Γλωσσικοί πόροι και Γλωσσική Τεχνολογία

Μαρία Γαβριηλίδου
Ινστιτούτο Επεξεργασίας Λόγου
ΕΚ «Αθηνά»

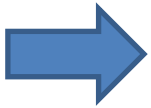
Η δύναμη των δεδομένων

- Τα (επιστημονικά) δεδομένα μεταμορφώνουν και βελτιώνουν με ταχύτατους ρυθμούς τη ζωή μας
  - δεδομένα βιολογικά, γεωγραφικά, δημογραφικά, μετεωρολογικά, οικονομικά, αριθμητικά…
  - σε αρχεία κειμένου, ήχου, εικόνας, βίντεο, πολυμεσικά, …
- τα δεδομένα αποκτούν ιδιαίτερη αξία όταν διαμοιράζονται και ανοίγονται
  - τόσο για ερευνητικούς και τεχνολογικούς σκοπούς
  - όσο και για την δημιουργία καινοτομικών εφαρμογών

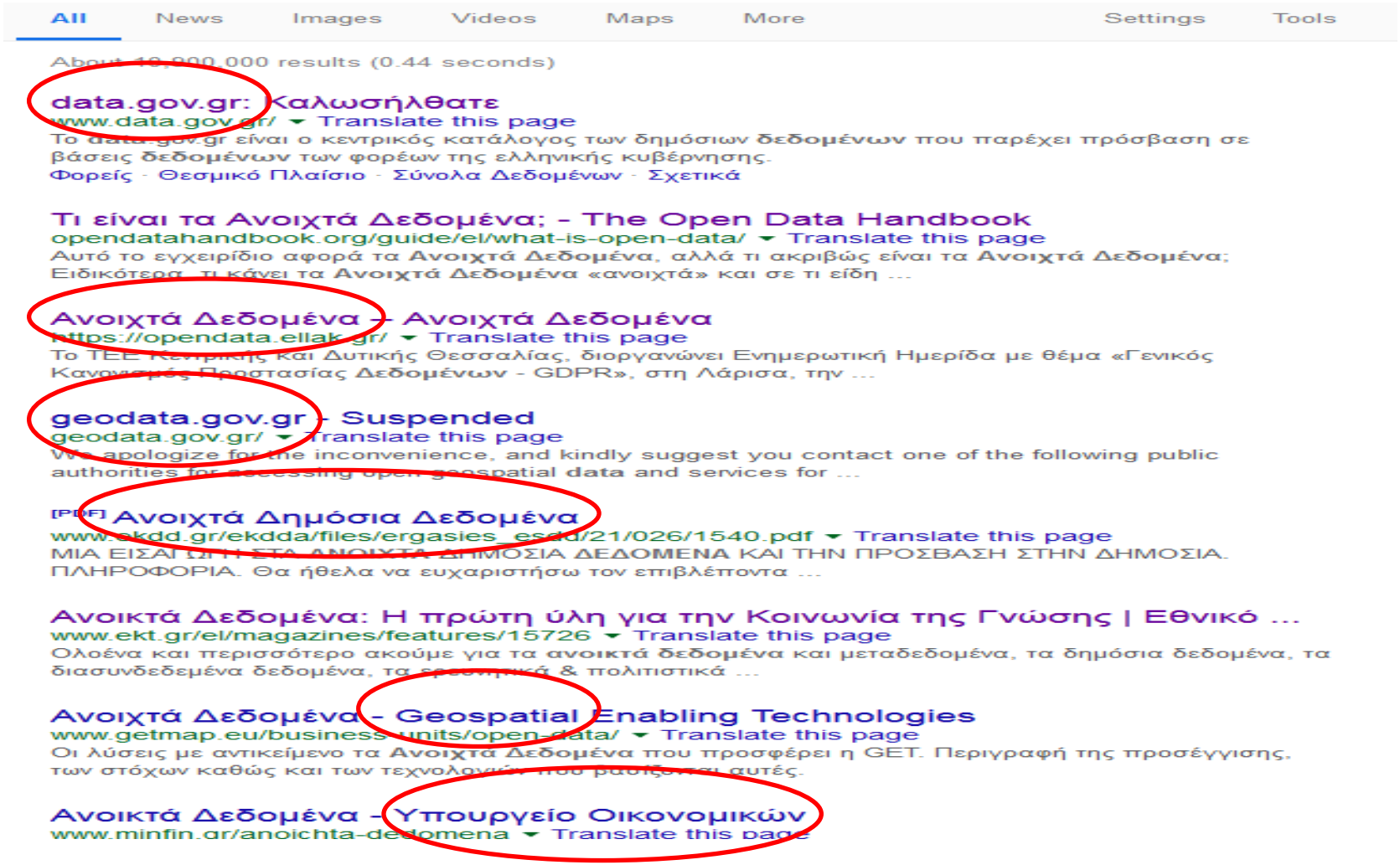
Η σημασία των δεδομένων

- Τα δεδομένα είναι πολύτιμα
  - για τη διατύπωση, τον έλεγχο και την επαλήθευση μιας ερευνητικής υπόθεσης / μεθόδου / πειραματικής διαδικασίας (και από τρίτους)
  - για την αναπαραγωγή ενός πειράματος
  - για την επαναχρησιμοποίησή τους σε άλλα πλαίσια, πιθανώς και άλλων επιστημών
- Τα δεδομένα στα οποία βασίστηκε μια έρευνα είναι εξίσου σημαντικά με τα αποτελέσματα της έρευνας

Μεγάλα δεδομένα

- αυξημένες δυνατότητες υπολογιστών (παραγωγή, επεξεργασία, αποθήκευση δεδομένων)
- αυτοματοποιημένες διαδικασίες
- μειωμένο κόστος αποθήκευσης
- αύξηση παραγωγής δεδομένων 
- **μεγάλα δεδομένα:** σύνολα δεδομένων με τεράστιο όγκο
