ανοιχτά γλωσσικά δεδομένα:
η υποδομή γλωσσικών πόρων και υπηρεσιών clarin:el

Σαράντος Καπιδάκης[1], Στέλιος Πιπερίδης[2], Πένυ Λαμπροπούλου[2], Μαρία Γαβριηλίδου[2]

([1]Ιόνιο Πανεπιστήμιο, [2]Ε.Κ. Αθηνά / ΙΕΛ)



Με τη χρηματοδότηση της Πράξης CLARIN-Αττική, Υποστήριξη και ανάπτυξη ελληνικών εταίρων για συμμετοχή στις Κοινοπραξίες των Ευρωπαϊκών Ερευνητικών Υποδομών στα πλαίσια του ESFRI/2006

το πρόβλημα

η πρόκληση

- Οι ψηφιακές τεχνολογίες φέρνουν κοντά μας πολλούς πόρους και σχετικούς τρόπους επεξεργασίας τους
- Όμως δεν ξέρουμε
  - Τι υπάρχει – και πόσο ποιοτικό είναι
  - Πώς το εντοπίζουμε
  - Με τι όρους διατίθεται
  - Αν συνδυάζεται με άλλους πόρους - συμβατότητα
  - Τι τεχνολογίες διατίθενται για πιθανή επεξεργασία τους

ανάγκες

- Να εντοπίζουμε και χρησιμοποιήσουμε τους ποικίλους γλωσσικούς πόρους που υπάρχουν για διδακτικούς, δημιουργικούς, ερευνητικούς, ... σκοπούς
- Να αξιοποιήσουμε αλλά και να επεξεργαστούμε και συνδυάσουμε γλωσσικούς πόρους, παράγοντας καινοτόμα αποτελέσματα
- Να εμπλουτίσουμε / ενοποιήσουμε τους γλωσσικούς πόρους με άλλους δικούς μας, για χρήση από εμάς ή άλλους
- Να προβάλουμε τον πλούτο της γλώσσας μας.

η λύση: CLARIN

στόχοι

- Το ευρωπαϊκό 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 όλων των επιστημών, και κυρίως των Κοινωνικών και Ανθρωπιστικών Επιστημών (ΚΑΕ)

όραμα

Ένας ερευνητής από το γραφείο του στην Κέρκυρα θα μπορεί:

- με μία εγγραφή (single sign-on, με το ήδη υπάρχον ακαδημαϊκό login) με πιστοποίηση (authentication)
- 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

δηλαδή…

- το CLARIN σκοπεύει να ενσωματώσει
  - **Γλωσσικούς Πόρους**: ψηφιακό περιεχόμενο κάθε είδους (κείμενο, ήχο, εικόνα, βίντεο), όπως
    - συλλογές κειμενικού και πολυμεσικού υλικού, πρωτογενούς και επισημειωμένου (π.χ. κείμενα με μορφοσυντακτική ανάλυση, συλλογές ταινιών με μεταδεδομένα για τις κινήσεις των συμμετεχόντων)
    - λεξικά, θησαυρούς, οντολογίες, ορολογικά γλωσσάρια,
    - υπολογιστικές γραμματικές κτλ.
  - **Εργαλεία Γλωσσικής Τεχνολογίας**: εργαλεία αναγνώρισης φωνής, λημματοποιητές, εργαλεία εξαγωγής περίληψης κτλ.
- σε ένα συστηματικά οργανωμένο **δίκτυο αποθετηρίων** (όπου συγκεντρώνονται και περιγράφονται οι πόροι)
- το οποίο θα είναι **διαθέσιμο σε ερευνητές** όλων των επιστημών

πώς;

- μέσα από **εθνικά υπο-δίκτυα οργανισμών** που μεριμνούν για την έρευνα και την ψηφιακή προσαρμογή και ετοιμότητα της γλώσσας (ή των γλωσσών) της εκάστοτε χώρας
- σήμερα στο CLARIN συμμετέχουν περισσότερα από 200 μέλη από 33 χώρες

