

Η ερευνητική υποδομή CLARIN-EL

Στέλιος Πιπερίδης,
Συντονιστής CLARIN-EL
Ινστιτούτο Επεξεργασίας του Λόγου
Ε.Κ. "Αθηνά"



Μερικά αριθμητικά στοιχεία

- Αφθονία περιεχομένου στο διαδίκτυο
- Δεκ. 2008 : 487 δις GB ψηφιακού περιεχομένου δημιουργημένου (πρωτογενές+ψηφιοποιούμενο) το 2008, δηλαδή ~3.892.179.868.480.350.000.000 bits, ~162T ψηφιακές εικόνες, ~19B blue-ray disks
- Το ψηφιακό σύμπαν αναμένεται να διπλασιάζεται κάθε 18 μήνες
- Ωστόσο, η αφθονία περιεχομένου ποικίλλει ανάλογα με τη γλώσσα (π.χ λιγότερο χρησιμοποιούμενες γλώσσες, μειονοτικές γλώσσες, διαλέκτους, κτλ.)
- Παντοκρατορία αγγλικού περιεχομένου μέχρι πρόσφατα

Μερικά αριθμητικά στοιχεία (2)

- Τα πράγματα αλλάζουν:
 - 1996 : 66% της ψηφιακής κοινότητας βρίσκονται στις ΗΠΑ
 - 2008 (Δεκ.) : 83% της ψηφιακής κοινότητας εκτός ΗΠΑ (28% ΕΕ, 42% Ασία)
- Διάφοροι τύποι διαδικτυακών δεδομένων (κανονικό κείμενο, λιγότερο ή περισσότερο «καλοσχηματισμένη» κειμενική επικοινωνία σε μπλογκ, chatrooms κτλ., εικόνες, βίντεο κτλ.)
- Κάποια από αυτά τα δεδομένα είναι στην πραγματικότητα επισημειώσεις άλλων πρωτογενών δεδομένων (π.χ. περιλήψεις, μεταγραφές/υπότιτλοι, λεζάντες εικόνων ή ακόμη και γνώμες)

Μερικά αριθμητικά στοιχεία(3)

- Τα κοινωνικά δίκτυα, το web2.0 και το επερχόμενο web3.0 παρέχουν νέες πηγές, αλλά και νέες προκλήσεις καθώς και νέες εφαρμογές, π.χ. εξόρυξη γνώμης (opinion mining), έλεγχος της αγοράς, ψηφιακή διπλωματία κτλ.
- Τα ευρυζωνικά δίκτυα διευκολύνουν τα παραπάνω να γίνονται διαρκώς πιο πολυμεσικά, με τη φυσική γλώσσα να παραμένει το κύριο μέσο
- ενόσω οι νέες αγορές και ο ανταγωνισμός ενισχύουν την πολυγλωσσική τους διάσταση
- Μεγάλο ποσοστό αυτού του περιεχομένου εμπίπτει σε αυτό που οριοθετείται και αυτοπροσδιορίζεται με την ευρεία έννοια ως πολιτιστικό περιεχόμενο

Μερικά αριθμητικά στοιχεία(4)

- Το μέγεθος του διαθέσιμου περιεχομένου, οι ανάγκες πρόσβασης σε αυτό, η επεξεργασία του για πολλούς διαφορετικούς σκοπούς, καθιστούν αναγκαία τη χρήση υπολογιστικών εργαλείων
- Με τη φυσική γλώσσα να αποτελεί το κύριο μέσο δημιουργίας/οργάνωσης/αναζήτησης/ανάλυσης/μετάφρασης/εξόρυξης/κλπ. η σημασία της γλωσσικής τεχνολογίας καθίσταται αυταπόδεικτη

Γλωσσική τεχνολογία

- Η εμπειρική στροφή στη ΓΤ, παρά τα όποια προβλήματα, οδηγεί σε αξιοσημείωτη πρόοδο καθημερινά (βλ. Google translate)
- Εντυπωσιακές βελτιώσεις στις υπολογιστικές μεθόδους και τεχνικές (κυρίως μηχανική μάθησης)
- Πολλές “quick and dirty” τεχνικές με πολύ ικανοποιητικά αποτελέσματα
- Αλλά πάντα με τη διαθεσιμότητα κατάλληλων δεδομένων και εργαλείων, δηλ. πόρων, *απαραίτητη προϋπόθεση* για κάθε τεχνολογική πρόοδο