  - αρκετά petabytes (1015 bytes)

Ανοιχτά δεδομένα



http://www.data.gov.gr/dataset

clarin:el

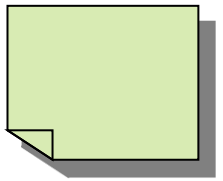
Τύποι δεδομένων 22/6/2018			
XLS (1849)	DOCX (86)	TIFF (31)	XHTML (8)
XLSX (764)	csv, xml, json (71)	ODS (31)	rar, shp, xml (7)
HTML (623)	RAR (70)	URL (25)	pdf, zip (7)
PDF (529)	KML (70)	url, xml (22)	WMS (7)
CSV (437)	application/msword (62)	xml, rar (19)	zip, doc (6)
XML (240)	url (59)	.ods (18)	rtf (6)
.xls (220)	JPEG (53)	.html (18)	RSS (6)
JSON (206)	TXT (44)	rar,xls (16)	KMZ (6)
.xlsx (190)	DOC (44)	msword, msexcel, pdf (15)	RAR,CSV (5)
SHP (110)	html, pdf (41)	dwg (11)	.XLSX (5)
zip, shp (109)	.pdf (39)	WORD (10)	ZIP, JPEG (4)
ZIP (105)	arcgis (35)	PNG (10)	.doc (4)

Γλωσσικά δεδομένα

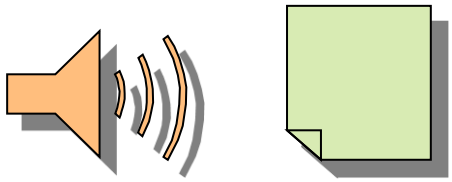
- αρχεία (files) που περιλαμβάνουν δεδομένα σε κάποια φυσική γλώσσα
- τα γλωσσικά δεδομένα είναι απαραίτητα στους μελετητές της γλώσσας
- αλλά και σε όποιους χρησιμοποιούν γλωσσικό υλικό ως βάση για την έρευνα/εργασία τους
- ειδικά στη γλωσσική τεχνολογία:
  - στατιστική επεξεργασία
  - γλωσσολογική επεξεργασία
  - δημιουργία γλωσσικών μοντέλων / γραμματικών
  - εκπαίδευση εργαλείων
  - αυτόματη κατασκευή δομημένων δεδομένων (λεξικών, ορολογικών λιστών…)
  - επαλήθευση υποθέσεων
  - κ.λπ.