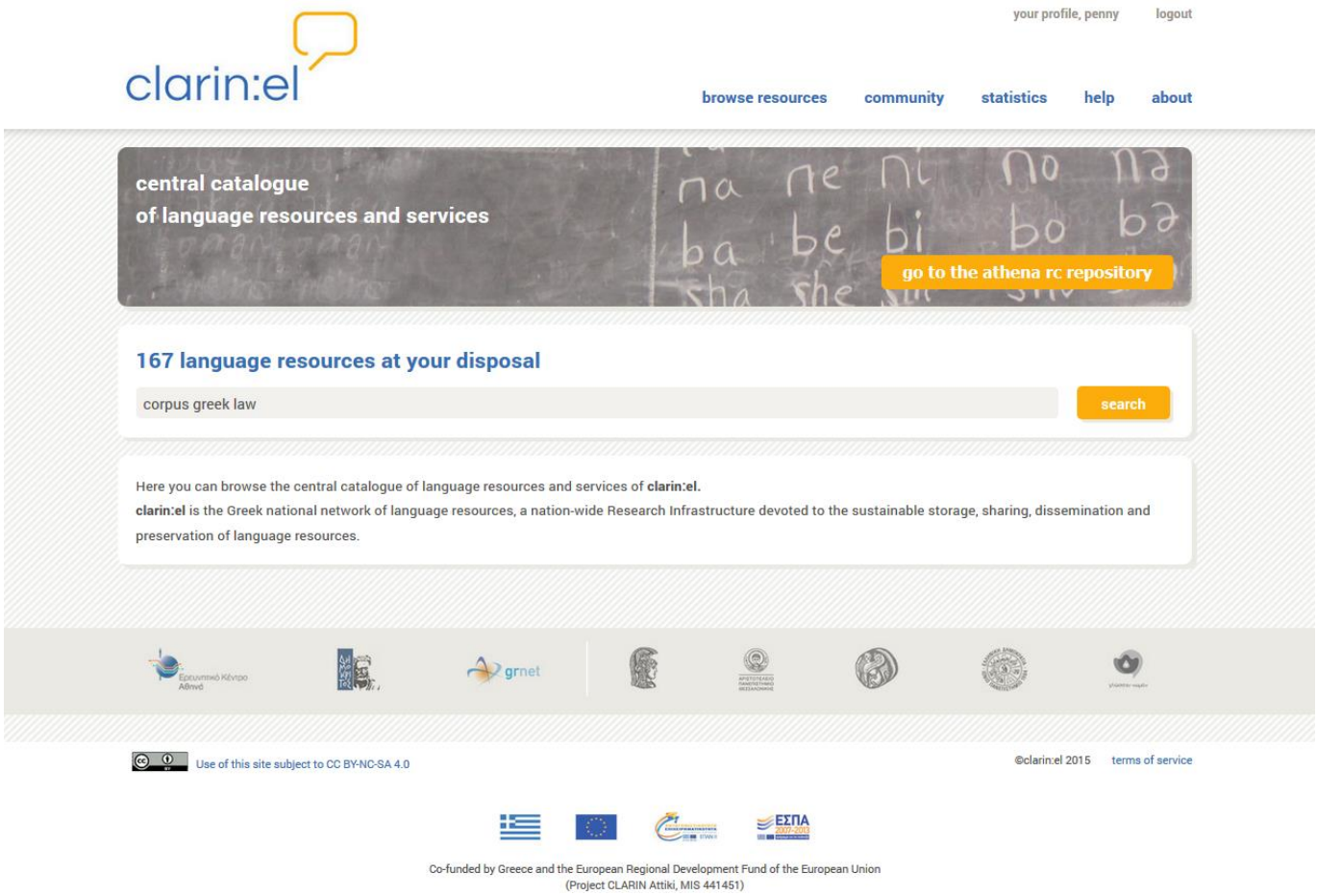
η ελληνική υποδομή clarin:el

το clarin στην Ελλάδα

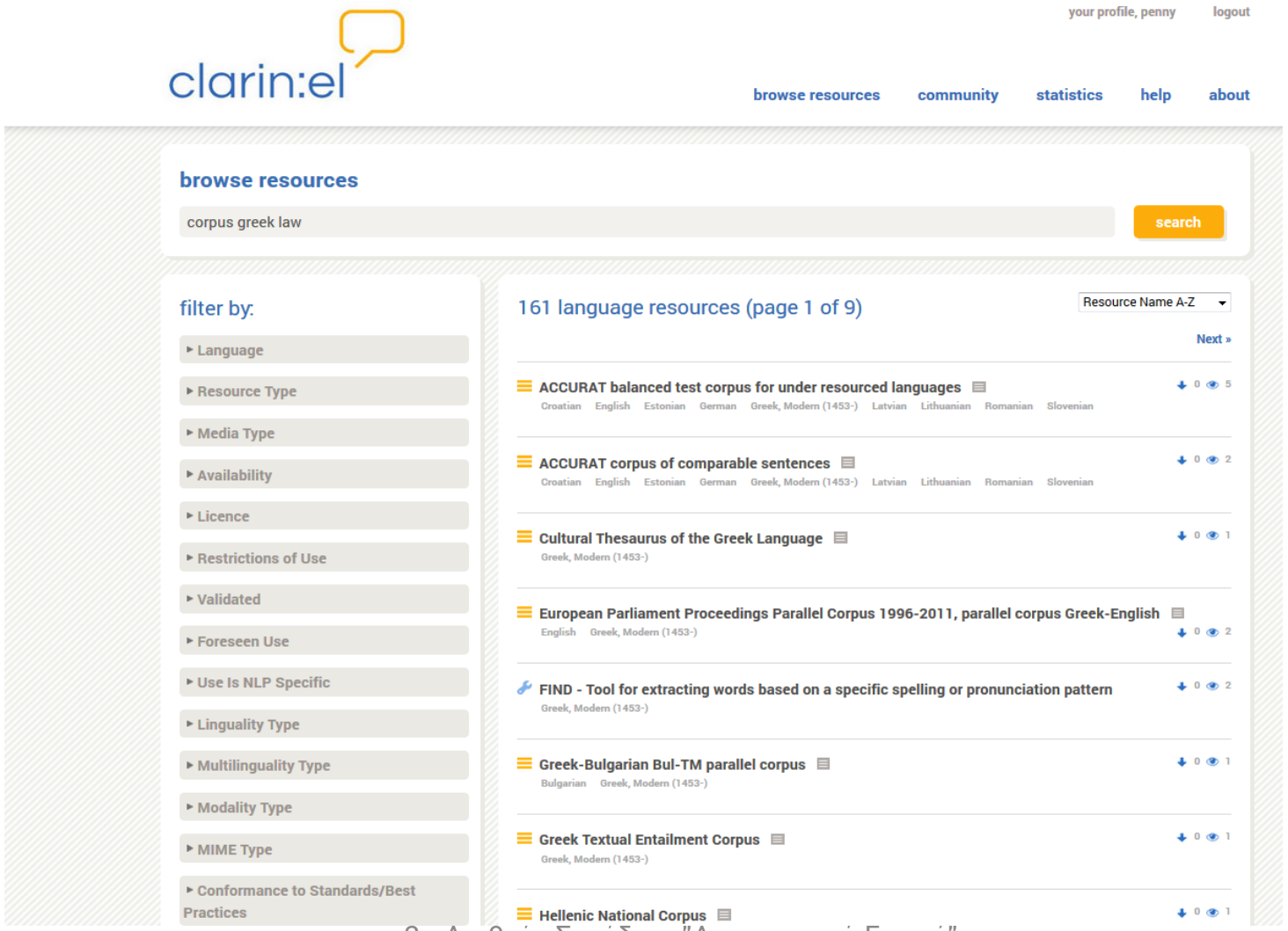
- διανύει την κατασκευαστική φάση (2012 – 2015)
- έγινε μέλος του CLARIN ERIC τον Φεβρουάριο του 2015
- διάρθρωση ελληνικής υποδομής: δύο στενά διασυνδεδεμένα υποσυστήματα:
  - **τεκμηρίωσης**, αποθήκευσης, **διαμοιρασμού** + αναζήτησης, ανάκτησης, **καταφόρτωσης** γλωσσικών πόρων (resources infrastructure)
  - **επεξεργασίας** γλωσσικών δεδομένων μέσω διαδικτυακών υπηρεσιών γλωσσικής επεξεργασίας και παραγωγή νέων δεδομένων (processing infrastructure)
- αρχικά, φιλοδοξεί να (προσ)φέρει τους πόρους που χρειάζεται ένας ερευνητής για την έρευνά του
  - σε 5 βήματα
  - και με ένα επιπρόσθετο 6ο βήμα, όταν αυτό είναι δυνατό...