Ο στόχος της ΕΥ CLARIN

- Η ΕΥ CLARIN (www.clarin.eu) στοχεύει να δημιουργήσει μία ολοκληρωμένη και διαλειτουργική ερευνητική υποδομή Γλωσσικών Πόρων και Τεχνολογιών
- καταπολεμώντας έτσι την ισχύουσα αποσπασματικότητα
- και προσφέροντας ένα σταθερό, συνεπές, εύχρηστο και επεκτάσιμο περιβάλλον πρόσβασης σε γλωσσικά δεδομένα
- στην υπηρεσία των Κοινωνικών και Ανθρωπιστικών Επιστημών (ΚΑΕ)

Το πρόβλημα

- πολλές αρχειακές συλλογές, ωστόσο πόροι και εργαλεία ΓΤ είναι γνωστά μόνο σε ορισμένες ερευνητικές κοινότητες
- συλλογές, πόροι και εργαλεία ασύνδετα μεταξύ τους καθιστώντας την αποτελεσματική (ορθή και πλήρη) αναζήτηση τους δύσκολη
- η εύρεση πόρων εξαρτάται από τη γλώσσα τεκμηρίωσης
- ακολουθούν διαφορετικά πρότυπα τεκμηρίωσης
- δεν υπάρχουν κίνητρα για διάθεση πόρων

Η αποστολή του CLARIN

- Τι;
 - να δημιουργήσει μια υποδομή που να διαθέσει Γλωσσικούς Πόρους και Τεχνολογίες στους ερευνητές των Κοινωνικών και Ανθρωπιστικών Επιστημών (ΚΑΕ)
- Πώς;
 - ενοποιώντας υφιστάμενες ψηφιακές αρχειακές συλλογές σε μία «ομοσπονδία αρχείων» με ενιαία διαδικτυακή πρόσβαση
 - παρέχοντας σχετικές διαδικτυακές υπηρεσίες με τη μορφή γλωσσικών υπολογιστικών εργαλείων που "τρέχουν" πάνω στα γλωσσικά δεδομένα

Δηλαδή;

- το CLARIN σκοπεύει να ενσωματώσει
 - Γλωσσικούς Πόρους: ψηφιακό περιεχόμενο κάθε είδους (κείμενο, ήχο, εικόνα, βίντεο), πρωτογενείς και επισημειωμένους (ηχογραφήσεις, μαγνητοσκοπήσεις ή κείμενα), λεξικά, οντολογίες, ορολογικά γλωσσάρια κτλ. και
 - εργαλεία Γλωσσικής Τεχνολογίας: εργαλεία αναγνώρισης φωνής, ληματοποιητές, συντακτικούς αναλυτές, εργαλεία αυτόματης εξαγωγής περίληψης, εργαλεία εξαγωγής πληροφορίας κτλ.
 - σε ένα συστηματικά οργανωμένο δίκτυο αποθετηρίων το οποίο θα είναι διαθέσιμο μέσω διαδικτυακών υπηρεσιών σε ερευνητές όλων των επιστημών

Το όραμα του CLARIN

- Ένας ερευνητής από το γραφείο του στην Κέρκυρα θα μπορεί:
 - με μία πιστοποιημένη εγγραφή (authenticated single sign-on)
 - να ψάξει, να βρει και να πάρει την έγκριση να χρησιμοποιήσει κείμενα
 - από την Οξφόρδη, το Μπέργκεν και το Λέιντεν
 - να επιλέξει το ακριβές σύνολο δεδομένων στα οποία θέλει να δουλέψει και να αποθηκεύσει την επιλογή του
 - να τρέξει πάνω στην επιλογή του εργαλεία σημασιολογικής ανάλυσης από την Αθήνα και
 - στατιστικά εργαλεία από τη Βουδαπέστη
 - να χρησιμοποιήσει την υπολογιστική ισχύ ενός άλλου υπολογιστικού κέντρου, όπου και όποτε απαιτείται
 - να αποθηκεύσει τη διαδικασία και τα αποτελέσματα της ανάλυσης και
 - να τα μοιραστεί με συνεργάτες του στο Παρίσι, στη Βιέννη και στο Ελσίνκι

Σενάριο χρήσης

● Επικοινωνία/MME

- Εργασία: Ανάλυση πολιτικού λόγου
- Πώς:
 - 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.)
 - κανονικοποίηση σε μορφή συμβατή με τα εργαλεία
 - 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 κτλ.

