να προσθέσουμε και την περιγραφή σε όποια άλλη γλώσσα επιθυμούμε και κρίνουμε ότι θα ήταν χρήσιμο.
- **Αναγνωριστικό γλωσσικού πόρου** (LRT identifier): Μπορούμε να το επιλέξουμε από μια λίστα προεπιλεγμένων τιμών ή αλλιώς, με τη δημοσίευση του πόρου δίδεται αυτόματα από την υποδομή ένα αναγνωριστικό handle.
- **Περιγραφή** (Description): δίνουμε την περιγραφή του πόρου υποχρεωτικά στα αγγλικά και έπειτα σε όποια άλλη γλώσσα επιθυμούμε. Η περιγραφή θα πρέπει να είναι όσο το δυνατόν πιο πλήρης και αντιπροσωπευτική του περιεχομένου του γλωσσικού πόρου.
- **Αριθμός εκδοχής του πόρου** (version): Τον επιλέγουμε εμείς, αλλιώς δίδεται αυτόματα από την εφαρμογή.
- **Πάροχος γλωσσικού πόρου** (LRT provider): προαιρετικό πεδίο όπου συμπληρώνουμε ποιος είναι ο πάροχος του γλωσσικού πόρου.
- **Δημιουργός γλωσσικού πόρου** (LRT creator): προαιρετικό πεδίο όπου συμπληρώνουμε ποιος είναι ο δημιουργός του γλωσσικού πόρου.
- **Μεταφόρτωση λογότυπου** (Upload a logo): Αν επιθυμεί κάποιος, μπορεί να προσθέσει ένα λογότυπο που θα σχετίζεται με το γλωσσικό πόρο, για καλύτερη παρουσίαση και θέαση.
- **Τομέας** (Domain): Οι τομείς στους οποίους αναφέρεται και έχει πρακτική εφαρμογή το υλικό, αν και προαιρετικό πεδίο, θα ήταν χρήσιμο να συμπληρωθεί γιατί βοηθάει στην ευκολότερη ευρεσιμότητα, ταξινόμηση και ευρετηρίαση του πόρου.

- **Γεωγραφική κάλυψη** (Geographic coverage): Στην περίπτωση που το υλικό αναφέρεται σε κάποια συγκεκριμένη γεωγραφική περιοχή, μπορεί να συμπληρωθεί το συγκεκριμένο πεδίο.
- **Χρονική κάλυψη** (Time coverage): Στην περίπτωση που το υλικό αναφέρεται σε κάποια συγκεκριμένη χρονική περίοδο, για να δώσουμε έμφαση, μπορούμε να επισημάνουμε ποια είναι αυτή.
- **Λέξεις κλειδιά** (keywords): Προσθέτουμε λέξεις κλειδιά που είναι αντιπροσωπευτικές του γλωσσικού πόρου και θα βοηθήσουν στον πιο εύκολο εντοπισμό, ταξινόμηση και ευρετηρίαση του.
- **Επαφή** (contact): Προσθέτουμε τα στοιχεία, όπως είναι το e-mail ή το ονοματεπώνυμο του ατόμου/ατόμων ή φορέα με τους οποίος μπορεί να έρθει κάποιος σε επαφή αναφορικά με κάποιο συγκεκριμένο γλωσσικό πόρο που έχει μεταφορτωθεί στο Clarin:el.
- **Προαιρετικά πεδία Documented in/ Previous versions/ Version of/ Part of/ Similar to/ Relation:** Τα συγκεκριμένα πεδία μπορούν να συμπληρωθούν αν κριθεί αναγκαίο, για έναν πόρο. Πρόκειται για περιπτώσεις όπου i)περιεχόμενο του υλικού αναφέρεται σε κάποια άλλη πηγή, ii)αποτελεί προηγούμενη εκδοχή κάποιου πόρου, iii)είναι μια εκδοχή σχετιζόμενη με άλλο γλωσσικό υλικό iv)είναι μέρος κάποιου άλλου πόρου v)είναι παρόμοιο με άλλο πόρο ή vi) έχει σχέση με κάποιον άλλο πόρο. Ο τεκμηριωτής χρησιμοποιεί κάποιο ή κάποια από τα προαιρετικά πεδία αυτά, ανάλογα με τις ανάγκες τεκμηρίωσης που προκύπτουν.
- **Περιλαμβάνονται προσωπικά ή ευαίσθητα προσωπικά δεδομένα** (personal/sensitive data included): Θα πρέπει να επισημανθεί αν ο γλωσσικός πόρος περιέχει ή όχι προσωπικά ή ευαίσθητα προσωπικά δεδομένα.
- **Ανωνυμοποίηση** (Anomymized): Η ανωνυμοποίηση πρέπει να γίνει στην περίπτωση που ένας γλωσσικός πόρος περιέχει προσωπικά ή ευαίσθητα δεδομένα που καλό θα ήταν να μην κοινοποιηθούν δημόσια. Θα μπορούσε πιθανόν να αφορά κάποια κείμενα συνεντεύξεων κτλ. Ο τεκμηριωτής είναι αυτός που θα κρίνει αν θα πρέπει να προβεί στη συγκεκριμένη ενεργεία ή όχι.
- **Όροι χρήσης και διανομή του πόρου** (licence terms): Συμπληρώνουμε με ποια άδεια διατίθεται στο κοινό ο πόρος. Οι άδειες χρήσης είναι Creative Commons. Επίσης επιλέγουμε αν το υλικό θα είναι μεταφορτώσιμο ή επεξεργάσιμο, ή αν πιθανόν θα μπορούν να το δούνε μόνο συγκεκριμένοι χρήστες και όχι το ευρύ

κοινό. Οι συγκεκριμένες ενέργειες είναι επιλογή του τεκμηριωτή και είναι σε συνάρτηση με την πολιτική του ακαδημαϊκού ιδρύματος, το δημιουργό και πάροχο του πόρου.

- **Μέγεθος** (amount/size unit): Από μια μακρά λίστα προεπιλεγόμενων τιμών, επιλέγουμε την ποσότητα του υλικού σε μετρήσιμη μονάδα καθώς και το μέγεθος του. Κάθε φορά προσέχουμε η τιμή να είναι η πιο ενδεδειγμένη. Για παράδειγμα, σε ένα λεξικό / εννοιολογικό πόρο, το μέγεθος μπορεί να μετρηθεί σε λήμματα.
- **Μορφότυπος** (data format): Από μια μακρά λίστα προεπιλεγόμενων τιμών, διαλέγουμε αυτή που αναφέρεται στο μορφότυπο που παρουσιάζεται το υλικό που έχουμε στη διάθεση μας.
- **Γλώσσα του πόρου** (language): Η τεκμηρίωση και διασύνδεση του υλικού πραγματοποιείται στην ελληνική πλατφόρμα της διεθνούς υποδομής του Clarin. Οπότε φυσικό είναι το υλικό να είναι στην ελληνική γλώσσα. Παρόλα αυτά, κάποιες σπάνιες φορές μπορεί να έχουμε υλικό σε άλλες γλώσσες ή ακόμα το φαινόμενο του δίγλωσσου / πολύγλωσσου πόρου είναι συχνό. Για παράδειγμα σε ένα δίγλωσσο γλωσσάρι, εκτός από την ελληνική γλώσσα, θα πρέπει να καταγράψουμε και την άλλη γλώσσα (ή γλώσσες σε πολύγλωσση περίπτωση) στην οποία βρίσκεται το υλικό.
- **Υποκατηγορία** (subclass): θα πρέπει επίσης να δηλώσουμε σε ποια υποκατηγορία εντάσσεται ο πόρος, ανάμεσα σε κάποιες επιλέξιμες τιμές (πχ text, audio, video, image, numerical text).
- **Επίπεδο κωδικοποίησης του πόρου** (encoding level): Όταν έχουμε ένα λεξικο-εννοιολογικούς πόρο, θα πρέπει να δώσουμε και το επίπεδο κωδικοποίησης που του ταιριάζει, ανάμεσα σε τιμές μιας λίστας (πχ phonetics, morphology, syntax, pragmatics, semantics....). Επίσης μπορούμε προαιρετικά να συμπληρώσουμε το πεδίο του **τύπου περιεχομένου** (content type) από μια ανάλογη λίστα τιμών (πχ audio type, lemma type, morphological type..).

3.9 Κάποιοι περιορισμοί του Clarin:el

Τέλος, κάποια μειονεκτήματα που θέτουν περιορισμούς και δυσκολίες σχετίζονται με κάποιες λειτουργίες του Clarin:el και καλό θα ήταν να διορθωθούν μελλοντικά ώστε να γίνει

πιο εύχρηστη η υποδομή. Μερικά σημαντικά θέματα που παρουσιάζει και αναφέρονται είτε στο σχεδιασμό της εφαρμογής, είτε σε εναλλακτικές του χρήστη, είναι τα παρακάτω:

- Η δύσχρηστη επικύρωση των πόρων από τη μεριά του υπεύθυνου του ακαδημαϊκού αποθετηρίου καθώς και των λοιπών ελεγκτών που εμπλέκονται στη δημοσίευση των πόρων, που θα πρέπει να κάνουν πολλά κλικ σε διαφορετικές οθόνες, χωρίς να μπορούν να προβούν σε μαζική επικύρωση περισσότερων του ενός πόρου τη φορά.
- Η αδυναμία να πραγματοποιήσει κάποιος διορθώσεις μόνος του στα μεταδεδομένα του πόρου του, εφόσον τον έχει μεταφορτώσει ήδη στο Clarin:el. Για παράδειγμα, αν ένας δημιουργός γλωσσικού πόρου αλλάξει κάτι στο έργο του, ενώ αυτό έχει ήδη δημοσιευτεί, για να αλλάξει τα μεταδεδομένα του θα πρέπει να απευθυνθεί στην ομάδα που διαχειρίζεται την υποδομή, αφού μόνος του είναι αδύνατον να το κάνει. Αυτό δημιουργεί καθυστερήσεις στο χρήστη που φαινομενικά θα πρέπει να είναι σίγουρος από την αρχή για τα στοιχεία που θα θελήσει να περάσει για την τεκμηρίωση ενός πόρου, πριν τη δημοσίευση του.
- Η ανεύρεση συγκεκριμένων πόρων μέσα στο ιδρυματικό αποθετήριο, με στόχο να γίνει κάποια τροποποίηση, δεν είναι εύκολη, αφού συχνά αναγκαζόμαστε να ανοίξουμε έναν έναν τους πόρους για να βρούμε αυτόν που ψάχνουμε. Επίσης συχνά η αναζήτηση με βάση κάποιο συγκεκριμένο μεταδεδομένο δεν δίνει ακριβή αποτελέσματα, παρότι το μεταδεδομένο με το οποίο αναζητάμε κάποιο συγκεκριμένο πόρο είναι συμπληρωμένο με την ανάλογη τιμή.
- Κάποιες φορές περνιούνται λάθος μεταδεδομένα από το χρήστη, λόγω μη κατανόησης του τι θα πρέπει να επιλέξει και συμπληρώσει σε κάθε πεδίο, αναφορικά με τα μεταδεδομένα και τεκμηρίωση ενός γλωσσικού πόρου. Αυτό αφορά περισσότερο την περιγραφή των γλωσσικών πόρων. Παρανοήσεις και λάθη μπορεί να γίνουν κυρίως με τις κατηγορίες του υλικού, το μέγεθος των πόρων (size) ή τη μορφή των δεδομένων (data format) αφού κάποιοι όροι προς επιλογή μοιάζουν και μπορεί να γίνει λάθος στην επιλογή τους. Πιθανόν θα πρέπει να υπάρχει καλύτερη και αναλυτικότερη περιγραφή για τον τεκμηριωτή. Ή ίσως η ομάδα του Clarin πρέπει να ελέγχει ενδεικτικά κάποιους αρχικούς πόρους και να επισημάνει τα λάθη τεκμηρίωσης στον τεκμηριωτή,

πριν αυτός προλάβει να διασυνδέσει με την εφαρμογή μεγάλο αριθμό γλωσσικών πόρων με μη σωστά μεταδεδομένα, που μετά θα είναι χρονοβόρο και δύσκολο να διορθωθούν εφόσον θα έχουν ήδη δημοσιευθεί.

- Επίσης θα ήταν χρήσιμο να μπορεί να εξαγάγει κάποιος μόνος του τα στοιχεία όλων των γλωσσικών πόρων, συγκεντρωμένους σε κάποιο έγγραφο excel ή άλλης εύχρηστης μορφής, έτσι ώστε να μπορεί να παρατηρήσει συγκριτικά όλους τους πόρους που έχει ανεβάσει και τα μεταδεδομένα τους και να προβεί πιο εύκολα στις απαραίτητες διορθώσεις όπου χρειάζονται. Τη συγκεκριμένη διαδικασία μπορεί να την πραγματοποιήσει η ομάδα του Clarin άλλα όχι ο απλός χρήστης.

Παρά τις όποιες δυσλειτουργίες της, θα πρέπει να επισημάνουμε παρόλα αυτά ότι γενικά η ελληνική εκδοχή της υποδομής του Clarin:el βελτιώνεται συνεχώς και εμπλουτίζεται, ενώ την υποστήριξη της έχει αναλάβει μια δυναμική ομάδα που υποστηρίζει άμεσα το χρήστη στα αιτήματα του. Επίσης, το *Ηλεκτρονικό Εγχειρίδιο Χρήσης* της Υποδομής διαθέτει πληθώρα από χρήσιμες πληροφορίες αναφορικά με την υποδομή, τις οποίες μπορεί κάποιος να συμβουλευτεί για τη μεταφόρτωση και τεκμηρίωση του υλικού του (https://clarin-platform-documentation.readthedocs.io/el/stable/).

