



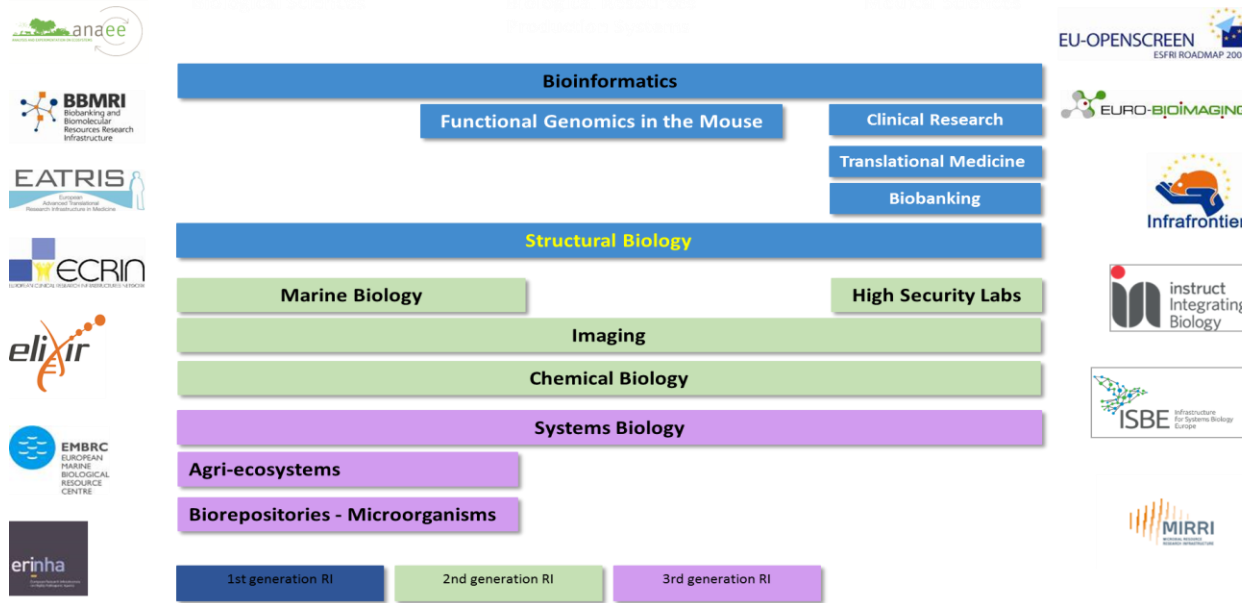
# Ο ρόλος των (επιστημονικών) δεδομένων

# Η δύναμη των δεδομένων

---

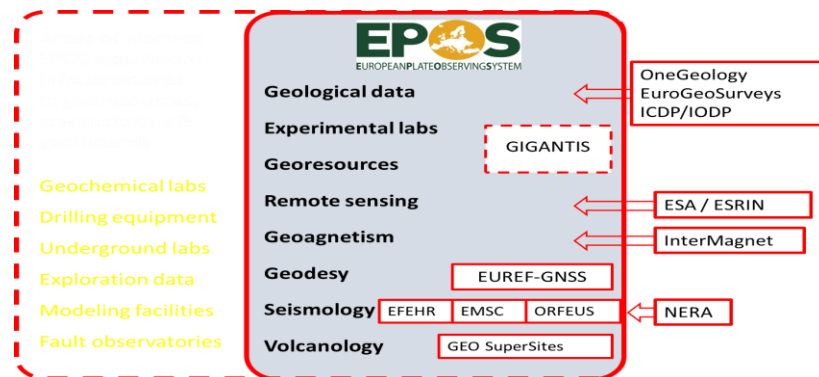
- ◆ Τα (επιστημονικά) δεδομένα μεταμορφώνουν ήδη και βελτιώνουν με ταχύτατους ρυθμούς τη ζωή μας
  - Βιολογικά, γεωγραφικά, μετεωρολογικά, οικονομικά, κ.ά. δεδομένα
- ◆ Η εμπειρία από πολλά επιστημονικά πεδία δείχνει ότι τα δεδομένα και οι υπηρεσίες επεξεργασίας τους αποκτούν ιδιαίτερη αξία όταν διαμοιράζονται 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