Οι 4 διαστάσεις του εγχειρήματος

- Η τεχνική διάσταση
- Η γλωσσική διάσταση
- Η χρηστική διάσταση
- Η διαχειριστική και νομική διάσταση

Η τεχνική διάσταση

- λειτουργικές και τεχνικές προδιαγραφές της υποδομής
- ανάπτυξη πρωτότυπου εφαρμογής
- με ενσωμάτωση μεγάλης ποικιλίας
 - Γλωσσών
 - Πόρων
 - Υπηρεσιών
- και ολοκλήρωση σε αυτήν υπάρχοντων πόρων και εργαλείων (με μετατροπές και προσαρμογές, όπου απαιτείται)
- «ομοσπονδία» υφιστάμενων αρχειακών συλλογών
- έμφαση στη διαλειτουργικότητα
- έμφαση στα σχετικά πρότυπα

Η γλωσσική διάσταση

- κάλυψη όλων των γλωσσών που μιλιούνται ή μελετώνται στα κράτη που συμμετέχουν στο έργο
- τα πρότυπα που προτείνει το έργο για την αναπαράσταση και την τεκμηρίωση των πόρων και των εργαλείων πρέπει να καλύπτουν όλες τις γλώσσες
- ορισμός και υλοποίηση του BLARK (Basic Language Resources Toolkit) για όλες τις γλώσσες
- καταγραφή υπαρχόντων εργαλείων και πόρων
- ενίσχυση της ανάπτυξης όσων δεν υπάρχουν

Η χρηστική διάσταση

- οι χρήστες της ΕΥ είναι επιστήμονες των ΚΑΕ – δεν είναι ειδικοί στη Γλωσσική Τεχνολογία (ΓΤ)
- γνωρίζουμε τις ανάγκες τους;
- γνωρίζουν τα οφέλη από τη χρήση της ΓΤ;
- υπάρχουν λίγα εργαλεία για διαδικτυακή επεξεργασία των δεδομένων
- και όταν υπάρχουν, τα εργαλεία ΓΤ συχνά είναι δύσχρηστα για τους μη ειδικούς
- απαραίτητα:
 - ανάλυση αναγκών των χρηστών
 - δημοσιοποίηση & διάχυση δυνατοτήτων της ΓΤ

Η νομική διάσταση

• ζητήματα πνευματικών δικαιωμάτων

- στόχος είναι οι πόροι να είναι ελεύθερα διαθέσιμοι
- παρέχεται μία απλή άδεια χρήσης με ξεκάθαρους όρους (συμβατή με Creative Commons)
- ωστόσο, υπάρχει μέριμνα για τον σεβασμό
 - των πνευματικών δικαιωμάτων υφισταμένων ή και μελλοντικών μη ελεύθερα διαθέσιμων πόρων
 - των διαφορετικών εθνικών νομοθεσιών

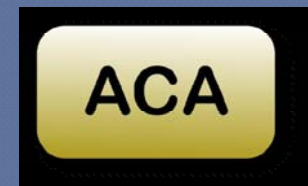
Κατηγορίες διάθεσης πόρων

- ελεύθερα διαθέσιμοι (Publicly Available)

- Protocol for Implementing Open Access Data (CC0) ή
- Open Database License (ODbL)



- για ακαδημαϊκή / ερευνητική χρήση



- χρήση υπό περιορισμούς

- πρόσθετοι περιορισμοί ηθικού χαρακτήρα ή
- σχετικοί με προστασία δεδομένων



Τι ΔΕΝ κάνει το CLARIN

- δεν δημιουργεί νέους πόρους – απλώς εντοπίζει υπάρχοντες και τους προσαρμόζει όπου είναι απαραίτητο
- δεν υλοποιεί εφαρμογές
- δεν εστιάζει στις μεγάλες γλώσσες – όλες είναι εξίσου σημαντικές
- δεν ενισχύει την ευρωπαϊκή βιομηχανία – οι χρήστες – στόχος είναι οι ερευνητές των ΚΑΕ αλλά και γενικότερα η επιστημονική κοινότητα

Γιατί μια ευρωπαϊκή υποδομή;

- μεγάλη αποσπασματικότητα στα έργα του χώρου
- σχετικές δράσεις παραμένουν άγνωστες
- απουσία διαλειτουργικότητας
- μικρή βιωσιμότητα αποτελεσμάτων
- ύπαρξη τεχνογνωσίας και εμπειρίας αλλά όχι σε όλα τα κράτη
- εργαλεία ανεξάρτητα γλώσσας μπορούν άμεσα να χρησιμοποιηθούν για άλλους πόρους
- εργαλεία εξαρτώμενα από γλώσσα συχνά μπορούν να προσαρμοστούν και σε άλλες γλώσσες
- τα περισσότερα κράτη δεν μπορούν να αναλάβουν μόνα τους το κόστος της προσπάθειας
- απουσία διακρατικού συντονισμού

