# Τύποι πόρων προς συγκέντρωση
# Σενάρια χρήσης

Μαρία Γαβριηλίδου
Πένυ Λαμπροπούλου
Γιούλη Μπακαγιάννη
**ΙΕΛ /ΕΚ «Αθηνά»**

14 Μαΐου 2015

# Τι τύπους πόρων συγκεντρώνουμε;

# Δεδομένα και μεταδεδομένα

- **Δεδομένα**: το πρωτογενές γλωσσικό υλικό
- **Μεταδεδομένα**: η τεκμηρίωση των δεδομένων με βάση ένα συγκεκριμένο σχήμα περιγραφής

- Το CLARIN EL συγκεντρώνει τόσο **δεδομένα** όσο και **μεταδεδομένα**.

# Οργάνωση του υλικού σε τρεις κατηγορίες

- Μεταδεδομένα και πόροι στο CLARIN EL
- Μεταδομένα στο CLARIN EL και οι πόροι διαθέσιμοι από αλλού
- Μεταδεδομένα στο CLARIN EL, ενώ οι πόροι δεν είναι ψηφιοποιημένοι ή διαθέσιμοι ηλεκτρονικά

# Μεταδεδομένα & πόροι στο CLARIN EL: η βασική κατηγορία

- το CLARIN EL παρέχει τον πόρο και την τεκμηρίωσή του
- ο χρήστης μπορεί
  - να ψάξει στον κατάλογο με βάση την τεκμηρίωση του πόρου,
  - να εντοπίσει και να καταφορτώσει τον πόρο ή
  - *να χρησιμοποιήσει κάποια υπηρεσία γλωσσικής επεξεργασίας και να πάρει τα αποτελέσματα στον υπολογιστή του*.
- Το CLARIN EL φέρει την ευθύνη για τη διάθεση του πόρου, σύμφωνα με τους όρους διάθεσής του όπως καθορίζονται από τον πάροχο.

# Μεταδομένα στο CLARIN EL, αλλά όχι οι πόροι

- το CLARIN EL παρέχει την τεκμηρίωση του πόρου αλλά όχι τον ίδιο τον πόρο (τον οποίο δεν έχει στα αποθετήριά του)
- παρέχει απλώς σύνδεσμο για το σημείο όπου βρίσκεται ο πόρος, ο οποίος παρέχεται από άλλη υποδομή ή φορέα
- το CLARIN EL δεν φέρει καμία ευθύνη για την διαθεσιμότητα του πόρου

# Χωριστός κατάλογος μόνο με μεταδεδομένα

- ο πόρος δεν είναι διαθέσιμος διαδικτυακά ή δεν είναι καν σε ψηφιακή μορφή, όμως είναι χρήσιμος
- το CLARIN EL παρέχει την τεκμηρίωση και παραπέμπει στον δημιουργό/διαθέτη του πόρου
- ΔΕΝ είναι προτεραιότητα του CLARIN EL η συλλογή τέτοιου τύπου υλικού
- αν κριθεί χρήσιμο να καταγραφεί η ύπαρξη του πόρου, γίνεται τεκμηρίωσή του σε χωριστό κατάλογο "συμπληρωματικού υλικού", ώστε να μην υπάρξει σύγχυση με τους πραγματικούς πόρους

# Προδιαγραφές συγκέντρωσης πόρων

# Το υλικό που συγκεντρώνει το CLARIN EL

- ψηφιοποιημένα κείμενα γραπτού ή προφορικού (μεταγραμμένου) λόγου
- ηχητικά αρχεία καταγεγραμμένου προφορικού λόγου
- βίντεο
- λεξικά/γλωσσάρια
- συλλογές κειμένων
- …

# Γλώσσα

- μονόγλωσσοι (Ελληνικά)
- δίγλωσσοι ή πολύγλωσσοι (που περιλαμβάνουν Ελληνικά)
- προτεραιότητα στα Νέα Ελληνικά
- δεν αποκλείονται παλαιότερες μορφές και διάλεκτοι, απλώς είναι σε χαμηλότερη προτεραιότητα
- [οι διαθέσιμες ΓΤ είναι για τα ΝΕ!]