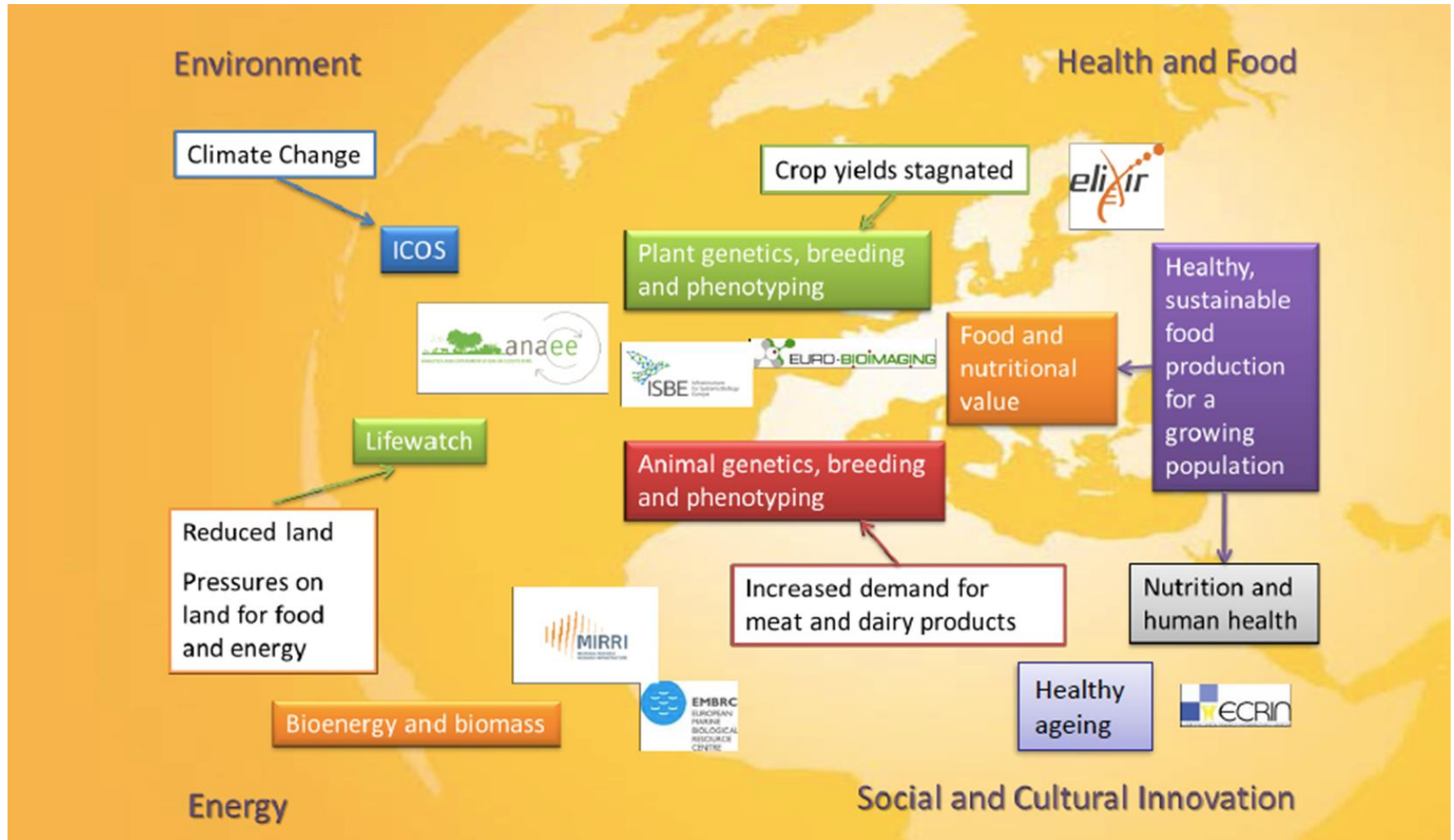


στις επιστήμες  
ζωής και υγείας

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 X, με επισημείωση σε επίπεδο Ψ
  - Σώμα παράλληλων κειμένων μεταξύ Ελληνικών και γλώσσας X,
  - 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](http://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;