Ποιοι είμαστε

- η κοινοπραξία CLARIN έχει
 - 36 εταίρους
 - από 22 χώρες
- το δίκτυο CLARIN έχει
 - 177 οργανισμούς - μέλη με 195 μονάδες / τμήματα
 - από 33 χώρες

Η ελληνική διάσταση: CLARIN-EL

- Σε προπαρασκευαστική φάση
- Κοινοπραξία του έργου
 - Ε.Κ. "Αθηνά"
 - Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ)
 - Ινστιτούτο Πληροφοριακών Συστημάτων και Προσομοίωσης (ΙΠΣΥΠ)
 - ΕΚΕΦΕ "ΔΗΜΟΚΡΙΤΟΣ" / Ινστιτούτο Πληροφορικής και Επικοινωνιών
 - Εθνικό Δίκτυο Έρευνας και Τεχνολογίας (ΕΔΕΤ) Α.Ε.
 - Ιόνιο Πανεπιστήμιο / Εργαστήριο Ψηφιακών Βιβλιοθηκών και Ηλεκτρονικής Δημοσίευσης
 - Εθνικό Ίδρυμα Ερευνών / Εθνικό Κέντρο Τεκμηρίωσης
 - Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών - Τμήμα Φιλολογίας / Τομέας Γλωσσολογίας
 - Κέντρο Ελληνικής Γλώσσας (ΚΕΓ)
 - Πανεπιστήμιο Αιγαίου / Τμήμα Μεσογειακών Σπουδών - Εργαστήριο Γλωσσολογίας

Δίκτυο CLARIN-EL

• το υπό σύσταση δίκτυο

- περιλαμβάνει οργανισμούς και φυσικά πρόσωπα από τις δύο εμπλεκόμενες ερευνητικές κοινότητες
 - της ΓΤ και
 - των ΚΑΕ
- με τη βοήθεια των μελών του δικτύου διεξάγεται η χαρτογράφηση του χώρου

Στόχος της χαρτογράφησης: η καταγραφή

- των φορέων και των ιδιωτών από τις κοινότητες των ΚΑΕ και των ΓΠΤ που μπορούν να αξιοποιήσουν τη ΓΤ στις ερευνητικές τους δράσεις
- των γλωσσικών πόρων και τεχνολογιών (ΓΠΤ) που υπάρχουν για την ελληνική γλώσσα, είτε πρόκειται για πρωτογενές υλικό από την περιοχή των ΚΑΕ είτε για ΓΠΤ που έχουν αναπτυχθεί στον χώρο της ΓΤ
- της ερευνητικής και αναπτυξιακής δραστηριότητας των φορέων και ιδιωτών που δραστηριοποιούνται στον χώρο της ΓΤ
- των αναγκών των χρηστών



Χαρτογράφηση

Στο πλαίσιο του CLARIN-EL, διενεργείται έρευνα με σκοπό τη χαρτογράφηση των δύο εμπλεκόμενων ερευνητικών κοινοτήτων, δηλαδή της Γλωσσικής Τεχνολογίας (ΓΤ) και των Κοινωνικών και Ανθρωπιστικών Επιστημών (ΚΑΕ). Η χαρτογράφηση αυτή έχει σκοπό να αποτυπώσει τα εξής στοιχεία:

- την ερευνητική και αναπτυξιακή δραστηριότητα των φορέων και ιδιωτών που δραστηριοποιούνται στον χώρο της ΓΤ,
- τους γλωσσικούς πόρους και τεχνολογίες (ΓΠΤ) που διατίθενται για την ελληνική γλώσσα, είτε πρόκειται για πρωτογενές υλικό από την περιοχή των ΚΑΕ είτε για ΓΠΤ που έχουν αναπτυχθεί στον χώρο της ΓΤ
- τους φορείς και ιδιώτες από τις κοινότητες των ΚΑΕ και των ΓΠΤ που μπορούν να αξιοποιήσουν τη ΓΤ στις ερευνητικές τους δράσεις.