Πρόσφατο, το Σεπτέμβριο του 2022, το Clarin:el αξιολογήθηκε επίσημα και πιστοποιήθηκε ως CLARIN B-Centre της CLARIN ERIC και ως αξιόπιστο αποθετήριο δεδομένων από την CoreTrustSeal, η οποία αποτελεί ένα διεθνή, μη κυβερνητικό και μη κερδοσκοπικό οργανισμό που προσφέρει στα ψηφιακά αποθετήρια δεδομένων μια πιστοποίηση βασικού επιπέδου με βάση τις Απαιτήσεις Βασικών Αξιόπιστων Αποθετηρίων Δεδομένων (https://www.clarin.gr/).

3.10 Ανακεφαλαίωση

Η συγκεκριμένη διπλωματική εργασία αποτελεί περιγραφή της έρευνας που πραγματοποιήσαμε με σκοπό τη συλλογή γλωσσικών πόρων που σχετίζονται με το Πανεπιστήμιο Δυτικής Αττικής και στη συνέχεια, τη διασύνδεση τους με την υποδομή Clarin:el. Η μεθοδολογία που ακολουθήθηκε περιγράφηκε διεξοδικά στο συγκεκριμένο κεφάλαιο. Στη συνέχεια θα εξετάσουμε τα ευρήματα που απορρέουν από την έρευνα.

Κεφάλαιο 4. Αποτελέσματα – Ευρήματα

Στο συγκεκριμένο κεφάλαιο περιγράφονται αναλυτικά τα αποτελέσματα της έρευνας μας αναφορικά με το υλικό που συλλέξαμε, τα χαρακτηριστικά του, την τεκμηρίωση του, τους παραγωγούς του ή άλλα παρεμφερή στοιχεία.

4.1 Κάποια κοινά χαρακτηριστικά στα μεταδεδομένα

Στην περίπτωση του υλικού που συλλέξαμε σε μορφή σώματος κειμένου ή λεξικο εννοιολογικού πόρου, θα είχε ενδιαφέρον να αναφέρουμε κάποια κοινά χαρακτηριστικά στα μεταδεδομένα που χρησιμοποιήσαμε στην περιγραφή και εκφράζουν τους περισσότερους πόρους:

- Οι περισσότεροι πόροι, εκτός από ελάχιστους που είναι στα αγγλικά, είναι μόνο στην ελληνική γλώσσα. Κάποιοι λίγοι περιέχουν δίγλωσσο υλικό, κυρίως αποδόσεις ορολογίας στα αγγλικά
- Τα σώματα κειμένου που συλλέξαμε αποτελούνται αποκλειστικά από πρωτογενή υλικό
- Το επίπεδο κωδικοποίησης των λεξικο-εννοιολογικών πόρων είναι σημασιολογία κατά πλειοψηφία ή μορφολογία
- Όλοι οι πόροι αποτελούνται από κείμενο (text part)
- Επιλέχθηκε οι πόροι να είναι όλοι καταφορτώσιμοι (downloadable) ώστε να μπορούν να τους μεταφορτώσουν οι ενδιαφερόμενοι
- Διαλέξαμε κανένας πόρος να μην είναι επεξεργάσιμος (processable) για αποφυγή αλλοίωσης τους και επίσης επειδή σε κάποιες μορφές μικτού υλικού, το οποίο περιλαμβάνει μη επεξεργάσιμους μορφότυπους, μαζί με τους επεξεργάσιμους από το Clarin:el, το υλικό είναι μερικώς μη επεξεργάσιμο στην παρούσα φάση, ούτως ή άλλως. Μελλοντικά θα μπορούσε να αλλάξει η κατάσταση κάποιων πόρων σε επεξεργάσιμους
- Επιλέχθηκε οι πόροι να είναι ανοιχτοί στο κοινό, οπότε έχουμε no private distribution στην οποία δεν θα είχαν όλοι πρόσβαση
- Κανένας από τους πόρους δεν περιέχει προσωπικά ή ευαίσθητα προσωπικά δεδομένα, οπότε δεν χρειάστηκε να ανωνυμοποιηθούν

- Η μορφή των δεδομένων (data format) είναι σε μορφότυπο TEXT, PDF, MS-WORD ή και PPT και Excel, τα τελευταία δύο μόνο ως μέρος μικτού πόρου που περιέχει και αρχείο στο συγκεκριμένο μορφότυπο μαζί με το υλικό σε αποδεκτούς από το Clarin τύπους
- Οι άδειες χρήσης των γλωσσικών πόρων που επιλέχθηκαν είναι ανοιχτής πρόσβασης Creative Commons

4.2 Αναλυτική παρουσίαση του υλικού

Στη συνέχεια θα παρουσιάσουμε αναλυτικά τη μορφή και την προέλευση των γλωσσικών πόρων που συλλέχθηκαν και διασυνδέθηκαν με το Clarin:el, στα πλαίσια της παρούσας διπλωματικής εργασίας.

Αναφορικά με κάποιους γλωσσικούς πόρους στους οποίους δύσκολα μπορούμε να διακρίνουμε το ίδρυμα προέλευσης τους, δηλαδή αν έχουν παραχθεί από το Πανεπιστήμιο Δυτικής Αττικής ή τους προκατόχους του που είναι το ΤΕΙ Αθηνών και το ΤΕΙ Πειραιά, όπως για παράδειγμα είναι τα συγγράμματα του Κάλλιππου, θα θεωρήσουμε ως ορόσημο το 2018 που είναι η ημερομηνία ιδρύσεως του ΠαΔΑ. Οπότε και τα επτά συγγράμματα του Κάλλιππου από διδάσκοντες στο πρώην τμήμα *Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*, που έχουμε ενδεικτικά συμπεριλάβει στο υλικό που μεταφορτώσαμε στο Clarin:el, θεωρούμε ότι ανήκουν σε υλικό που παράχθηκε από το ΤΕΙ Αθηνών, εφόσον έχουν δημοσιευθεί το 2015.

Συνολικά έχουμε διασυνδέσει 193 γλωσσικούς πόρους με το Clarin:el. Πιο αναλυτικά, διαπιστώνουμε ότι το μεγαλύτερο μέρος του υλικού που έχουμε μεταφορτώσει στο Clarin:el προέρχεται από το ΤΕΙ Πειραιά (80 γλωσσικοί πόροι), έπειτα από το ΤΕΙ Αθηνών (68 γλωσσικοί πόροι) και τέλος από το Πανεπιστήμιο Δυτικής Αττικής (45 γλωσσικοί πόροι). Αναφορικά με το ΤΕΙ Αθηνών, έχουμε βρει συλλογικά μεγαλύτερο αριθμό γλωσσικών πόρων απ’ ότι στο ΤΕΙ Πειραιά. Παρόλα αυτά, οι περισσότεροι από τους διαθέσιμους γλωσσικούς πόρους που παράχθηκαν από το ΤΕΙ Αθηνών και καταφέραμε να συλλέξουμε βρίσκονται σε μορφότυπους οι οποίοι δεν μπορούν να υποστηριχθούν από το Clarin:el (όπως για παράδειγμα μεγάλο μέρος των Ανοιχτών Ακαδημαϊκών Μαθημάτων του ΤΕΙ Αθηνών βρίσκονται σε μορφή PPT). Οπότε οι συγκεκριμένοι πόροι δεν μπορούν να διασυνδεθούν με την εφαρμογή. Το Πανεπιστήμιο Δυτικής Αττικής είναι αντίστοιχα μια σχετικά πρόσφατη δομή. Οπότε φυσικό είναι το υλικό που έχει παράγει στα πλαίσια τη εκπαιδευτικής του διαδικασίας να είναι μικρότερο σε αριθμό.

Προέλευση πόρου ανά ακαδημαϊκό ίδρυμα	Αριθμός γλωσσικών πόρων
ΤΕΙ Πειραιά	80
ΤΕΙ Αθηνών	68
Πανεπιστήμιο Δυτικής Αττικής	45

Πίνακας 2. Προέλευση πόρου ανά ακαδημαϊκό ίδρυμα

Το είδος του υλικού που ενσωματώσαμε στην υποδομή του Clarin:el αποτελείται κατά μεγάλη πλειοψηφία από Ανοιχτά Ακαδημαϊκά Μαθήματα του ΤΕΙ Αθηνών και του ΤΕΙ Πειραιά. Έπειτα, έχουμε κάποιο αριθμό γλωσσικών πόρων οι οποίοι μας αποστάλθηκαν απευθείας από τους δημιουργούς τους, που είναι διδάσκοντες στο Πανεπιστήμιο Δυτικής Αττικής, έπειτα από επικοινωνία μαζί τους. Πρόκειται για άρθρα ή διάφορες δημοσιεύσεις τους. Ενδεικτικά έχουμε ενσωματώσει οκτώ εργασίες φοιτητών του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* (επτά διπλωματικές εργασίες του παρόντος Προγράμματος Μεταπτυχιακών Σπουδών *«Διαχείριση Πληροφοριών σε Βιβλιοθήκες, Αρχεία, Μουσεία»* και μία πτυχιακή εργασία), επτά συγγράμματα από τον Κάλλιππο με δημιουργούς καθηγητές του Τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* και έξι περιγραφές μαθημάτων, που στάλθηκαν από τη διδάσκουσα καθηγήτρια τους. Φυσικά το Clarin:el θα μπορούσε να εμπλουτιστεί μελλοντικά με επιπλέον διπλωματικές και πτυχιακές εργασίες καλής ποιότητας και συγγράμματα του Κάλλιππου και από άλλες σχολές και τμήματα του Πανεπιστημίου Δυτικής Αττικής.

Είδος γλωσσικού πόρου	Αριθμός γλωσσικών πόρων
Ανοιχτά Ακαδημαϊκά Μαθήματα	148
Υλικό που στάλθηκε απευθείας από τους διδάσκοντες του ΠαΔΑ	24
Εργασίες φοιτητών του ΠαΔΑ	8
Κάλλιππος	7
Περιγραφές μαθημάτων	6

Πίνακας 3. Είδος γλωσσικού πόρου

Στη συνέχεια θα εξετάσουμε τους παρεχόμενους γλωσσικούς πόρους ανά σχολή και τμήμα, τόσο του Πανεπιστημίου Δυτικής Αττικής, όσο και των δύο Τεχνολογικών Ιδρυμάτων, των ΤΕΙ Αθηνών και Πειραιά δηλαδή.

Αναφορικά με τα δύο ΤΕΙ, Αθηνών και Πειραιά, παρατηρούμε ότι όλο το υλικό που διαθέσαμε στο Clarin:el προέρχεται από τα Ανοιχτά Ακαδημαϊκά Μαθήματα που έχουν δημιουργηθεί και δημοσιευθεί με άδειες ανοιχτής πρόσβασης Creative Commons. Πιο συγκεκριμένα, διασυνδέσαμε στη γλωσσική υποδομή 80 μαθήματα από το ΤΕΙ Πειραιά και 68 από το ΤΕΙ Αθηνών.

Προέλευση Ανοιχτού Ακαδημαϊκού Μαθήματος	Αριθμός γλωσσικών πόρων
ΤΕΙ Αθηνών	68
ΤΕΙ Πειραιά	80

Πίνακας 4. Προέλευση Ανοιχτών Ακαδημαϊκών Μαθημάτων

Ταξινομώντας τους παρεχόμενους πόρους ανά σχολή του ΤΕΙ Αθηνών, διαπιστώνουμε ότι ο μεγαλύτερος αριθμός προέρχεται από τη σχολή Τεχνολογικών Εφαρμογών (36 πόροι), ακολουθούμενης από τη σχολή Καλλιτεχνικών Σπουδών (11 πόροι), τη σχολή Επαγγελμάτων Υγείας και Πρόνοιας (9 πόροι), τη σχολή Τεχνολογίας Τροφίμων και Διατροφής (7 πόροι) και τέλος τη σχολή Διοίκησης και Οικονομίας (5 πόροι).

Υλικό ανά σχολή του ΤΕΙ Αθηνών	Αριθμός γλωσσικών πόρων
Σχολή Διοίκησης και Οικονομίας	5
Σχολή Επαγγελμάτων Υγείας και Πρόνοιας	9
Σχολή Καλλιτεχνικών Σπουδών	11
Σχολή Τεχνολογίας Τροφίμων και Διατροφής	7
Σχολή Τεχνολογικών Εφαρμογών	36

Πίνακας 5. Υλικό ανά σχολή του ΤΕΙ Αθηνών

Εξετάζοντας τους γλωσσικούς πόρους του ΤΕΙ Αθηνών ανά τμήμα που τους παράγει, οι περισσότεροι προέρχονται από το τμήμα *Ναυπηγών Μηχανικών* (19 πόροι), ακολουθούμενο από τα τμήματα *Οινολογίας και Τεχνολογίας Ποτών*, *Μηχανικών Πληροφορικής* και *Φωτογραφίας και Οπτικοακουστικών Τεχνών* (από 6 πόρους το καθένα). Πιο αναλυτικά παρουσιάζονται οι παρεχόμενοι γλωσσικοί πόροι ανά τμήμα, στον παρακάτω πίνακα.