χρήση της υποδομής από καταναλωτές πόρων

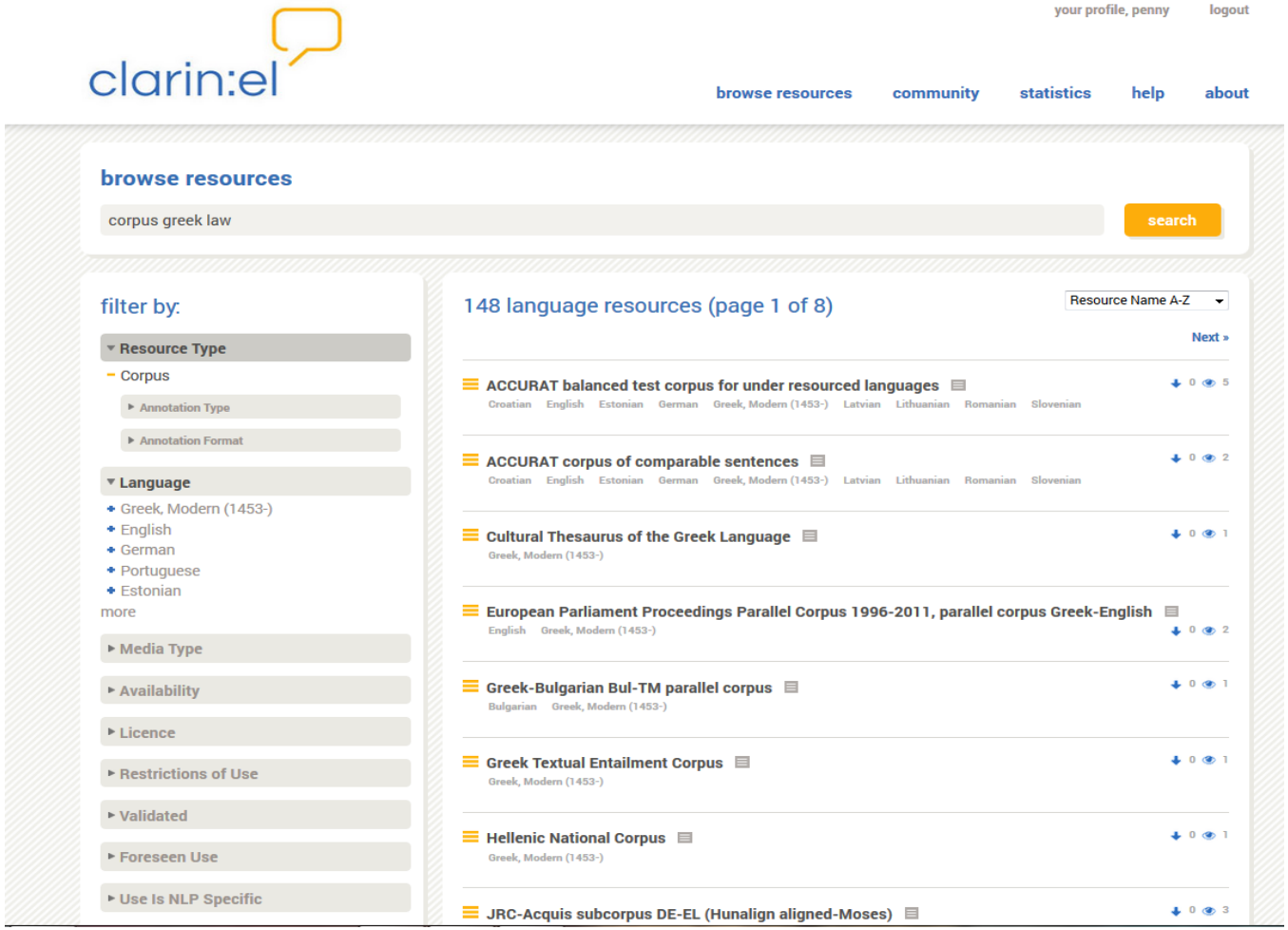
1. αναζήτηση με λέξεις κλειδιά, π.χ. ελληνικό σώμα κειμένων στη θεματική περιοχή "νομικά"



