CLARIN-EL
Common Language Resources and Technology Infrastructure

Η Ευρωπαϊκή Υποδομή CLARIN και η Εθνική Υποδομή CLARIN EL

Στέλιος Πιπερίδης,
Πένυ Λαμπροπούλου, Γιούλη Μπακαγιάννη

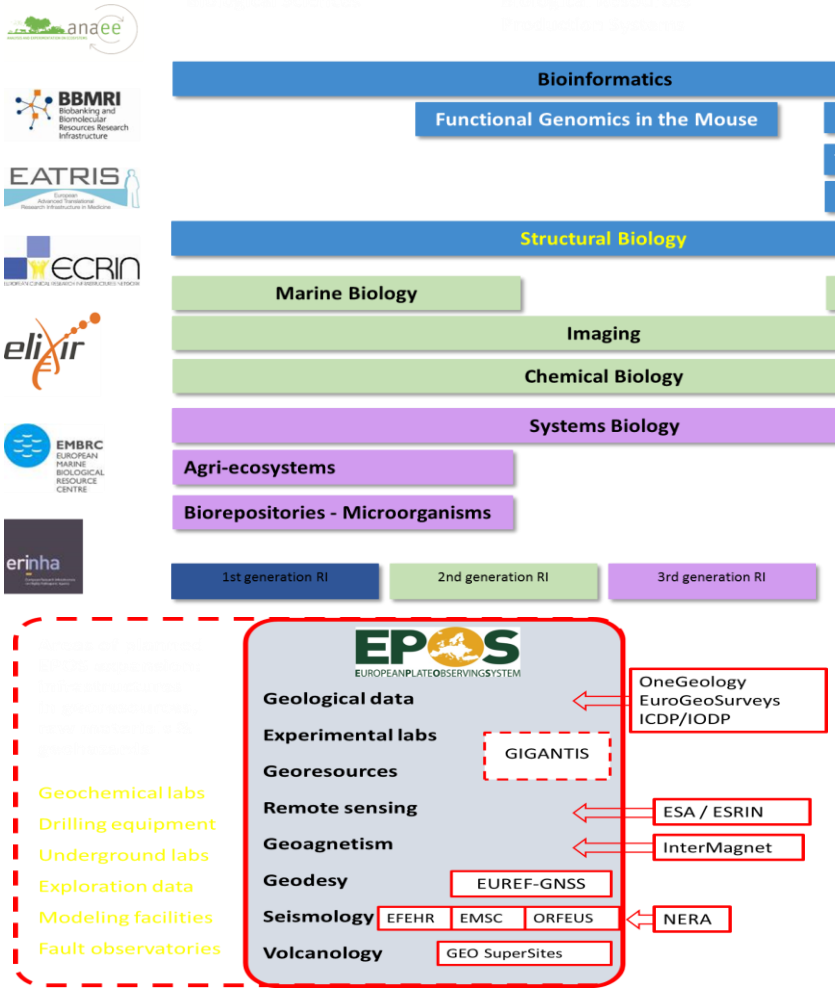
Ινστιτούτο Επεξεργασίας Λόγου
ΕΚ «Αθηνά»

Ο ρόλος των (επιστημονικών) δεδομένων

Η δύναμη των δεδομένων

- Τα (επιστημονικά) δεδομένα μεταμορφώνουν ήδη και βελτιώνουν με ταχύτατους ρυθμούς τη ζωή μας
  - Βιολογικά, γεωγραφικά, μετεωρολογικά, οικονομικά, κ.ά. δεδομένα
- Η εμπειρία από πολλά επιστημονικά πεδία δείχνει ότι τα δεδομένα και οι υπηρεσίες επεξεργασίας τους αποκτούν ιδιαίτερη αξία όταν διαμοιράζονται και ανοίγονται
  - τόσο για ερευνητικούς και τεχνολογικούς σκοπούς, συμπεριλαμβανόμενης της αξιολόγησης αυτών
  - όσο και για την δημιουργία καινοτομικών εφαρμογών
- Η πρόσβαση στα αποτελέσματα του Human Genome Project «γέννησε» ~ 3 δις ευρώ σε επενδύσεις E&A, ~ 500 δις ευρώ σε οικονομική δραστηριότητα
- Η συγκέντρωση γενετικών δεδομένων ασθενών με Alzheimer's οδήγησε στην ανακάλυψη 5 νέων γονιδίων σχετιζόμενων με την θεραπεία της νόσου
- "Τα επιστημονικά δεδομένα είναι πολύτιμα για να είναι κλειδωμένα στο συρτάρι"

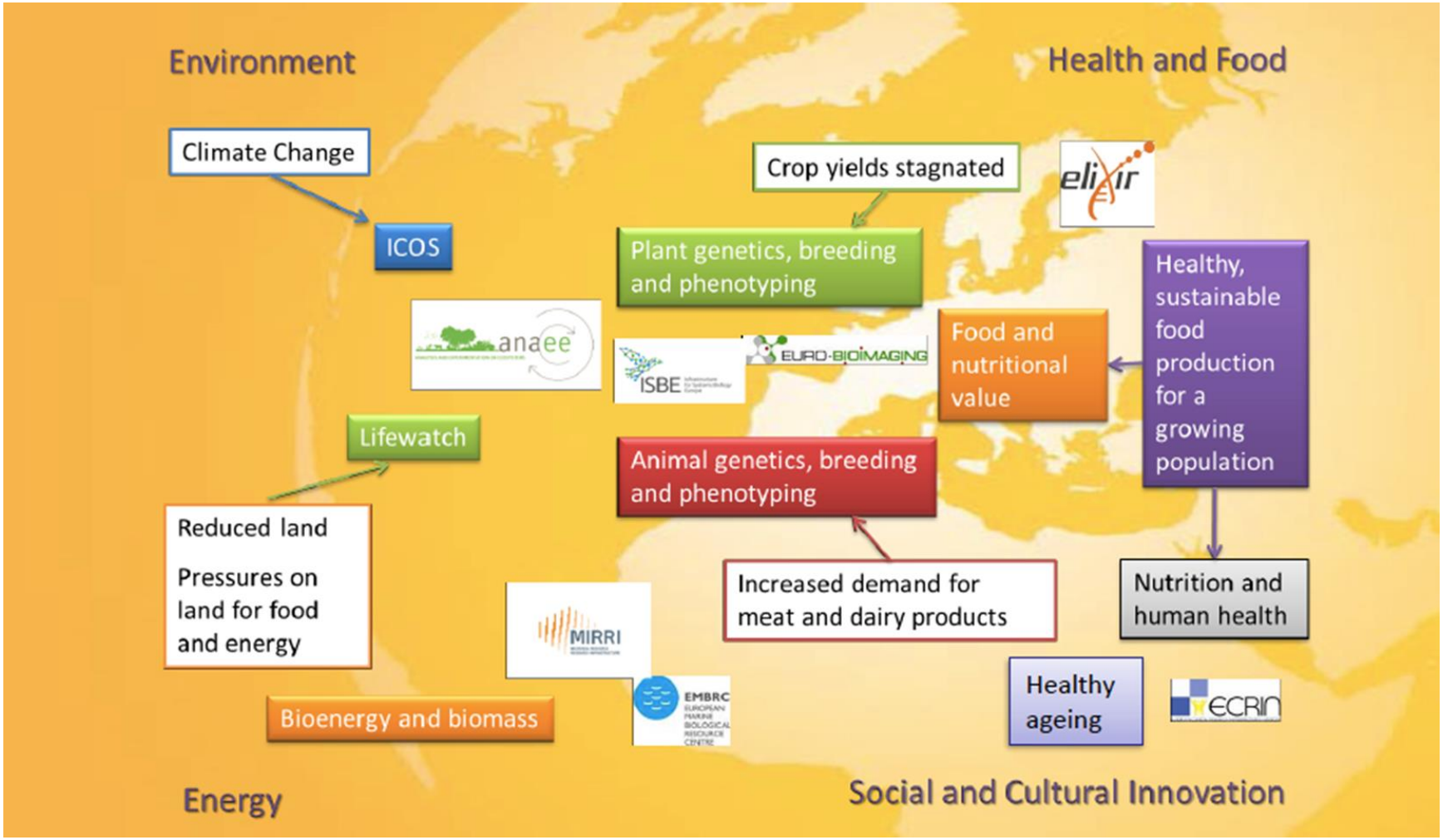
Άνοιγμα και διασύνδεση δεδομένων



στις επιστήμες ζωής και υγείας

στις επιστήμες της γης

Διασύνδεση δεδομένων διαφόρων πεδίων



Γλωσσικά δεδομένα και υπηρεσίες

Γλωσσικά δεδομένα και υπηρεσίες

- Τι γίνεται όμως με τη γλώσσα και τα γλωσσικά δεδομένα;
- Με τη γλωσσική τεχνολογία που χωρίς να το συνειδητοποιούμε την χρησιμοποιούμε καθημερινά για να επεξεργαστούμε περιεχόμενο, να αλληλεπιδράσουμε με ανθρώπους και μηχανές;
- Πολύ συχνά στις ηλεκτρονικές λίστες βλέπουμε ερωτήματα/αιτήματα για δεδομένα/υπηρεσίες
  - Σώμα ελληνικών κειμένων περιόδου Χ, με επισημείωση σε επίπεδο Ψ
  - Σώμα παράλληλων κειμένων μεταξύ Ελληνικών και γλώσσας Χ,
  - Στοιχισμένο και …
  - Εργαλεία μορφοσυντακτικής ανάλυσης ή αναγνώρισης γλώσσας, κ.λπ.
  - Εργαλεία επεξεργασίας video, κ.λπ.

Εξεύρεση Γλωσσικών Πόρων; Διαθεσιμότητα;

- Σύμφωνα με παλαιότερες αλλά και πρόσφατες μελέτες μόνο ένα μικρό τμήμα των γλωσσικών πόρων (LRs) είναι γνωστό/ ανακοινώνεται/ γίνεται προσβάσιμο/ ανταλλάσσεται/ ...
- ... παρόλο που η συλλογή δεδομένων και η επιμέλειά τους (καθαρισμός, φροντίδα και συντήρηση) καθώς και η επισημείωσή τους είναι δαπανηρές διαδικασίες
- Για να επιτευχθεί πρόοδος, να διευκολυνθεί η ανάπτυξη χρήσιμων εφαρμογών, χρειαζόμαστε όλους αυτούς τους επιστημονικούς, τεχνικούς, νομικούς, οργανωτικούς, κοινωνικούς μηχανισμούς που θα επιτρέψουν την πρόσβαση, ανακύκλωση και επαναπροσδιορισμό της χρήσης των απαραίτητων πόρων.

Το πρόβλημα

- πολλές αρχειακές συλλογές, πόροι, εργαλεία και υπηρεσίες ΓΤ είναι γνωστά μόνο σε ορισμένες ερευνητικές κοινότητες
- συλλογές, πόροι, εργαλεία και υπηρεσίες είναι ασύνδετα μεταξύ τους και αυτό κάνει την αναζήτηση δύσκολη
- η εύρεση πόρων, υπηρεσιών, κλπ εξαρτάται από τη γλώσσα τεκμηρίωσης
- ακολουθούν διαφορετικά πρότυπα τεκμηρίωσης
- συχνά γλωσσικά δεδομένα και υπηρεσίες/εργαλεία είναι ασύμβατα μεταξύ τους
- ο ερευνητής (πάροχος ή χρήστης) αντιμετωπίζει ένα νομικό κυκεώνα σχετικά με τα δικαιώματα και τις υποχρεώσεις του
- δεν υπάρχουν κίνητρα για διαμοιρασμό και διάθεση πόρων

Η Ερευνητική Υποδομή CLARIN

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 όλων των επιστημών και κυρίως των Κοινωνικών και Ανθρωπιστικών Επιστημών (ΚΑΕ)

Η αποστολή του CLARIN

- **Τι;**
- να δημιουργήσει μια υποδομή που να διαθέσει Γλωσσικούς Πόρους και Τεχνολογίες στους ερευνητές όλων των επιστημών

- **Πώς;**
- ενοποιώντας 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, λημματοποιητές, συντακτικούς αναλυτές, εργαλεία αυτόματης εξαγωγής περίληψης, εργαλεία εξαγωγής πληροφορίας κτλ.
- σε ένα συστηματικά οργανωμένο **δίκτυο αποθετηρίων** το οποίο θα είναι διαθέσιμο σε ερευνητές όλων των επιστημών
- που διαθέτει **εθνικά υπο-δίκτυα** που μεριμνούν για την έρευνα και την ψηφιακή προσαρμογή και ετοιμότητα των διαφόρων γλωσσών

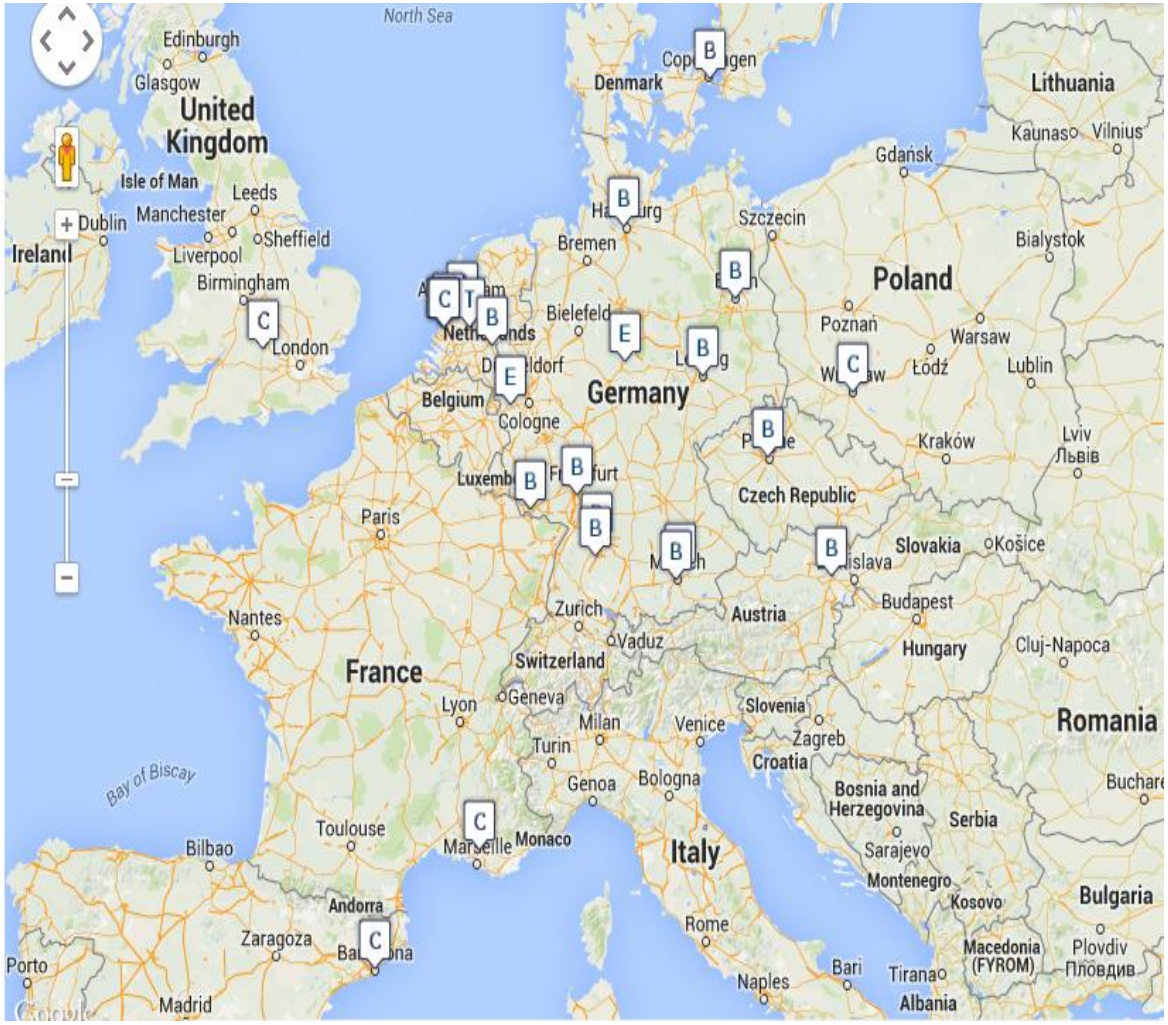
Γιατί μια ευρωπαϊκή υποδομή;

- 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

Το όραμα του CLARIN

- Ένας ερευνητής από το γραφείο του στην Κέρκυρα θα μπορεί:
- με μία εγγραφή (single sign-on) 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

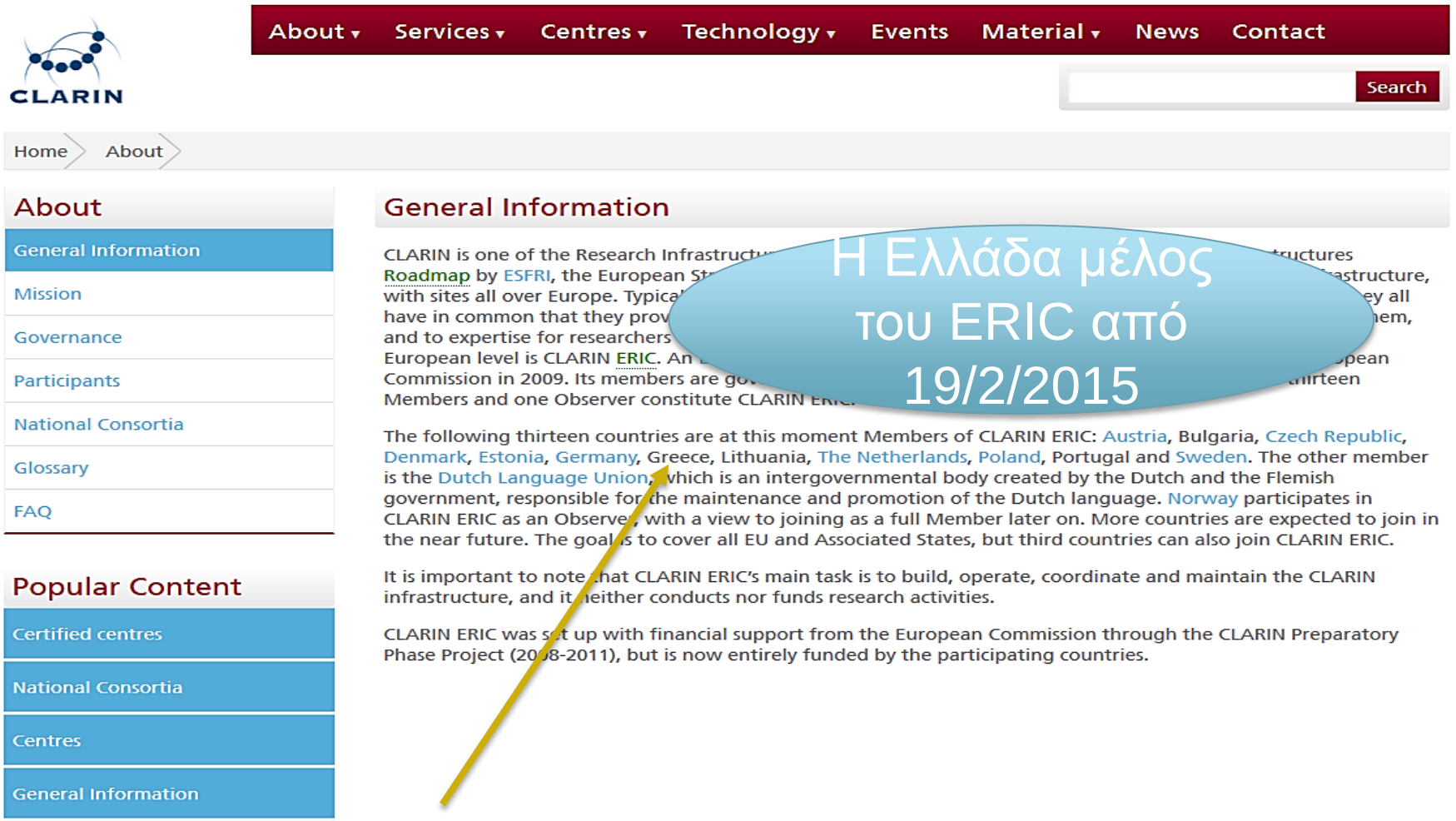
Πιστοποιημένα κέντρα CLARIN



14 κράτη

28 πιστο-ποιημένα κέντρα

CLARIN ERIC


CLARIN
About
Services
Centres
Technology
Events
Material
News
Contact
Search
Home
About
About
General Information
Mission
Governance
Participants
National Consortia
Glossary
FAQ
Popular Content
Certified centres
National Consortia
Centres
General Information
Material
General Information
Η Ελλάδα μέλος του ERIC από 19/2/2015
The following thirteen countries are at this moment Members of CLARIN ERIC: Austria, Bulgaria, Czech Republic, Denmark, Estonia, Germany, Greece, Lithuania, The Netherlands, Poland, Portugal and Sweden. The other member is the Dutch Language Union, which is an intergovernmental body created by the Dutch and the Flemish government, responsible for the maintenance and promotion of the Dutch language. Norway participates in CLARIN ERIC as an Observer, with a view to joining as a full Member later on. More countries are expected to join in the near future. The goal is to cover all EU and Associated States, but third countries can also join CLARIN ERIC.
It is important to note that CLARIN ERIC's main task is to build, operate, coordinate and maintain the CLARIN infrastructure, and it neither conducts nor funds research activities.
CLARIN ERIC was set up with financial support from the European Commission through the CLARIN Preparatory Phase Project (2008-2011), but is now entirely funded by the participating countries.