Υλικό ανά τμήμα του ΤΕΙ Αθηνών	Αριθμός γλωσσικών πόρων
Ναυπηγών Μηχανικών	19
Οινολογίας και Τεχνολογίας Ποτών	6
Μηχανικών Πληροφορικής	6
Φωτογραφίας και Οπτικοακουστικών Τεχνών	6
Πολιτικών Μηχανικών και Μηχανικών Τοπογραφίας και Γεωπληροφορικής	5
Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης	4
Μηχανικών Ενεργειακής Τεχνολογίας	4
Γραφιστικής	3
Αισθητικής και Κοσμετολογίας	2
Οπτικής και Οπτομετρίας	2
Ραδιολογίας και Ακτινολογίας	2
Μηχανικών Βιοιατρικής Τεχνολογίας	2
Συντήρησης Αρχαιοτήτων και Έργων Τέχνης	1
Δημόσιας Υγείας και Κοινοτικής Υγείας	1
Οδοντικής Τεχνολογίας	1
Ιατρικών Εργαστηρίων	1
Εσωτερικής Αρχιτεκτονικής, Διακόσμησης και Σχεδιασμού Αντικειμένων	1
Τεχνολογίας Τροφίμων	1
Διοίκησης Επιχειρήσεων - Τουριστικών Επιχειρήσεων και Επιχειρήσεων Φιλοξενίας	1

Πίνακας 6. Υλικό ανά τμήμα του ΤΕΙ Αθηνών

Αντίστοιχα, θα αναλύσουμε το παρεχόμενο υλικό ανά σχολή και τμήμα του ΤΕΙ Πειραιά. Το περισσότερο υλικό προέρχεται από τη σχολή Τεχνολογικών Εφαρμογών (57 πόροι). Ενώ η σχολή Διοίκησης και Οικονομίας διαθέτει 23 πόρους.

Υλικό ανά σχολή του ΤΕΙ Πειραιά	Αριθμός γλωσσικών πόρων
Τεχνολογικών Εφαρμογών	57
Διοίκησης και Οικονομίας	23

Πίνακας 7. Υλικό ανά σχολή του ΤΕΙ Πειραιά

Το τμήμα *Διοίκησης Επιχειρήσεων* είναι ο δημιουργός των περισσότερων παρεχόμενων πόρων στο Clarin:el (16 πόροι), ακολουθούμενο από τα τμήματα *Μηχανολόγων Μηχανικών* (14 πόροι), *Μηχανικών Αυτοματισμού* (13 πόροι) ή *Ηλεκτρονικών Μηχανικών* (12 πόροι). Στον παρακάτω πίνακα παρουσιάζονται οι παρεχόμενοι γλωσσικοί πόροι ανά τμήμα του ΤΕΙ Πειραιά πιο αναλυτικά.

Υλικό ανά τμήμα του ΤΕΙ Πειραιά	Αριθμός γλωσσικών πόρων
Διοίκηση Επιχειρήσεων	16
Μηχανολόγων Μηχανικών	14
Μηχανικών Αυτοματισμού	13
Ηλεκτρονικών Μηχανικών	12
Μηχανικών Ηλεκτρονικών Υπολογιστών Συστημάτων	8
Λογιστικής και χρηματοοικονομικής	7
Ηλεκτρολόγων Μηχανικών	6
Πολιτικών Μηχανικών	4

Πίνακας 8. Υλικό ανά τμήμα του ΤΕΙ Πειραιά

Το Πανεπιστήμιο Δυτικής Αττικής αποτελεί το συνεχιστή των δύο τεχνολογικών ιδρυμάτων, Αθηνών και Πειραιά. Οι προερχόμενοι από το ακαδημαϊκό ίδρυμα γλωσσικοί πόροι, οι οποίοι έχουν διασυνδεθεί με τη γλωσσική εφαρμογή αποτελούνται κατά πλειοψηφία από υλικό που δόθηκε απευθείας από τους διδάσκοντες του ΠαΔΑ, έπειτα από επικοινωνία μαζί τους (άρθρα, προϊόν έρευνας τους, δημοσιεύσεις κτλ.). Στη συνέχεια διαθέτουμε κάποια επιλεγμένα συγγράμματα από τον Κάλλιππο, εργασίες φοιτητών, καθώς και περιγραφές μαθημάτων.

Είδος υλικού από το ΠαΔΑ	Αριθμός γλωσσικών πόρων
Υλικό που δόθηκε απευθείας από τους διδάσκοντες του ΠαΔΑ	24
Συγγράμματα από τον Κάλλιππο	7
Διπλωματικές εργασίες	7
Περιγραφές μαθημάτων	6
Πτυχιακή εργασία	1

Πίνακας 9. Είδος γλωσσικού πόρου του ΠαΔΑ

Ταξινομώντας το υλικό που προέρχεται από το ΠαΔΑ ανά σχολή, διαπιστώνουμε ότι ο μεγαλύτερος αριθμός γλωσσικών πόρων προέρχεται από τη σχολή Διοικητικών Οικονομικών και Κοινωνικών Επιστημών. Δεδομένου ότι η έρευνα διεξάγεται στα πλαίσια του Προγράμματος Μεταπτυχιακών σπουδών τμήματος της συγκεκριμένης σχολής και της οικειότητας με τους διδάσκοντες, οι οποίοι ανταποκρίθηκαν πιο εύκολα στην παρούσα έρευνα, δικαιολογείται η μεγαλύτερη συλλογή υλικού από τη συγκεκριμένη σχολή και αντίστοιχα το τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης. Πιθανόν να υπάρχει και άλλο γλωσσικό υλικό, σε άλλες σχολές και τμήματα του ΠαΔΑ, το οποίο δεν κατέστη δυνατόν να ανακτηθεί. Έπειτα, το υλικό προέρχεται από τη σχολή Μηχανικών, ενώ διαθέτουμε και δύο γλωσσικούς πόρους από διδάσκοντες των σχολών Εφαρμοσμένων Τεχνών και Πολιτισμού και Επιστημών Υγείας και Πρόνοιας.

Υλικό ανά σχολή του ΠαΔΑ	Αριθμός γλωσσικών πόρων
Διοικητικών Οικονομικών και Κοινωνικών Επιστημών	35
Μηχανικών	8
Εφαρμοσμένων Τεχνών και Πολιτισμού	1
Επιστημών Υγείας και Πρόνοιας	1

Πίνακας 10. Υλικό ανά σχολή του ΠαΔΑ

Αναλύοντας το γλωσσικό υλικό που διασυνδέσαμε με το Clarin:el, όπως ειπώθηκε ήδη και εξηγήθηκε παραπάνω, το μεγαλύτερο μέρος, με 29 γλωσσικούς πόρους, προέρχεται

από το τμήμα *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*. Έπειτα, υπάρχουν 8 πόροι από το τμήμα *Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής*, 6 από το τμήμα *Αγωγής και Φροντίδας στην Πρώιμη Παιδική Ηλικία* και από ένας πόρος από τα τμήματα *Γραφιστικής και Οπτικής Τεχνολογίας* και *Βιοιατρικών Επιστημών*. Θα ήταν χρήσιμο να επισημάνουμε ότι ο πόρος από το τελευταίο τμήμα εκφράζει υλικό αναφορικά με πεδίο προσωπικού ενδιαφέροντος διδάσκοντος του τμήματος και δεν έχει σχέση με το αντικείμενο σπουδών πάνω στις βιοιατρικές επιστήμες.

Υλικό ανά τμήμα του ΠαΔΑ	Αριθμός γλωσσικών πόρων
Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης	29
Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής	8
Αγωγής και Φροντίδας στην Πρώιμη Παιδική Ηλικία	6
Γραφιστικής και Οπτικής Τεχνολογίας	1
Βιοιατρικών Επιστημών	1

Πίνακας 11. Υλικό ανά τμήμα του ΠαΔΑ

Αναφορικά με τον τύπο του διαθέσιμου στο Clarin:el γλωσσικού πόρου, η μεγάλη πλειοψηφία είναι σώματα κειμένου (188 γλωσσικοί πόροι), ενώ υπάρχουν και 5 λεξικό – εννοιολογικοί πόροι. Υλικό από εργαλεία ή υπηρεσίες γλωσσικής επεξεργασίας καθώς και από γλωσσικές περιγραφές και υπολογιστικά μοντέλα δεν έχει συλλεχθεί καθόλου.

Τύπος γλωσσικού πόρου	Αριθμός γλωσσικών πόρων
Σώμα κειμένου	188
Λεξικό – εννοιολογικοί πόροι	5

Πίνακας 12. Τύπος γλωσσικού πόρου

Όλο το υλικό των 193 γλωσσικών πόρων που συλλέξαμε και διασυνδέσαμε με τη υποδομή του Clarin:el είναι επίσης σε μορφή κειμένου (text).

Μορφή πόρου	Αριθμός γλωσσικών πόρων
Text	193

Πίνακας 13. Μορφή των γλωσσικών πόρων

Αναφορικά με τη γλώσσα στην οποία είναι γραμμένοι οι γλωσσικοί πόροι που σωρευτήκαν στο Clarin:el, η ελληνική γλώσσα συναντάται 190 φορές, η αγγλική 6 φορές ενώ έχουμε και ένα γλωσσάρι σε διάλεκτο της Μεσσηνίας, με τη μετάφραση τους στα κοινά νέα ελληνικά.

Γλώσσα πόρων	Αριθμός γλωσσικών πόρων
Ελληνικά	190
Αγγλικά	6
Μεσσηνιακή διάλεκτος	1

Πίνακας 14. Γλώσσα των πόρων

Οι 189 γλωσσικοί πόροι είναι μονόγλωσσοι, στη συντριπτική τους πλειοψηφία στην ελληνική γλώσσα (άλλωστε το Clarin:el έχει ως αποστολή κυρίως να συλλέξει υλικό στα ελληνικά). Παρόλα αυτά, έχουμε και κάποιους λίγους, μεμονωμένους πόρους στα αγγλικά, οι οποίοι αποτελούν δημοσιεύσεις καθηγητών του ΠαΔΑ στη γλώσσα αυτή. Οι δίγλωσσοι πόροι είναι 4. Οι 3 στα ελληνικά και τα αγγλικά και ο ένας στα κοινά νέα ελληνικά και το γλωσσικό ιδίωμα της Μεσσηνίας.

Γλωσσικός τύπος	Αριθμός γλωσσικών πόρων
Μονόγλωσσοι	189
Δίγλωσσοι	4

Πίνακας 15. Γλωσσικός τύπος

Στον παρακάτω συγκεντρωτικό πίνακα διακρίνουμε μια ανασκόπηση του γλωσσικού υλικού του Πανεπιστημίου Δυτικής Αττικής, που συλλέξαμε και διασυνδέσαμε με το Clarin:el, αναφορικά με τη μορφή και το είδος του:

	Τύπος		Μορφή	Γλώσσα			Γλωσσικός τύπος	
	Σώμα κειμένου	Λεξικό / εννοιολογικός πόρος	Κείμενο (text)	Ελληνικά	Αγγλικά	Μεσσηνιακή διάλεκτος	Μονόγλωσσο	Δίγλωσσο
Ανοιχτά ακαδημαϊκά μαθήματα	148		148	148			148	
Υλικό που στάλθηκε από τους διδάσκοντες	19	5	24	22	6	1	20	4
Εργασίες φοιτητών	8		8	7	1		8	
Συγγράμματα από τον Κάλλιππο	7		7	7			7	
Περιγραφές μαθημάτων	6		6	6			6	

Πίνακας 16. Συγκεντρωτικός πίνακας του υλικού σε αριθμητικές τιμές

4.3 Κυριότερα ευρήματα/ αποτελέσματα

Στο προηγούμενο κεφάλαιο παρουσιάσαμε αναλυτικά και ομαδοποιημένα τους γλωσσικούς πόρους που παράχθηκαν από το Πανεπιστήμιο Δυτικής Αττικής, καθώς και από τα δύο τεχνολογικά ιδρύματα, το ΤΕΙ Αθήνας και το ΤΕΙ Πειραιά, με βάση τη μορφή και τα χαρακτηριστικά τους. Όπως διαπιστώσαμε, η προέλευση των περισσότερων πόρων που διασυνδέσαμε με το Clarin:el ήταν από το ΤΕΙ Πειραιά. Η πλειοψηφία του υλικού προέρχεται από τα Ανοιχτά Ακαδημαϊκά Μαθήματα. Ενώ οι σχολές και τα τμήματα των μηχανικών και των τεχνολογικών εφαρμογών γενικά, τόσο στα δύο τεχνολογικά ιδρύματα, όσο και στην περίπτωση του Πανεπιστημίου Δυτικής Αττικής, υπερτερούν ως παραγωγοί των πόρων. Η πλειοψηφία του υλικού που συλλέχθηκε, δεν δόθηκε απευθείας από τους διδάσκοντες αλλά αποτελείται από γλωσσικούς πόρους, αποτέλεσμα του διδακτικού έργου του ΠαΔΑ και κυρίως των προκατόχων του, το οποίο έχει διατεθεί και δημοσιευθεί με άδειες ανοιχτής πρόσβασης. Πρόκειται πλειοψηφικά για υλικό αποτελούμενο από σώματα κειμένου, σε ελληνική γλώσσα και μονόγλωσσο, σε τύπο text.