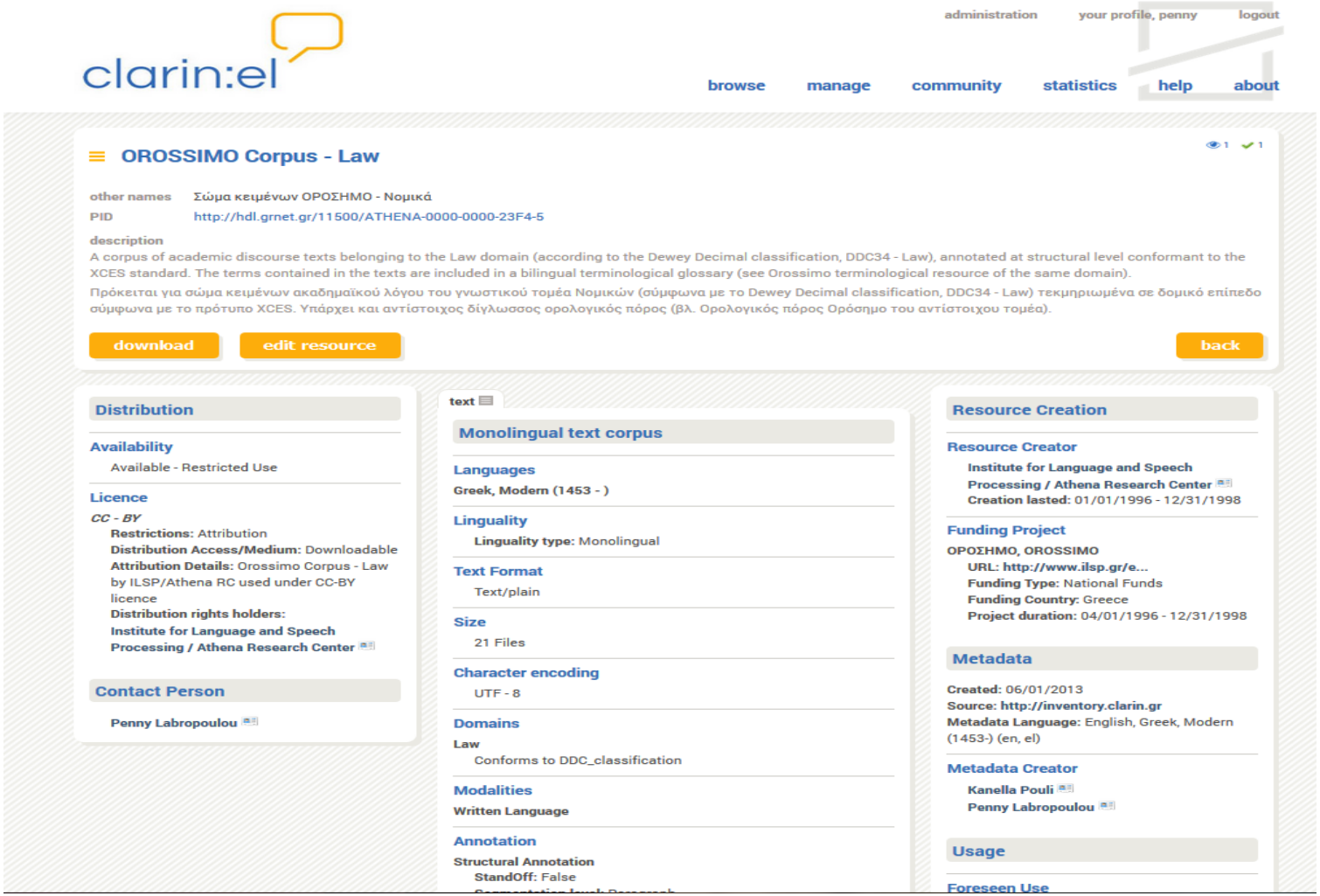
2. φυλλομέτρηση αποτελεσμάτων



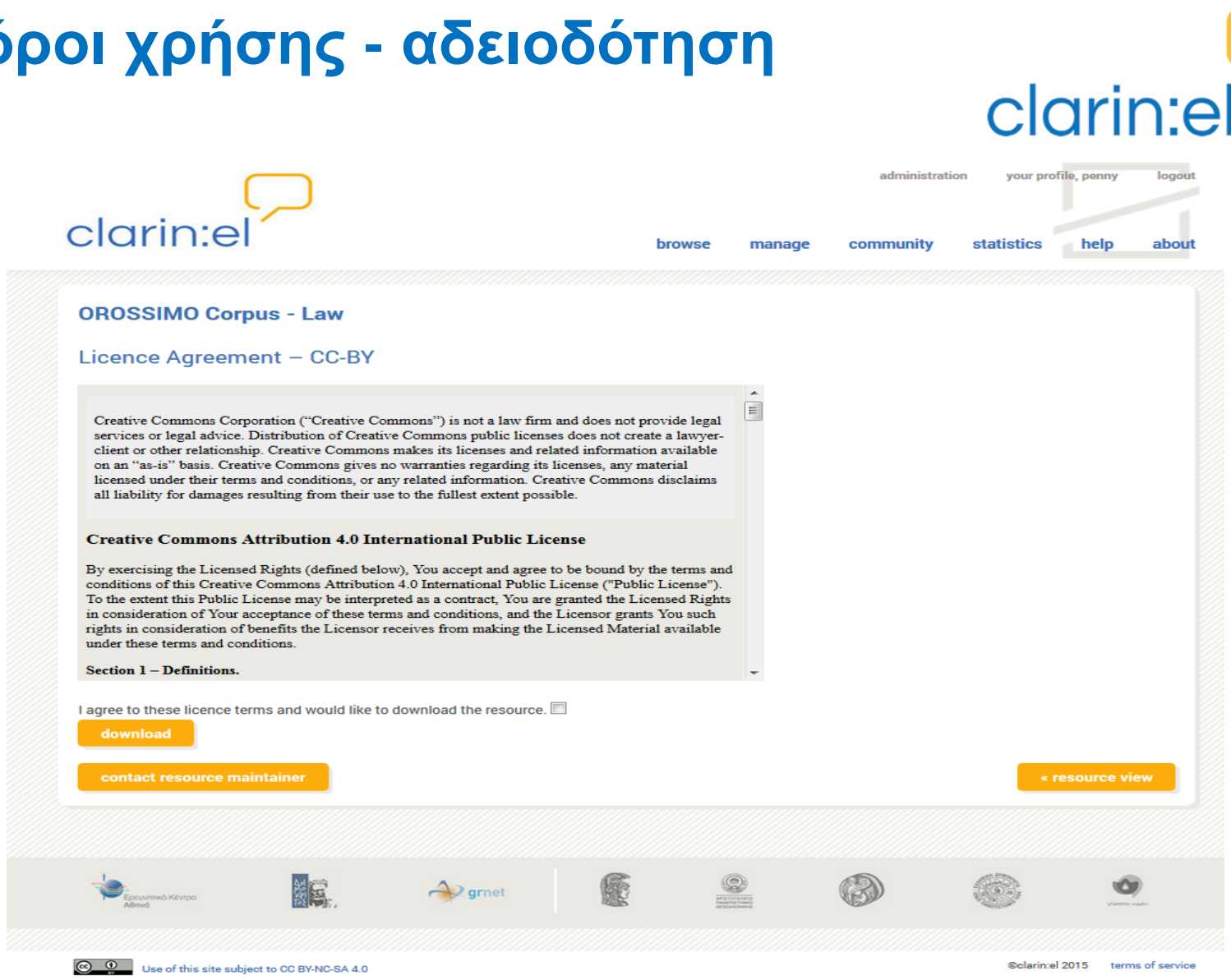
2α. αναζήτηση με χρήση φίλτρων



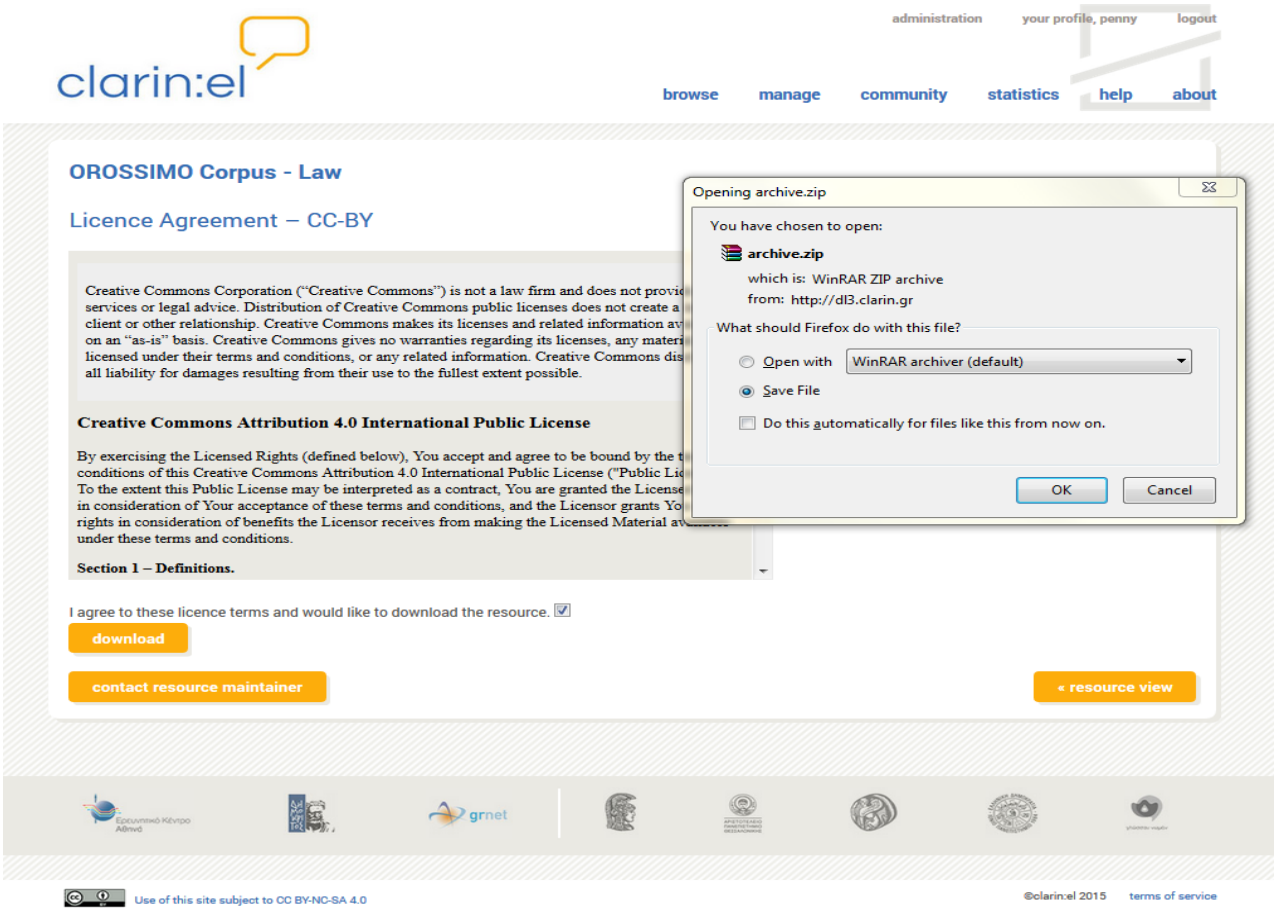
3. θέαση επιλεγμένου πόρου



4. όροι χρήσης - αδειοδότηση

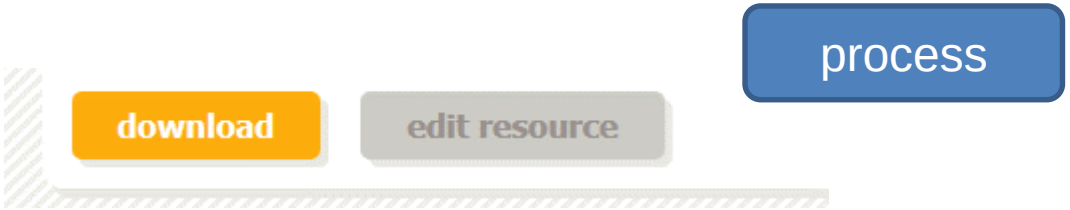


5. καταφόρτωση του πόρου



6. γλωσσική επεξεργασία πόρων

clarin:el