Η Ελληνική Υποδομή CLARIN

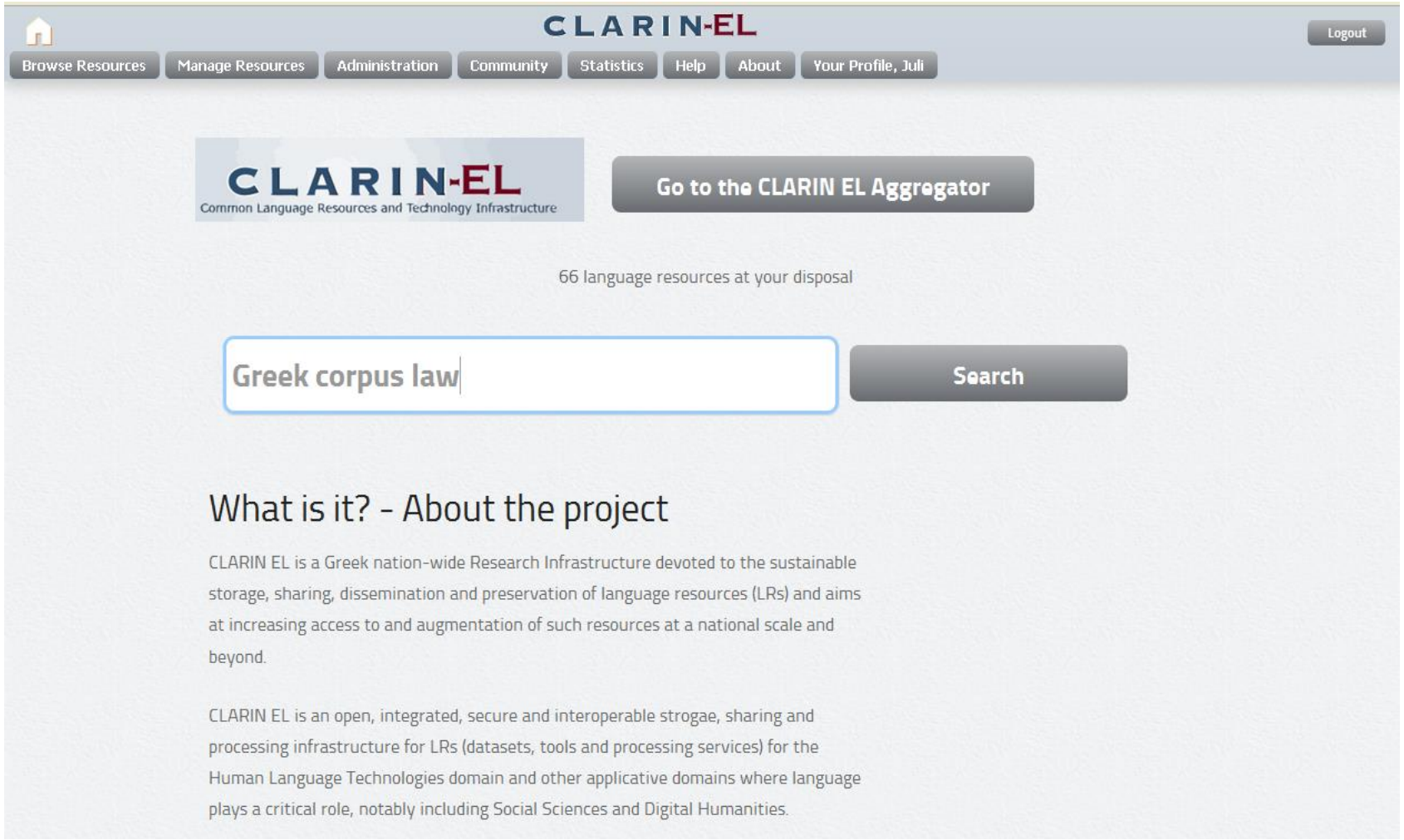
Ο στόχος μας

- να ταιριάξουμε τις ανάγκες και τις προσδοκίες των παρόχων γλωσσικών πόρων και των καταναλωτών/χρηστών βελτιώνοντας το κατά πόσο ένας πόρος είναι ορατός (visibility), τεκμηριωμένος (documentation), αναγνωρίσιμος (identification), διαθέσιμος και καλά συντηρημένος (preservation), είτε πρόκειται για γλωσσικά δεδομένα είτε για (βασικά) γλωσσικά εργαλεία.
- να ταιριάξουμε δεδομένα και εργαλεία/υπηρεσίες και να βελτιώσουμε την δυνατότητα εύκολης χρήσης υπηρεσιών γλωσσικής επεξεργασίας με σκοπό την οργανική ανάπτυξη της υποδομής
- Είναι ένα μακροπρόθεσμο πολυδιάστατο εγχείρημα βάσει του οποίου οι γλωσσικοί πόροι θα συνεισφέρουν στην προαγωγή της έρευνας, της τεχνολογίας και της καινοτομίας καθώς θα είναι διαθέσιμοι σε ευρεία κλίμακα, συγκεντρωμένοι και ανοικτά προσβάσιμοι για κοινή χρήση

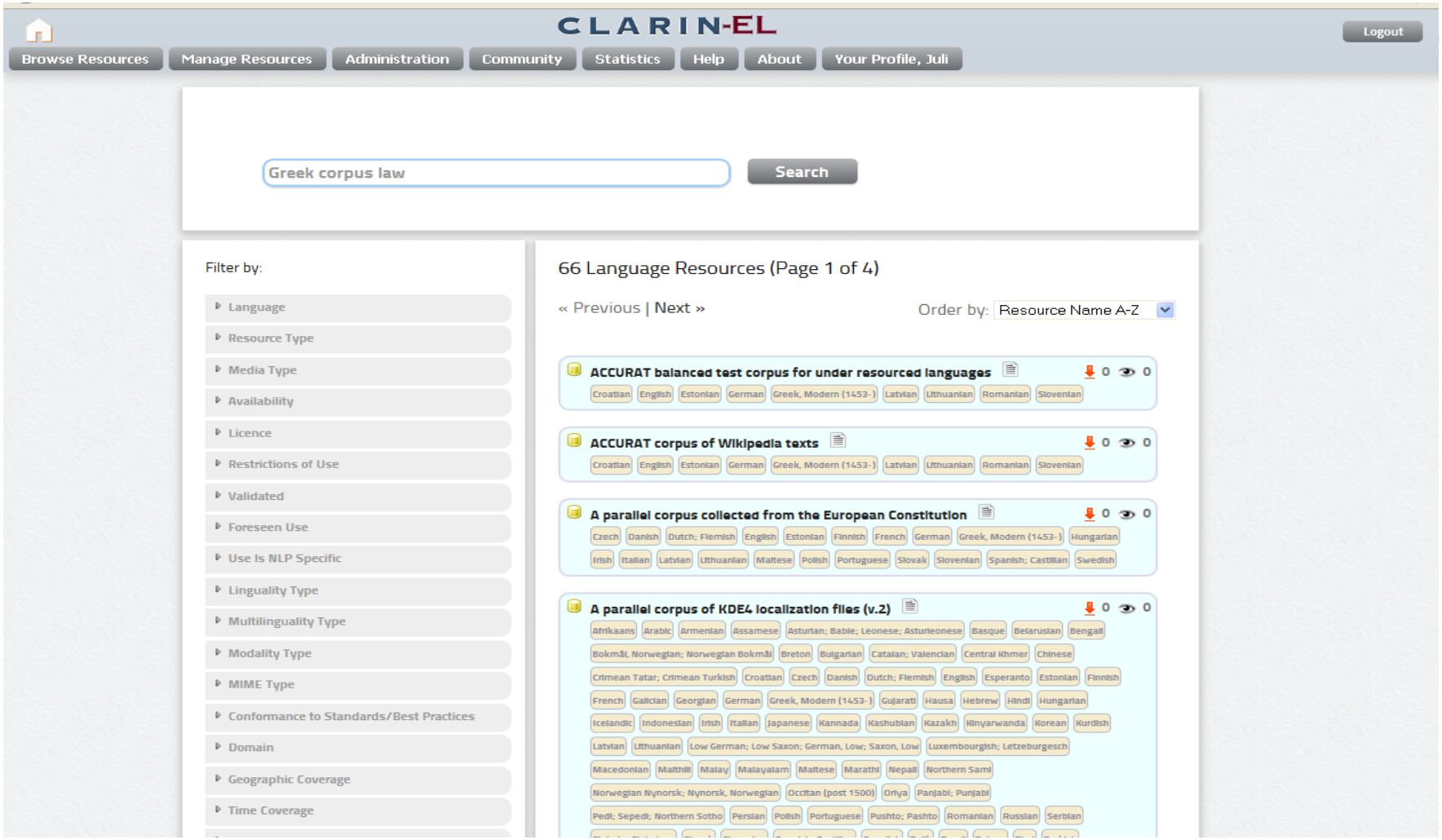
Διάρθρωση του στόχου

- 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)
- Σαν πρώτο βήμα, φιλοδοξεί να (προσ)φέρει τους πόρους που χρειάζεται ένας ερευνητής για την έρευνά του
- σε 5 βήματα, κατά το δυνατόν…

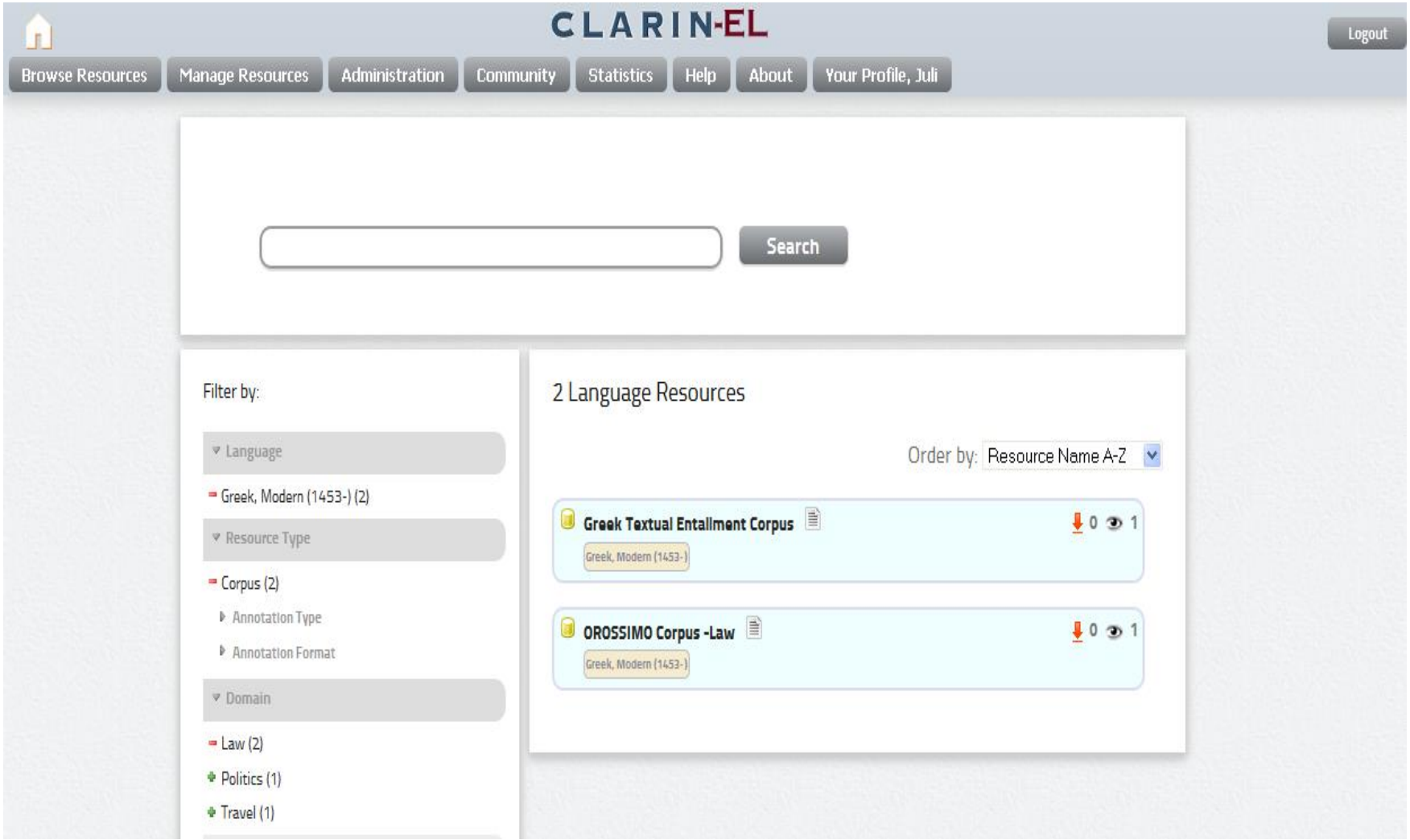
1. Αναζήτηση με λέξεις κλειδιά, π.χ. Ελληνικό σώμα κειμένων σχετικό με τη θεματική περιοχή «νομικά»