Μορφές γλωσσικών δεδομένων

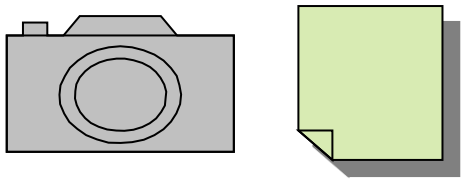
- Μέσο: κείμενο, ήχος, εικόνα, βίντεο



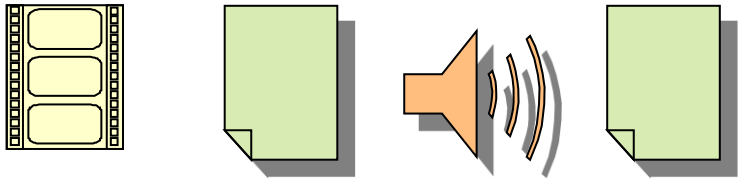
γραπτά γλωσσικά δεδομένα



ηχητικά γλωσσικά δεδομένα + μεταγραφή

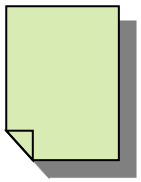


εικόνες + λεζάντες

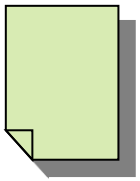


βίντεο (πολυμεσικά δεδομένα) + υπότιτλοι + σενάριο/μεταγραφή

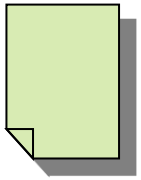
που όλοι περιέχουν γραπτό υλικό



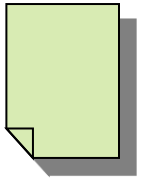
γραπτά γλωσσικά δεδομένα

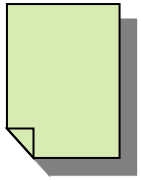


ηχητικά γλωσσικά δεδομένα



εικόνες





βίντεο (πολυμεσικά δεδομένα)

Τι είναι γλωσσικοί πόροι

- σύνολα ψηφιακών γλωσσικών δεδομένων
- συγκεντρωμένων και δομημένων σύμφωνα με ορισμένα κριτήρια
- τεκμηριωμένων σύμφωνα με ορισμένο σχήμα μεταδεδομένων
- που χρησιμοποιούνται
  - στη μελέτη της γλώσσας
  - στη δημιουργία, αξιολόγηση και βελτίωση συστημάτων γλωσσικής τεχνολογίας
  - στις ψηφιακές εκδόσεις
  - στην εκπαίδευση
  - σε κάθε επιστήμη, ως πρωτογενές ή δευτερογενές υλικό

Κατηγορίες γλωσσικών πόρων (1)

- πόροι πρωτογενούς περιεχομένου
  - πόροι ψηφιακού / ψηφιοποιημένου λόγου: γραπτά κείμενα (ψηφιοποιημένα βιβλία, κείμενα διαδικτύου, εφημερίδες, κτλ.), ηχογραφήσεις προφορικού λόγου (συνεντεύξεις, ραδιοφωνικές εκπομπές κτλ.),
  - βιντεοσκοπήσεις (τηλεοπτικές εκπομπές, συλλογές με εκφράσεις προσώπου, χειρονομίες κτλ.)
  - εικόνες (ψηφιακές ή ψηφιοποιημένες φωτογραφίες με τις λεζάντες τους)
- πόροι επεξεργασμένου υλικού
  - διαφόρων ειδών επισημειώσεις σε κείμενα, ήχο και πολυμεσικά δεδομένα (μορφοσυντακτικά επισημειωμένα κείμενα, μεταγραφές ηχητικών αρχείων, επισημειωμένα αρχεία βίντεο κτλ.)
- πόροι αναφοράς
  - διάφοροι τύποι δομημένης γλωσσικής γνώσης (λίστες λέξεων, λεξικά, θησαυροί κτλ.) → για οργάνωση, επεξεργασία και μελέτη πόρων πρωτογενούς περιεχομένου