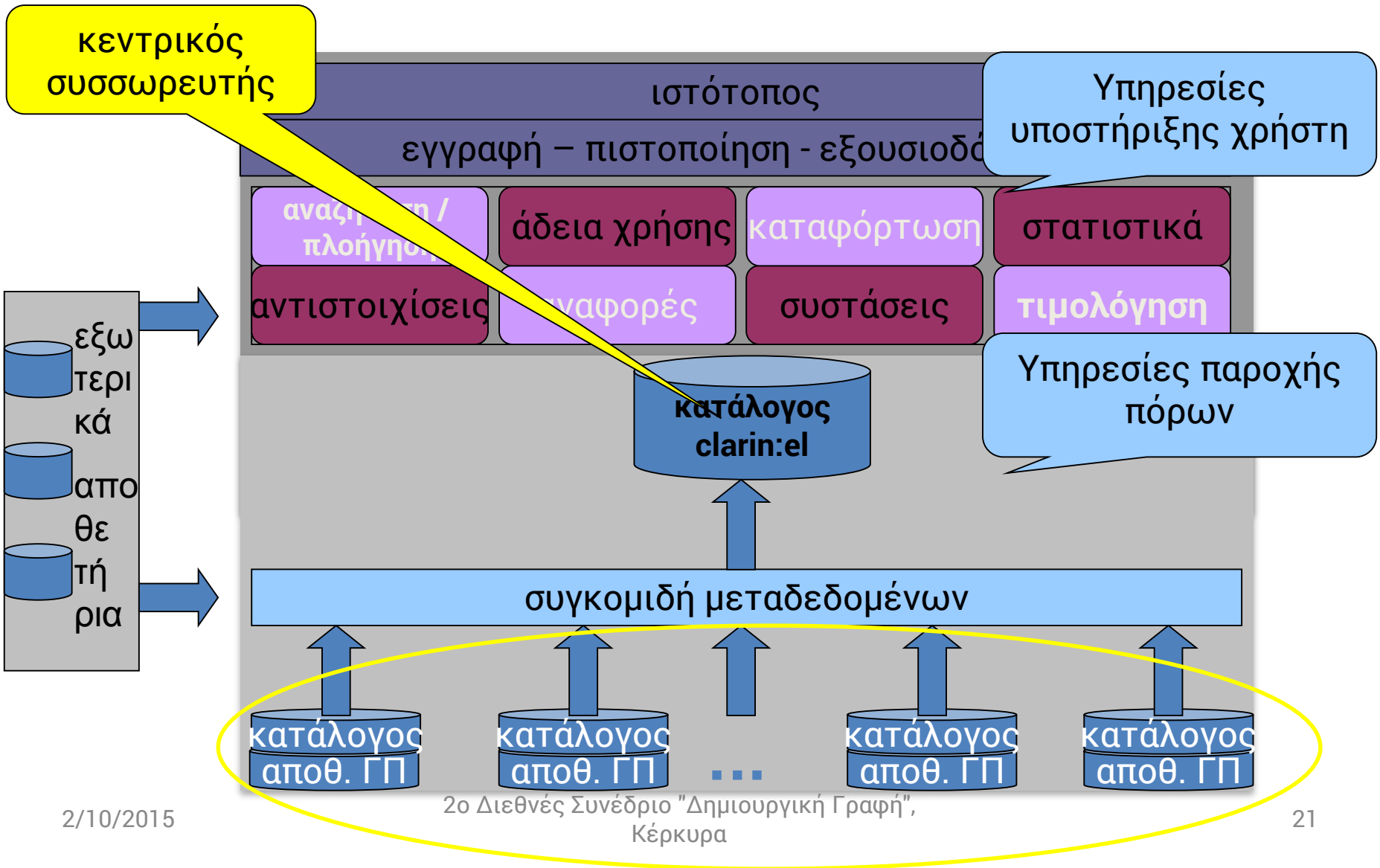


- υπηρεσίες επεξεργασίας μονογλωσσικών σωμάτων κειμένων
  - αναγνώριση λέξεων και προτάσεων (tokenisation and sentence splitting)
  - μορφοσυντακτική ανάλυση (part of speech tagging)
  - λημματοποίηση (lemmatisation)
  - συντακτική ανάλυση (syntactic parsing)
  - αναγνώριση και εξαγωγή ορολογίας (term recognition and extraction)