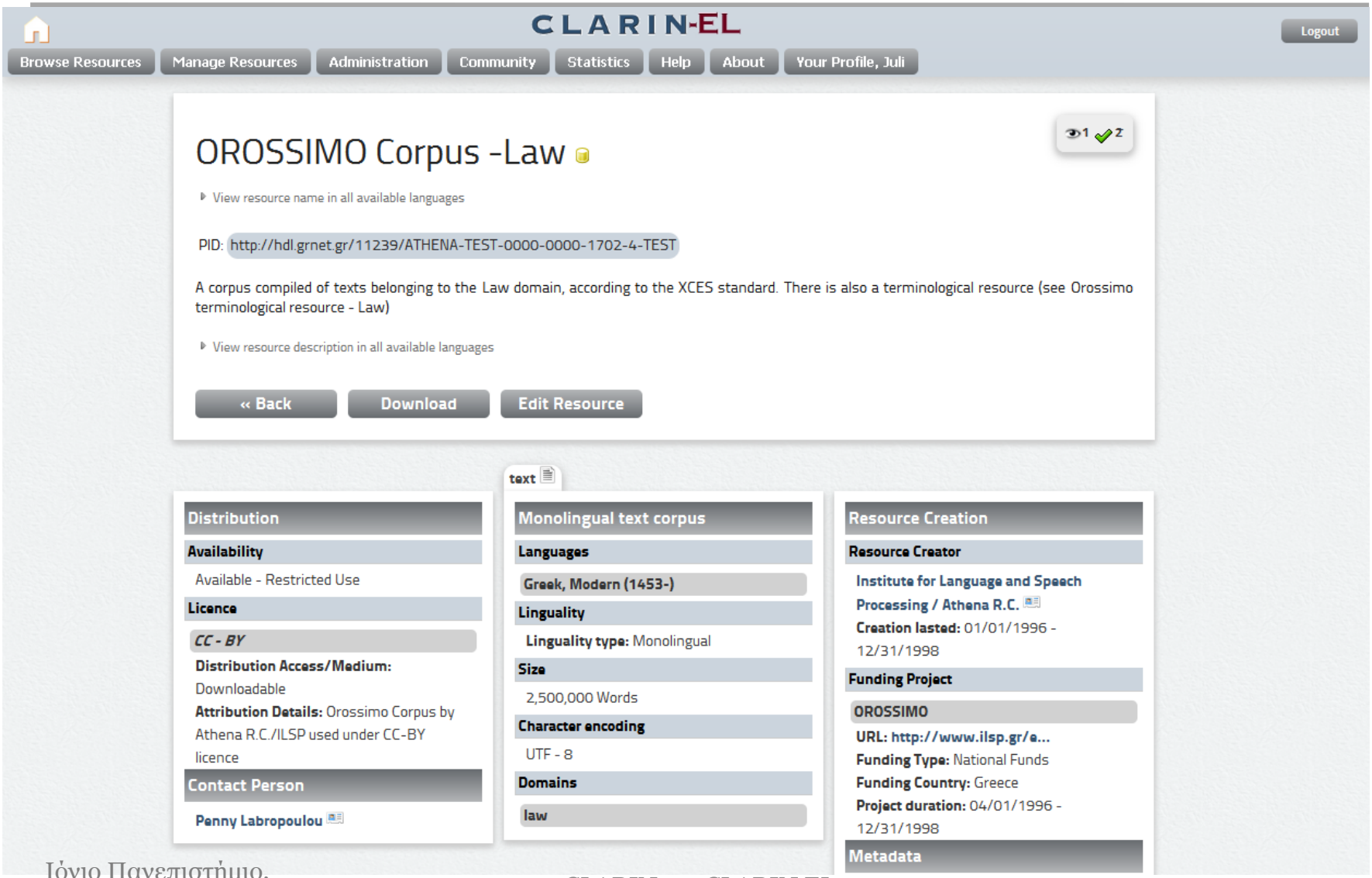
2. Φυλλομέτρηση αποτελεσμάτων



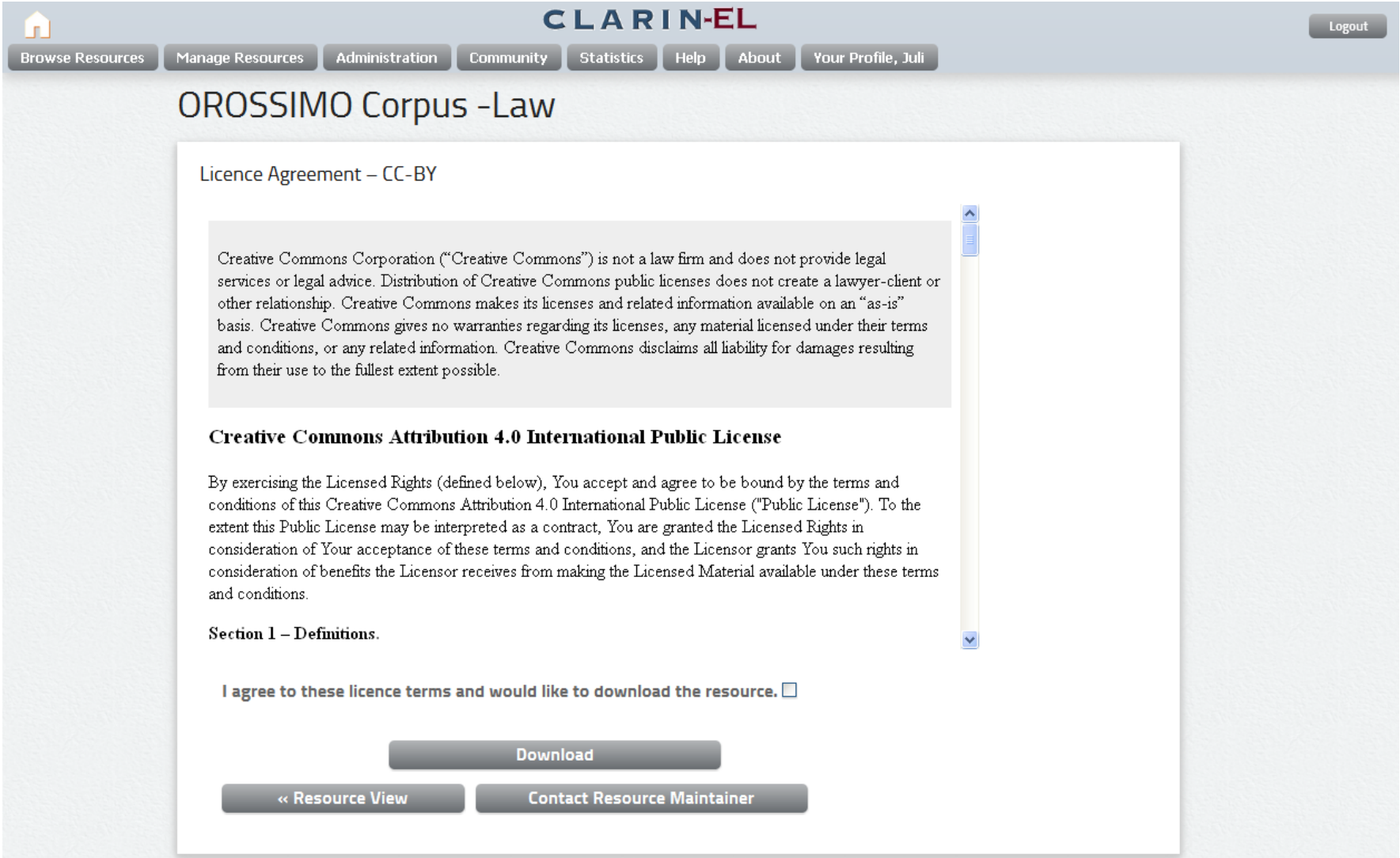
2α. Αναζήτηση με χρήση φίλτρων



3. Επιλογή του πόρου προτίμησης

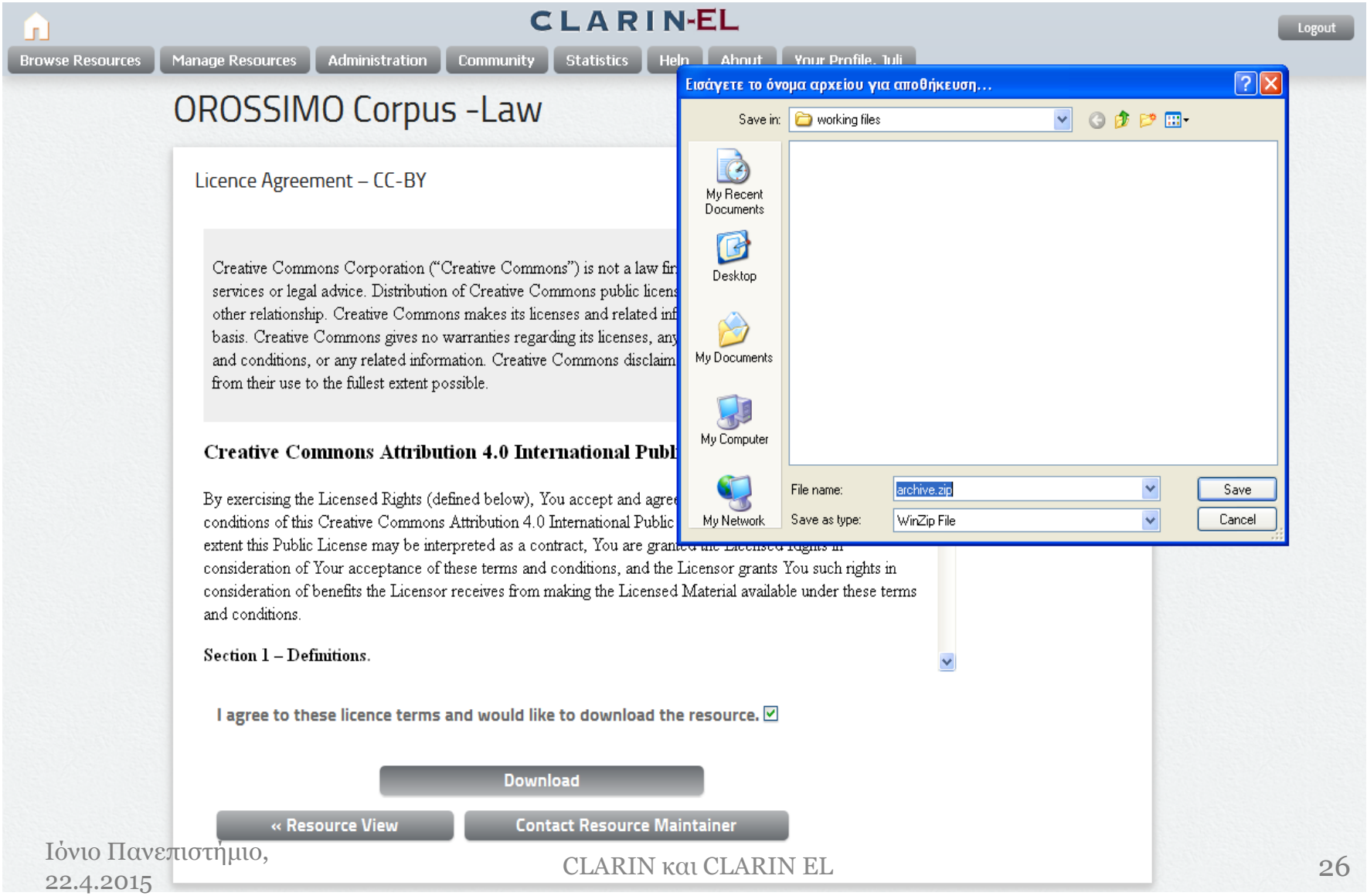


4. Όροι χρήσης - αδειοδότηση



Co-funded by the Greek CLARIN Attiki project (MIS 441451), Support for ESFRI/2006 Research Infrastructures, of the Greek Government.

5. Καταφόρτωση του πόρου



Πίσω από όλα,
ένα μοντέλο περιγραφής του κόσμου των γλωσσικών δεδομένων και υπηρεσιών

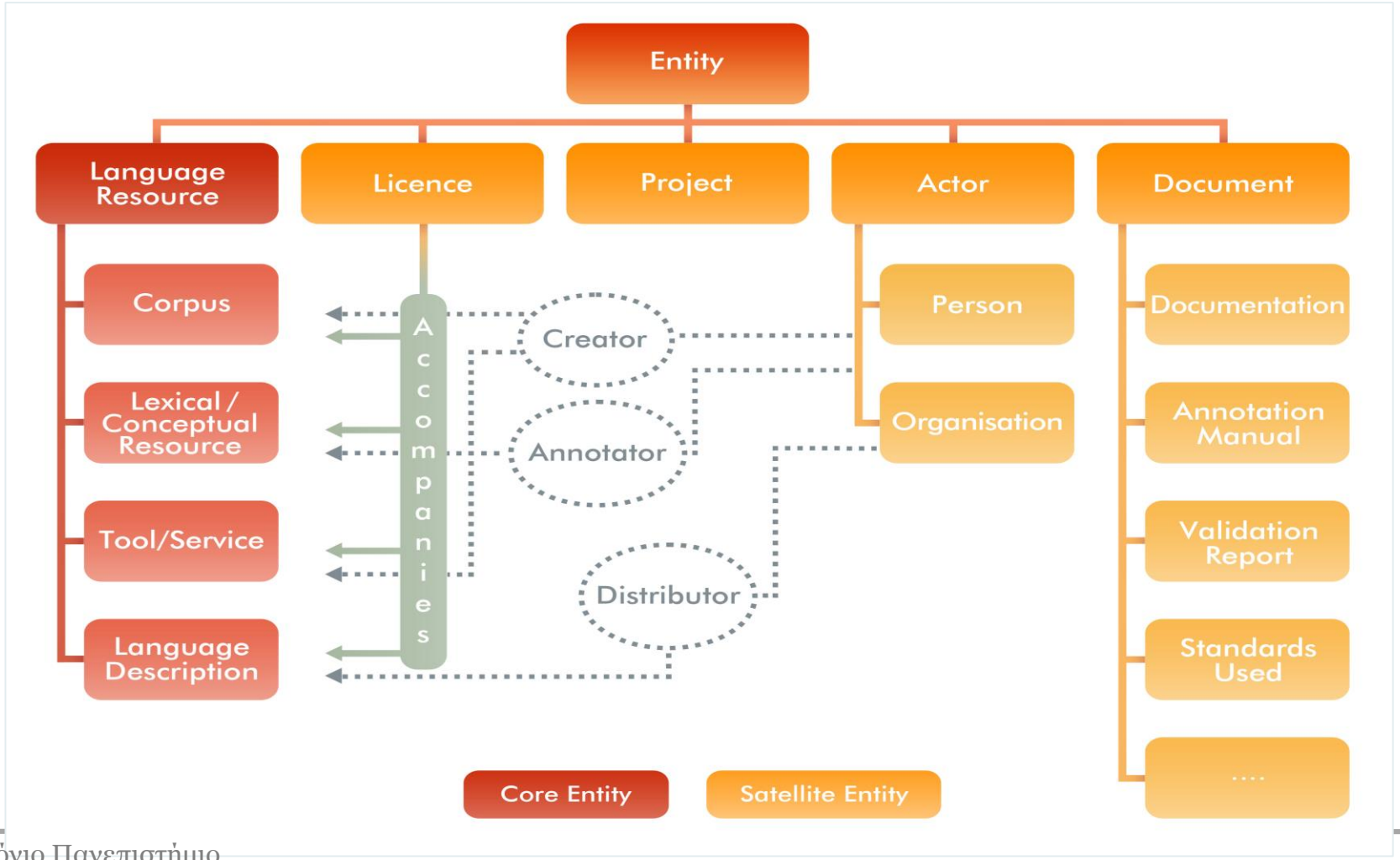
Στόχος

- Να υποστηρίξει τους χρήστες σε όλες τις ενέργειες τους
  - τεκμηρίωση πόρων + υπηρεσιών (δημιουργία, επικαιροποίηση)
  - αναζήτηση και ανάκτηση,
  - περιήγηση,
  - φόρτωση και καταφόρτωση πόρων
  - συγκομιδή μεταδεδομένων,
  - εποπτεία χρήσης κ.λπ.

Σχήμα μεταδεδομένων - οντολογία

- Περιγραφόμενες οντότητες
  - Κύρια οντότητα: γλωσσικός πόρος (δεδομένα-εργαλεία/υπηρεσίες)
  - Δορυφορικές οντότητες: σχετικά αντικείμενα, π.χ.
    - *δράστης*: *άτομα* και *οργανισμοί* που εμπλέκονται ως δημιουργοί πόρων, χρηματοδότες, διανομείς, κλπ.
    - *έγγραφο*: έγγραφα αναφοράς, όπως δημοσιεύσεις που περιγράφουν τον πόρο, αναφορές, εγχειρίδια, κ.λπ.
    - *έργο: έργα που χρηματοδότησαν τη δημιουργία* του πόρου, ή στο πλαίσιο του οποίου χρησιμοποιήθηκε ο πόρος κ.λπ.
    - *άδεια:* η άδεια διάθεσης του πόρου

Οντολογία (απόσπασμα)

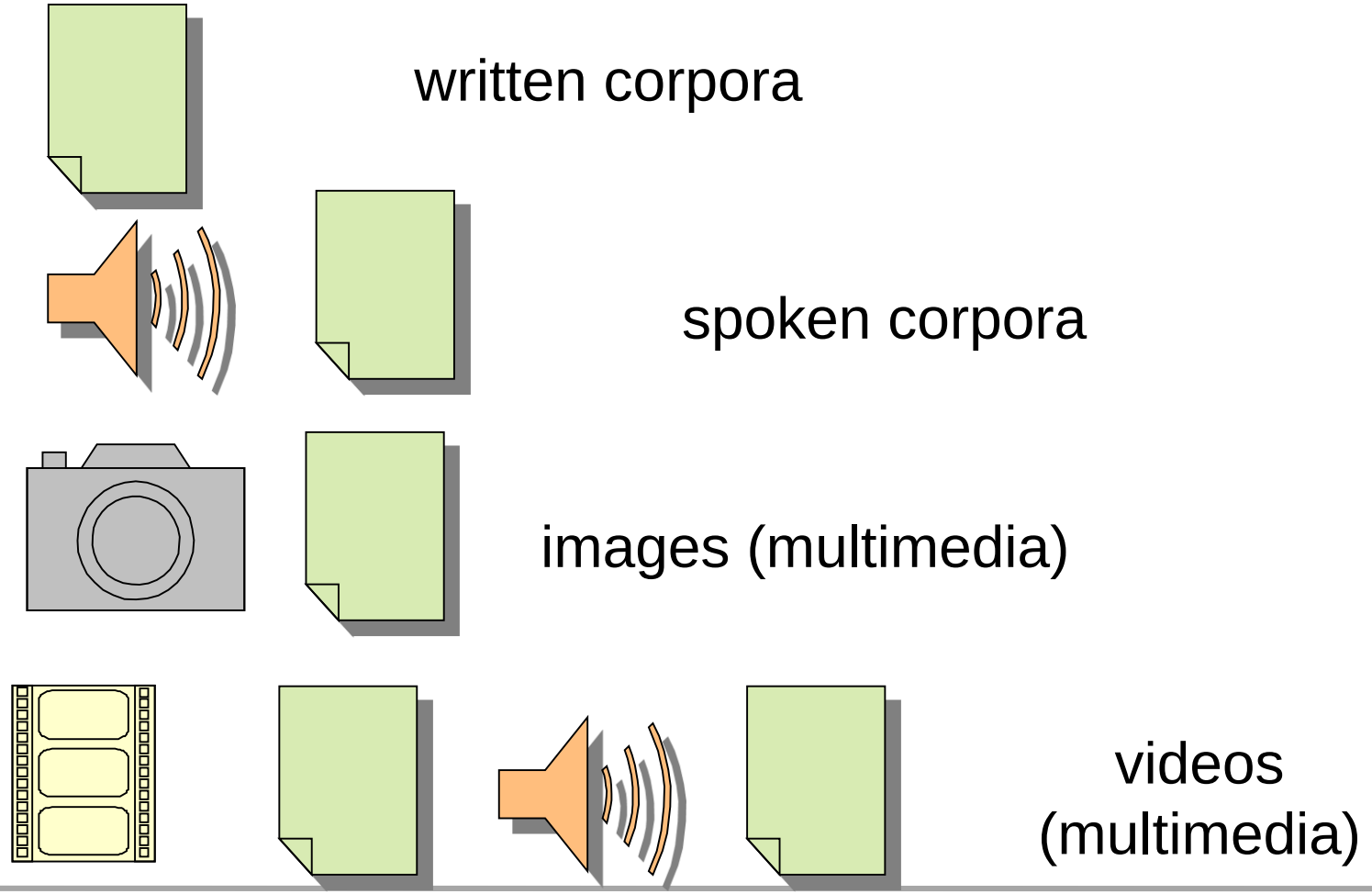


Τυπολογία πόρων (1)

- Τυπολογία πόρων
  - 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, τυπολογική βάση, κτλ.)
    - ***εργαλείο / τεχνολογία*** (π.χ. λημματοποιητής, επισημειωτής, εργαλείο αυτόματης μετάφρασης κτλ.)

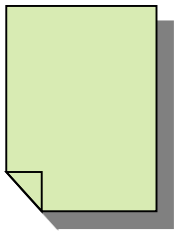
Τυπολογία πόρων (2)

- mediaType: text, audio, image, video

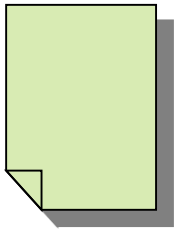


Τυπολογία πόρων (3)

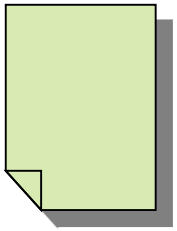
- search for text



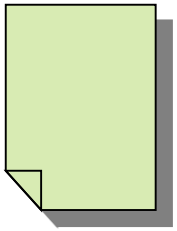
written corpora

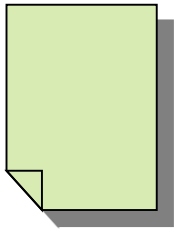


spoken corpora



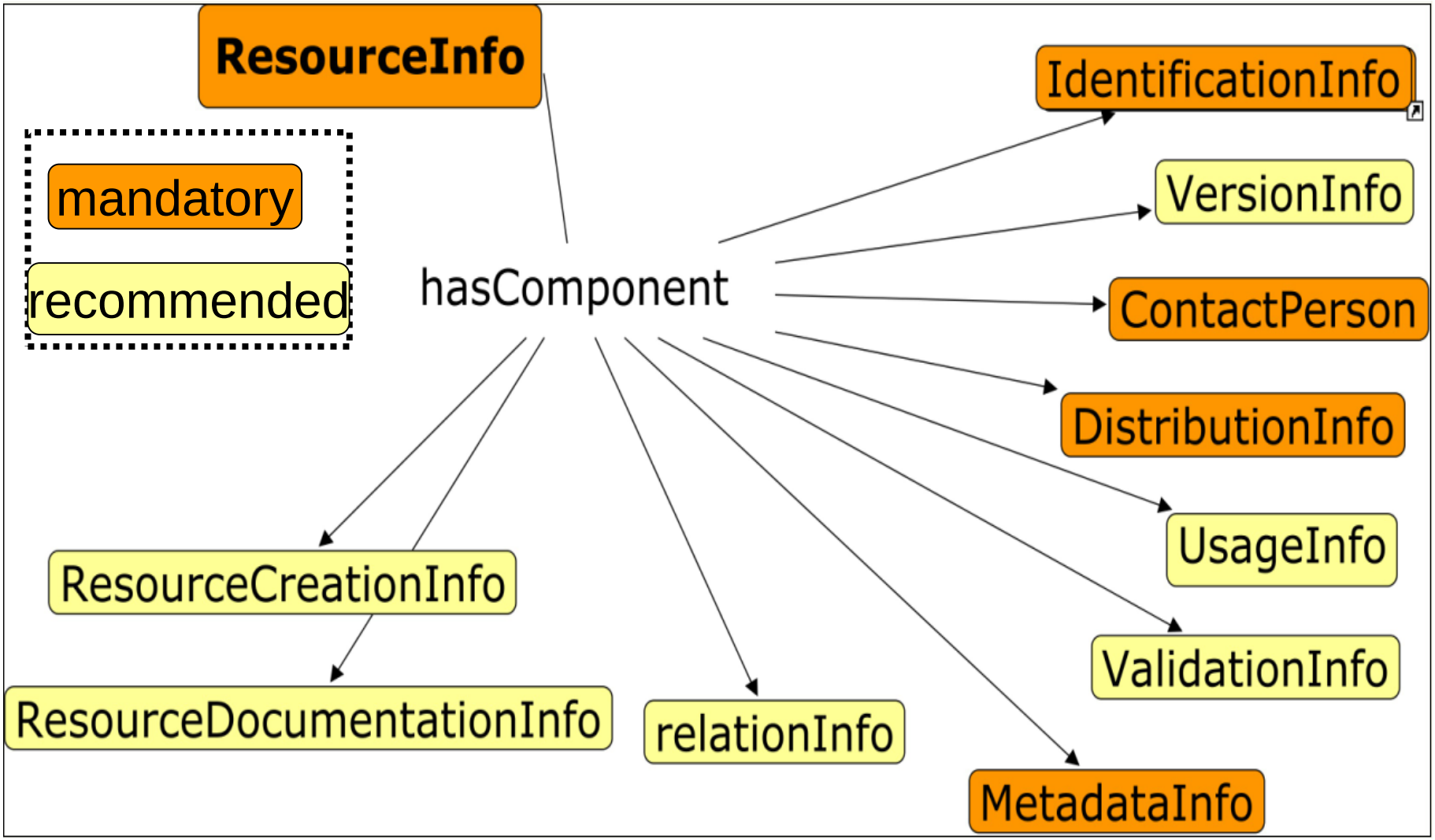
images (multimedia)



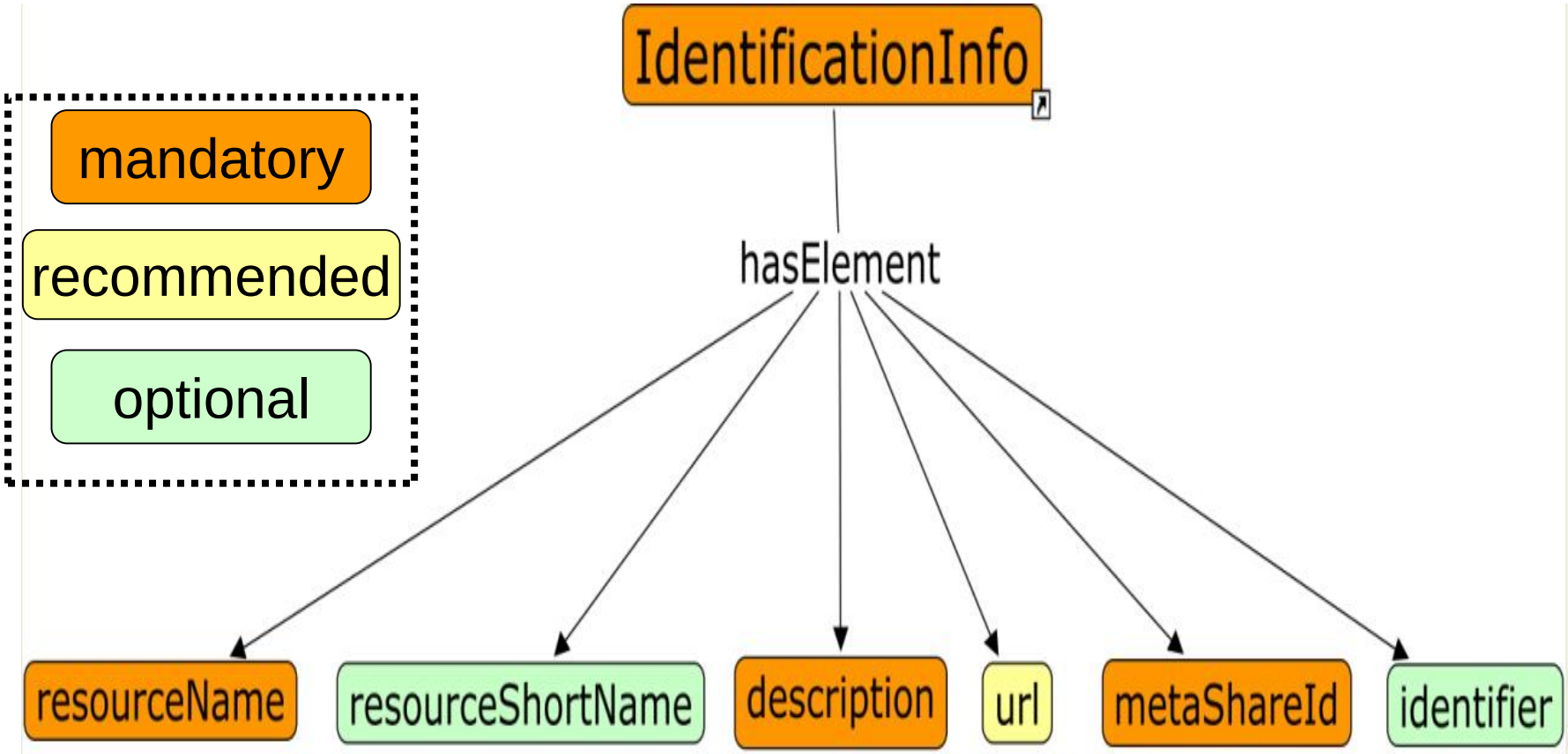


videos (multimedia)