# Μέσο

- Κείμενο
- Ήχος
- Βίντεο
- Εικόνα

# Κειμενικό είδος / επίπεδο λόγου

- σε οποιοδήποτε τομέα, επίπεδο λόγου ή κειμενικό είδος
  - λογοτεχνία /επιστήμες / τύπος …
  - οικείο/επίσημο ύφος
  - άρθρα / μυθιστορήματα /δημόσια έγγραφα / συγγράμματα / επιστολές / αναρτήσεις σε κοινωνικά μέσα δικτύωσης …

# Μορφότυποι κειμενικού υλικού

- txt
- xml
- html
- doc/rtf
- pdf
- ppt

- τελευταία προτεραιότητα: κείμενο σε μορφή εικόνας (tiff, jpg)
- δακτυλόγραφο & χειρόγραφο → ΟΧΙ

# Μορφότυποι ηχητικού υλικού & πολυμεσικού υλικού (βίντεο)

- Ηχητικό υλικό
  - ηχητικό/ακουστικό αρχείο wav file
  - Mp3
  - απλή ορθογραφική μεταγραφή από απομαγνητοφώνηση
  - φωνητική μεταγραφή σε IPA
- Πολυμεσικό υλικό
  - Mp4
  - avi

# Κωδικοποίηση

- UTF-8

# Εικόνα

- jpg
- tiff
- pdf (όχι σκαναρισμένα)

- φωτογραφίες σκέτες → ΟΧΙ
- φωτογραφίες με λεζάντες → ΝΑΙ

# Λεξικά/θησαυροί

- βέλτιστη επιλογή: το περιεχόμενό τους (=τα λήμματα), ώστε να είναι δυνατόν να αποθηκεύονται και να καταφορτώνονται από το CLARIN
  - σε αυτή την περίπτωση
    - LMF, tbx
    - txt, csv
    - xml, SKOS / RDF
- λιγότερο επιθυμητή επιλογή: διαθέσιμα μέσω διεπαφής, οπότε τα τεκμηριώνουμε και κάνουμε απλή παραπομπή στο url τους
- έντυπα: ΟΧΙ, εκτός αν είναι εύκολα μετατρέψιμα σε ηλεκτρονικά επεξεργάσιμη μορφή

# Βάσεις δεδομένων

- Οι βάσεις δεδομένων αυτές καθαυτές δεν μπορούν να ενταχθούν στο CLARIN EL ως γλωσσικοί πόροι. Ωστόσο, το περιεχόμενό τους μπορεί (κατά περίπτωση).
- Αν είναι βάσεις δεδομένων με κειμενικό υλικό (π.χ. βιβλιοκριτικές / βιβλιοπαρουσιάσεις) → ΝΑΙ, εφόσον μπορούμε να εξαγάγουμε τα κείμενα και να δημιουργήσουμε ένα ενιαίο dataset
  - π.χ. *το corpus των βιβλιοπαρουσιάσεων*

  στην περίπτωση αυτή, ισχύουν οι προδιαγραφές για το κειμενικό υλικό.

# Ευαίσθητα δεδομένα

- Πόροι με προσωπικά δεδομένα: δεκτά
  - αν έχει γίνει ή αν μπορεί να γίνει ανωνυμοποίηση