4.4 Περιορισμοί

Στα πλαίσια της συγκεκριμένης ερευνητικής προσπάθειας ήρθαμε αντιμέτωποι με κάποιους περιορισμούς οι οποίοι δυσκόλεψαν το έργο μας. Πιο συγκεκριμένα, η έρευνα και οι σκοποί της αντιμετωπίστηκαν κάποιες φορές με ερωτήσεις όπως «που θα χρησιμοποιηθεί το υλικό», «τι είναι η συγκεκριμένη εφαρμογή και ποιοι είναι οι βαθύτεροι λόγοι που συλλέγει γλωσσικό υλικό» κτλ. που έδειχναν κάποια καχυποψία.

Τα ενημερωτικά μηνύματα ηλεκτρονικού ταχυδρομείου προς αναζήτηση υλικού δεν απαντήθηκαν, λόγω φόρτου εργασίας των διδασκόντων, παρά μόνο από μια μικρή μειοψηφία. Και γενικά κρίθηκε αναγκαία η αυτοπρόσωπη επιδιωκόμενη αυθόρμητη συνάντηση δια ζώσης, στους χώρους της σχολής, στο γραφείο τους ή τους διαδρόμους και αίθουσες του Πανεπιστημίου Δυτικής Αττικής, με τους πιθανολογούμενους παραγωγούς γλωσσικών πόρων, διδάσκοντες του Πανεπιστημίου Δυτικής Αττικής. Η συγκεκριμένη ενέργεια έγινε κατά κύριο λόγο με την περίπτωση των διδασκόντων στο τμήμα *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης* που ήταν γνωστοί και πιο οικείοι στη συγγραφέα της διπλωματικής, και σε μικρότερο ποσοστό με διδάσκοντες άλλων σχολών και τμημάτων.

Η έρευνα μας οδήγησε κυρίως σε γλωσσικούς πόρους που παράχθηκαν από τους διδάσκοντες και το τμήμα *«Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης»*, με τους οποίους υπήρχε μεγαλύτερη οικειότητα. Ενώ από τα υπόλοιπα τμήματα του ΠαΔΑ, μεμονωμένα, κάποιοι διδάσκοντες παρείχαν υλικό, καθώς επίσης είχαμε πρόσβαση σε υλικό που βρέθηκε μέσω διαδικτύου και ψηφιακών αποθετηρίων.

Υπήρχε ακόμα σημαντικός χρονικός περιορισμός που περιόρισε την ανεύρεση του υλικού αφού η διπλωματική ολοκληρώνεται μέσα σε συγκεκριμένα χρονικά περιθώρια και η αναζήτηση υλικού, μέσω επικοινωνίας με τους παραγωγούς πόρων, μπορεί να είναι εξαιρετικά χρονοβόρα. Έχοντας όμως διασυνδέσει πλέον το Πανεπιστήμιο Δυτικής Αττικής με τη διαδικτυακή υποδομή του Clarin, έχουν τεθεί οι βάσεις για τη διαρκή τροφοδότηση του αποθετηρίου με νέο υλικό και τη συνεχόμενη συμπλήρωση του αποθετηρίου της σχολής με τους ανάλογους γλωσσικούς πόρους, κάθε φορά που θα παράγεται κάτι καινούριο.

Όσον αφορά το υλικό από τον Κάλλιππο, όπως είδαμε διασυνδέσαμε με την υποδομή μόνο συγγράμματα από καθηγητές του τμήματος *«Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης»*. Θα μπορούσε να προστεθεί υλικό και από άλλους διδάσκοντες σε άλλες σχολές και τμήματα του ΠαΔΑ, έχοντας λάβει αντίστοιχα τη συγκατάθεση και άδεια τους. Ο χρονικός περιορισμός που είχαμε δεν μας έδωσε το

περιθώριο να ζητήσουμε και να λάβουμε τις αντίστοιχες άδειες από παραγωγούς πόρων άλλων σχολών και τμημάτων.

Οι εργασίες φοιτητών που διασυνδέθηκαν με το Clarin είναι ενδεικτικές και περιορισμένες σε αριθμό. Πιθανόν θα μπορούσαν να διασυνδεθούν και άλλες εργασίες μελλοντικά, οι οποίες θα χαρακτηρίζονται από υψηλή ποιότητα και ενδιαφέρον για την ερευνητική κοινότητα. Διδάσκοντες στο ΠαΔΑ θα μπορούσαν να μας επισημάνουν ανάλογες πτυχιακές που περιέχουν υλικό με επιστημονικό ενδιαφέρον.

Άλλοι σημαντικοί περιορισμοί που αντιμετωπίσαμε σχετίζονται με τη μορφή των μη υποστηριζόμενων από την υποδομή μορφότυπων. Ή επίσης ότι κάποιο από το υλικό που βρέθηκε βρίσκεται σε μορφή που δεν υποστηρίζεται πλέον, όπως είναι για παράδειγμα το adobe flash player που δεν μπορούμε ούτε καν να ανοίξουμε και να συμβουλευτούμε το υλικό. Το συγκεκριμένο υλικό θα πρέπει να υποστεί επεξεργασία και συνεχή αναβάθμιση σε πιο σύγχρονη εκδοχή, με update.

Άλλος σημαντικός περιορισμός που κληθήκαμε να αντιμετωπίσουμε ήταν αυτός των πνευματικών δικαιωμάτων ή των δικαιωμάτων διάθεσης του υλικού. Παραγωγοί γλωσσικών πόρων μας ενημέρωσαν ότι έχουν σχετικό υλικό άλλα δεν έχουν τα ανάλογα πνευματικά δικαιώματα πάνω του, ώστε να μπορούν να το διαθέσουν. Η ακόμα διδάσκοντες που εργάζονταν προηγουμένως σε διαφορετικό πανεπιστημιακό ίδρυμα, είχαν ήδη διασυνδέσει τους γλωσσικούς πόρους που διέθεταν με το ιδρυματικό αποθετήριο στο Clarin του προηγούμενου πανεπιστημίου τους.

Επίσης, δεν θα πρέπει να παραβλέψουμε τις δυσκολίες που δημιουργούνται στο έργο μας λόγω κάποιων περιορισμών του Clarin:el, οι οποίοι περιεγράφηκαν ήδη πιο πάνω.

Η διαδικασία της αναζήτησης του υλικού, της επικοινωνίας με τους δημιουργούς του, της τεκμηρίωσης και της διασύνδεσης του με την υποδομή του Clarin:el είναι ιδιαίτερα χρονοβόρα. Θα πρέπει να γίνει σωστή διαχείριση του χρόνου που θα έχει κάποιος στη διάθεση του για την ολοκλήρωση του έργου. Στην υλοποίηση της διπλωματικής εργασίας, επιχειρήθηκε αξιολόγηση και ιεράρχηση των δράσεων, ώστε να πραγματοποιηθεί έγκαιρα η διπλή δράση της θεωρητικής και πρακτικής εφαρμογής που ήταν αναγκαία. Παρόλα αυτά, κάποιες στιγμές χάθηκε πολύτιμος χρόνος στην προσπάθεια κατανόησης σε βάθος του θέματος της διπλωματικής εργασίας. Βέβαια θα πρέπει να επισημανθεί ότι η σε βάθος κατανόηση ενός θέματος είναι πάντα μια σημαντική και απαραίτητα διεργασία, έτσι ώστε να παραχθεί το σωστό και άρτιο επιδιωκόμενο αποτέλεσμα.

Η επικοινωνία με τους δημιουργούς εμπεριείχε επίσης δυσκολία, λόγω κάποιας προσωπικής συστολής ή του γεγονότος ότι δεν ήταν εύκολα προσβάσιμοι ή διαθέσιμοι. Ενώ

και η τεκμηρίωση του υλικού, με την επιλογή των σωστών μεταδεδομένων και δεδομένων περιγραφής εμπεριέχει δυσκολίες, λόγω κάποιων όρων του Clarin:el που δεν είναι εύκολα διακριτό πότε πρέπει να προτιμηθούν έναντι άλλων. Αν ξεκινούσαμε στην παρούσα φάση τη συγκεκριμένη έρευνα, με την εμπειρία που υπάρχει πλέον αναφορικά με το θέμα και τις προεκτάσεις του, θα επιδιωκόταν ακόμα πιο πιστή εφαρμογή των χρονοδιαγραμμάτων των σταδίων υλοποίησης της δράσης και εξοικονόμησης χρόνου, όπως και πρότερης κατανόησης όλων των διαστάσεων του θέματος, πριν την αναζήτηση του υλικού και την επικοινωνία με τους δημιουργούς του. Γενικά, στην παρούσα διπλωματική εργασία επιχειρήθηκε να εξηγηθούν όσο το δυνατόν καλύτερα οι δυσκολίες, επιλογές ή εναλλακτικές αναφορικά με τη συλλογή, ταξινόμηση και τεκμηρίωση των γλωσσικών πόρων, έτσι ώστε να καταλάβουν καλύτερα την όλη διαδικασία οι όποιοι μελλοντικοί χρήστες και τεκμηριωτές.

4.5 Συμβουλές για βέλτιστες πρακτικές

Το στήσιμο ενός ψηφιακού αποθετηρίου για λογαριασμό κάποιου ακαδημαϊκού ή ερευνητικού ιδρύματος, καθώς και η συλλογή γλωσσικών πόρων, η διασύνδεση και η τεκμηρίωση τους μέσα στην υποδομή του Clarin:el είναι μια εξαιρετικά σύνθετη, επίπονη και χρονοβόρα διαδικασία. Αν κάποιος ξεκινούσε να κάνει την ίδια εργασία και διαδικασίες για λογαριασμό κάποιου άλλου φορέα, με βάση την παρούσα εμπειρία μας, θα προτεινόταν να ακολουθήσει κάποιες βέλτιστες πρακτικές που θα διευκόλυναν το έργο του. Πρόκειται για τις παρακάτω:

- Να μελετήσει την ανάλογη βιβλιογραφία και να κατανοήσει καλύτερα τόσο τους όρους που πραγματεύεται η συγκεκριμένη εργασία, όσο και όλες τις διαστάσεις του θέματος.
- Να κατανοήσει τι υλικό από γλωσσικούς πόρους θα πρέπει να συλλέξει καθώς και σε ποιους μορφότυπους θα πρέπει αυτό να βρίσκεται ούτως ώστε να είναι αξιοποιήσιμο από το Clarin:el.
- Μέσω τις κατανόησης όλων των διαστάσεων της επιδιωκόμενης εργασίας, να αποκτήσει σιγουριά και να καταφέρει να πείσει τους δημιουργούς των γλωσσικών πόρων, στο στάδιο της επικοινωνίας μαζί τους, η οποία θα είναι πιο αποτελεσματική, έτσι ώστε να διαθέσουν όσο το δυνατόν περισσότερους γλωσσικούς πόρων, σε αξιοποιήσιμους μορφότυπους προς διασύνδεση με το Clarin:el.

- Να προσπαθήσει, με διάφορους τρόπους και μεθόδους, να συλλέξει όσο το δυνατόν περισσότερο και ποικίλο γλωσσικό υλικό που έχει παραχθεί από το φορέα του, από όσο το δυνατόν περισσότερους παραγωγούς. Αν για παράδειγμα ο συγκεκριμένος φορέας είναι κάποιο ακαδημαϊκό ίδρυμα, καλό θα είναι το υλικό να προέρχεται από όσο το δυνατόν περισσότερες σχολές, τμήματα ή διδάσκοντες του.
- Εκτός από το υλικό που ενδεχομένως θα συλλέξει απευθείας από τους παραγωγούς του, καλό θα ήταν να αναζητήσει και διάφορους άλλους γλωσσικούς πόρους που ανταποκρίνονται στο φορέα και μπορούν να διατεθούν με όρους ανοιχτής πρόσβασης και άδειες Creative Commons.
- Ορθό θα είναι να ζητηθεί η άδεια από τους δημιουργούς των γλωσσικών πόρων σχετικά με το αν επιθυμούν τη διασύνδεση του υλικού τους με το Clarin:el.
- Να μελετήσει σε βάθος τις οδηγίες σχετικά με την τεκμηρίωση των γλωσσικών πόρων και των μεταδεδομένων της περιγραφής τους στο Clarin:el στο *Ηλεκτρονικό εγχειρίδιο χρήσης Clarin:el.*
- Δεδομένου ότι όταν γίνει οι επικύρωση των γλωσσικών πόρων και η δημοσίευση τους στην υποδομή, έπειτα δεν είναι δυνατή η διόρθωση όποιων λανθασμένων επιλογών, παρά μόνον αν τις κάνει unpublish και αναιρέσει τη δημοσίευση τους και τις ξαναδημιουργήσει από την αρχή, το άτομο που μεταφορτώνει το υλικό στο Clarin:el θα πρέπει να περάσει πρώτα σωστά όλα τα μεταδεδομένα και να βεβαιωθεί ότι δεν έχει κάνει λάθη, πριν την δημοσίευση τους.
- Στην περίπτωση που κάποιος γλωσσικός πόρος βρίσκεται σε μορφότυπο μη επεξεργάσιμο από το Clarin:el αλλά παρόλα αυτά η μετατροπή του σε επεξεργάσιμο μορφότυπο είναι εύκολη και δεν αλλοιώνεται η μορφή του υλικού, πιθανόν θα μπορούσε να πραγματοποιηθεί από τον τεκμηριωτή των πόρων η συγκεκριμένη μετατροπή.