Σας καλούμε να συμμετάσχετε στην έρευνα αυτή, συμπληρώνοντας όποιο/α από τα ερωτηματολόγια που ακολουθούν ταιριάζει/ουν στην περίπτωση σας:

Ερωτηματολόγιο για φυσικά πρόσωπα	Συμπληρώνεται υποχρεωτικά από κάθε ενδιαφερόμενο/η να συμμετάσχει στο δίκτυο CLARIN-EL, είτε ως ιδιώτης είτε ως εργαζόμενος/η σε οργανισμό. Στην περίπτωση που συμμετέχει ως εκπρόσωπος ενός οργανισμού, θα πρέπει να συμπληρώσει ξεχωριστά και το ερωτηματολόγιο για οργανισμούς. Μέλη του ίδιου οργανισμού μπορούν επίσης να συμμετάσχουν στο δίκτυο συμπληρώνοντας κανονικά μόνο το ερωτηματολόγιο για φυσικά πρόσωπα.
Ερωτηματολόγιο για οργανισμούς	Συμπληρώνεται μία φορά για κάθε οργανισμό - στην περίπτωση αυτή δηλώνεται ένα άτομο ως υπεύθυνος επικοινωνίας εκ μέρους του οργανισμού. Το άτομο αυτό πρέπει να συμπληρώσει επίσης το ερωτηματολόγιο για φυσικά πρόσωπα. Τα υπόλοιπα μέλη του ίδιου οργανισμού που επιθυμούν να συμμετάσχουν στο δίκτυο συμπληρώνουν μόνο το ερωτηματολόγιο για φυσικά πρόσωπα.
Ερωτηματολόγιο για ερευνητικά/αναπτυξιακά έργα Γλωσσικής Τεχνολογίας	Συμπληρώνεται από τον/την υπεύθυνο/η του έργου (αφού συμπληρώσει το ερωτηματολόγιο για φυσικά πρόσωπα και, εφόσον είναι και εκπρόσωπος οργανισμού, το αντίστοιχο για οργανισμούς) και αποκλειστικά για έργα από τον χώρο της Γλωσσικής Τεχνολογίας.
	Συμπληρώνεται από τον/την υπεύθυνο/η του πόρου (αφού συμπληρώσει το ερωτηματολόγιο για φυσικά πρόσωπα και, εφόσον είναι και εκπρόσωπος οργανισμού, το αντίστοιχο για οργανισμούς). Ως πόρος πρωτογενούς υλικού ορίζεται κάθε πόρος ψηφιακού / ψηφιοποιημένου λόγου , όπως

Προσεχή Γεγονότα

Δεν υπάρχουν γεγονότα

< Σεπτέμβριος 2010 >						
Κ	Δ	Τ	Τ	Π	Π	Σ
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2

Σύνδεσμοι

- ▶ Ευρωπαϊκό CLARIN
- ▶ Ολλανδικό CLARIN
- ▶ Ισπανικό CLARIN

Είσοδος

Όνομα Χρήστη

Κωδικός

Να με θυμάσαι ☐

Σύνδεση

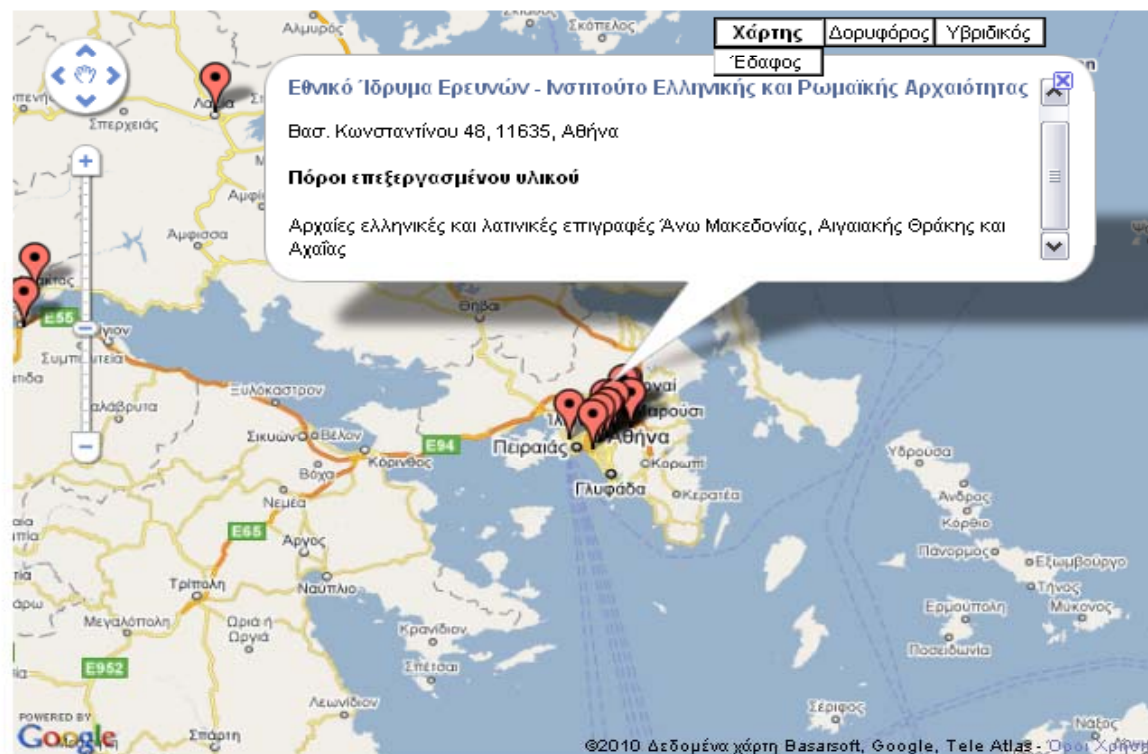
- Ξεχάσατε τον κωδικό σας;
- Ξεχάσατε το όνομα χρήστη;
- Δημιουργία λογαριασμού