  ή
  - αν υπάρχει η γραπτή συγκατάθεση των συμμετεχόντων

# Υπάρχουσες συλλογές κειμένων

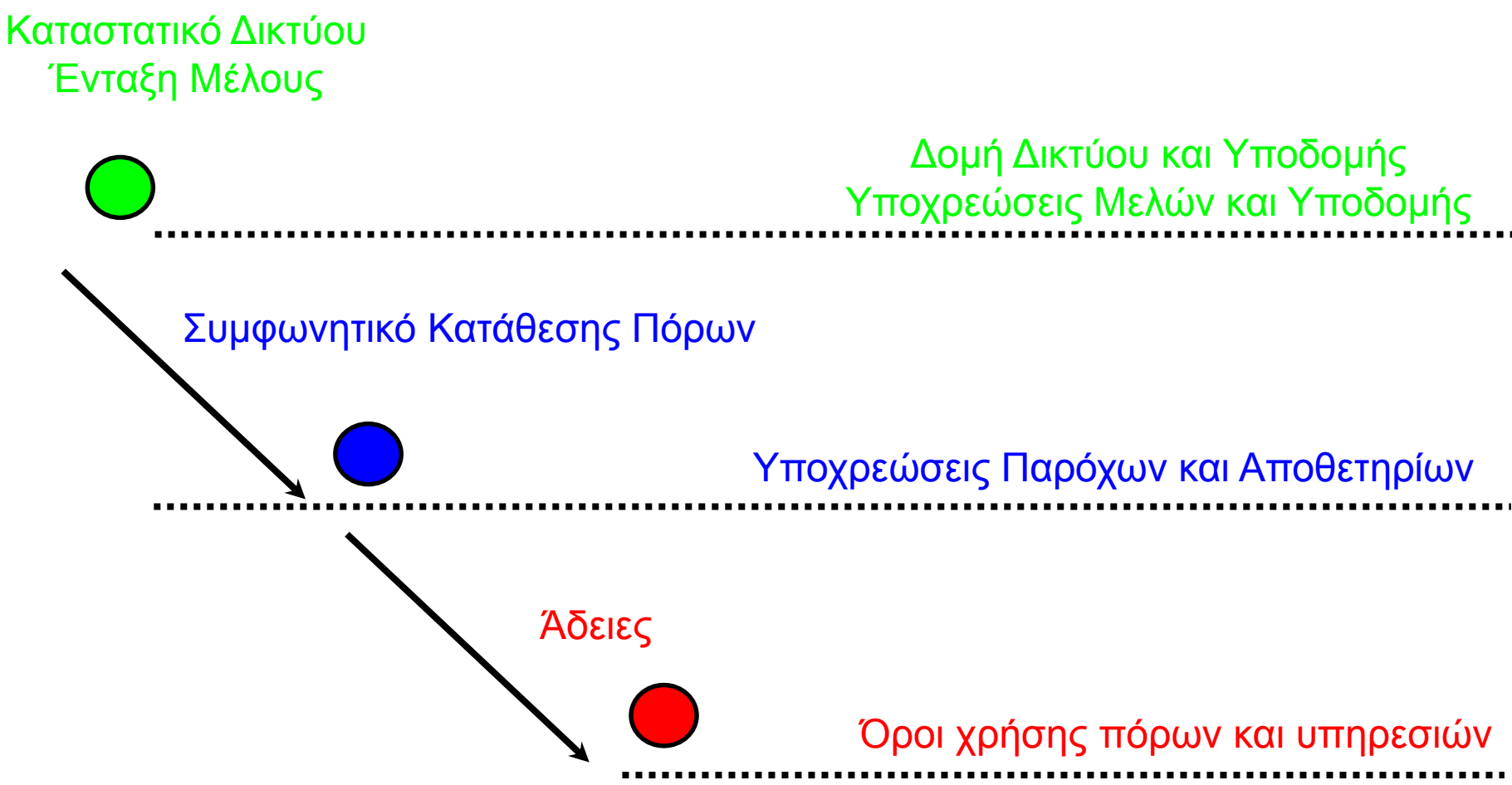
- Προτιμητέα λύση: να τεκμηριωθεί και αποθηκευτεί στο CLARIN EL το περιεχόμενό τους ως ενιαίος (και επεξεργάσιμος από την υποδομή) πόρος, π.χ.
  - *Οι δημοσιεύσεις του Χ Τμήματος ή*
  - *Οι δημοσιεύσεις πάνω στο Χ θέμα*
- Ο αρχικός πόρος εξακολουθεί να διατίθεται όπως και πριν
- Λιγότερο προτιμητέα λύση: τεκμηρίωση και παραπομπή στο url τους