Κεφάλαιο 5. Συζήτηση – Συμπεράσματα – Μελλοντικές επεκτάσεις

Η παρούσα διπλωματική εργασία επιδίωξε να συλλέξει υλικό που αποτελείται από γλωσσικούς πόρους, οι οποίοι έχουν παραχθεί στα πλαίσια του διδακτικού έργου του Πανεπιστημίου Δυτικής Αττικής, από τους διδάσκοντες, τους φοιτητές ή άλλους εμπλεκόμενους άμεσα με το πανεπιστήμιο. Το υλικό ταξινομήθηκε, τεκμηριώθηκε και διασυνδέθηκε με την ψηφιακή υποδομή γλωσσικών πόρων του Clarin:el. Στην υποδομή δημιουργήθηκε το ψηφιακό αποθετήριο του ΠαΔΑ.

Στο τελευταίο κεφάλαιο της έρευνας μας, θα επισημανθούν τα συμπεράσματα που εξήχθησαν καθώς και ποιες θα μπορούσαν να είναι οι μελλοντικές προεκτάσεις που απορρέουν από τη συγκεκριμένη έρευνα.

5.1 Ανακεφαλαίωση

Οι βασικοί έμπρακτοι στόχοι της διπλωματικής εργασίας ήταν η διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με την ψηφιακή υποδομή του Clarin:el. Το ΠαΔΑ έγινε μέλος του ελληνικού παραρτήματος μιας διεθνούς υποδομής και αποτελεί πλέον κομμάτι μιας διεθνούς κοινότητας που διαμοιράζεται γλωσσικούς πόρους και εργαλεία γλωσσικής τεχνολογίας, με στόχο την προώθηση της έρευνας στην Ελλάδα και το εξωτερικό.

Η ερευνητική προσέγγιση που χρησιμοποιήθηκε ήταν η έρευνα πεδίου με στόχο τη συλλογή του ζητούμενου υλικού. Πραγματοποιήθηκε δράση επικοινωνίας με τους δημιουργούς των γλωσσικών πόρων οι οποίοι είναι διδάσκοντες στο ΠαΔΑ, γραπτά, με μήνυμα ηλεκτρονικού ταχυδρομείου, με ενημερωτική συνάντηση μέσω της πλατφόρμας Teams ή και δια ζώσης. Η ανταπόκριση ήταν σχετικά μικρή αλλά παρόλα αυτά, συγκεντρώθηκε κάποιος ικανοποιητικός αριθμός γλωσσικού υλικού. Στη συνέχεια συλλέξαμε άλλους πόρους που είχαν ανοιχτές άδειες Creative Commons. Πρόκειται για συγγράμματα από τις ακαδημαϊκές ψηφιακές εκδόσεις του Κάλλιππου. Επίσης συλλέχτηκαν κείμενα που ήταν σε υποστηριζόμενους μορφότυπους από το Clarin:el, από τα *Ανοιχτά Ακαδημαϊκά Μαθήματα* του ΤΕΙ Αθηνών και του ΤΕΙ Πειραιά, τα οποία αποτελούν την προηγούμενη νομική μορφή που συγχωνευόμενα δημιούργησαν το Πανεπιστήμιο Δυτικής Αττικής. Τέλος, ενδεικτικά συμπεριλήφθηκαν στην έρευνα διπλωματικές εργασίες μικρού αριθμού φοιτητών του τμήματος *Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*, και του

συγκεκριμένου μεταπτυχιακού προγράμματος σπουδών, έπειτα από ερώτηση αν θα επιθυμούσαν το υλικό τους να διασυνδεθεί με το Clarin:el. Στη συνέχεια, το υλικό, με την κατάλληλη επεξεργασία και τεκμηρίωση, διασυνδέθηκε με το Clarin:el.

Ανακεφαλαιώνοντας τα αποτελέσματα της έρευνας μας, οι γλωσσικοί πόροι σε σώματα κειμένου σε αμιγώς ελληνική γλώσσα υπερτερούν.

5.2 Συζήτηση / Συμπεράσματα

Τα οφέλη της διασύνδεσης του Πανεπιστήμιου Δυτικής Αττικής με το Clarin:el είναι σημαντικά. Το ΠαΔΑ γίνεται μέλος μιας διεθνούς κοινότητας διαμοιρασμού πόρων και ανταλλαγής γνώσης και τεχνογνωσίας. Οι δημιουργοί των πόρων κερδίζουν επίσης προβολή και πιθανόν αναγνώριση του έργου τους.

Η υποδομή του Clarin:el παρέχει σημαντικό έργο, ενώ η ομάδα του είναι πάντα πρόθυμοι να συνδράμει το χρήστη στη μεταφόρτωση και τεκμηρίωση των γλωσσικών του πόρων.

5.3 Αξιοποίηση / Πρακτικές προεκτάσεις της έρευνας

Μια σημαντική πρακτική προέκταση της συλλογής υλικού του Πανεπιστημίου Δυτικής Αττικής και της διασύνδεσης του με το Clarin:el αποτελεί η συγκέντρωση σε ένα μέρος όλου του υλικού το οποίο έχει παραχθεί από το ΠαΔΑ και η ευκολότερη αναζήτηση και ανεύρεση του με στόχο τη βέλτιστη αξιοποίηση των γλωσσικών του πόρων. Θα μπορούσαμε να φανταστούμε το αποθετήριο του Πανεπιστημίου Δυτικής Αττικής στο Clarin:el ως μια μορφή “ψηφιακής βιβλιοθήκης» που περιλαμβάνει το υλικό που έχει παραχθεί στα πλαίσια του εκπαιδευτικού έργου του ακαδημαϊκού ιδρύματος.

5.4 Μελλοντικές επεκτάσεις / Πρακτικές Προεκτάσεις της Έρευνας

Την παρούσα στιγμή, στο αποθετήριο του ΠαΔΑ στο Clarin:el δεν περιλαμβάνεται όλο το υλικό που έχει παραχθεί στα πλαίσια της εκπαιδευτικής του διαδικασίας. Περιλαμβάνεται μόνον αυτό που καταφέραμε να συγκεντρώσουμε, που έχουμε την άδεια να δημοσιεύσουμε, που δεν προστατεύεται από δεσμευτικούς όρους πνευματικών δικαιωμάτων ή που έχει

μορφότυπους επιθυμητούς ή αξιοποιήσιμους από το Clarin:el. Μελλοντικά θα ήταν επιθυμητό να συνεχιστεί η συστηματική τροφοδότηση της υποδομής με γλωσσικούς πόρους του Πανεπιστημίου Δυτικής Αττικής. Ίσως το Clarin:el αναπτύξει στο μέλλον εργαλεία που θα καθιστούν αξιοποιήσιμο υλικό σε μη υποστηριζόμενους στην παρούσα φάση μορφότυπους. Οπότε τότε θα είναι δυνατή η συμπλήρωση του ψηφιακού αποθετηρίου και με επιπλέον γλωσσικούς πόρους που παραλήφθηκαν προς το παρόν, λόγω μη επιθυμητών μορφότυπων.

5.5 Γλωσσικό υλικό που μπορεί να προστεθεί μελλοντικά

Έχοντας καταγράψει όλες τις ενέργειες μας σχετικά με τη διασύνδεση του Πανεπιστημίου Δυτικής Αττικής με την υποδομή του Clarin:el και τον εμπλουτισμό του αποθετηρίου του με υλικό, τις προκλήσεις και τους περιορισμούς που αντιμετωπίσαμε, τελειώνοντας, θα παραθέσουμε κάποιο υλικό που θα μπορούσε να αναζητηθεί μελλοντικά και να διασυνδεθεί με την εφαρμογή. Πρόκειται για το παρακάτω υλικό:

- Επιπλέον υλικό που θα μπορούσε να συλλεχθεί από τους διδάσκοντες του ΠαΔΑ
- Υλικό που θα προέρχεται από τους φοιτητές του ΠαΔΑ, όπως αξιόλογες πτυχιακές και διπλωματικές εργασίες
- Συγγράμματα από τον Κάλλιππο
- Άρθρα που δημοσιεύθηκαν
- Υλικό από συμμετοχή σε ερευνητικά προγράμματα
- Υλικό σε μορφότυπους που δεν υποστηρίζονται στην παρούσα φάση από το Clarin:el αλλά θα μπορούσαν να υποστηριχθούν στο μέλλον
- Επιπλέον ανοιχτά ακαδημαϊκά μαθήματα
- Μελλοντική παραγωγή γλωσσικών πόρων του ΠαΔΑ, στα πλαίσια της ερευνητικής του λειτουργίας

Βιβλιογραφικές Αναφορές

Ελληνόγλωσση βιβλιογραφία

Αθηνά. https://www.athenarc.gr/. [Ανάκτηση 2/10/2022].

Ανοιχτά Ακαδημαϊκά Μαθήματα στο ΤΕΙ Αθήνας. https://ocp.teiath.gr. [Ανάκτηση 2/9/2022].

Ανοιχτά Ακαδημαϊκά Μαθήματα στο ΤΕΙ Πειραιά. https://opencourses.gr/results.xhtml?ln=el&uni=TEI+Πειραιά. [Ανάκτηση 13/9/2022].

Απολλωνίς. https://apollonis-infrastructure.gr/. [Ανάκτηση 4/9/2022].

Αρτέμη, Π., Ευαγόρου, Α., Ζέρβας, Μ., Ντίνη-Κουνάδη, Α. (2010) Η πορεία προς την ανοιχτή πρόσβαση μέσω των Creative Commons – Περίπτωση Κτίσις. Στο Κακάλη, Κ., [επιμέλεια] *19ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών – Επιστημονικές κοινότητες και βιβλιοθήκες στον κόσμο της κοινωνικής δικτύωσης και συνέργειας.* Αθήνα : Βιβλιοθήκη – Υπηρεσία Πληροφόρησης Παντείου Πανεπιστημίου.

Clarin:el. https://www.clarin.gr/. [Ανάκτηση 1/7/2022].

Clarin:el. (2020). *Ηλεκτρονικό εγχειρίδιο χρήσης Clarin:el*. Ανακτήθηκε στις 2/9/2022, από https://clarin-platform-documentation.readthedocs.io/el/stable/.

Γαβριηλίδου, Μ., & Πιπερίδης, Σ., (2021). *Τεχνητή Νοημοσύνη, Γλωσσική Τεχνολογία και Γλωσσικές υποδομές.* Αθήνα: Περιοδικό «Οικονομικός Ταχυδρόμος»,21/10/2021. https://www.ot.gr/2021/10/19/academia/texniti-noimosyni-glossiki-texnologia-kai-glossikes-ypodomes/.[Ανάκτηση 29/9/2022].

Γούτσος, Δ., & Φραγκάκη, Γ. (2015). *Εισαγωγή στη γλωσσολογία σωμάτων κειμένων* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/1932.

Dacos, M. (2011). *Διακήρυξη των ψηφιακών ανθρωπιστικών σπουδών*. https://www.tcp.hypothesees.org/495. [Ανάκτηση 9/9/2022].

Dariah-gr/ ΔΥΑΣ. https://dyas-net.gr/. [Ανάκτηση 6/9/2022].

Δημητρούλια, Ξ., & Τικτοπούλου, Α. (2015). *Ψηφιακές λογοτεχνικές σπουδές* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/5827.

Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα. https://www.kallipos.gr. [Ανάκτηση 2/9/2022].

Κανελλοπούλου-Μπότη, Μ. (2004). *Το δίκαιο της πληροφορίας*. Αθήνα : Νομική Βιβλιοθήκη.

Καπιδάκης, Σ. (2014). *Εισαγωγή στις ψηφιακές βιβλιοθήκες*. Αθήνα : Δίσιγμα Εκδόσεις.

Καπιδάκης, Σ., Λαζαρίνης, Φ., & Τοράκη, Κ. (2015). *Θέματα βιβλιοθηκονομίας και επιστήμης των πληροφοριών* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/1674.

Κυριάκη-Μάνεση, Δ., & Κουλούρης, Α. (2015). *Διαχείριση ψηφιακού περιεχομένου* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/2496.

Μαρκόπουλος, Γ. (2006). *Ζητήματα υπολογιστικής γλωσσολογίας: Prolog και μορφολογική ανάλυση*. Αθήνα: Περιοδικό «Παρουσία», Παράρτημα 69.

Ν. 4521/2018 (ΦΕΚ Α΄, 38/2-3-2018): «Ίδρυση Πανεπιστημίου Δυτικής Αττικής».

Πανεπιστήμιο Δυτικής Αττικής. https://www.uniwa.gr. [Ανάκτηση 2/9/2022].

Τάντος, Α., Μαρκαντωνάτου, Σ., Αναστασιάδη-Συμεωνίδη, Ά., & Κυριακοπούλου, Π. (2015). *Υπολογιστική γλωσσολογία* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/2205.

Ξενόγλωσση βιβλιογραφία

Clarin. https://www.clarin.eu/. [Retrieved 1/7/2022].

Creative Commons. https://creativecommons.org/. [Retrieved 4/9/2022].

Dariah.eu. https://www.dariah.eu/. [Retrieved 4/9/2022].

Dipper, S. (2008). Theory-driven and corpus-driven computational linguistics, and the use of corpora. In A.Lüdeling & M. Kytö (eds) *Corpus Linguistics: An International Handbook*. Berlin: Walter deGruyter, 68-96.

Europeana.eu. https://www.europeana.eu/. [Retrieved 4/9/2022].

Schreibman, S., Siemens R. & Unsworth J. (eds.). (2004). *A Companion to Digital Humanities*. Oxford: Blackwell, http://www.digitalhumanities.org/companion. [Retrieved 9/9/2022].