  - αναγνώριση οντοτήτων (named entity recognition)
- υπηρεσίες επεξεργασίας πολυγλωσσικών σωμάτων κειμένων
  - για Ελληνικά - Αγγλικά
    - οι παραπάνω για κάθε γλώσσα του παράλληλου ΣΚ
  - για Ελληνικά - Χ (=γλώσσα της ΕΕ)
    - στοίχιση σε επίπεδο πρότασης, …

οι θεμέλιοι λίθοι της υποδομής

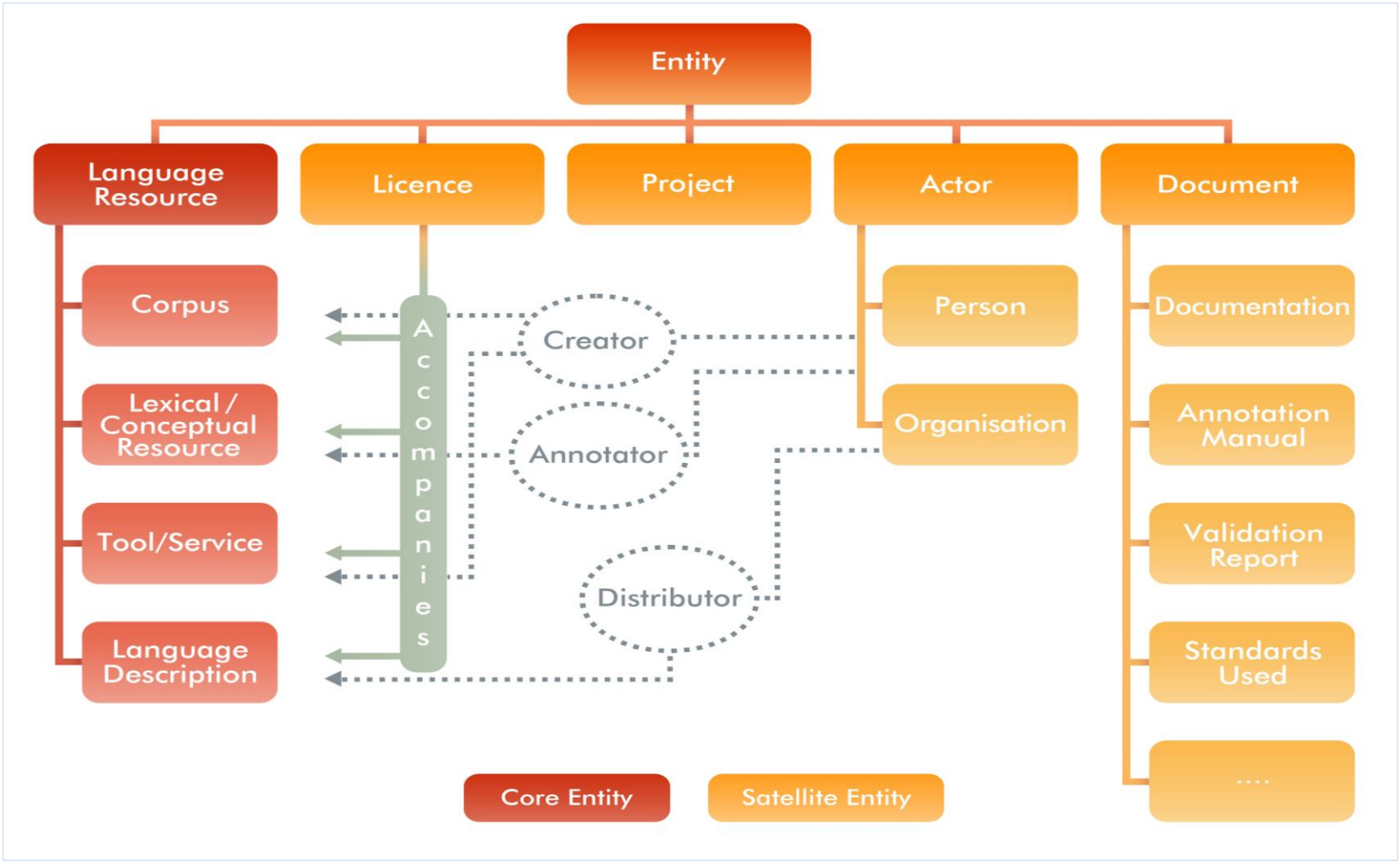
αρχιτεκτονική υποδομής & σχήμα τεκμηρίωσης