Κατηγορίες γλωσσικών πόρων (2)

clarin:el

- εργαλεία/εφαρμογές γλωσσικής τεχνολογίας
  - εργαλεία και ολοκληρωμένες εφαρμογές γλωσσικής επεξεργασίας
    - στοίχιση πολύγλωσσων κειμένων,
    - μορφολογική επισημείωση,
    - λημματοποίηση,
    - συντακτική ανάλυση,
    - εξόρυξη γνώσης κτλ.
- εργαλεία παρουσίασης/προβολής δεδομένων
  - προβολής κειμένων,
  - συλλογών πολυμεσικών δεδομένων,
  - αποτελεσμάτων επεξεργασίας κτλ.

Είδη γλωσσικών πόρων

- σώματα κειμένων
  - γραπτά και προφορικά σώματα κειμένων, πολυμεσικά και πολυτροπικά σώματα
- λεξικοί / εννοιολογικοί πόροι
  - λεξικά, λίστες λέξεων, σημασιολογικά λεξικά, ορολογικά γλωσσάρια, οντολογίες κτλ
- γλωσσικές περιγραφές
  - γραμματικές, τυπολογικές βάσεις δεδομένων, γλωσσικά μοντέλα κτλ
- εργαλεία / υπηρεσίες
  - εργαλεία, εφαρμογές και διαδικτυακές υπηρεσίες επεξεργασίας δεδομένων, π.χ. λημματοποιητής, επισημειωτής, εργαλείο αυτόματης μετάφρασης κτλ

τι ΔΕΝ είναι γλωσσικός πόρος

- μια **τυχαία** συλλογή ψηφιακών κειμένων
- μια **ατεκμηρίωτη** συλλογή ψηφιακών κειμένων
- μια συλλογή **λίγων** ψηφιακών κειμένων
- μια συλλογή κειμένων σε μορφή **εικόνας**
- ένας **διαδικτυακός τόπος**

Μεταδεδομένα τεκμηρίωσης

- μεταδεδομένα: δεδομένα που χρησιμοποιούνται για να περιγράψουν δεδομένα
- αποτελούν δομημένη πληροφορία που αποτυπώνει την τεκμηρίωση των δεδομένων
- γνωστά παραδείγματα μεταδεδομένων:
  - δημιουργός
  - έτος δημιουργίας
  - περιγραφή
  - θεματική ταξινόμηση …
- χρησιμεύουν στον εντοπισμό και τη διαχείριση των δεδομένων (από ανθρώπους ή μηχανές)