Καταγραφή ΓΠΤ (1)

- **πόροι δεδομένων:** **ψηφιακά** δεδομένα σε κάθε **μέσο** (κείμενο, ήχο, εικόνα, βίντεο) που μπορούν να χρησιμοποιηθούν
 - ως υλικό για μελέτη από ερευνητές ΚΑΕ (π.χ. σώμα γραπτών κειμένων / ηχογραφημένου λόγου/βίντεο για μελέτη από γλωσσολόγους, συλλογή βιβλίων και τεκμηρίων με θέμα τον μινωικό πολιτισμό για αρχαιολόγους κτλ.)
 - ως υλικό που θα χρησιμοποιηθεί για την καλύτερη πρόσβαση στο υπό μελέτη υλικό (π.χ. λεξικά, θησαυροί, οντολογίες κτλ. που επιτρέπουν την πρόσβαση στο υλικό με βάση το λήμμα, λέξη-κλειδί κτλ.)
 - ως υλικό για την ανάπτυξη και βελτίωση εργαλείων ΓΤ (π.χ. γλωσσολογικά επισημειωμένο σώμα κειμένων για την εκπαίδευση εργαλείου)
- **εργαλεία/εφαρμογές ΓΤ:** εργαλεία που επεξεργάζονται τα γλωσσικά δεδομένα (π.χ. μορφοσυντακτικός επισημειωτής, λημματοποιητής, εργαλείο αναγνώρισης ονοματικών οντοτήτων κτλ.)

Παρουσίαση αποτελεσμάτων καταγραφής

Clarín-EL Map



Επιλογές εμφάνισης

- ☐ Πρόσωπα
- ☐ Ερευνητικά έργα
- ☒ Πρωτογενείς πόροι
- ☒ Πόροι επεξεργασμένου υλικού
- ☒ Πόροι αναφοράς
- ☐ Εργαλεία/Εφαρμογές

Καταγραφή αναγκών των χρηστών

- μελετώνται οι απαιτήσεις των διαφορετικών κλάδων των ΚΑΕ
- ώστε να συνδεθούν οι απαιτήσεις αυτές με συγκεκριμένες εφαρμογές που οι γλωσσικές τεχνολογίες μπορούν να προσφέρουν και
- να εντοπιστούν οι πόροι και τα δεδομένα που απαιτούνται για να προσφερθούν οι υπηρεσίες αυτές

Βασικές Αρχές δικτύου CLARIN-EL(1)

- Ανοικτό και κατανεμημένο
- Με τη βέλτιστη γεωγραφική κατανομή
- Πόροι, εργαλεία και διαδικτυακές υπηρεσίες στην υπηρεσία χρηστών κατόπιν εγγραφής τους στο δίκτυο
- Πόροι, εργαλεία και διαδικτυακές υπηρεσίες παραμένουν, προσφέρονται, ελέγχονται, βελτιώνονται, κλπ. από τους συνιστώντες κόμβους
- Αναζήτηση πόρων, εργαλείων και υπηρεσιών μέσω περιγραφών της ύπαρξής τους σε κεντρικό συσσωρευτή
- Πρόσβαση μέσω αυθεντικοποίησης-εξουσιοδότησης και σύμφωνα με τα δικαιώματα χρήσης

Βασικές Αρχές δικτύου CLARIN-EL(2)

- Προγραμματίζει την κάλυψη κενών τόσο σε πόρους όσο και σε τεχνολογία σύμφωνα με τις ανάγκες των χρηστών και τις αρχές των BLARK και ELARK προσαρμοσμένων στις ΚΑΕ
- Προωθεί τη χρήση διεθνών προτύπων και βέλτιστων πρακτικών (όπου υπάρχουν) με στόχο τη διαλειτουργικότητα πόρων και εργαλείων τόσο του ελληνικού δικτύου όσο και του ευρωπαϊκού
- Παρέχει διαρκή εκπαίδευση στη χρήση και υποστηρίζει σε όλα τα στάδια έρευνας, από μετασχηματισμό δεδομένων μέχρι την παρουσίασή τους
- Στηρίζεται και προωθεί τις αρχές των ανοικτών δεδομένων και του ανοικτού λογισμικού