αρχιτεκτονική clarin:el



οντολογία – οντότητες περιγραφής

clarin:el

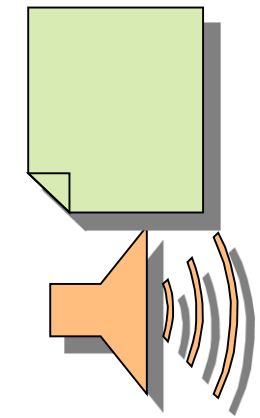


τυπολογία πόρων (1)

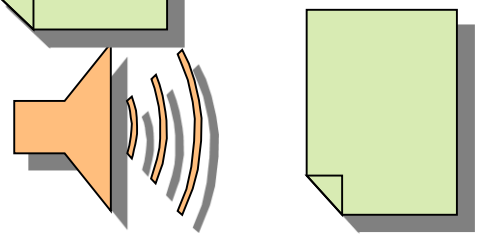
- ταξινόμηση βάσει 2 κύριων κριτηρίων: "τύπος πόρου" (resource type) & "μέσο πόρου" (media type)
- τύπος πόρου
  - ***σώμα κειμένων*** (γραπτού / προφορικού λόγου, πολυτροπικών/πολυμεσικών κειμένων)
  - ***λεξικός / εννοιολογικός πόρος*** (π.χ. λεξικό, ορολογικός πόρος, κατάλογος λέξεων, οντολογία κτλ.)
  - ***γλωσσική περιγραφή*** (π.χ. γλωσσικό μοντέλο, υπολογιστική γραμματική, κτλ.)
  - ***εργαλείο / τεχνολογία*** (π.χ. λημματοποιητής, επισημειωτής, εργαλείο αυτόματης μετάφρασης κτλ.)

τυπολογία πόρων (2)

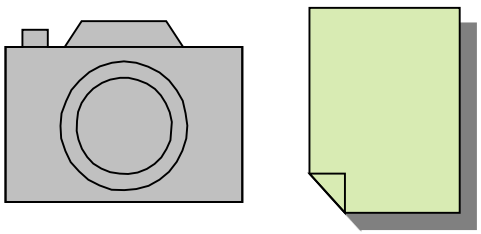
- τύπος μέσου: κείμενο, ήχος, εικόνα, βίντεο



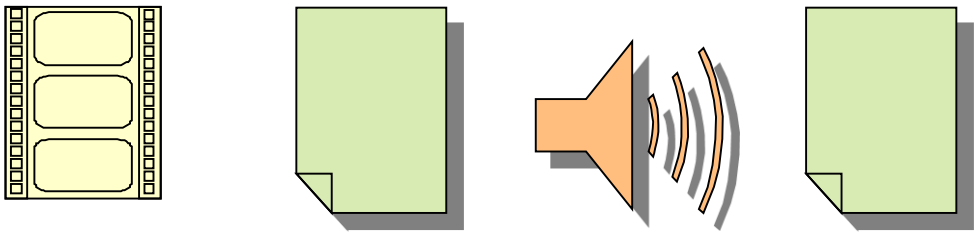
σώματα γραπτού λόγου



σώματα προφορικού λόγου



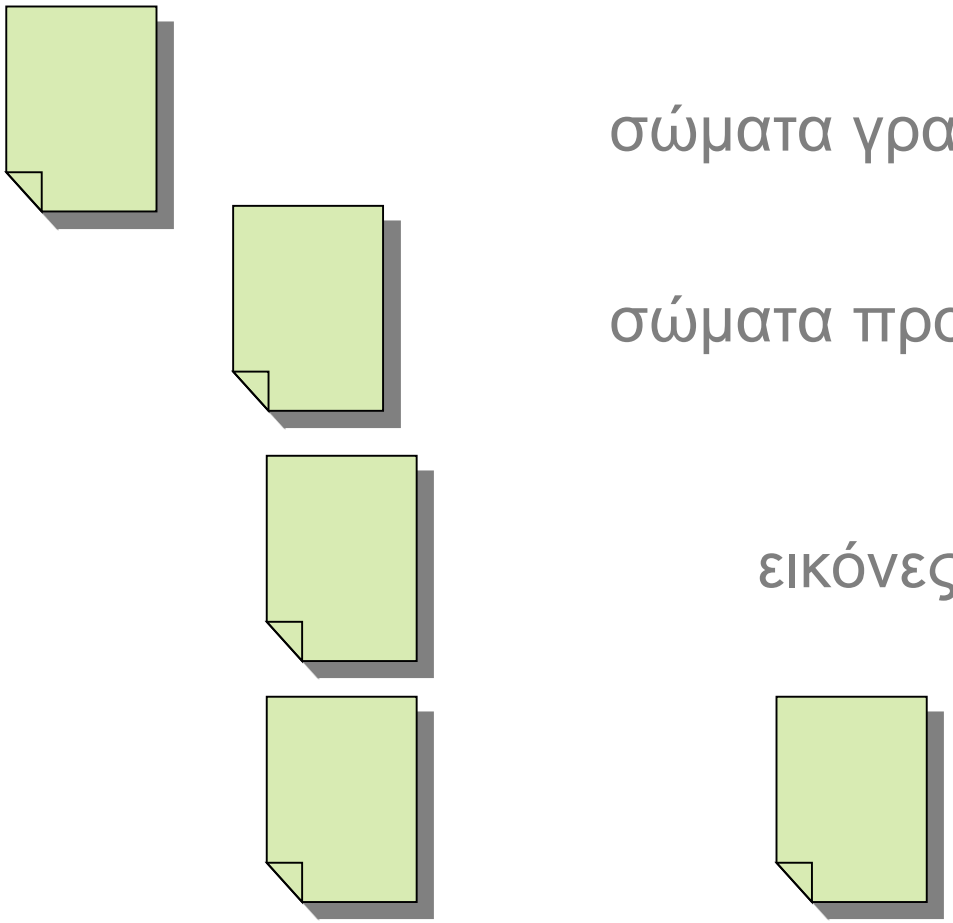
εικόνες



βίντεο

τυπολογία πόρων (3)