Shirley, W. (2005). Leung, International conference on developing digital institutional repositories: experiences and challenges. *Library high tech news, no. 2*, pp. 14-15.

Πρόσθετη Βιβλιογραφία, χωρίς παραπομπές στο κείμενο

Ανδριωτάκης, Μ. (2022). *Τεχνητή νοημοσύνη για όλους*. Αθήνα : Ψυχογιός Εκδόσεις.

Βαλεοντής, Κ., Κριμπάς, Π., Πανταζάρα, Μ., Τοράκη, Κ. & Τσιάμας, Γ. (2019). *Ελληνική Γλώσσα και Ορολογία - Ανακοινώσεις 12ου Συνεδρίου*. Αθήνα : Ελληνική Εταιρεία Ορολογίας.

Βαλεοντής, Κ., Κουτσουμπάρη, Ν. & Σδούκου, Χ., (2011). Ορολογία και Λεξικογραφία: Δύο θεματικά πεδία με συμπληρωματικούς στόχους, αλλά με εννοιολογικές και μεθοδολογικές διαφορές – σημεία σύγκλισης και προτάσεις περαιτέρω εναρμόνισης. In *ΕΛΕΤΟ–8ο Συνέδριο Ελληνική Γλώσσα και Ορολογία. Αθήνα. Από http://www. eleto. gr/download/Conferences/8th% 20Conference/Papers/8th_01-30_Koutsoubari_Sdoukou_Valeontis_Paper_V06. pdf,* (Vol. 30, No. 10, p. 2019). [Ανάκτηση 2/9/2022].

Βαλεοντής, Κ., & Μάντζαρη, Ε. (2006). Η γλωσσική διάσταση της Ορολογίας: Αρχές και μέθοδοι σχηματισμού των όρων. *1ο Διεθνές Συνέδριο μετάφρασης και διερμηνείας*, 13-14.

Καπιδάκης, Σ., Πιπερίδης, Σ., Λαμπροπούλου, Π. & Γαβριηλίδου, Μ. (2015). Ανοιχτά γλωσσικά δεδομένα: η υποδομή γλωσσικών πόρων και υπηρεσιών CLARIN:EL, *2ο Διεθνές Συνέδριο "Δημιουργική Γραφή"*. Κέρκυρα. [Conference paper].

Παπατσικουράκης, Χρ., Σίτας, Α.. [Επιμέλεια]. (2005). *Από τη βιβλιοθηκονομία στην επιστήμη της πληροφόρησης*. Αθήνα : Τυπωθήτω – Γιώργος Δάρδανος.

Πιπερίδης, Σ., Λαμπροπούλου, Π. & Γαβριηλίδου, Μ. (2015). Clarin:el: Δημιουργώ, επεξεργάζομαι, μοιράζομαι, *10ο Συνέδριο "Ελληνική Γλώσσα και Ορολογία"*. Αθήνα. [Conference paper].

Πιπερίδης, Σ., Λαμπροπούλου, Π. & Γαβριηλίδου, Μ. (2015). Clarin:el: μια υποδομή τεκμηρίωσης, διαμοιρασμού και επεξεργασίας γλωσσικών δεδομένων,*12ο Συνέδριο Ελληνικής Γλωσσολογίας*. Βερολίνο. [Conference paper].

Πουλή, Κ., Τσιούλη, Η. & Λαμπροπούλου, Π. (2017). Ορολογικοί πόροι ΟΡΟΣΗΜΟ: επιμέλεια, ταξινόμηση και αποτελέσματα,*11ο Συνέδριο "Ελληνική Γλώσσα και Ορολογία"*. Αθήνα. [Conference paper].

Τοράκη, Κ., (2014). Χημική ορολογία, τυποποίηση και η συμβολή της Επιτροπής ΕΛΟΤ/ΤΕ21. *1ο πανελλήνιο συνέδριο χημικής ονοματολογίας και ορολογίας, Αθήνα*, 22

Τοράκη, Κ., Χατζημάρη, Σ.,Τσάφου, Σ., & Γεωργάκη, Δ. (2008). *Δημιουργία Ελληνικού Θησαυρού Επιστημονικών όρων.*, 2008 . In 17ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών, Ιωάννινα (GR), 24-26 Σεπτεμβρίου. [Conference paper]

Τσάφου, Σ., & Χατζημαρή, Σ. (2001). *Θησαυροί και θεματική ευρετηρίαση στις Ελληνικές βιβλιοθήκες.*. In 10ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών, Θεσσαλονίκη (GR), 2001. [Conference paper]

Χρόνη, Α. (2012) *Θησαυροί και Ελληνική πραγματικότητα: Η χρήση των θησαυρών από τις βιβλιοθήκες στην Ελλάδα*. [PhD Thesis].

Ibekwe-SanJuan, F., & SanJuan, E. (2003). TermWatch: variations terminologiques et veille scientifique. *Actes du congrés International Society for Knowledge Organization, ISKO 2003*, 1-11.

Iordanidou, A., Pantazara, M., Mantzari, E., Ophanos, G., Vagelatos, A., & Papapanagiotou, V. About recognition of Greek multi-word complex terms in the biomedical domain.

Patry, A., & Langlais, P. (2005, August). Corpus-based terminology extraction. In *Terminology and Content Development–Proceedings of 7th International Conference on Terminology and Knowledge Engineering, Litera, Copenhagen*.

Yoon, T. S. (2004). Design and Implementation of an Ontology-based Knowledge Management System. In *Proceedings of the CALSEC Conference* (pp. 107-111). Society for e-Business Studies.

Παράρτημα 1 – Συχνές ερωταποκρίσεις

Στα πλαίσια της εκπόνησης της παρούσας διπλωματικής εργασίας, και για την καλύτερη ενημέρωση προς τους παραγωγούς των γλωσσικών πόρων, αναφορικά με τους στόχους και σκοπούς της έρευνας μας, συντάξαμε τις παρακάτω συνήθεις ερωταποκρίσεις.

Συνήθεις ερωταποκρίσεις (Συχνές Ερωτήσεις | CLARIN-EL)

- ***Πείτε μου ορισμένα παραδείγματα γλωσσικών πόρων που πιθανόν να έχω και θα μπορούσα να διαθέσω στο Clarin***

Το Clarin φιλοξενεί μεγάλη ποικιλία γλωσσικών πόρων. Πιθανόν κάποιος να διαθέτει τους ανάλογους πόρους, χωρίς να το ξέρει. Θα μπορούσε να είναι κάποιο μεγάλο κείμενο, κάποιο γλωσσάρι, αντιστοιχίες όρων με ορισμούς, εργασίες, άρθρα, κάποια γραπτή καταγραφή βιντεοσκοπημένου ή ηχογραφημένου υλικού, βιβλία, περιγραφές μαθημάτων, λεξικά (μονόγλωσσα ή δίγλωσσα / πολύγλωσσα), εγχειρίδια, θησαυροί, λογισμικά που σχετίζονται με τη γλωσσική επεξεργασία ή γενικά οποιασδήποτε μορφής δομημένο ή αδόμητο κείμενο.

- ***Τι είναι οι γλωσσικοί – ορολογικοί πόροι;***

Γλωσσικός / ορολογικός πόρος είναι οποιοδήποτε σύνολο δεδομένων, σε κάθε μορφή, δομημένο ή αδόμητο, που συνδέεται με τη γλώσσα. Μπορεί να είναι πόροι με **πρωτογενές περιεχόμενο** (λόγος σε ψηφιακή ή ψηφιοποιημένη μορφή, βιβλία, κείμενα, διάφορες σημειώσεις, σημειώσεις από τα μαθήματα, σώματα κειμένων, κείμενα που προέρχονται από το διαδίκτυο, εφημερίδες, συνεντεύξεις, εκπομπές, βιντεοσκοπημένο υλικό, περιγραφές μαθημάτων κτλ.), **επεξεργασμένοι πόροι** (υποσημειώσεις που έχουν δημιουργηθεί αυτόματα ή από το χρήστη, μεταγραφές ηχογραφημένων ή βιντεοσκοπημένων αρχείων, ηλεκτρονικά εγχειρίδια κτλ.), **πόροι οργανωμένης μορφής της γνώσης** (μονόγλωσσα ή πολύγλωσσα λεξικά, γλωσσάρια, λίστες λέξεων, θησαυροί κτλ.) ή **διάφορες εφαρμογές και εργαλεία γλωσσικής τεχνολογίας** (εργαλεία λογισμικού σε κείμενα, εργαλεία εξόρυξης γνώσης, λημματοποίησης, παρουσίασης δεδομένων κτλ.).

- ***Τι είναι οι γλωσσικές τεχνολογίες;***

Οι γλωσσικές τεχνολογίες είναι διάφορα υπολογιστικά εργαλεία γλωσσικής ανάλυσης με τα οποία μπορούν να πραγματοποιηθούν ανάλυση, επισημείωση, επεξεργασία και τροποποίηση των διαφόρων γλωσσικών / ορολογικών δεδομένων.

- ***Τι είναι οι υπηρεσίες γλωσσικής επεξεργασίας;***

Οι υπηρεσίες γλωσσικής επεξεργασίας επιτρέπουν τη χρήση των γλωσσικών πόρων, τεχνολογιών και εφαρμογών μέσω του διαδικτύου.

- ***Ποιοι είναι οι παραγωγοί γλωσσικών πόρων;***

Ως παραγωγό / πάροχο γλωσσικών πόρων εννοούμε το ακαδημαϊκό ή ερευνητικό ίδρυμα ή τον ιδιώτη που διαθέτει τους γλωσσικούς πόρους που έχει παράγει στην υποδομή του CLARIN:EL, προς χρήση από την ερευνητική κοινότητα, τους επαγγελματίες της γλώσσας, λοιπούς επιστήμονες και ερευνητές, τον απλό πολίτη, το ευρύ κοινό κτλ.

- ***Τι είναι το CLARIN;***

Το Clarin είναι μια ευρωπαϊκή διαδικτυακή υποδομή η οποία συγκεντρώνει γλωσσικούς πόρους, τεχνολογίες και υπηρεσίες, σε διάφορες γλώσσες, με σκοπό να τους διαθέσει προς την ερευνητική κοινότητα ή και τον απλό ιδιώτη. Είναι οργανωμένο σε κατά τόπους και γλώσσα υποδομές του Clarin. Το υλικό που συσσωρεύει τεκμηριώνεται κατάλληλα και μπορεί να είναι επεξεργάσιμο μέσω γλωσσικών τεχνολογιών. Παρέχει επίσης εκπαίδευση αναφορικά με τις γλωσσικές τεχνολογίες.

- ***Τι είναι το CLARIN:EL;***

Το CLARIN:EL είναι το ελληνικό παράρτημα της ευρωπαϊκής υποδομής Clarin.

- ***Τι ορολογικούς / γλωσσικούς πόρους μπορεί να έχει το ΠΑΔΑ;***

Το Πανεπιστήμιο Δυτικής Αττικής, μέσα από τη λειτουργία του, το εκπαιδευτικό και το ερευνητικό του έργο, παράγει πληθώρα γλωσσικών πόρων. Οι άμεσοι παραγωγοί του γλωσσικού υλικού μπορεί να είναι τόσο οι διδάσκοντες του, όσο και οι φοιτητές των σχολών και τμημάτων του. Το παραγόμενο υλικό μπορεί να είναι ποικιλόμορφο: γλωσσάρια, λεξικά μονόγλωσσα ή πολύγλωσσα διαφόρων όρων, λίστες δεδομένων, θησαυροί, κείμενα, εργασίες, άρθρα, εγχειρίδια κτλ.

- ***Γιατί είναι χρήσιμο να συγκεντρωθούν και να καταγραφούν οι γλωσσικοί πόροι που ΠΑΔΑ;***

Το ΠΑΔΑ, με τη συγκέντρωση και καταγραφή των γλωσσικών του πόρων, καθώς και τη διασύνδεση τους με το Clarin, γίνεται μέλος μίας κοινότητας που περιλαμβάνει τα ερευνητικά κέντρα και πανεπιστήμια της Ελλάδας, με διασύνδεση και με τα αντίστοιχα του εξωτερικού. Δημιουργεί ένα ιδρυματικό αποθετήρια στην υποδομή, μέσω της οποίας το υλικό του προβάλλεται, διαμοιράζεται και γίνεται επεξεργάσιμο από την ερευνητική κοινότητα. Επιπλέον, προσφέρεται προβολή σε καθέναν από τους δημιουργούς των γλωσσικών πόρων προσωπικά.

- ***Γιατί να το διαθέσω εγώ προσωπικά του γλωσσικούς μου πόρους στο Clarin;***

Η διάθεση των γλωσσικών πόρων που παράχθηκαν από κάποιον δημιουργό στο Clarin, τον καθιστά μέλος ενός δικτύου μιας εθνικής και διεθνούς επιστημονικής κοινότητας, προσφέροντας του προβολή, και παράλληλα την ικανοποίηση ότι συνεισφέρει στην εξέλιξη και ανάπτυξη των γλωσσικών εργαλείων και τεχνολογιών, με γνώμονα ότι οι ψηφιακές ανθρωπιστικές επιστήμες είναι μια σημαντική νέα διάσταση στην επιστήμη η οποία έχει προκύψει τα τελευταία χρόνια, με συνδυασμό των ανθρωπιστικών και των ψηφιακών / υπολογιστικών επιστημών. Οι γλωσσικοί πόροι του δημιουργού γίνονται πιο εύκολα προσβάσιμοι στο χρήστη και την ερευνητική κοινότητα.