CLARIN (-EL) και άλλες
συνεργατικές πρωτοβουλίες

FLARENET

Fostering Language Resources Network

www.flarenet.eu

Η αποστολή

- Το FLaReNet είναι ένα διεθνές δίκτυο επιστημόνων γλωσσικής τεχνολογίας από όλο τον κόσμο και όλους τους πιθανούς κλάδους και χώρους (ακαδημαϊκό, βιομηχανικό, κλπ) με σκοπό τον
 - προσδιορισμό επιστημονικών και τεχνολογικών προτεραιοτήτων
 - την διαμόρφωση συστάσεων προς τις εμπλεκόμενες κοινότητες σχετικά με πρότυπα γλωσσικής τεχνολογίας και βέλτιστες πρακτικές

Τι κάνει το FLaReNet

- Καταρτίζει τον διεθνή χάρτη γλωσσικών πόρων και γλωσσικών εργαλείων με σκοπό την ανάδειξη των κενών και ελλείψεων σε επίπεδο γλωσσών και εφαρμογών
- Καταρτίζει τον διεθνή χάρτη μεθοδολογιών που χρησιμοποιούνται για την ανάπτυξη των γλωσσικών πόρων (π.χ. Wordnet, σημασιολογικά επισημειωμένα ΣΚ)
- Κατασκευάζει την wikipedia για την γλωσσική τεχνολογία
- Σχεδιάζει την οντολογία του χώρου της γλωσσικής τεχνολογίας

META-NET
Technology Alliance for Multilingual
Europe
META-SHARE
Open Resource Infrastructure

META-SHARE (πρώτα βήματα)

- ανοιχτή, ολοκληρωμένη, ασφαλής και διαλειτουργική υποδομή ΓΠ και ΓΤ για τον τομέα των Τεχνολογιών Ανθρώπινου Λόγου (ΤΑΛ) και άλλων τομέων (πχ. ψηφιακές βιβλιοθήκες, γνωσιακά συστήματα, ρομποτική κτλ.), που θα παρέχει
 - συνεχώς εξελισσόμενες, κλιμακούμενες υπηρεσίες ΓΠ/ΓΤ,
 - σύγχρονα και αναδυόμενα δεδομένα (datasets), εργαλεία και τεχνολογίες
 - βάσει κατανεμημένων δικτυωμένων (networked) αποθετηρίων και κέντρων δεδομένων προσβάσιμων μέσω κοινών διεπαφών
 - που θα είναι συμβατά με τα διεθνή πρότυπα, θα ξεπερνούν διαφορές μορφότυπων, ορολογικές ή σημασιολογικές, θα επιτρέπουν / διευκολύνουν την παροχή υπηρεσιών και στατικές ή δυναμικές συνθέσεις, π.χ. workflows
 - Συμβατών με νομικούς και σχετικούς με την ασφάλεια περιορισμούς

ΓΠ και ΓΤ του META-SHARE

- Γλωσσικά δεδομένα (γραπτά και προφορικά: σώματα κειμένων, λεξικά, γραμματικές, οντολογίες/ορολογίες, κτλ.)
- δεδομένα σχετιζόμενα με τη γλώσσα (που περιλαμβάνουν ή σχετίζονται με άλλα μέσα και τροπικότητες)
- εργαλεία και τεχνολογίες επεξεργασίας και επισημείωσης,
- υπηρεσίες με χρήση γλωσσικών εργαλείων και τεχνολογιών,
- ροές εργασιών (workflows) συνδυάζοντας διαλειτουργικές υπηρεσίες, εργαλεία, μετρικές και πρωτόκολλα αξιολόγησης, υπηρεσίες αξιολόγησης (assessment and evaluation)