Γλωσσική τεχνολογία

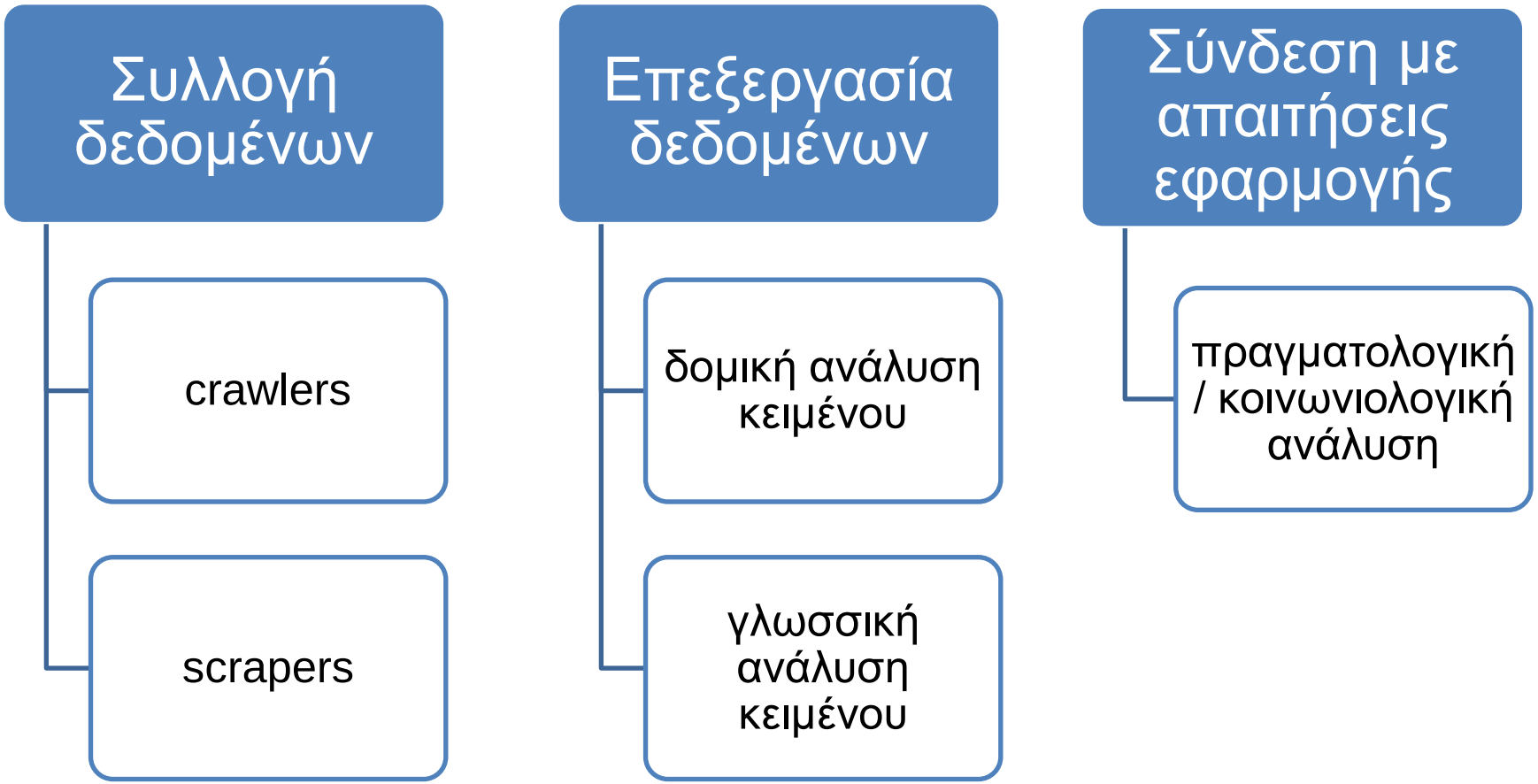
clarin:el

- καθημερινά χρησιμοποιούμε υπηρεσίες γλωσσικής τεχνολογίας
  - για να επεξεργαστούμε περιεχόμενο
  - να επικοινωνήσουμε με ανθρώπους και μηχανές
- παραδείγματα
  - ορθογραφικοί και συντακτικοί διορθωτές
  - αναζήτηση / φωνητική αναζήτηση στο google
  - «έξυπνο» πληκτρολόγιο στα κινητά τηλέφωνα και βοηθοί smartphones

Εφαρμογές ΓΤ

- αυτόματη ανάλυση και επεξεργασία λόγου
  - μορφολογική, συντακτική και σημασιολογική ανάλυση κειμένου
- εξόρυξη πληροφορίας από κείμενα, π.χ.
  - αναγνώριση ονοματικών οντοτήτων (ονόματα ανθρώπων, τοπωνύμια, ονόματα οργανισμών κ.λπ.),
  - αναγνώριση γεγονότων και των σχέσεων που τα συνδέουν
- μηχανική μετάφραση
- αυτόματη περίληψη
- αυτόματη εξαγωγή ορολογίας
- αυτόματη κατηγοριοποίηση κειμένων
- εξόρυξη γνώμης / ανάλυση συναισθήματος
- σύνθεση / αναγνώριση φωνής
- εκπαιδευτική τεχνολογία…