- ***Ποιοι θα έχουν πρόσβαση στους πόρους μου στο Clarin;***

Πρόσβαση στους γλωσσικούς πόρους που συσσωρεύονται στο Clarin μπορεί να έχει οποιοσδήποτε χρήστης τους αναζητήσει στην υποδομή. Οι απλοί επισκέπτες (μη εγγεγραμμένοι χρήστες) μπορούν να καταφορτώσουν (download) μόνο τους ανοιχτούς πόρους σε όλους, σύμφωνα με τους όρους που ορίζει ο πάροχος / δημιουργός του πόρου. Οι εγγεγραμμένοι χρήστες της υποδομής μπορούν να καταφορτώσουν (download) όλους του διαθέσιμους γλωσσικούς πόρους, σύμφωνα με τους όρους που ορίζει ο πάροχος / δημιουργός του πόρου. Κάποιοι πόροι είναι επεξεργάσιμοι, ενώ κάποιοι άλλοι είναι μόνο για θέαση. Συνήθως αυτό εξαρτάται από τα μορφότυπα του κάθε πόρου και αν αυτά είναι συμβατά με την υποδομή.

- ***Ποιοι είναι οι όροι για τη διαθεσιμότητα των γλωσσικών πόρων στο κοινό;***

Οι γλωσσικοί πόροι διατίθενται ελεύθερα από το Clarin, για ερευνητικούς σκοπούς, με ανοιχτές άδειες χρήσης Creative Commons (CC). Στο Clarin οι πόροι διατίθενται είτε με

ακριβώς την ίδια άδεια με τον πρωτότυπο πόρο, είτε με κάποια διαφορετική άδεια που επιθυμεί να τον διαθέσει ο παραγωγός του, με αναφορά στον πρωτότυπο πόρο (https://creativecommons.org/licenses/by/3.0/legalcode).

- ***Πως μεταφορτώνονται οι γλωσσικοί πόροι στο Clarin;***

Για να μεταφορτώθει κάποιος γλωσσικός πόρος στο Clarin και έπειτα να αποθηκευτεί στην υποδομή και να μπορεί να διαμοιραστεί στην κοινότητα, θα πρέπει ο πάροχος του να είναι εγγεγραμμένος χρήστης στο Clarin:EL και να εκπροσωπεί ή να είναι μέλος και να ανήκει σε κάποιο από τα Ιδρυματικά Αποθετήρια ή το Αποθετήριο Φιλοξενούμενων Πόρων που φιλοξενεί η υποδομή. Ο κάθε πόρος τεκμηριώνεται κατάλληλα και τα μεταδεδομένα του αποθηκεύονται. Το Ηλεκτρονικό Εγχειρίδιο Χρήσης της Υποδομής διαθέτει όλες τις ανάλογες χρήσιμες πληροφορίες. (https://clarin-platform-documentation.readthedocs.io/el/).

- ***Αν μεταφορτωθεί κάποιος γλωσσικός πόρος στο Clarin θα μπορώ να το αλλάζω έπειτα;***

Ο δημιουργός ή πάροχος του κάθε γλωσσικού πόρου είναι δυνατόν να τον τροποποιήσει και, ανάλογα με το πόσο μεγάλες θα είναι οι τροποποιήσεις, είτε θα προστεθεί μια νέα εκδοχή του πόρου, είτε θα γίνουν απλώς οι διορθώσεις στον ήδη υφιστάμενο πόρο. Είναι δυνατόν ακόμα και να τον αποσύρει από την υποδομή έπειτα από απλό αίτημα.

- ***Πώς θα μπορούν να χρησιμοποιηθούν οι πόροι μου στο Clarin;***

Οι διατιθέμενοι πόροι μέσω της υποδομής του Clarin μπορούν να χρησιμοποιηθούν από την ερευνητική κοινότητα και το ευρύ κοινό με καταφόρτωση (download) ή θέαση για ερευνητικούς λόγους, εξέλιξη της επιστήμης, ανάπτυξη εργαλείων και λογισμικών γλωσσικής τεχνολογίας, απλή ενημέρωση κτλ.

- ***Έχω γράψει ένα μεγάλο ελληνικό κείμενο (βιβλίο, άρθρο, …). Μπορώ να το κάνω διαθέσιμο στο Clarin;***

Η υποδομή Clarin φιλοξενεί γλωσσικούς πόρους που περιλαμβάνουν σώματα κειμένων. Οπότε, κάποιο βιβλίο, άρθρο ή γενικά μεγάλου μεγέθους κείμενο εμπίπτει ακριβώς στην κατηγορία του υλικού που αποθηκεύεται προς επεξεργασία στην υποδομή.

- ***Έχω φτιάξει ένα λεξικό (δίγλωσσο, πολύγλωσσο, ερμηνευτικό, κτλ…). Μπορώ να το κάνω διαθέσιμο στο Clarin;***

Τα λεξικά ή οποιαδήποτε μορφή οργάνωσης της γνώσης (πχ θησαυροί, λίστες ορολογικών πόρων, λίστα δεδομένων κτλ.) είναι δυνατόν να χρησιμοποιηθούν κατάλληλα και να υποστούν επεξεργασία από το χρήστη. Οπότε πρόκειται για υλικό το οποίο επιθυμεί να φιλοξενήσει η υποδομή Clarin.

- ***Διαθέτω κάποιο εργαλείο/ λογισμικό / εφαρμογή γλωσσικής τεχνολογίας. Μπορώ να το κάνω διαθέσιμο στο Clarin;***

Η υποδομή Clarin φιλοξενεί εργαλεία, λογισμικά ή εφαρμογές γλωσσικής τεχνολογίας. Ο πάροχος / δημιουργός τους μπορεί να τα διαθέσει, ούτως ώστε να μπορούν να χρησιμοποιηθούν από τους ενδιαφερόμενους, μέσω της υποδομής.

- ***Ποια είναι η δομή του Clarin;***

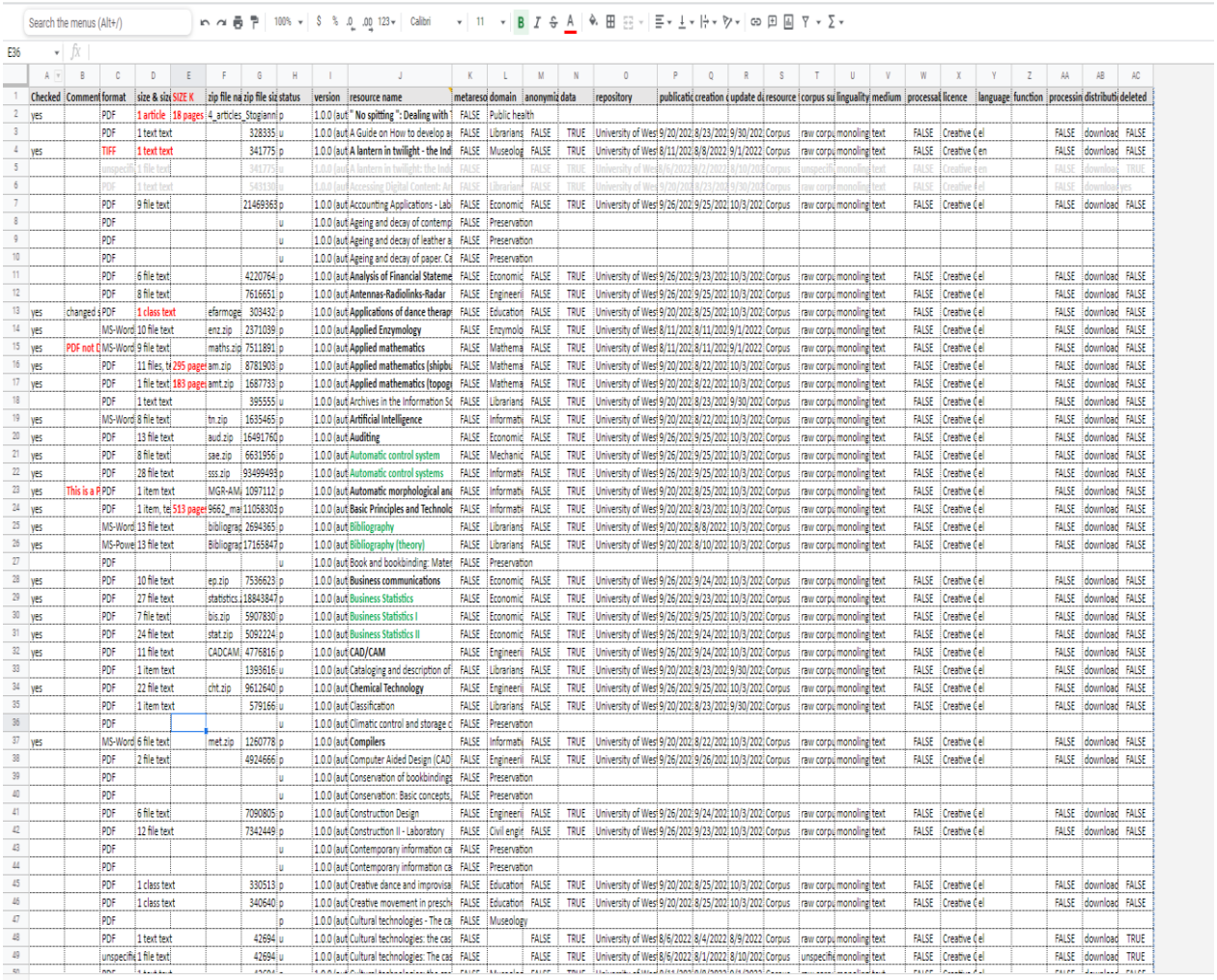
Το Clarin περιλαμβάνει τον Κεντρικό Κατάλογο (ή Κεντρικό Συσσωρευτή) , ο οποίος συγκεντρώνει τα μεταδεδομένα τεκμηρίωσης και την περιγραφή των γλωσσικών πόρων, υπηρεσιών και εργαλείων. Είναι οργανωμένα ανά αποθετήριο, το οποίο αντιπροσωπεύει κάποιο ίδρυμα. Υπάρχει και ένα αποθετήριο φιλοξενούμενων πόρων που σωρεύει πόρους που δεν συσχετίζονται με κάποιο συγκεκριμένο ίδρυμα. Τα δεδομένα που φιλοξενούνται, συγκεντρώνονται, τεκμηριώνονται, προβάλλονται και διατίθενται στην ερευνητική κοινότητα και το ευρύ κοινό, προς χρήση και επεξεργασία.

- ***Πώς μπορώ να μάθω περισσότερα για το Clarin;***

Για περισσότερες πληροφορίες σχετικά με το Clarin, ο ενδιαφερόμενος δημιουργός, πάροχος ή χρήστης μπορεί να απευθυνθεί τόσο στον ιστότοπο της διεθνούς υποδομής του Clarin (https://www.clarin.eu/), όσο και στο ελληνικό παράρτημα Clarin:EL (https://www.clarin.gr/el). Επίσης, αναφορικά με την εισαγωγή γλωσσικών πόρων σχετιζόμενων με το Πανεπιστήμιο Δυτικής Αττικής, οι ενδιαφερόμενοι μπορούν να απευθυνθούν στη φοιτήτρια Αγγελική Μπαμνιώτη (a.bamnioti@gmail.com) ή τον επιβλέποντα καθηγητή της παρούσας πτυχιακής εργασίας Σαράντο Καπιδάκη (sarantos@uniwa.gr).

Παράρτημα 2– Παράδειγμα εξαγωγής δεδομένων γλωσσικών πόρων (Πηγή Clarin:el)

Στο κυρίως κείμενο αναφέρθηκε ότι ένας από τους περιορισμούς του Clarin:el είναι ότι ο χρήστης δεν είναι δυνατόν να κάνει μόνος του συγκεντρωτική εξαγωγή των μεταδεδομένων των γλωσσικών του πόρων, η οποία θα μπορούσε να τον βοηθήσει σε σύγκριση, αντιπαράθεση των στοιχείων του και πιθανή βελτίωση τους. Στο παράρτημα αυτό παραθέτουμε μια εξαγωγή των δεδομένων μας που ζητήσαμε να γίνει από την ομάδα του Clarin:el, με στόχο να γίνουν κάποιες διορθώσεις σε λανθασμένα μεταδεδομένα.