αναζήτηση *"κειμένου"*



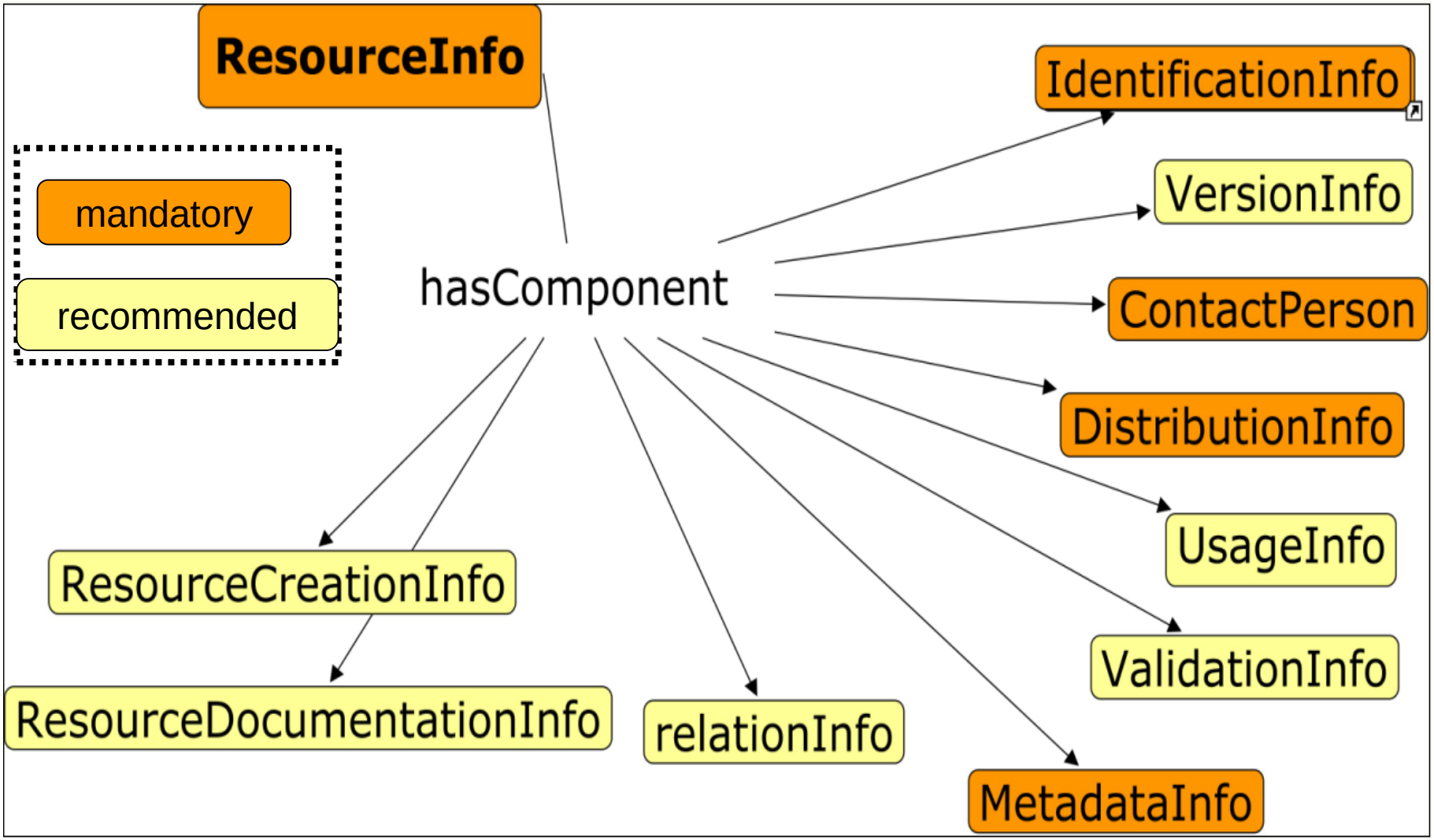
σώματα γραπτού λόγου

σώματα προφορικού λόγου

εικόνες

βίντεο

σχήμα τεκμηρίωσης: υποχρεωτικά και προαιρετικά συστατικά πληροφορίας



δίκτυο clarin:el

χρήστες υποδομής

- πίσω από την υποδομή, το **περιεχόμενο**!
- χρήστες της υποδομής: **καταναλωτές** και **πάροχοι**
- γιατί να διαθέσει κανείς τους πόρους του;
- διαμοιρασμός, εμπλουτισμός, αξιοποίηση, πολλαπλασιασμός

δίκτυο clarin:el

- αποκλειστικά πάροχοι Γλωσσικών Πόρων
- μέλη του δικτύου μπορούν να γίνουν
  - φορείς-μέλη (με ή χωρίς ιδρυματικά αποθετήρια)
  - συνεργαζόμενοι ερευνητές (στο Αποθετήριο Φιλοξενούμενων Πόρων)
- ποιοι είναι ήδη μέλη - Κατασκευαστική Φάση (μέχρι 31/12/2015)

- Πιλοτική Φάση (από 1/1/2016)
  - Όλοι οι υπόλοιποι ακαδημαϊκοί φορείς (πανεπιστήμια, ερευν. κέντρα)

περισσότερες πληροφορίες

- www.clarin.gr
- http://inventory.clarin.gr

info@clarin.gr

clarin.gr

@CLARIN_el

https://www.linkedin.com/grp/home?gid=8309819

Ευχαριστώ πολύ!