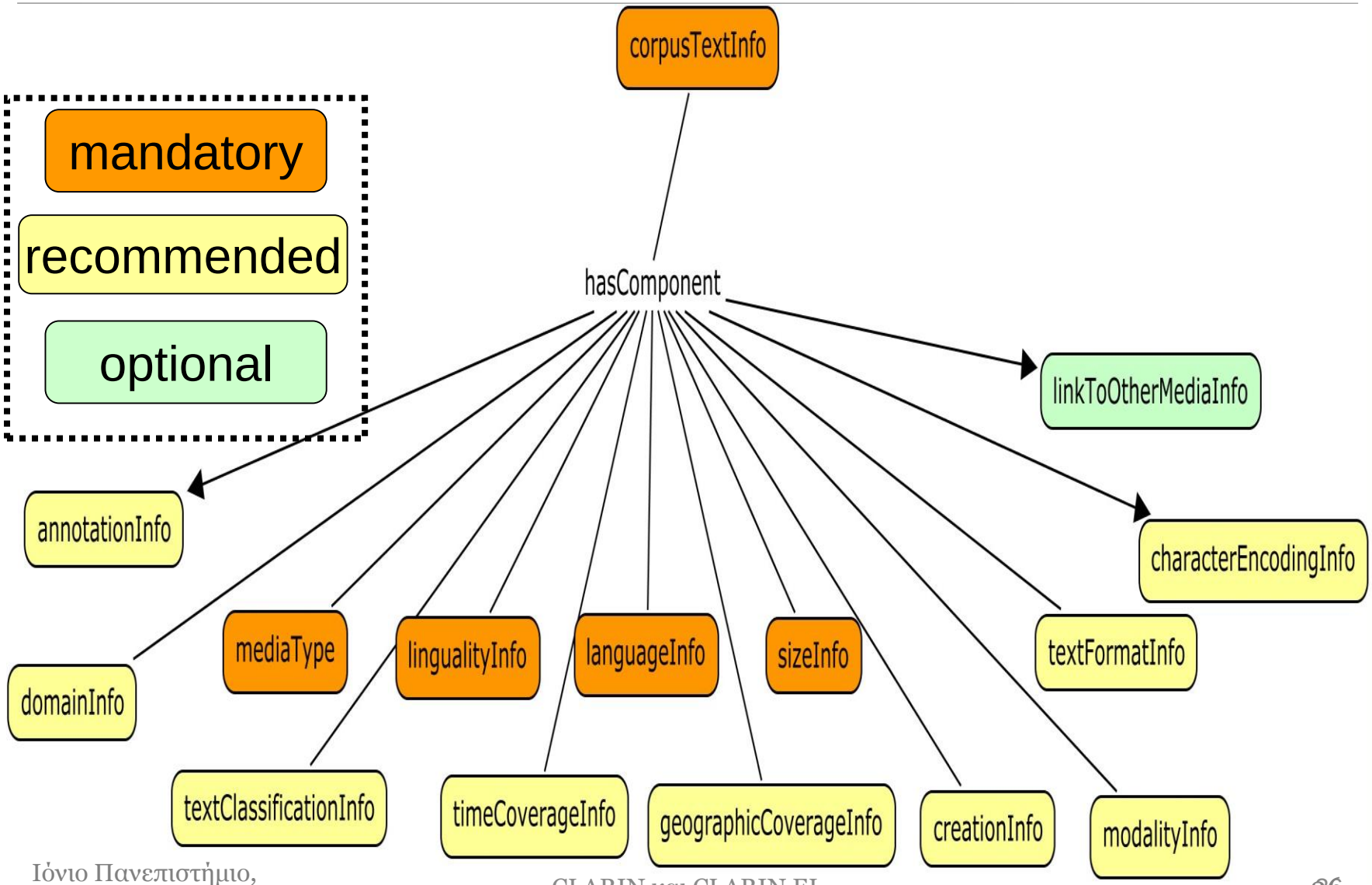
Το βασικό συστατικό περιγραφής



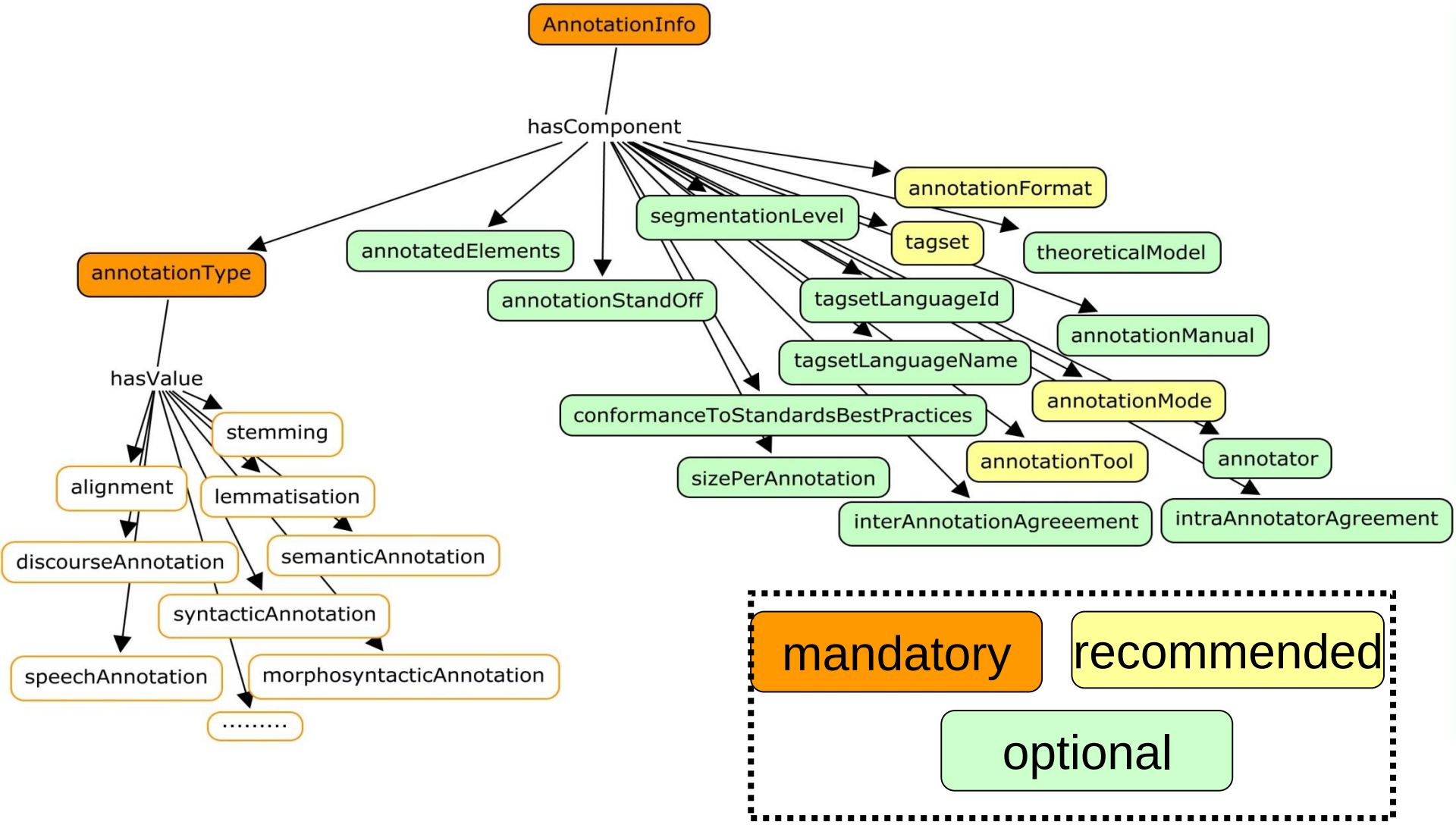
Identification component



corpusTextInfo

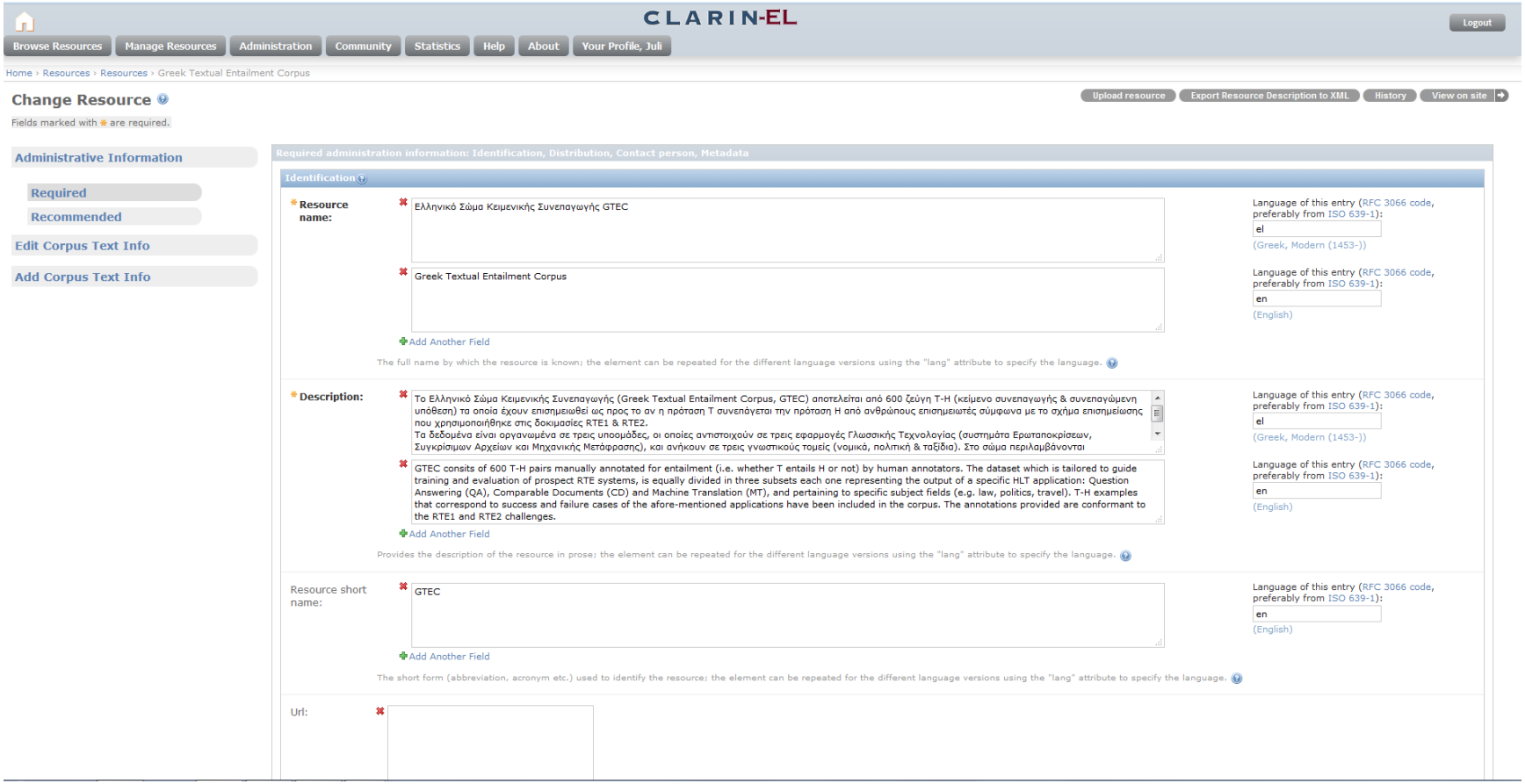


Annotation Component

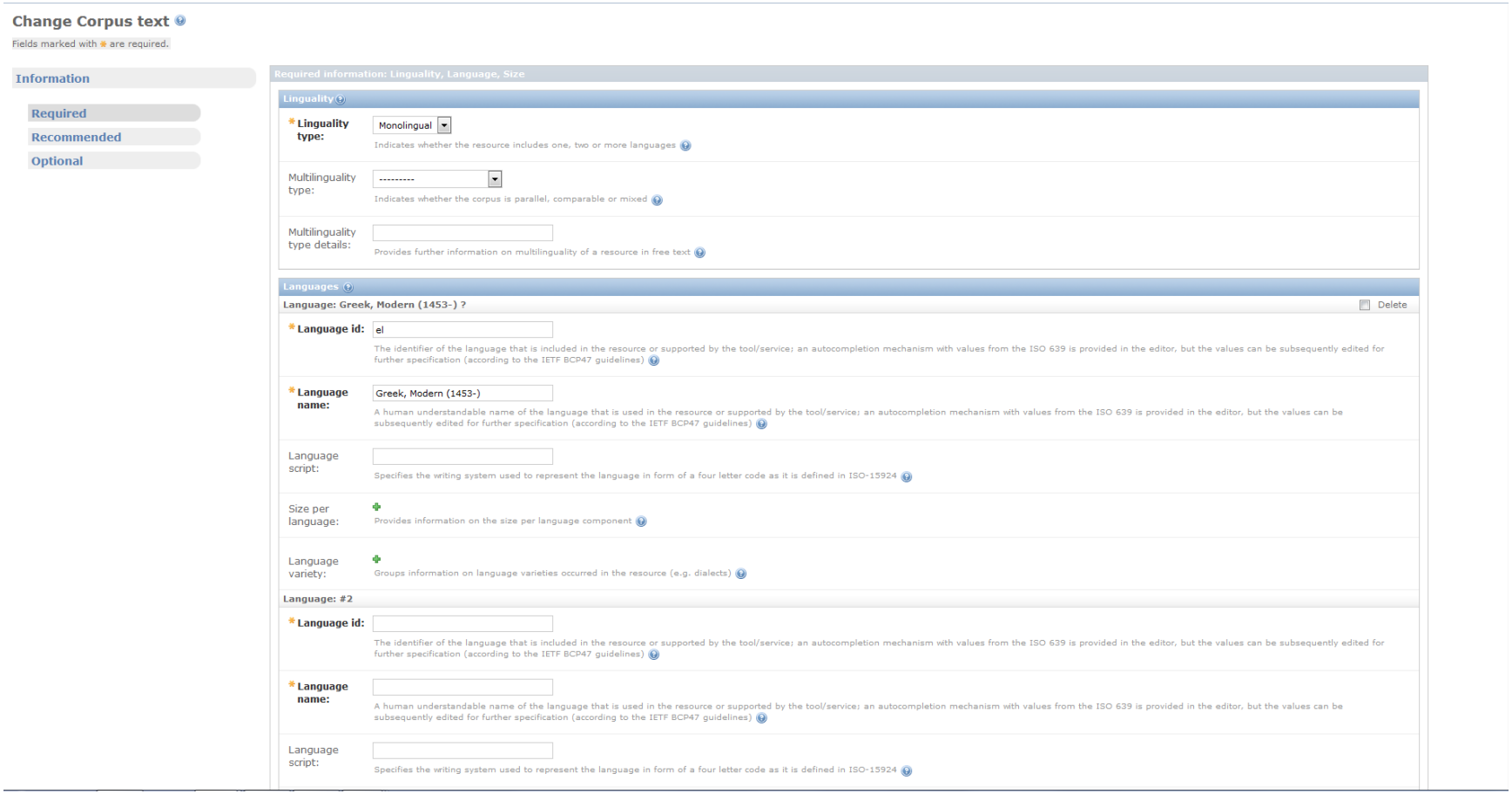


…και σε πιο εύχρηστη μορφή

Τεκμηρίωση μονογλωσσικού ΣΚ

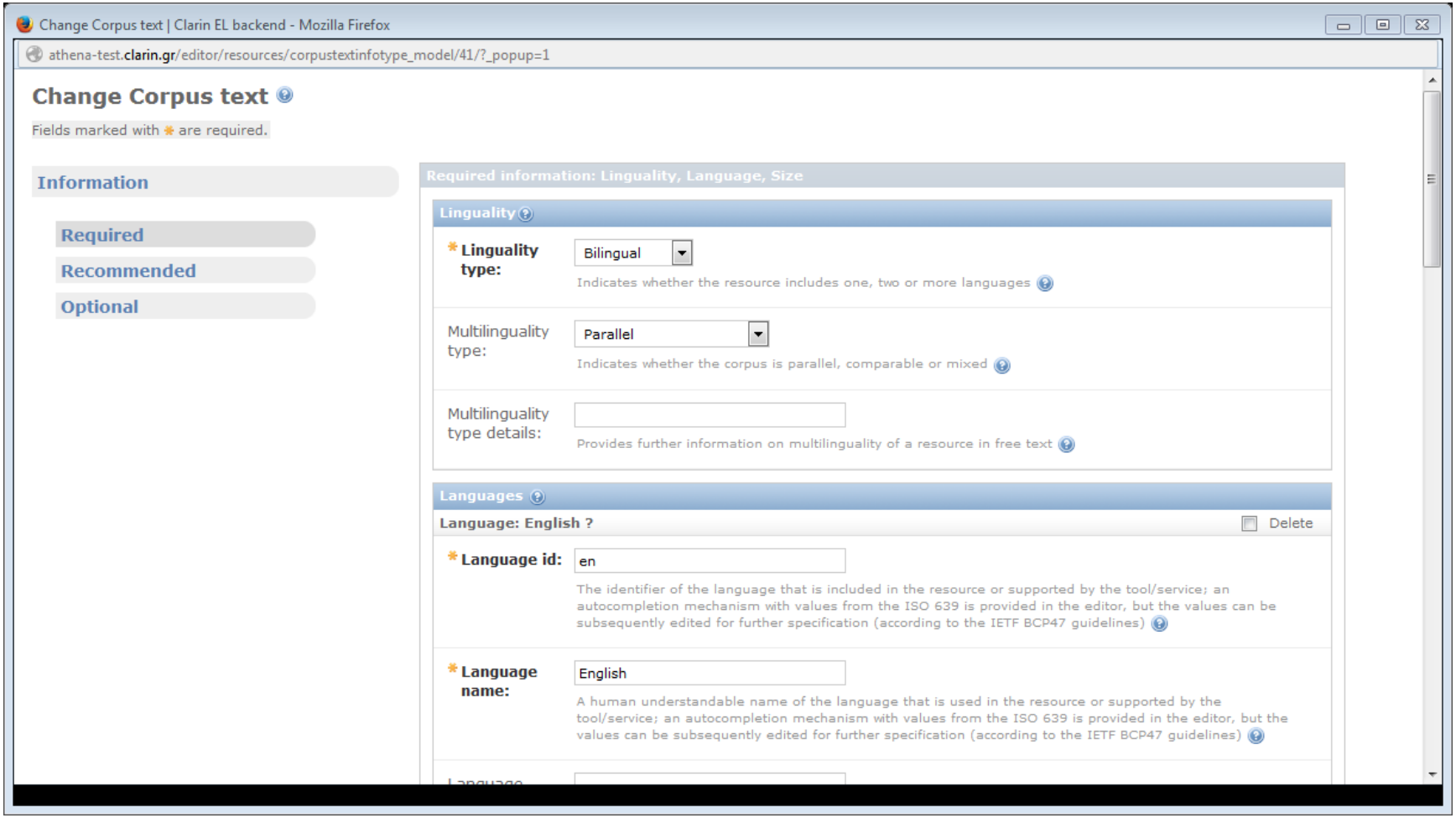


Τεκμηρίωση μονογλωσσικού ΣΚ (2)