# Άλλου τύπου υλικό

- υλικό που δεν εμπίπτει σε αυτές τις κατηγορίες αλλά αξιολογείται ως χρήσιμο
  - μπορεί να εντάσσεται στην κατηγορία των πόρων που έχουν μόνο μεταδεδομένα στο CLARIN με παραπομπή στο url τους
  - ή στην κατηγορία του απλού καταλόγου πληροφοριακού υλικού

# Νομικά ζητήματα

# Ροή νομικών διαδικασιών



# Καταστατικό Δικτύου

- Υπογράφεται
  - μεταξύ των Μελών του Δικτύου
- Ορίζει
  - τον σκοπό, τις αρχές και τη δομή του Δικτύου
- Ρυθμίζει
  - τους τύπους, τα δικαιώματα και τις υποχρεώσεις των Μελών
- Προβλέπει
  - το σχήμα διοίκησης και τα όργανα του Δικτύου

# Συμφωνητικό Κατάθεσης Πόρων

- νομικό  κείμενο το οποίο υπογράφεται μεταξύ Παρόχου και Υποδομής CLARIN EL
- με το οποίο ο Πάροχος δηλώνει ότι έχει (ή ότι έχει ξεκαθαρίσει) τα δικαιώματα διανοητικής ιδιοκτησίας και διάθεσης του πόρου
- εξουσιοδοτεί την Υποδομή να συμπεριλάβει τον πόρο στα Αποθετήριά της και να τον διαθέτει σύμφωνα με το σχήμα αδειοδότησης που ο ίδιος έχει επιλέξει
- και η Υποδομή δεσμεύεται για την απρόσκοπτη διάθεση του πόρου με τους ενδεδειγμένους και συμφωνημένους τρόπους.

# Ζητήματα διάθεσης

- κάθε πόρος που διατίθεται μέσω CLARIN EL πρέπει να έχει σαφές και ρητό καθεστώς διανοητικής ιδιοκτησίας
- ο πάροχος του πόρου πρέπει να κατέχει τα δικαιώματα διάθεσης (τα οποία βεβαίως διατηρεί) και
- ορίζει τους όρους με τους οποίους ο πόρος θα διατίθεται μέσω CLARIN EL

# Αρχές διάθεσης

- στόχος είναι οι πόροι να είναι ελεύθερα διαθέσιμοι
- για κάθε πόρο παρέχεται μία άδεια χρήσης με ξεκάθαρους όρους
- ωστόσο, υπάρχει μέριμνα για τον σεβασμό
  - των υφισταμένων πνευματικών δικαιωμάτων των πόρων
  - των διαφορετικών απαιτήσεων των παρόχων

# Κατηγοριοποίηση πόρων CLARIN ανάλογα με τη χρήση

- ελεύθερα διαθέσιμοι (Publicly Available)
  - Creative Commons 0 ή
  - Open Database License (ODbL)


- για ακαδημαϊκή / ερευνητική χρήση


- χρήση υπό περιορισμούς
  - πρόσθετοι περιορισμοί ηθικού χαρακτήρα ή
  - σχετικοί με προστασία δεδομένων