Checked	Comment	format	size & size	SIZE K	zip file na	zip file siz	status	version	resource name	metaresou	domain	anonymiz	data	repository	publicatic	creation c	update da	resource	corpus su	linguality	medium	processat	licence	language	function	processin	distributi	deleted
yes		PDF	1 article	18 pages	4_articles_	5709	p	1.0.0	" No spitting ": Dealing with	FALSE	Public health																	
		PDF	1 text text			328335	u	1.0.0	A Guide on How to develop a	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/23/202	9/30/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		TIFF	1 text text			341775	p	1.0.0	A lantern in twilight - the Ind	FALSE	Museolog	FALSE	TRUE	University of Wes	8/11/202	8/8/2022	9/1/2022	Corpus	raw corpu	monoling	text	FALSE	Creative C	en		FALSE	download	FALSE
		unspecifi	1 file text			541775	u	1.0.0	A lantern in twilight: the Ind	FALSE		FALSE	TRUE	University of Wes	8/6/2022	8/7/2022	8/10/202	Corpus	unspecifi	monoling	text	FALSE	Creative C	en		FALSE	download	TRUE
		PDF	1 text text			5431980	u	1.0.0	Accessing Digital Content: A	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/23/202	9/30/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	en		FALSE	download	yes
		PDF	9 file text			21469363	p	1.0.0	Accounting Applications - Lab	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF					u	1.0.0	Ageing and decay of contemp	FALSE	Preservation																	
		PDF					u	1.0.0	Ageing and decay of leather a	FALSE	Preservation																	
		PDF					u	1.0.0	Ageing and decay of paper. Ca	FALSE	Preservation																	
		PDF	6 file text			4220764	p	1.0.0	Analysis of Financial Stateme	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/23/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	8 file text			7616651	p	1.0.0	Antennas-Radiolinks-Radar	FALSE	Engineeri	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes	changed s	PDF	1 class text		efarmoge	303432	p	1.0.0	Applications of dance therap	FALSE	Education	FALSE	TRUE	University of Wes	9/20/202	8/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		MS-Word	10 file text		enz.zip	2371039	p	1.0.0	Applied Enzymology	FALSE	Enzymolo	FALSE	TRUE	University of Wes	8/11/202	8/11/202	9/1/2022	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes	PDF not D	MS-Word	9 file text		maths.zip	7511891	p	1.0.0	Applied mathematics	FALSE	Mathema	FALSE	TRUE	University of Wes	8/11/202	8/11/202	9/1/2022	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	11 files, te	295 page	am.zip	8781903	p	1.0.0	Applied mathematics [shipbu	FALSE	Mathema	FALSE	TRUE	University of Wes	9/20/202	8/22/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	1 file text	183 page	amt.zip	1687733	p	1.0.0	Applied mathematics [topogi	FALSE	Mathema	FALSE	TRUE	University of Wes	9/20/202	8/22/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	1 text text			395555	u	1.0.0	Archives in the Information Se	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/23/202	9/30/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		MS-Word	8 file text		tn.zip	1635465	p	1.0.0	Artificial Intelligence	FALSE	Informati	FALSE	TRUE	University of Wes	9/20/202	8/22/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	13 file text		aud.zip	16491760	p	1.0.0	Auditing	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	8 file text		sae.zip	6631956	p	1.0.0	Automatic control system	FALSE	Mechanic	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	28 file text		sss.zip	93499493	p	1.0.0	Automatic control systems	FALSE	Informati	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes	This is a P	PDF	1 item text		MGR-AM	1097112	p	1.0.0	Automatic morphological ana	FALSE	Informati	FALSE	TRUE	University of Wes	9/20/202	8/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	1 item, te	513 page	9662_mai	11058303	p	1.0.0	Basic Principles and Technolo	FALSE	Informati	FALSE	TRUE	University of Wes	9/20/202	8/23/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		MS-Word	13 file text		bibliograp	2694365	p	1.0.0	Bibliography	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/8/2022	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		MS-Powe	13 file text		Bibliograp	17165847	p	1.0.0	Bibliography (theory)	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/10/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF					u	1.0.0	Book and bookbinding: Mater	FALSE	Preservation																	
yes		PDF	10 file text		ep.zip	7536623	p	1.0.0	Business communications	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/24/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	27 file text		statistics.j	18843847	p	1.0.0	Business Statistics	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/23/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	7 file text		bis.zip	5907830	p	1.0.0	Business Statistics I	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	24 file text		stat.zip	5092224	p	1.0.0	Business Statistics II	FALSE	Economic	FALSE	TRUE	University of Wes	9/26/202	9/24/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	11 file text		CADCAM	4776816	p	1.0.0	CAD/CAM	FALSE	Engineeri	FALSE	TRUE	University of Wes	9/26/202	9/24/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	1 item text			1393616	u	1.0.0	Cataloging and description of	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/23/202	9/30/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
yes		PDF	22 file text		cht.zip	9612640	p	1.0.0	Chemical Technology	FALSE	Engineeri	FALSE	TRUE	University of Wes	9/26/202	9/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	1 item text			579166	u	1.0.0	Classification	FALSE	Librarians	FALSE	TRUE	University of Wes	9/20/202	8/23/202	9/30/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF					u	1.0.0	Climatic control and storage c	FALSE	Preservation																	
yes		MS-Word	6 file text		met.zip	1260778	p	1.0.0	Compilers	FALSE	Informati	FALSE	TRUE	University of Wes	9/20/202	8/22/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	2 file text			4924666	p	1.0.0	Computer Aided Design (CAD	FALSE	Engineeri	FALSE	TRUE	University of Wes	9/26/202	9/26/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF					u	1.0.0	Conservation of bookbindings	FALSE	Preservation																	
		PDF					u	1.0.0	Conservation: Basic concepts.	FALSE	Preservation																	
		PDF	6 file text			7090805	p	1.0.0	Construction Design	FALSE	Engineeri	FALSE	TRUE	University of Wes	9/26/202	9/24/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	12 file text			7342449	p	1.0.0	Construction II - Laboratory	FALSE	Civil engir	FALSE	TRUE	University of Wes	9/26/202	9/23/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF					u	1.0.0	Contemporary information ca	FALSE	Preservation																	
		PDF					u	1.0.0	Contemporary information ca	FALSE	Preservation																	
		PDF	1 class text			330513	p	1.0.0	Creative dance and improvisa	FALSE	Education	FALSE	TRUE	University of Wes	9/20/202	8/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF	1 class text			340640	p	1.0.0	Creative movement in presch	FALSE	Education	FALSE	TRUE	University of Wes	9/20/202	8/25/202	10/3/202	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	FALSE
		PDF					p	1.0.0	Cultural technologies - The ca	FALSE	Museology																	
		PDF	1 text text			42694	u	1.0.0	Cultural technologies: the cas	FALSE		FALSE	TRUE	University of Wes	8/6/2022	8/4/2022	8/9/2022	Corpus	raw corpu	monoling	text	FALSE	Creative C	el		FALSE	download	TRUE
		unspecifi	1 file text			42694	u	1.0.0	Cultural technologies: The cas	FALSE		FALSE	TRUE	University of Wes	8/6/2022	8/1/2022	8/10/202	Corpus	unspecifi	monoling	text	FALSE	Creative C	el		FALSE	download	TRUE

Εικόνα 12. Εξαγωγή δεδομένων προς διόρθωση, από το Clarin

Παράρτημα 3 – Υποδείγματα του κειμένου ηλεκτρονικού ταχυδρομείου που στάλθηκε στους πιθανούς παραγωγούς γλωσσικών πόρων

Κείμενο που στάλθηκε στους διδάσκοντες του τμήματος Αρχειονομίας, *Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης*

Γειά σας

Στα πλαίσια του μεταπτυχιακού μου, αναζητούμε διάφορες μορφές κειμένου (πχ κείμενα, γλωσσάρια, μονόγλωσσα ή πολύγλωσσα λεξικά, θησαυροί, σώματα κειμένων, εργαλεία λογισμικού σε κείμενα, διάφορες σημειώσεις, σημειώσεις από τα μαθήματα, περιγραφές μαθημάτων, ηλεκτρονικά εγχειρίδια κ.λπ.) που έχουν παραχθεί από τους διδάσκοντες στο ΠΑΔΑ και θα μπορούσαν να ενσωματωθούν στην ελληνική version του Clarin, το οποίο είναι μια διεθνής υποδομή γλωσσικών πόρων και τεχνολογιών. Πολλά πανεπιστήμια και ερευνητικά κέντρα της Ελλάδος έχουν ήδη διασυνδεθεί με την υποδομή, η οποία ασχολείται με τις λεγόμενες digital humanities. Το ΠαΔΑ όχι ακόμα πλήρως. Οπότε, αν τυχόν έχετε κάποιου είδους παρόμοιο υλικό, θα μπορούσατε πιθανόν να με βοηθήσετε στη συγκέντρωση υλικού για τη διπλωματική μου εργασία. Με τον τρόπο αυτό προβάλλεται το έργο του Πανεπιστημίου Δυτικής Αττικής και φυσικά αυτό των δημιουργών του! Θα μπορούσαμε να κανονίσουμε κάποια ενημερωτική επικοινωνία μέσω teams, όποτε μπορείτε.

Σας ευχαριστώ για την προσοχή σας

Αγγελική Μπαμνιώτη

CLARIN:EL

Κείμενο που στάλθηκε στους υπόλοιπους διδάσκοντες του ΠαΔΑ

Γειά σας!

Ονομάζομαι Αγγελική Μπαμνιώτη και είμαι μεταπτυχιακή φοιτήτρια στο πρόγραμμα μεταπτυχιακών σπουδών: *«Διαχείριση Πληροφοριών σε Βιβλιοθήκες, Αρχεία, Μουσεία».* Η πτυχιακή μου εργασία έχει ως θέμα *«Συγκέντρωση και καταγραφή ελληνικών γλωσσικών πόρων του ΠαΔΑ»* και εκπονείται υπό την επίβλεψη του καθηγητή του Τμήματος Αρχειονομιας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης, κ. Σαράντου Καπιδάκη.

Στόχος της εργασίας είναι να διερευνήσει τι γλωσσικοί πόροι έχουν παραχθεί από τα μέλη του ΠαΔΑ, ποιοι από αυτούς είναι διαθέσιμοι και με τι όρους, και να καταγραφούν και πιθανώς να αναφορτωθούν (ή συνδεθούν) με το αντίστοιχο Ιδρυματικό αποθετήριο του ΠαΔΑ στο Clarin.

Στα πλαίσια του μεταπτυχιακού μου, αναζητούμε διάφορες μορφές κειμένου (πχ κείμενα, γλωσσάρια, μονόγλωσσα ή πολύγλωσσα λεξικά, θησαυροί, σώματα κειμένων, εργαλεία λογισμικού σε κείμενα, διάφορες σημειώσεις, σημειώσεις από τα μαθήματα, περιγραφές μαθημάτων, ηλεκτρονικά εγχειρίδια κ.λπ.) που έχουν παραχθεί από τους διδάσκοντες στο ΠΑΔΑ και θα μπορούσαν να ενσωματωθούν στην ελληνική version του Clarin, το οποίο είναι μια διεθνής υποδομή γλωσσικών πόρων και τεχνολογιών. Πολλά πανεπιστήμια και ερευνητικά κέντρα της Ελλάδος έχουν ήδη διασυνδεθεί με την υποδομή, η οποία ασχολείται με τις λεγόμενες digital humanities. Το ΠαΔα όχι ακόμα. Οπότε, αν τυχόν έχετε κάποιου είδους παρόμοιο υλικό, θα μπορούσατε πιθανόν να με βοηθήσετε στη συγκέντρωση υλικού για τη διπλωματική μου εργασία. Με τον τρόπο αυτό προβάλλεται το έργο του Πανεπιστημίου Δυτικής Αττικής και φυσικά αυτό των δημιουργών του!

Το Clarin είναι η εθνική Υποδομή Γλωσσικών πόρων και Τεχνολογιών στην Ελλάδα. Αποστολή του είναι: η συλλογή, η τεκμηρίωση, η συντήρηση και ο διαμοιρασμός ψηφιακών γλωσσικών πόρων, εργαλείων γλωσσικής τεχνολογίας καθώς και πιστοποιημένων διαδικτυακών υπηρεσιών γλωσσικής επεξεργασίας. Υποστηρίζει τους ερευνητές, ακαδημαϊκούς, φοιτητές, επαγγελματίες στον τομέα της γλώσσας, πολίτες-επιστήμονες και το ευρύ κοινό, που δραστηριοποιούνται στους τομείς των Γλωσσικών Σπουδών, των Ψηφιακών Ανθρωπιστικών και Κοινωνικών Επιστημών, της Πολιτιστικής Κληρονομιάς, της Γλωσσικής Τεχνολογίας και της Τεχνητής Νοημοσύνης, της

Πληροφορικής, των Γνωστικών Επιστημών, κλπ. Η Υποδομή CLARIN:EL συμμετέχει στον Εθνικό Οδικό Χάρτη Ερευνητικών Υποδομών και είναι το ελληνικό σκέλος της Ευρωπαϊκής Υποδομής CLARIN ERIC.

Έχω συλλέξει ήδη υλικό με δημιουργούς καθηγητές του Τμήματος "Αρχειονομιας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης". Αλλά θεωρώντας ότι θα πρέπει να εκπροσωπηθεί όλο το Πανεπιστήμιο Δυτικής Αττικής, οι σχολές και τα τμήματα του στο ιδρυματικό αποθετήριο του ΠΑΔΑ στο Clarin, θα ήθελα να σας ρωτήσω μήπως τυχόν έχετε στη διάθεση σας κάποιο γλωσσάρι (ή γλωσσικό εργαλείο εφόσον υπάρχουν αρκετά τμήματα σχετιζόμενα με την πληροφορική στο ΠΑΔΑ) το οποίο είναι αντιπροσωπευτικό της επιστήμης σας και θα μπορούσε πιθανόν να διαμοιραστεί στην υποδομή γλωσσικών πόρων του Clarin.

Η πτυχιακή εργασία και τα αποτελέσματά της θα είναι στη διάθεσή σας

Σας ευχαριστώ πολύ για τη συμμετοχή σας και είμαι στη διάθεση σας για όποια πληροφορία και επικοινωνία.

Αγγελική Μπαμνιώτη
Μεταπτυχιακή φοιτήτρια
Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας και Συστημάτων Πληροφόρησης
Πανεπιστήμιο Δυτικής Αττικής