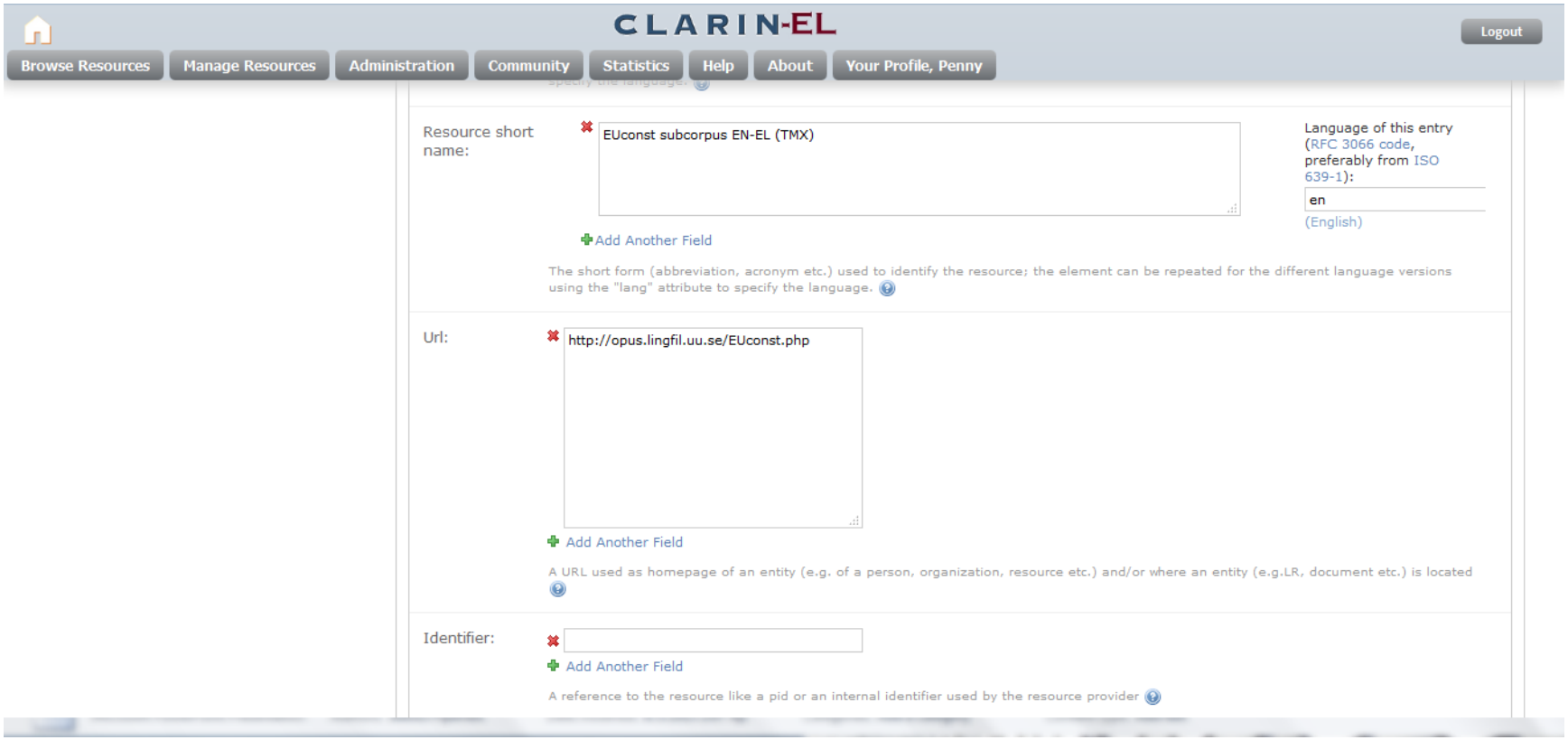


Τεκμηρίωση πολυγλωσσικού ΣΚ

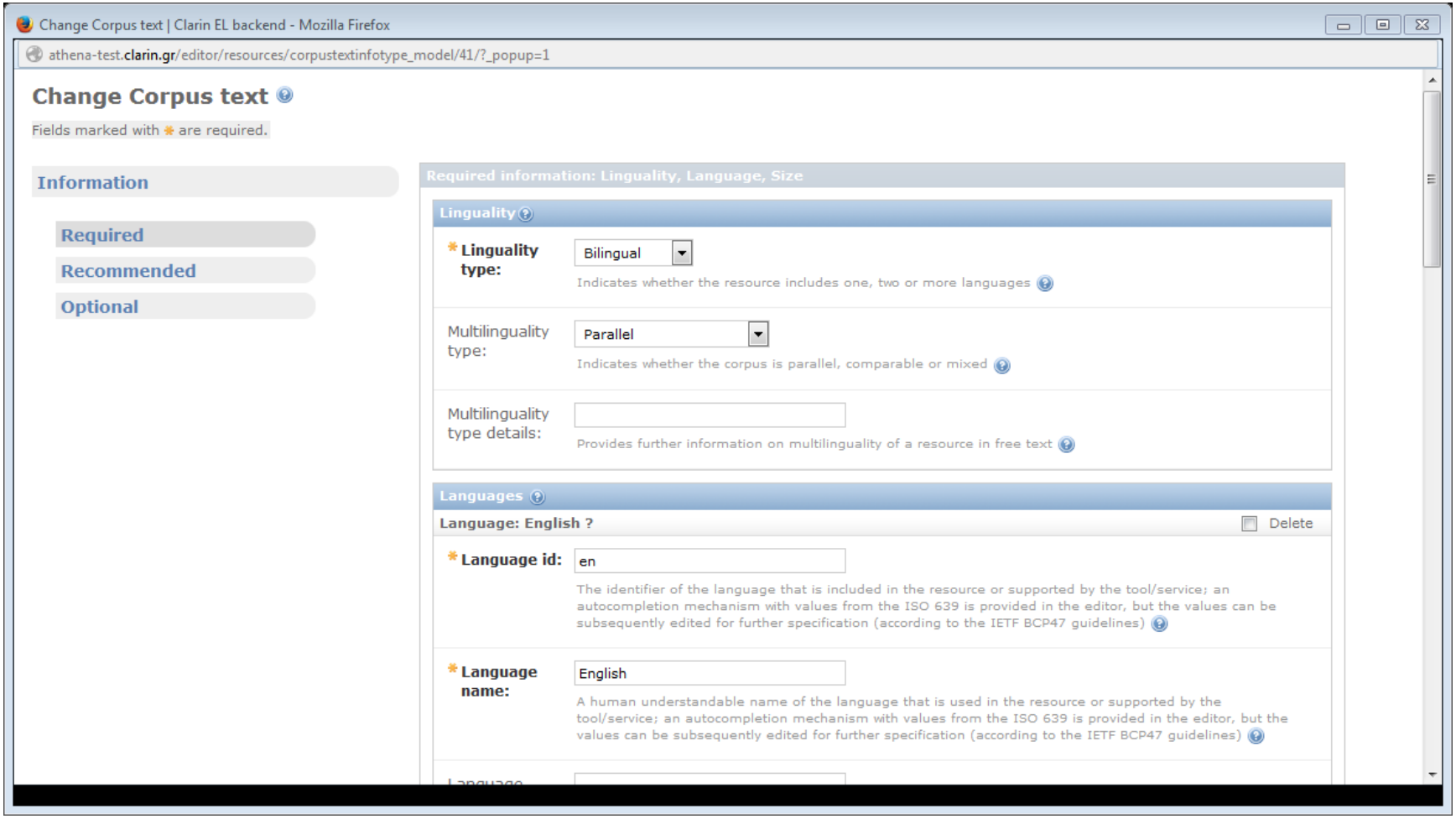
Τεκμηρίωση πολυγλωσσικού ΣΚ (2)



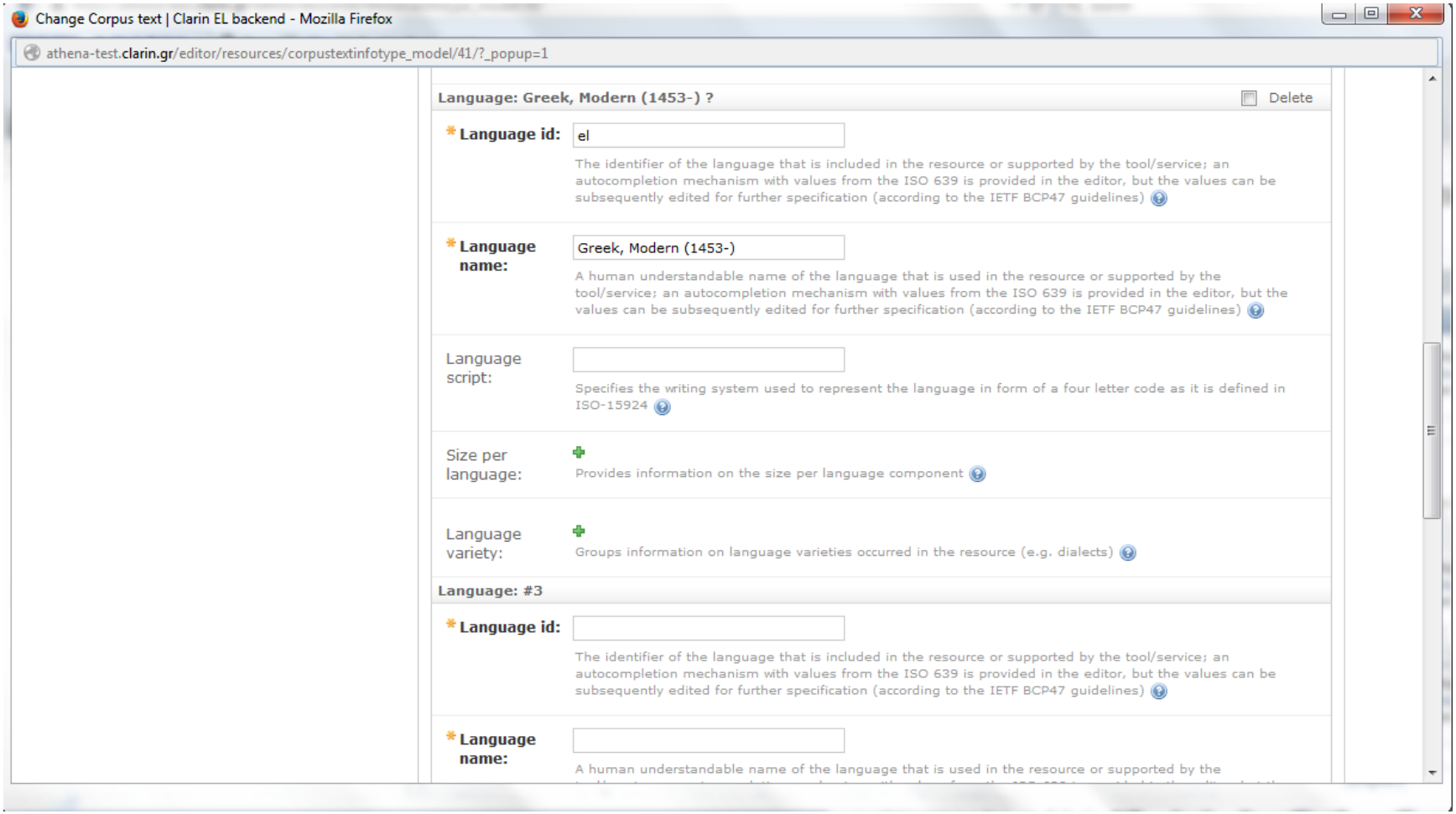
Τεκμηρίωση πολυγλωσσικού ΣΚ (3)



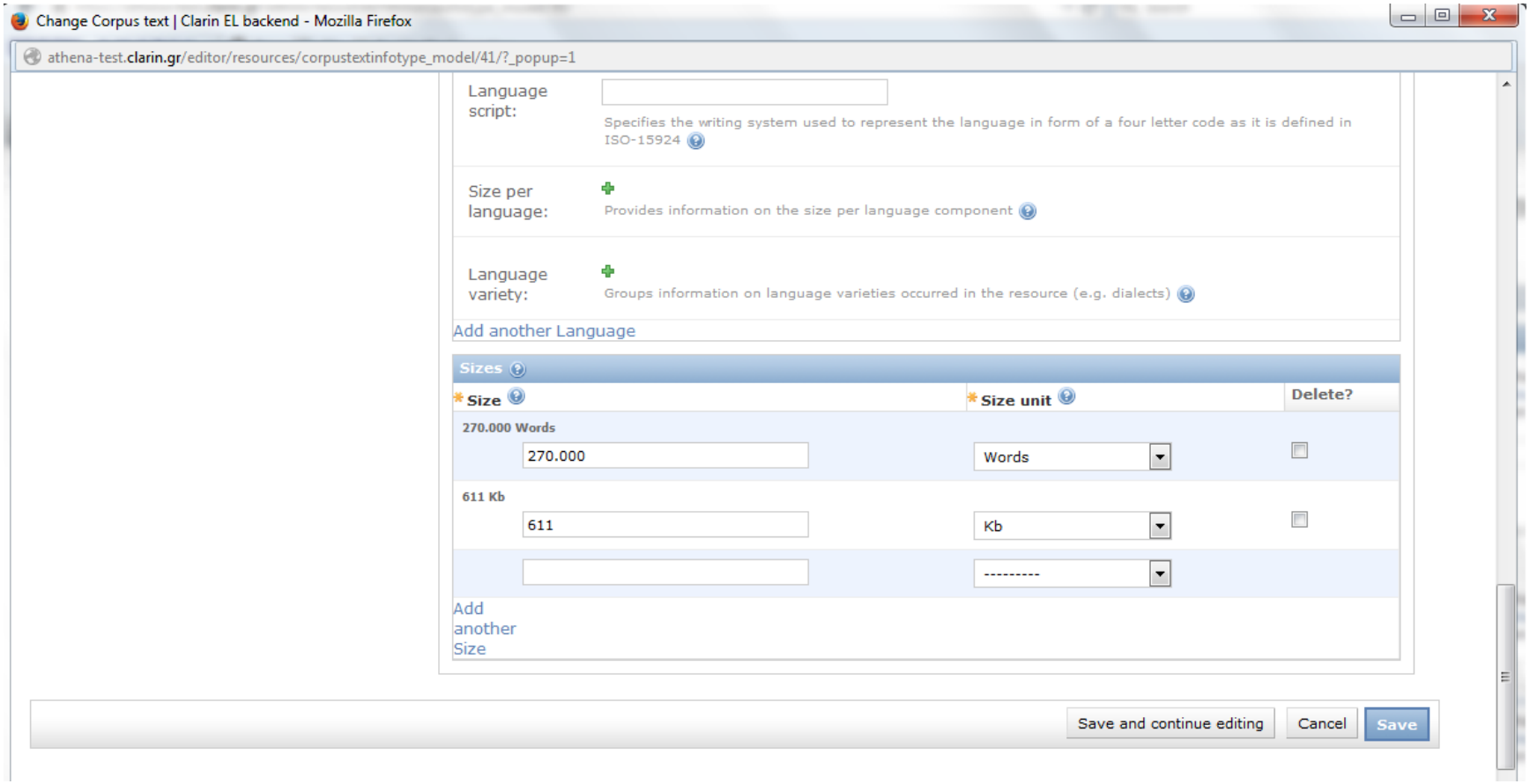
Τεκμηρίωση πολυγλωσσικού ΣΚ (4)



Τεκμηρίωση πολυγλωσσικού ΣΚ (5)

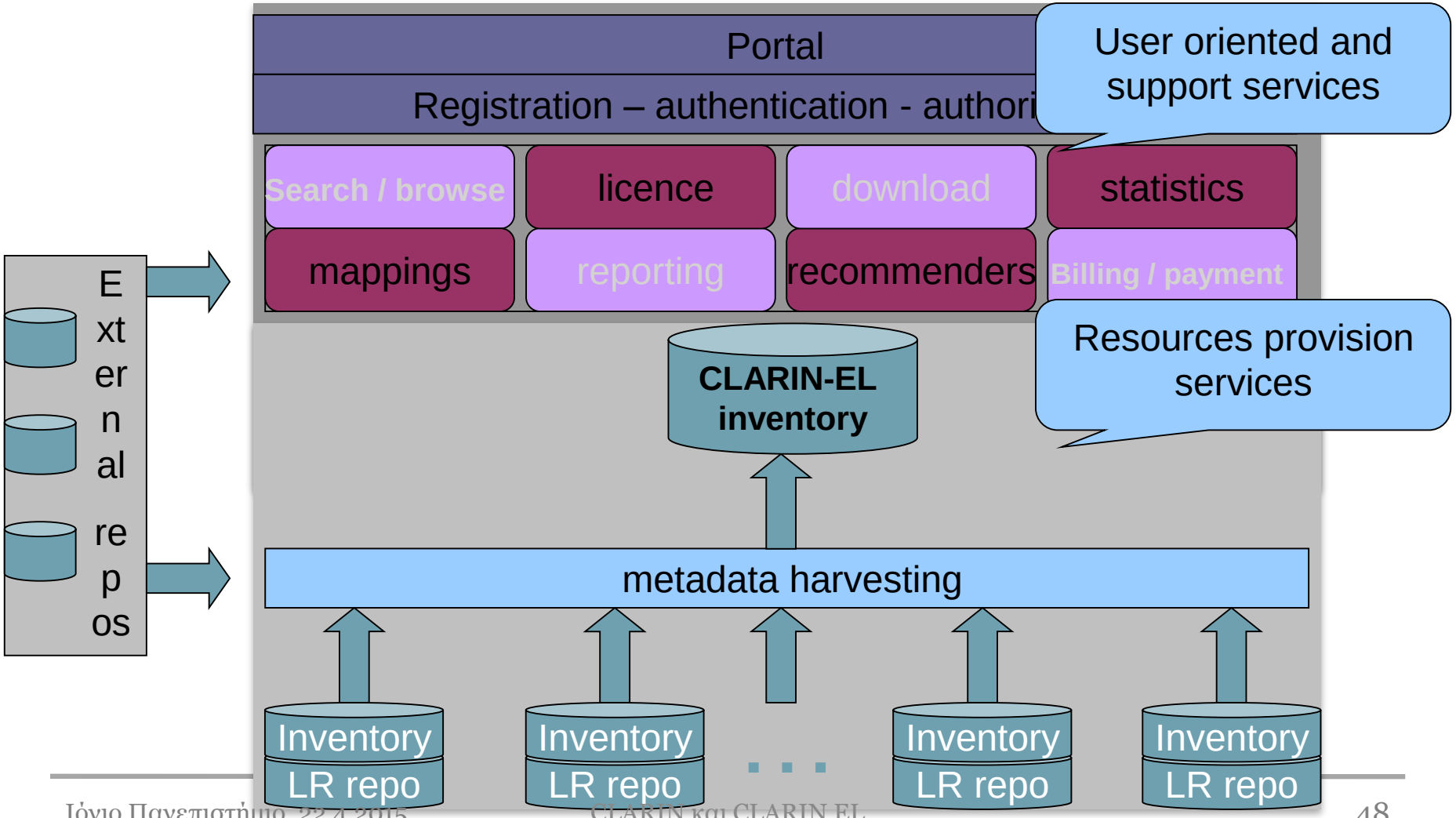


Τεκμηρίωση πολυγλωσσικού ΣΚ (6)