# Ζητήματα διανοητικής ιδιοκτησίας και αδειοδότησης

- Οι πρότυπες άδειες χρήσης του CLARIN EL περιλαμβάνουν
  - **άδειες Creative Commons** (ξεκινώντας από τις Creative Commons Zero)
  - **άδειες META-SHARE "No Redistribution"**, που επιτρέπουν την χρήση και αξιοποίηση πόρων, δίνοντας στον κάτοχο του πόρου πλήρη έλεγχο μέσω της απαγόρευσης της αναδιανομής
  - άδειες CLARIN ERIC
  - άδειες ανοικτού λογισμικού για εργαλεία και υπηρεσίες
- Θα λαμβάνονται υπόψη οι υπάρχουσες άδειες διάθεσης

# Σενάρια χρήσης γλωσσικών πόρων, τεχνολογιών και υπηρεσιών γλωσσικής επεξεργασίας

# Σενάρια γλωσσικής έρευνας

- Εντοπισμός φωνημάτων και των συνδυασμών τους σε corpus προφορικού λόγου, είτε στην αρχική του μορφή (ηχητικό υλικό) είτε στην ορθογραφική ή φωνητική του μεταγραφή.
- Σύγκριση της στατιστικής συχνότητας εμφάνισης κυρίων ονομάτων (τοπωνυμίων, ανθρωπωνυμίων, κτλ.) σε διαφορετικούς πόρους.
- Εντοπισμός συντακτικών σχημάτων / δομών σε κείμενα.

# Όμως …

**Οι γλωσσικοί πόροι και η γλωσσική τεχνολογία δεν είναι πολύτιμα μόνο για τους γλωσσολόγους …**

# Σενάριο 1

- **Θέμα**: η μελέτη της άποψης διαφόρων πολιτικών για την κρίση
- **Υλικό**: επιλογή πόρου κειμένων που προέρχεται από τον Τύπο
- **Αναζήτηση**: δηλώσεων / συνεντεύξεων / άρθρων
  - των συγκεκριμένων πολιτικών
  - με θέμα την κρίση
- **Εντοπισμός** του υλικού: με τη βοήθεια γλωσσικών εργαλείων που αναγνωρίζουν λέξεις σχετικές με το θέμα
- **Ανάλυση** του υλικού:
  - ποσοτική ανάλυση με στατιστικά εργαλεία
  - ποιοτική ανάλυση με γλωσσικά εργαλεία

# Σενάριο 2

- **ανάλυση πολιτικού λόγου:** συγκέντρωση πολιτικών κειμένων, ηχογραφήσεων & βίντεο από διάφορες πηγές (π.χ. αρχεία πολιτικών ιδρυμάτων, Αρχείο ΕΡΤ, ΕΟΑ, κτλ.)
- **προπαρασκευή**
  - για ηχητικά αρχεία: μεταγραφή σε γραπτό λόγο
  - για βίντεο: επισημείωση κινήσεων/χειρονομιών/εκφράσεων λόγου & σχέσεων μεταξύ τροπικοτήτων
  - για κείμενα: λημματοποίηση, μορφοσυντακτική επισημείωση, συντακτική ανάλυση, σημασιολογική επισημείωση (π.χ. λέξεις με συναισθηματικό φόρτο κτλ.)
- **ανάλυση/μελέτη**
  - στατιστική επεξεργασία (π.χ. συχνότητα λημμάτων, γραμματικών φαινομένων)
  - μελέτη τροπικοτήτων (π.χ. σύνδεση συγκεκριμένων χειρονομιών με ορισμένη στάση του ομιλητή)
- δυνατότητα μελέτης διαχρονικά, ανά ομιλητή, σε άλλες γλώσσες ...