Ποιοι συμμετέχουν στο META-SHARE

- Δημόσιοι και ιδιωτικοί φορείς E&A καθώς και βιομηχανικοί φορείς / πάροχοι και χρήστες ΤΑΛ:
 - Ακαδημαϊκά ιδρύματα, ερευνητικοί οργανισμοί, πανεπιστήμια
 - Ερευνητές και φοιτητές
 - Βιομηχανικοί οργανισμοί και ΜΜΕ
 - Εθνικές κυβερνήσεις, οργανισμοί ΕΕ και ιδιωτικοί επενδυτές.
- Συγκεκριμένα
 - Παραγωγοί και πάροχοι ΓΠ και ΓΤ,
 - Χρήστες ΓΠ και ΓΤ και technology integrators,
 - Αποθετήρια ΓΠ και ΓΤ και
 - Διαμορφωτές πολιτικής για τη ΓΤ και άλλοι χρηματοδότες και χορηγοί ΓΠ και ΓΤ. (βλ. “automatic, highly accurate and real-time translation between the major languages of the world — greatly lowering the barriers to international commerce and collaboration” Strategy for American Innovation, President Obama

Υπηρεσίες META-SHARE

- Υπηρεσίες εγγραφής, καταλογογράφησης, περιγραφής και φόρτωσης (uploading) πόρων / εργαλείων, θέασης και πρόσβασης, αξιολόγησης και διαχείρισης, αρχειοθέτησης, συντήρησης και διανομής, διαχείριση νομικών θεμάτων και θεμάτων Πνευματικής Ιδιοκτησίας
- Υπηρεσίες συστάσεων προς τους χρήστες, στατιστικά στοιχεία ανά γλώσσα, τύπο δεδομένων, εφαρμογή, κλπ.
- Υπηρεσίες σύνδεσης πόρων και εργαλείων μέσα από ένα πλαίσιο συντακτικής και σημασιολογικής διαλειτουργικότητας με σκοπό τη δυνατότητα δημιουργίας, δυναμικά (on the fly), ροών εργασιών και συστημάτων μεγαλύτερης πολυπλοκότητας

-
- "The vision I have for the Web is about **anything** being potentially **connected with anything**. It is a vision that provides us with new freedom, and allows us to grow faster than we ever could. . . . it brings the **workings of society** closer to the **workings of our minds**."

Tim Berners-Lee : Weaving the Web, 2000

Πληροφορίες

- ◉ CLARIN Website: <http://www.clarin.eu>
- ◉ CLARIN Office: clarin@clarin.eu
- ◉ Ελληνικός ιστότοπος: <http://www.clarin.gr>
- ◉ Επικοινωνία: info.clarin-el@ilsp.gr

Σας ευχαριστώ
πολύ!

Η ΕΥ CLARIN θα είναι

- ολοκληρωμένη: τα κέντρα που παρέχουν υπηρεσίες θα συνδέονται μέσω τεχνολογίας **Grid** και αποτελούν έναν **and form a virtually integrated domain**
- διαλειτουργική: οι πόροι και οι υπηρεσίες θα βασίζονται σε τεχνολογίες Σημασιολογικού Ιστού, προκειμένου να αντιμετωπίσουν τις υπάρχουσες διαφορές μορφής, δομής και ορολογίας
- σταθερή: οι πόροι και οι υπηρεσίες πρόκειται να παρέχονται συνεχώς και σε βάθος χρόνου, ώστε να αποτελούν αξιόπιστη πηγή για έρευνα
- προσπελάσιμη: η υποδομή θα είναι προσπελάσιμη μέσω διαδικτύου· θα προσφέρονται ποικίλοι τρόποι πρόσβασης και σχετικές εκπαιδευτικές διαδικασίες, ανάλογα με τις διαφορετικές κοινότητες χρηστών
- επεκτάσιμη: η υποδομή είναι ανοιχτή στην απρόσκοπτη προσθήκη νέων πόρων και υπηρεσιών

Σενάρια χρήσης – παραδείγματα (1)

• Λεξικογραφία

• Εργασίες:

- κατάρτιση λεμματολογίου
- κωδικοποίηση λήματος

• Πώς:

- δημιουργία σώματος κειμένων

- επιλογή από υφιστάμενα Σώματα Κειμένων αναζήτηση ΣΚ, επιλογή κειμένων ανάλογα με κριτήρια
- συγκέντρωση νέων κειμένων από διαδίκτυο web crawler, downloading
- οδηγίες για ψηφιοποίηση κειμένων

- διαχείριση κειμένων μετατροπή στην ίδια μορφή, «προσωρινή» αποθήκευση σε διαχειριστικό περιβάλλον, ταξινόμηση κειμένων βάσει πηγής

- επεξεργασία κειμένων στατιστικά εργαλεία μέτρησης συχνοτήτων, λεμματοποιητές, εργαλεία συμφραστικών πινάκων, εργαλεία για ημι-αυτόματη εύρεση συντακτικών πλαισίων υποκατηγοριοποίησης, εργαλεία για εντοπισμό πολυλεκτικών κτλ.