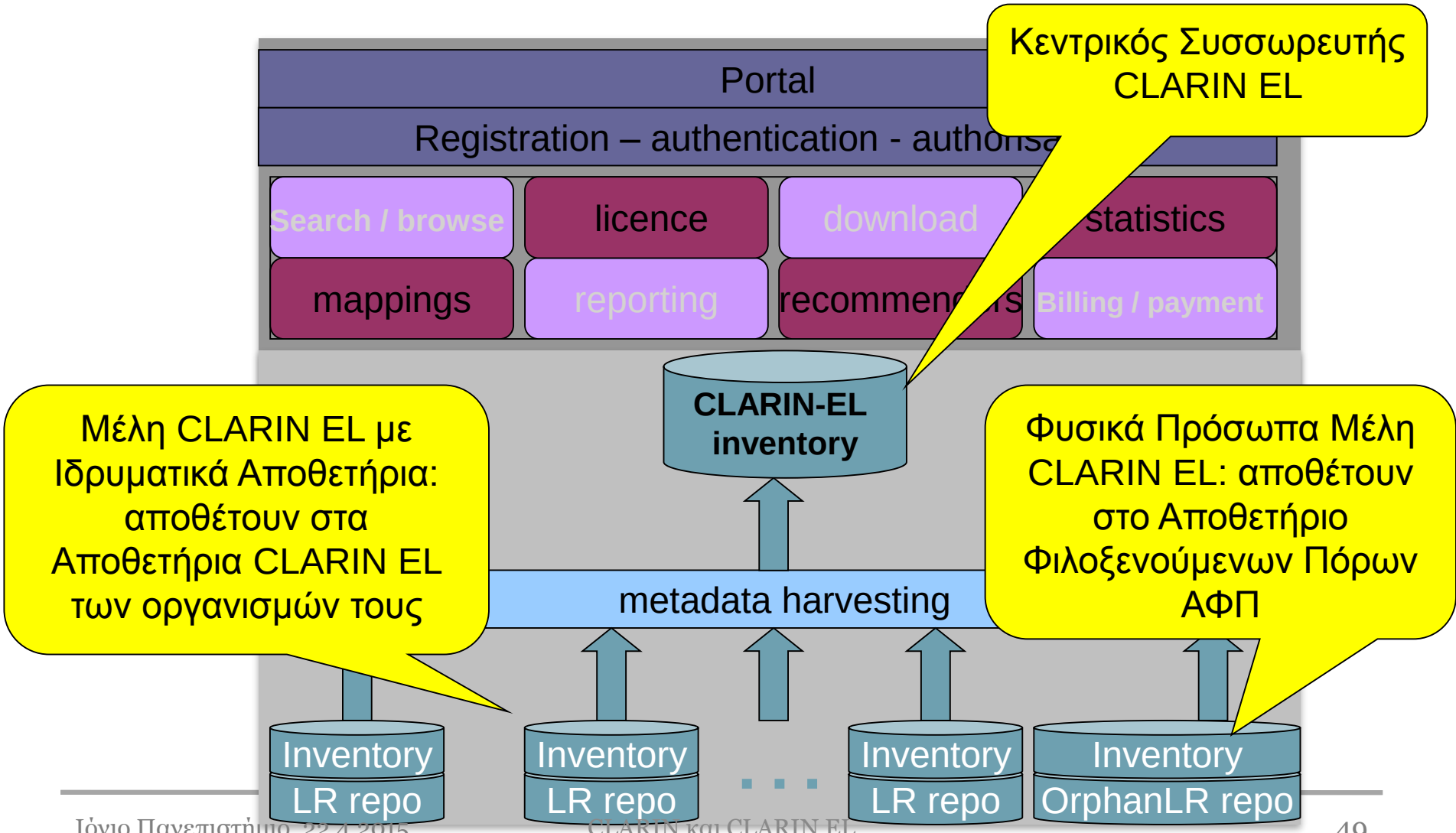


Αρχιτεκτονική δικτύου και υποδομής

Αρχιτεκτονική resources infrastructure



Αρχιτεκτονική – Σύσταση Δικτύου



Μέλη Δικτύου

- Κατασκευαστική Φάση (μέχρι 30/9/2015)
  - Υλοποίηση Υποδομής
    - ΕΚ «Αθηνά»
    - ΕΚΕΦΕ «Δημόκριτος»
    - Εθνικό Δίκτυο Ερευνας Τεχνολογίας (ΕΔΕΤ) ΑΕ
  - Χρήση και εμπλουτισμός
    - ΕΚ Πανεπιστήμιο Αθηνών
    - Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
    - Ιόνιο Πανεπιστήμιο
    - Πανεπιστήμιο Αιγαίου
    - Κέντρο Ελληνικής Γλώσσας
- Πιλοτική Φάση (από 1/1/2016)
  - Όλοι οι υπόλοιποι ακαδημαϊκοί φορείς (πανεπιστήμια, ερευν. κέντρα)

Τεχνικά συστατικά

- META-SHARE+ repository building software
- Πρωτόκολλα συγκομιδής (OAI-PMH, META-SHARE harvesting/syncing protocol)
- Metadata schemas (META-SHARE, CMDI (CLARIN), DC)
  - Μετασχηματισμοί μεταξύ META-SHARE-CMDI, META-SHARE-DC
- Αλυσίδες γλωσσικής επεξεργασίας & πρότυπα επισημείωσης και αναπαράστασης
  - http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
  - http://www.panacea-lr.eu/
  - http://qt21.metashare.ilsp.gr/

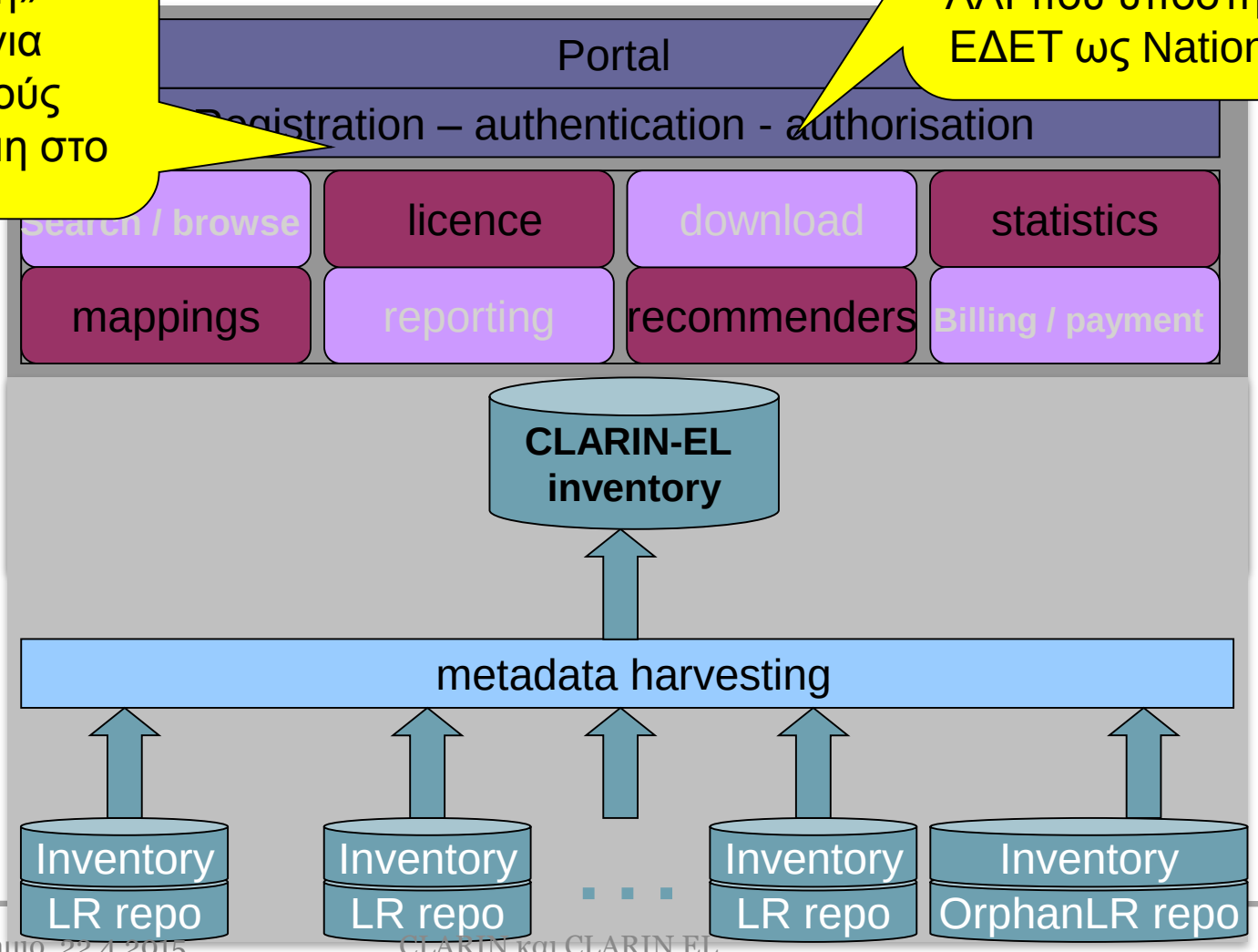
Υπολογιστική Υποδομή

- Όλα τα συνιστώντα υποσυστήματα προσαρμόζονται και αναπτύσσονται στην υπολογιστική υποδομή της ΕΔΕΤ
- Okeanos = cyclades & pithos
- cyclades : virtual machines and networking
- Pithos : virtual storage
- Βέλτιστη διαθεσιμότητα και δυναμική χρήση υπολογιστικών πόρων
- Ασφάλεια και διατήρηση πόρων, τόσο πρωτογενών (στο περιβάλλον του pithos, όσο και δευτερογενών (metadata databases στο περιβάλλον των cyclades)

Αρχιτεκτονική – Πρόσβαση (1)

«Κλασσική» εγγραφή για ακαδημαϊκούς χρήστες και μη στο ΑΦΠ

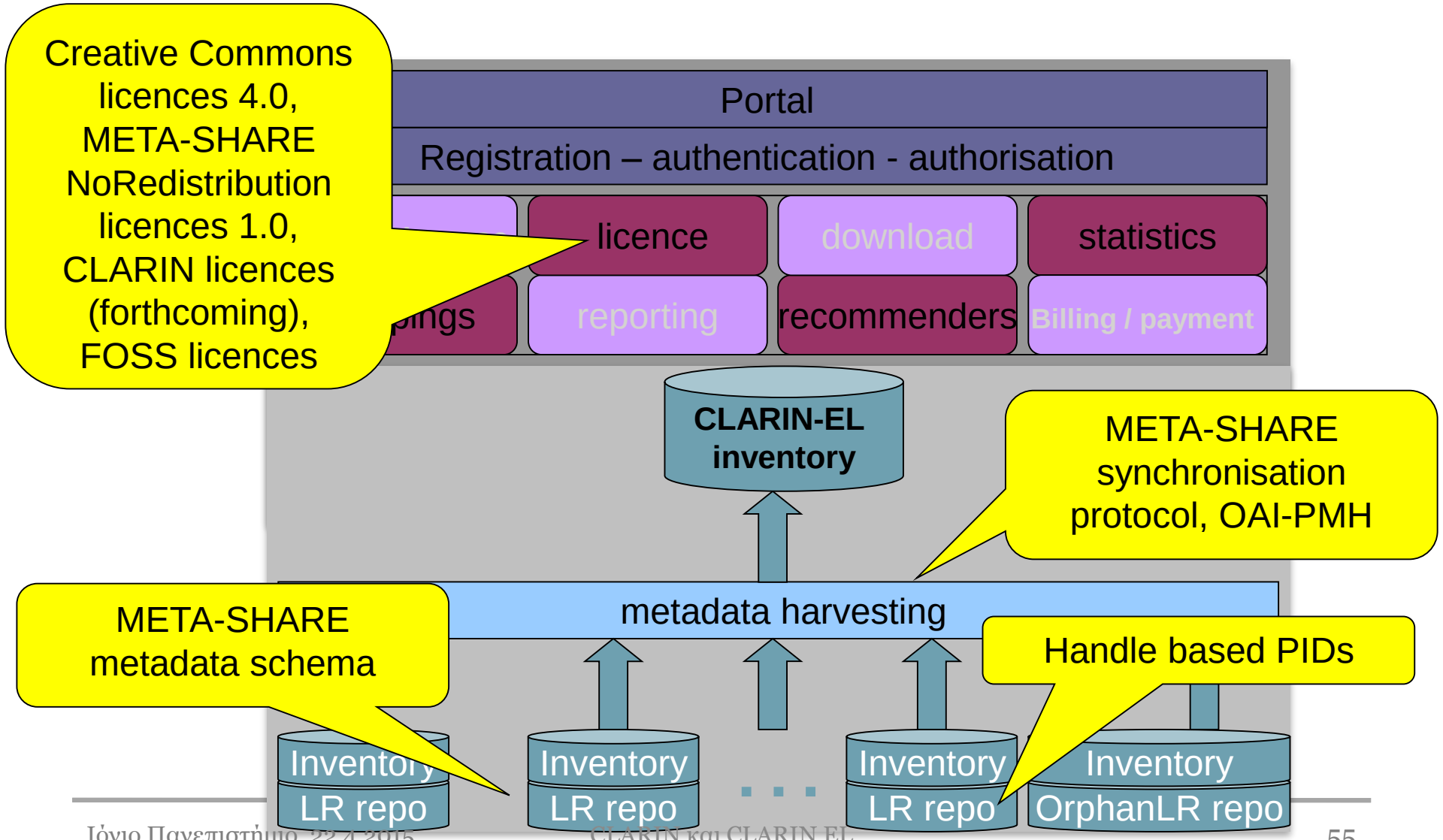
Είσοδος με ακαδημαϊκό λογαριασμό (SSO) μέσω της ομοσπονδίας AAI που υποστηρίζει η ΕΔΕΤ ως National IdP



Αρχιτεκτονική – Πρόσβαση (2)

- Όλοι οι συνεργαζόμενοι πάροχοι (ΕΚΠΑ, ΑΠΘ, ΙΟΝΙΟ, ΑΙΓΑΙΟ, ΚΕΓ) έχουν δικαίωμα ως ακαδημαϊκοί φορείς να είναι μέλη της ΑΑΙ ομοσπονδίας.
- ΕΚΠΑ, ΑΠΘ, ΙΟΝΙΟ, ΑΙΓΑΙΟ είναι ήδη μέλη της ομοσπονδίας
- Όλοι οι δυνάμει χρήστες των ΕΚΠΑ, ΑΠΘ, ΙΟΝΙΟ, ΑΙΓΑΙΟ θα έχουν δυνατότητα πρόσβασης με το ακαδημαϊκό τους login.
- Και μόνον με αυτό!

Αρχιτεκτονική – Πρότυπα



Πρόσβαση – Διαχείριση χρηστών

- Βασικές διχοτομήσεις
  - Πάροχοι
  - Καταναλωτές

- **Πάροχοι**
  - Ακαδημαϊκοί χρήστες -> Αποθετήρια ακαδημαϊκών οργανισμών (ιδρυματικά αποθετήρια)
  - Μη ακαδημαϊκοί χρήστες -> Αποθετήριο φιλοξενούμενων πόρων

- **Καταναλωτές**
  - Ακαδημαϊκοί χρήστες και χρήση -> Αναζήτηση, καταφόρτωση και επεξεργασία υλικού χωρίς περιορισμούς ή με τους ελάχιστους δυνατούς
  - Μη ακαδημαϊκοί χρήστες -> Αναζήτηση και καταφόρτωση σύμφωνα με τους περιορισμούς αδειοδότησης και επεξεργασία με περιορισμούς όγκου υλικού

Πρόσβαση – Διαχείριση παρόχων και πόρων

- Βασικοί ρόλοι παρόχων
  - Διαχειριστής αποθετηρίου (τεχνικός-επιστημονικός)
  - Τεκμηριωτές & ομάδες τεκμηριωτών
  - Συντονιστές ομάδων τεκμηριωτών

- Βασικές καταστάσεις πόρων
  - Εσωτερικής δημοσίευσης
  - Περιορισμένης δημοσίευσης
  - Πλήρους δημοσίευσης

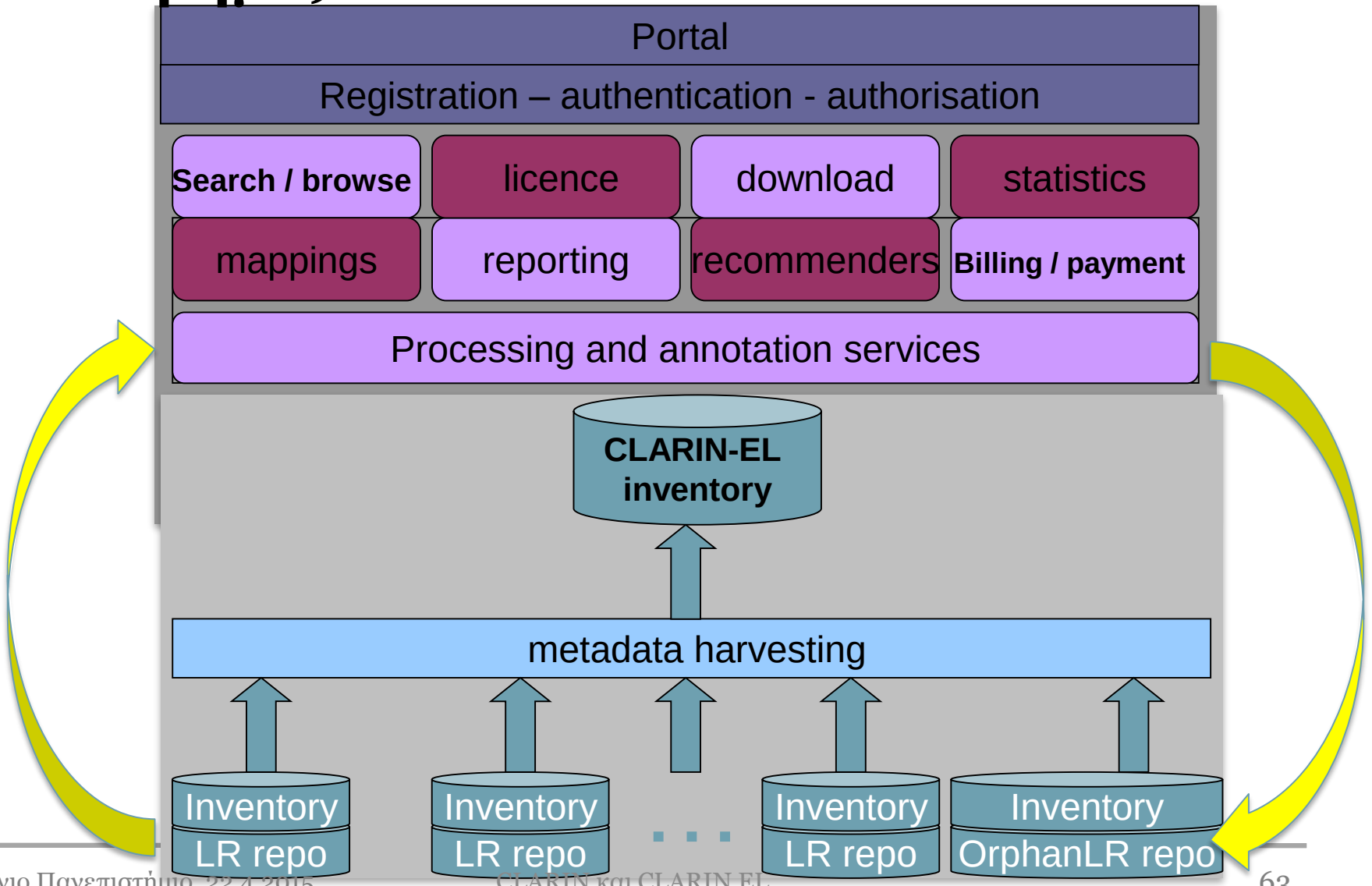
Υποστήριξη

- Για τη δημιουργία αποθετηρίου CLARIN EL
  - επικοινωνία στο clarin-el-coordinator@clarin.gr, και clarin-el-techadmin@clarin.gr
  - αυτοματοποιημένη διαδικασία, με σύντομη συνεργασία του Διαχειριστή Αποθετηρίου με τον CLARIN EL Τεχνικό Διαχειριστή της Υποδομής
- Γραφεία υποστήριξης σχετικά με
  - Νομικά θέματα → legal-helpdesk@clarin.gr
  - Τεχνικά θέματα → technical-helpdesk @clarin.gr
  - Θέματα τεκμηρίωσης → metadata–helpdesk @clarin.gr

Εθνικό δίκτυο CLARIN-EL

τι δουλεύουμε τώρα

Αρχιτεκτονική – Υπηρεσίες επεξεργασίας (το 6ο βήμα)



Υπηρεσίες γλωσσικής επεξεργασίας

- Υπηρεσίες επεξεργασίας μονογλωσσικών σωμάτων κειμένων (ΣΚ)
  - 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, …