Διαδικασίες και εργαλεία



Εργαλεία γλωσσικής τεχνολογίας

clarin:el

από το αδόμητο στο δομημένο περιεχόμενο

- Tokenizer
  - χωρίζει τις λέξεις του κειμένου και αναγνωρίζει λέξεις, ημερομηνίες, αριθμούς, κλπ.
- Sentence segmenter
  - χωρίζει τις προτάσεις του κειμένου
- Part-of-Speech Tagger
  - αναγνωρίζει και επισημειώνει το μέρος του λόγου κάθε λέξης
- Chunker
  - αναγνωρίζει και επισημειώνει τοπικές δομές της πρότασης (ονοματικές / ρηματικές / προθετικές / επιρρηματικές φράσεις)
- GrNE tagger
  - αναγνωρίζει και επισημειώνει τις ονοματικές οντότητες του κειμένου
- Dependency parser
  - αναγνωρίζει και επισημειώνει τη συντακτική δομή κάθε πρότασης του κειμένου

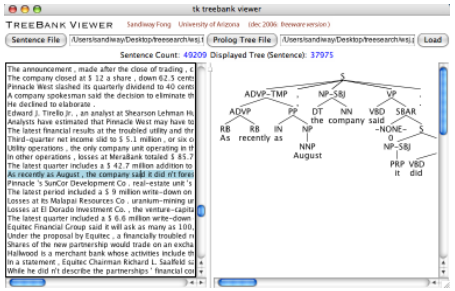
Άλλοι τύποι εργαλείων

- επισημειωτές άλλου τύπου
  - για χρήση από ανθρώπους
  - με ετικέτες επισημείωσης ελεύθερα ορισμένες από τον χρήστη
  - μη-γλωσσική επισημείωση
- στατιστικά εργαλεία
- εργαλεία οπτικοποίησης
- κ.ά.

Υποδομές γλωσσικών πόρων

Το πρόβλημα

clarin:el



Απαιτήσεις

- τι χρειάζεται;
  - γλωσσικά δεδομένα
    - σώματα κειμένων (κείμενα των πολιτικών αλλά και σώμα κειμένων γενικής χρήσης)
    - λεξικά / θεματικά λεξιλόγια / σημασιολογικά λεξικά
  - εργαλεία
    - επεξεργασίας και επισημείωσης
    - ανάλυσης αποτελεσμάτων
- πού θα τα βρει;
  - πρωτότυπες πηγές
  - αρχειακές συλλογές
  - άλλα πανεπιστήμια & ερευνητικά ιδρύματα
  - συρτάρια συναδέλφων

σε τι κατάσταση;

- μόνο κείμενο – ιδανικά, ήδη σε ψηφιακή επεξεργάσιμη μορφή
- επισημειωμένα

επιπλέον απαιτήσεις

- τεχνικές γνώσεις
- απαιτήσεις σε υπολογιστική δύναμη
- νομικά προβλήματα

Τι χρειάζεται;

- χρειαζόμαστε μηχανισμούς
  - επιστημονικούς
  - τεχνικούς
  - νομικούς
  - οργανωτικούς και
  - κοινωνικούς
- που θα επιτρέψουν την πρόσβαση, τον διαμοιρασμό και τον επαναπροσδιορισμό της χρήσης των πόρων
- ώστε να επιτευχθεί πρόοδος και να διευκολυνθεί η ανάπτυξη χρήσιμων εφαρμογών

αυτό το κενό σκοπεύει να καλύψει…

σε ευρωπαϊκό επίπεδο η Ερευνητική Υποδομή CLARIN και στην Ελλάδα το clarin:el