# Σενάριο 3

- **Θέμα**: μελέτη της θέση της γυναίκας στη μεσαιωνική εποχή με βάση τα ιπποτικά έπη και τα ακριτικά τραγούδια.
- Με τη βοήθεια
  - του εργαλείου εξαγωγής κυρίων ονομάτων ο ερευνητής επιλέγει όλα τα κύρια ονόματα θηλυκού γένους
  - του μορφοσυντακτικού αναλυτή επιλέγει τα επίθετα και τις μετοχές που τα συνοδεύουν
  - του εργαλείου εξαγωγής συμφραστικών πινάκων αντλεί όλες τις προτάσεις που περιλαμβάνουν αυτά τα πρόσωπα.
  - στατιστικών εργαλείων αναλύει ποσοτικά αυτό το σώμα υλικού
- Αξιολόγηση και ερμηνεία των αποτελεσμάτων των εργαλείων γλωσσικής τεχνολογίας από τον ειδικό

# Σενάριο 4

- **Θέμα**: Μελέτη των κινημάτων του 20ού αιώνα από υλικό εφημερίδων.
- Ο μελετητής
  - αναζητά τις οντότητες που εμφανίζονται στα κείμενα (άτομα, κόμματα, οργανώσεις κ.λπ.) με το εργαλείο αναζήτησης ονοματικών οντοτήτων
  - ζητά από την εφαρμογή να του φέρει όλα τα σημεία που εμφανίζονται αυτές οι οντότητες
  - ζητά από την εφαρμογή να κάνει ποσοτική ανάλυση (στατιστική συχνότητα) της εμφάνισης των οντοτήτων αυτών
  - ζητά από την εφαρμογή να κάνει συντακτική ανάλυση των προτάσεων αυτών για να δει τα γεγονότα με τα οποία συνδέεται η κάθε οντότητα (parsing & event extraction)

# Σενάριο 5

- Λεξικογραφία: κατάρτιση λημματολογίου & κωδικοποίηση λήμματος
- επιλογή γλωσσικού υλικού
  - επιλογή από υφιστάμενους πόρους
  - συγκέντρωση νέων κειμένων από διαδίκτυο, για δημιουργία νέου πόρου
- επεξεργασία με εργαλεία υποδομής
  - λημματοποιητές και μορφοσυντακτικοί επισημειωτές
  - στατιστικά εργαλεία μέτρησης συχνοτήτων
  - εργαλεία δημιουργίας συμφραστικών πινάκων
  - εργαλεία συντακτικής ανάλυσης
  - εργαλεία εξαγωγής ορολογίας από κείμενα…

# Σενάριο 6

- Εξαγωγή ορολογίας από κείμενα
  - Επιλογή πόρου στον συγκεκριμένο τομέα
  - Λημματοποίηση και μορφοσυντακτική επισημείωση
  - Χρήση εργαλείου εξαγωγής ορολογίας
  - Επεξεργασία καταλόγου όρων από ειδικούς του τομέα για επαλήθευση

# Σενάριο 7

- Αυτόματη εξαγωγή πληροφορίας από επιστημονικές δημοσιεύσεις στην ιατρική ή τη νομική επιστήμη
  - γονίδιο Χ σχετίζεται με ασθένεια Υ
  - νόμος Χ τροποποιεί νόμο Υ
- Αναζήτηση και ανάκτηση σχετικών δημοσιεύσεων με βάση οντολογίες, θησαυρούς ή λέξεις κλειδιά
- Αναγνώριση οντοτήτων και σχέσεων μεταξύ τους με χρήση εξειδικευμένων εργαλείων
  - Αναγνώρισης και εξαγωγής ορολογίας
  - Αναγνώρισης και εξαγωγής ονομάτων
  - Αναγνώρισης και εξαγωγής γεγονότων
  - Υπολογισμού των συντακτικών σχέσεων μεταξύ τους

# Άλλα σενάρια;

- Παρουσιάσαμε κάποια από τα σενάρια που έχουν ήδη διατυπωθεί από ερευνητές χρήστες γλωσσικών πόρων και τεχνολογιών
- Να τα εμπλουτίσουμε σύμφωνα με τις δικές σας ανάγκες;
  - Ποιος είναι ο πόρος που χρειάζομαι; → **κίτρινο**
  - Ποια είναι η υπηρεσία που χρειάζομαι; → **ροζ**
  - Ποιο είναι το σενάριό μου; → **πορτοκαλί**