Επεξεργάσιμα δεδομένα

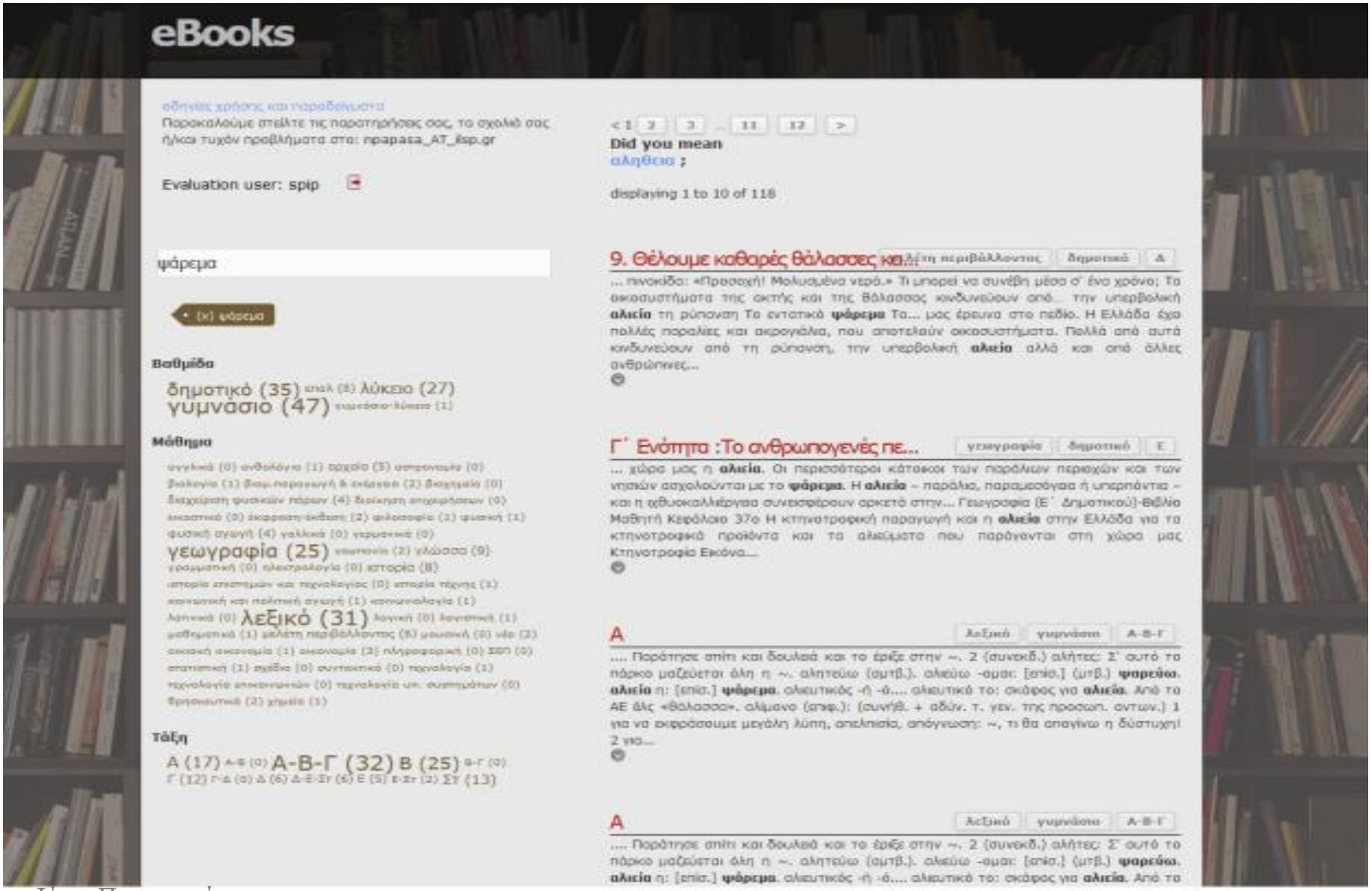
- Εκείνα που ικανοποιούν τις προδιαγραφές δεδομένων και υπηρεσιών
  - ως προς γλώσσα (Ελληνικά, Αγγλικά, …)
  - αναμενόμενο μορφότυπο (format) εισόδου
  - αναμενόμενη προεπεξεργασία
    - ένα ήδη μορφοσυντακτικά επισημειωμένο σύνολο δεδομένων μπορεί να
      - αναλυθεί συντακτικά, κλπ.
  - ανοικτά δικαιώματα χρήσης
    - που επιτρέπουν παράγωγα «έργα», δηλ. δημιουργία παράγωγων δεδομένων
  - ανοικτά διατιθέμενες υπηρεσίες επεξεργασίας
- Στην περίπτωση πιο σύνθετων εργασιών
  - Πλήρης αυτοματοποίηση όλης της ροής επεξεργασίας (workflow)

Επεξεργάσιμα δεδομένα (2)

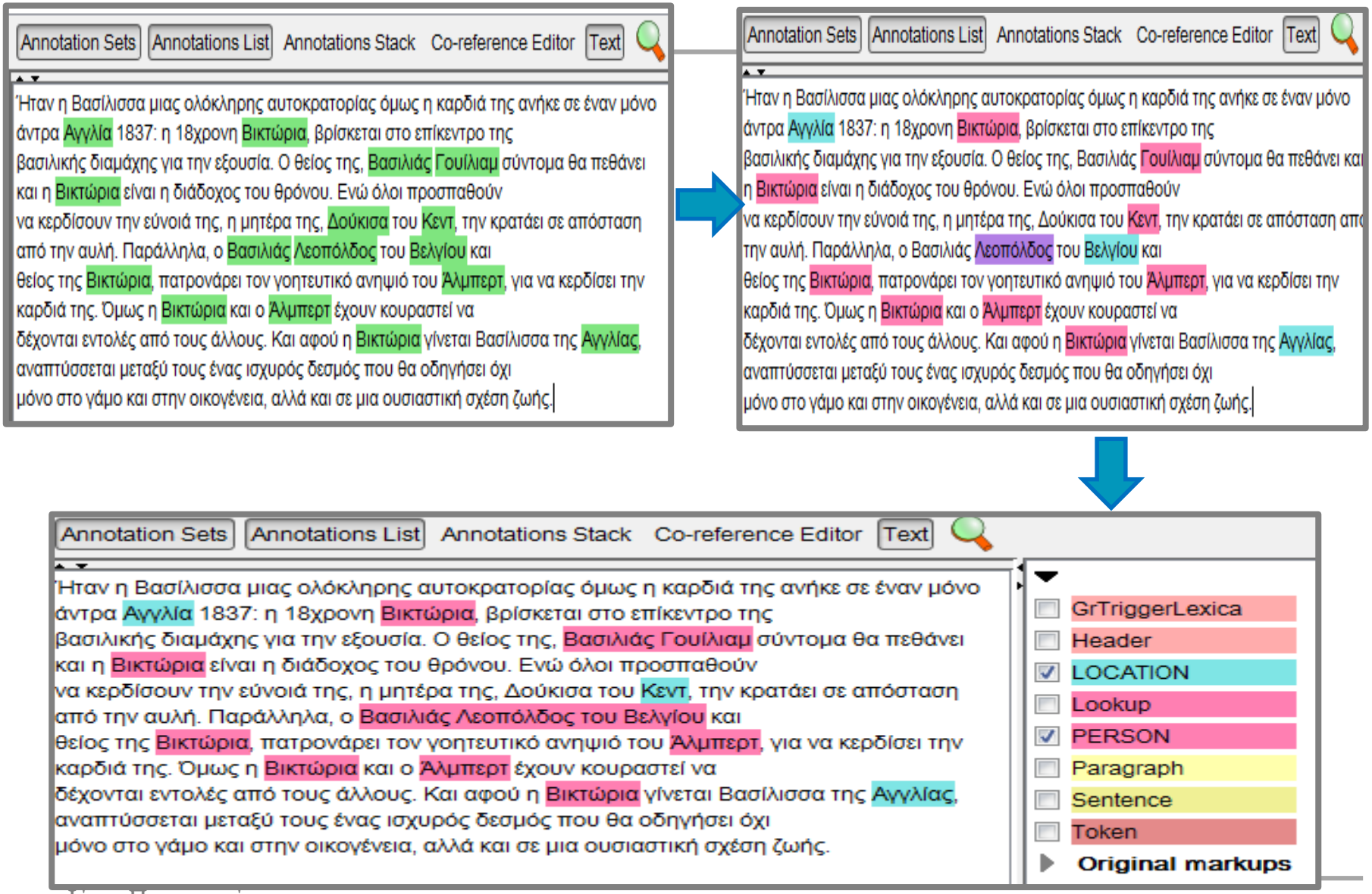
- τα προς επεξεργασία δεδομένα μπορούν να
  - βρίσκονται ήδη αποθηκευμένα σε κάποιο Ιδρυματικό Αποθετήριο, ή στο Αποθετήριο Φιλοξενούμενων Πόρων
  - φορτωθούν στην Υποδομή από τον χρήστη
    - προσωρινά
    - να ελεγχθούν ως προς την συμβατότητά τους με τις προδιαγραφές
- τα επεξεργασμένα δεδομένα συνιστούν νέα δεδομένα
  - που αποθηκεύονται στο ίδιο Ιδρυματικό Αποθετήριο με τα πρωτογενή δεδομένα, και συσχετίζονται οι περιγραφές τους
  - αν τα πρωτογενή τους δεδομένα φορτώθηκαν από τον χρήστη,
    - παραμένουν για 48 ώρες και στη συνέχεια διαγράφονται από την Υποδομή
    - αποθηκεύονται μόνιμα εφόσον ο χρήστης γίνει μέλος

Με όπλο τις υποδομές αυτές…

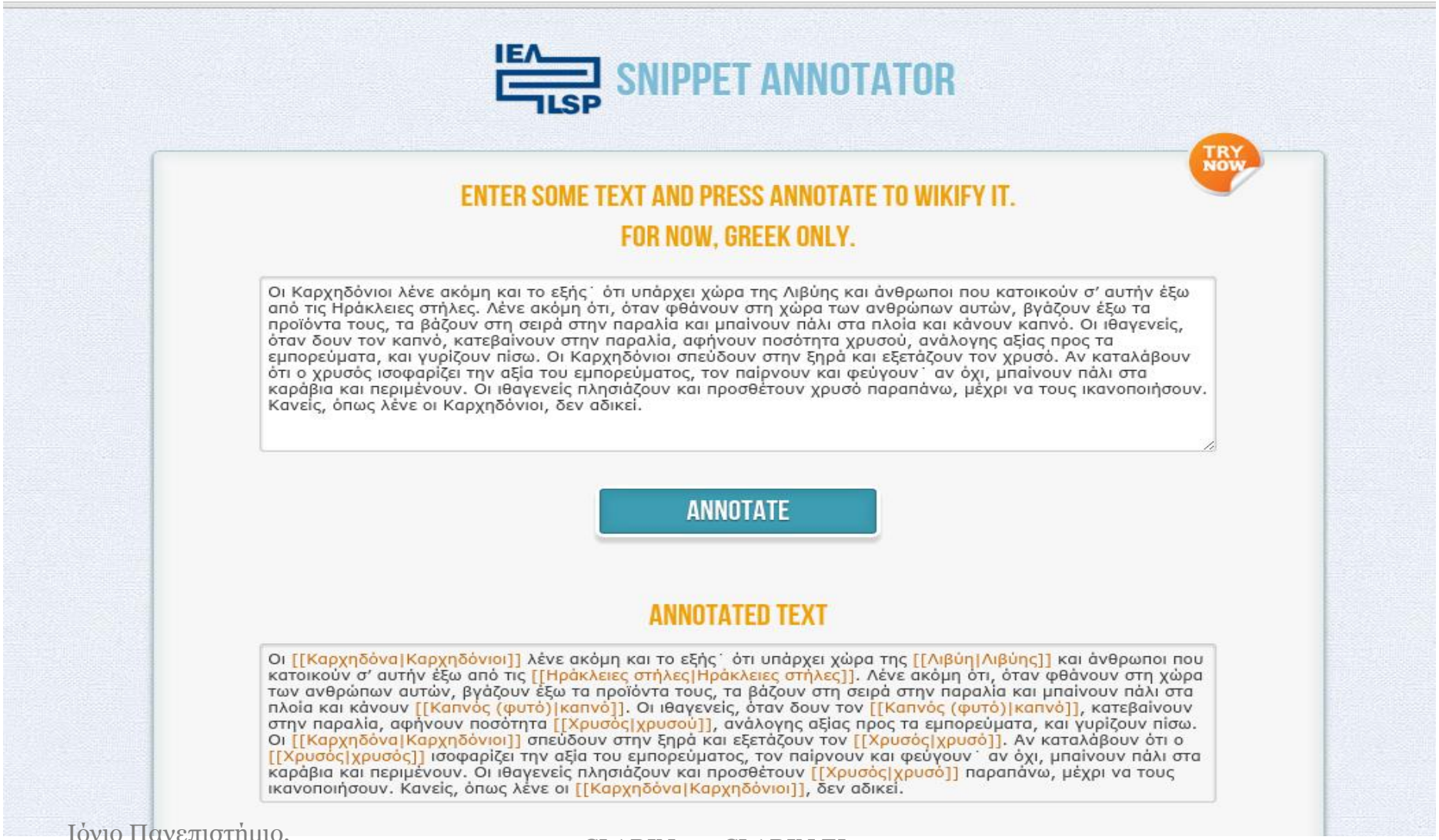
Κάνουμε καλύτερη την αναζήτηση…



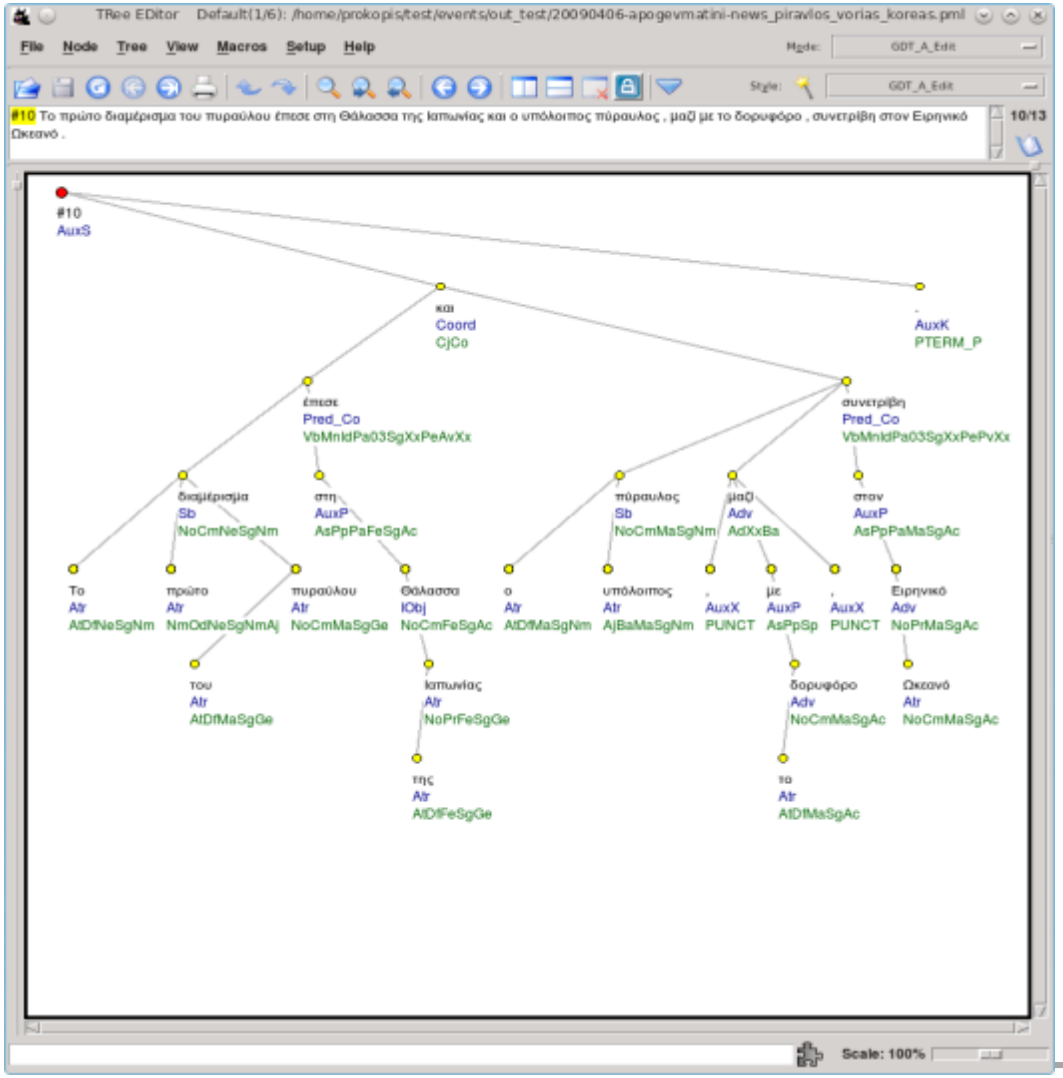
Αναγνωρίζουμε οντότητες και τα ονόματά τους

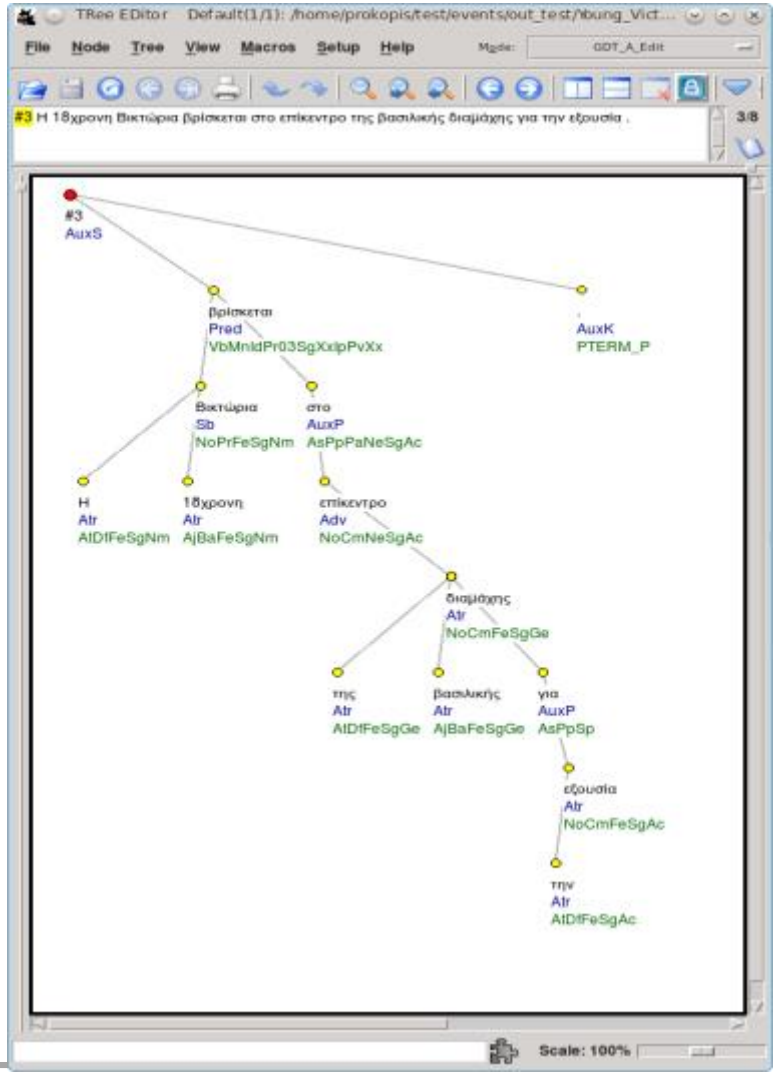


Συνδέουμε οντότητες και έννοιες με την wikipedia



Αναλύουμε συντακτικά προτάσεις ...





Για να αναγνωρίσουμε γεγονότα στα κείμενα

Ανεμοστρόβιλοι χτυπούν τις νοτιοανατολικές πολιτείες των ΗΠΑ.

Event=χτυπούν
Tense: PRESENT
Arguments=
1:(Sb) Ανεμοστρόβιλοι
2:(Obj) τις νοτιοανατολικές πολιτείες των ΗΠΑ

Από τους σίφωνες επλήγησαν οι πολιτείες Ιλλινόις , Ιντιάνα , Κεντάκυ , Τενεσί , Οχάιο , Τζόρτζια , Αλαμπάμα , Μισισίπι και βόρεια και νότια Καρολίνα .

Event=επλήγησαν
Tense: PAST
Arguments=
1:(OTHER) Από τους σίφωνες
2:(Sb) οι πολιτείες Ιλλινόις , Ιντιάνα , Κεντάκυ , Τενεσί , Οχάιο , Τζόρτζια , Αλαμπάμα , Μισισίπι και βόρεια και νότια Καρολίνα

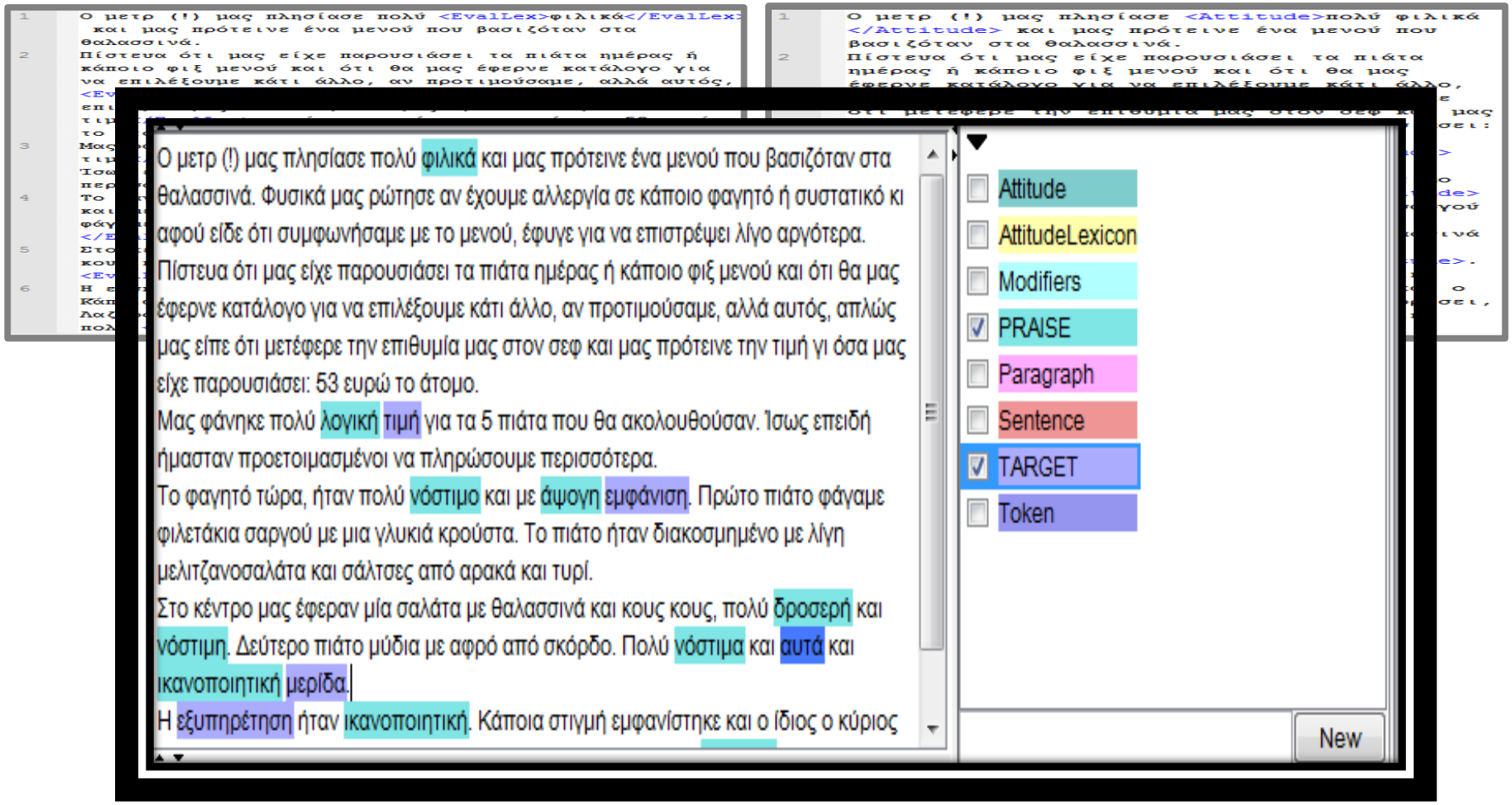
…και πιο πολύπλοκες δομές

Ο Πρόεδρος των ΗΠΑ , Μπαράκ Ομπάμα , δήλωσε ότι « η Ομοσπονδιακή Υπηρεσία Αντιμετώπισης Εκτάκτων Καταστάσεων θα προσφέρει βοήθεια σε κάθε πολιτεία εάν αυτό θεωρηθεί αναγκαίο » .

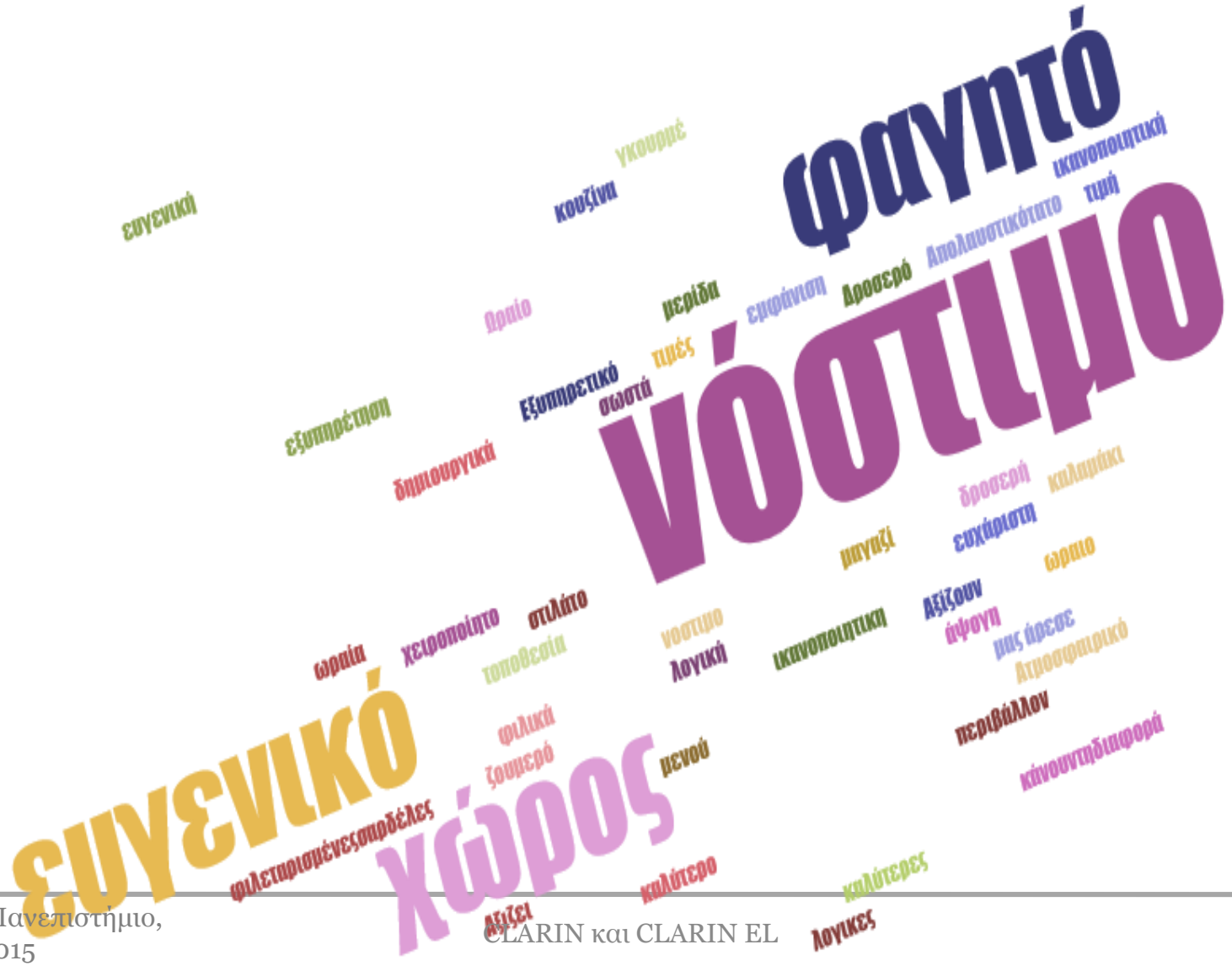
Event=δήλωσε
Tense: PAST
Arguments=
1:(Sb) Ο Πρόεδρος των ΗΠΑ , Μπαράκ Ομπάμα ,
2:(Obj) ότι « η Ομοσπονδιακή Υπηρεσία Αντιμετώπισης Εκτάκτων Καταστάσεων θα προσφέρει βοήθεια σε κάθε πολιτεία εάν αυτό θεωρηθεί αναγκαίο »

Event=προσφέρει
Tense: FUTURE
Arguments=
1:(Sb) η Ομοσπονδιακή Υπηρεσία Αντιμετώπισης Εκτάκτων Καταστάσεων
2:(Obj) βοήθεια
3:(OTHER) σε κάθε πολιτεία
4:(OTHER) εάν αυτό θεωρηθεί αναγκαίο

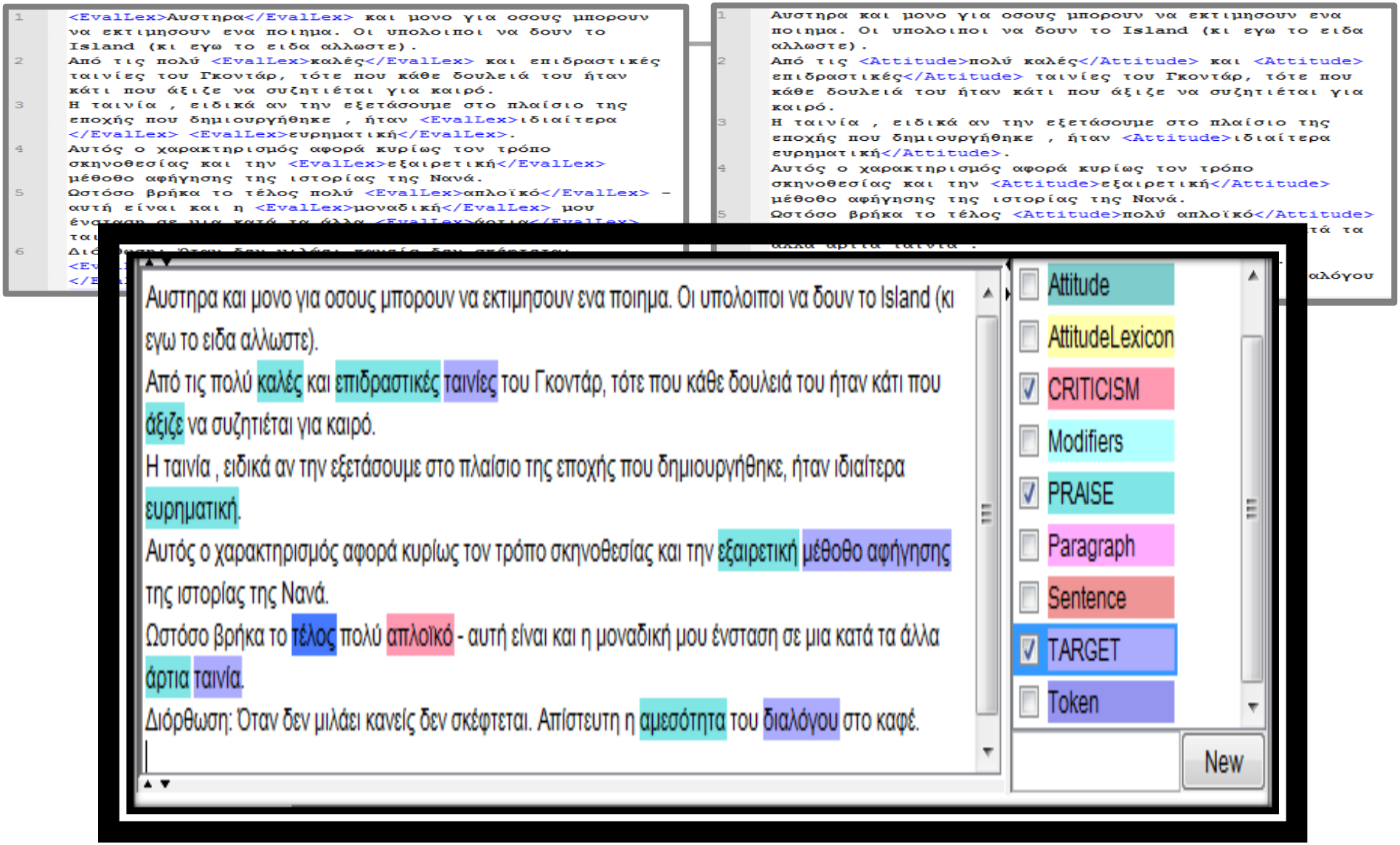
Εντοπίζουμε απόψεις -
π.χ. πόσο καλό είναι ένα εστιατόριο;



...ή πιο παραστατικά



ή πόσο καλή ήταν μια ταινία



Κι αν πάμε να μεταφράσουμε…

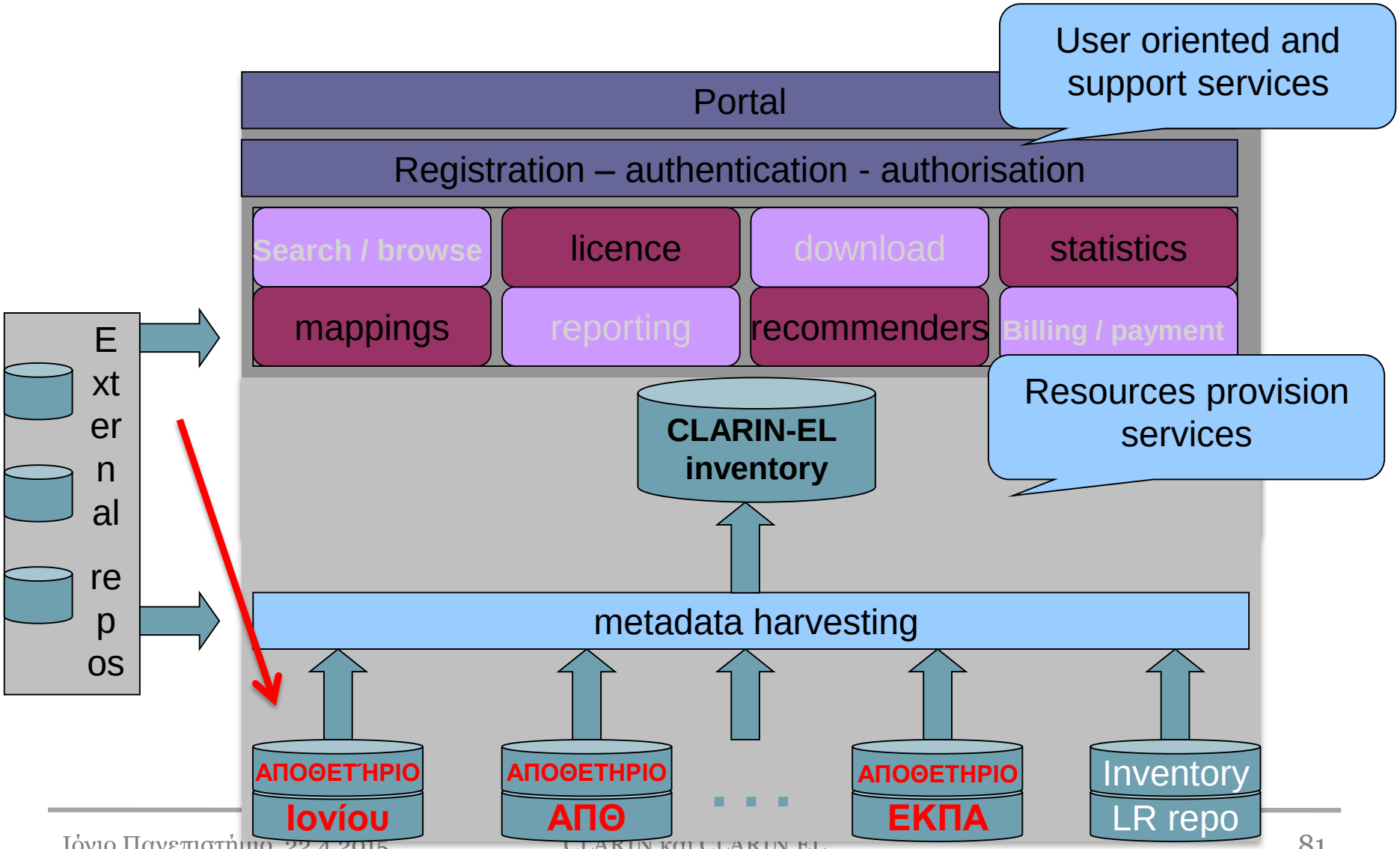
...προς τα ελληνικά

Αγγλικά	Ελληνικά	OK ?
The aim of the Summit is to look at what has been achieved so far in making consumer policy fit for the digital age – including looking at national best practices - and what remains to be done to tackle emerging challenges.	στόχος της συνόδου είναι να δούμε τι έχει επιτευχθεί μέχρι στιγμής για την πολιτική των καταναλωτών για την ψηφιακή εποχή - συμπεριλαμβανομένων των εθνικών ορθών πρακτικών - και αυτό που μένει να γίνει για να αντιμετωπίσει νέες προκλήσεις.	√
Provided that rules on the protection of personal data, when applicable, are complied with, data, once recorded, can be re-used many times without loss of fidelity.	υπό την προϋπόθεση ότι οι κανόνες για την προστασία των δεδομένων προσωπικού χαρακτήρα, κατά περίπτωση, να τηρούνται, τα δεδομένα, μια φορά, να μπορούν να επαναχρησιμοποιηθούν πολλές φορές χωρίς απώλεια της πιστότητας.	X

...προς τα αγγλικά

Ελληνικά	Αγγλικά	OK?
Η Κυβέρνηση απολαύει της εμπιστοσύνης της Βουλής αν η πρόταση εμπιστοσύνης εγκριθεί από την απόλυτη πλειοψηφία των παρόντων Βουλευτών, η οποία όμως δεν επιτρέπεται να είναι μικρότερη από τα δύο πέμπτα (2/5) του όλου αριθμού των Βουλευτών.	the government has the confidence of the house if the motion of confidence is approved by the absolute majority of the members present, but this may not be less than two fifths (2 / 5) of the total number of mps.	√
Κάθε ανεξάρτητη αρχή, συνταγματικά κατοχυρωμένη ή συσταθείσα με νόμο, υποβάλλει στον Πρόεδρο της Βουλής, μέχρι την 31η Μαρτίου κάθε έτους, έκθεση πεπραγμένων για το έργο της κατά το προηγούμενο έτος.	each of the independent authority, constitutionally guaranteed or created by law, •the president of the house, until 31 march • each year, a report on the activities for the work of the previous year.	X

Το Ιόνιο Πανεπιστήμιο στο CLARIN EL


User oriented and support services
Portal
Registration – authentication - authorisation
Search / browse
licence
download
statistics
mappings
reporting
recommenders
Billing / payment
Resources provision services
CLARIN-EL inventory
External repos
metadata harvesting
ΑΠΟΘΕΤΗΡΙΟ Ιονίου
ΑΠΟΘΕΤΗΡΙΟ ΑΠΘ
ΑΠΟΘΕΤΗΡΙΟ ΕΚΠΑ
Inventory LR repo

Εμπλουτισμός του Αποθετηρίου του Ιονίου Παν/μίου

Στοιχεία ταυτότητας πόρου

όνομα	
ακρωνύμιο (αν υπάρχει)	
περιγραφή (σύντομη περιγραφή σε ελεύθερο κείμενο)	
μέγεθος πόρου (σε όποια μονάδα εξυπηρετεί: bytes, λέξεις, λεπτά...)	
γλώσσες που περιλαμβάνονται	
μορφότυπος αρχείων (txt, doc, xml, tmx, mp4, avi, ...)	

Τύπος πόρου

γραπτά κείμενα	
προφορικός λόγος (μεταγραμμένος)	
ήχος	
βίντεο	
λεξικό / γλωσσάρι / θησαυρός	
άλλο (να διευκρινιστεί)	

Στοιχεία διάθεσης πόρου

διαθεσιμότητα (π.χ. ελεύθερα διαθέσιμος, διαθέσιμος με περιορισμούς (ποιους))	
τύπος άδειας (εάν έχει ήδη, π.χ. CC-BY...)	
όνομα κατόχου πνευματικών δικαιωμάτων	

Στοιχεία υπεύθυνου για επικοινωνία

ονοματεπώνυμο	
θέση / ιδιότητα	
σχολή / τμήμα / φορέας	
τηλέφωνο	
email	

Το CLARIN EL στο παρόν και μελλοντικό οικοσύστημα υποδομών

Apollonis

- Από τον Νοέμβριο 2014, το CLARIN EL έχει ενταχθεί, με τη νέα ονομασία CLARITAS στον Εθνικό Χάρτη Ερευνητικών Υποδομών 2014-2020
- από κοινού με την ΕΥ DARIAH-ΔΥΑΣ, τη εθνική υποδομή για Έρευνα στις Ανθρωπιστικές Επιστήμες, σχημάτισε την κοινοπραξία Apollonis
- με στόχο την ενίσχυση της διεπιστημονικής έρευνας, την υποστήριξη της μετάβασης των Κοινωνικών και Ανθρωπιστικών επιστημών στην ψηφιακή έποχή
- Την υποστήριξη του πολιτισμού και της πολιτιστικής κληρονομιάς

Το CLARIN EL στο οικοσύστημα ΕΥ

- Το CLARIN EL αποτελεί μέρος ενός ευρύτερου οικοσυστήματος ερευνητικών υποδομών και συνεργάζεται στενά με
  - META-NET
  - OPENAIRE
  - OPENMINTED
  - LANGUAGE GRID
  - LAPPS GRID

Προκλήσεις σήμερα και αύριο…

- Πρότυπα κωδικοποίησης και αναπαράστασης γλωσσικών πόρων και γλωσσικής γνώσης με σκοπό τη μεγιστοποίηση της διαλειτουργικότητας μεταξύ τους
- Προτυποποίηση και κατά το δυνατόν απλοποίηση του κύκλου εκκαθάρισης πνευματικών δικαιωμάτων
- Μηχανισμοί διάδοσης και ενίσχυσης της κουλτούρας διαμοιρασμού πόρων, εργαλείων και ερευνητικών αποτελεσμάτων
- Οργανωτικά, επιχειρησιακά και επιχειρηματικά μοντέλα για τη βιωσιμότητα των εθνικών και ευρωπαϊκών υποδομών γλωσσικής τεχνολογίας

E/A

Ευχαριστώ πολύ!