---

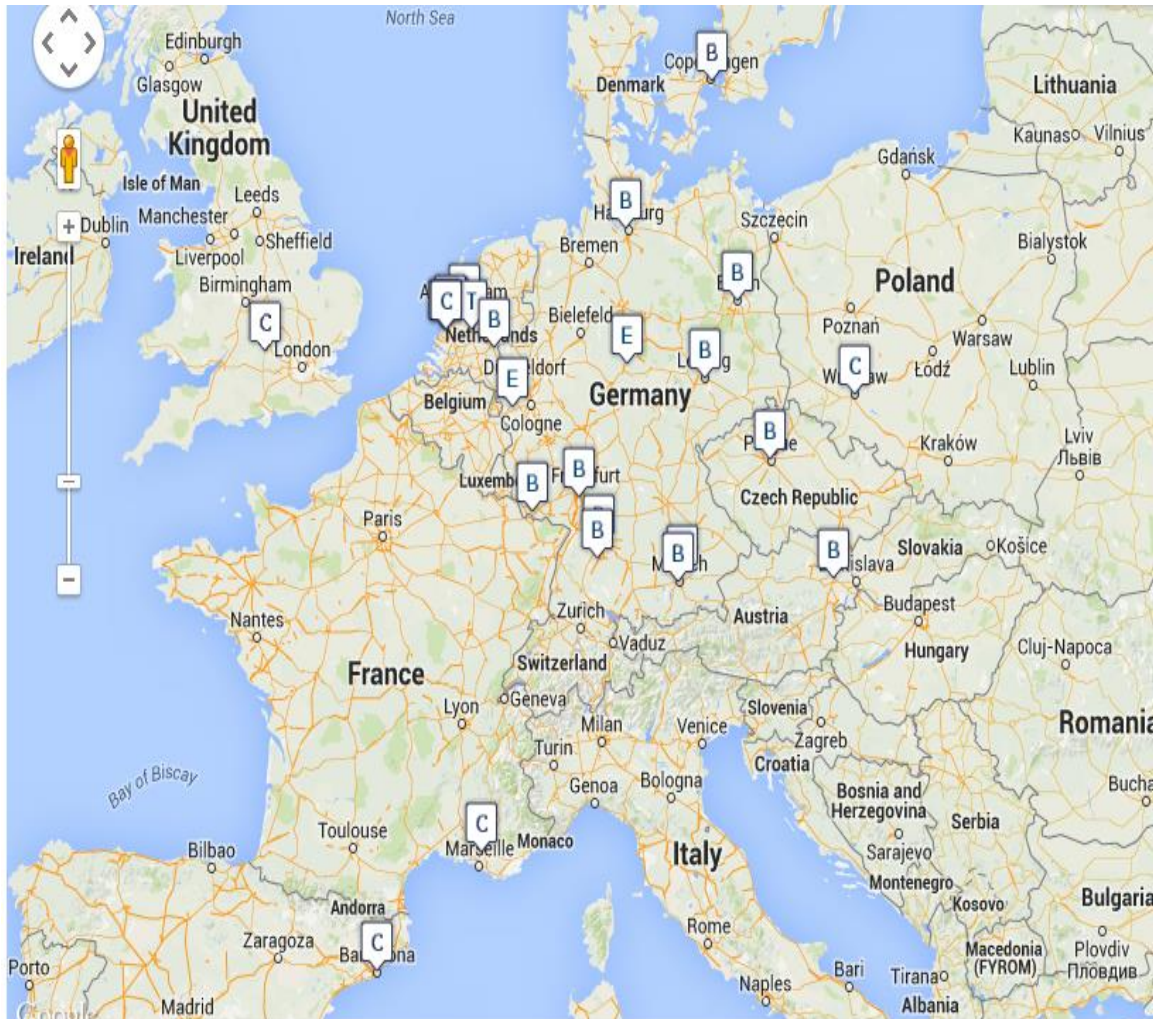
- ◆ μεγάλη αποσπασματικότητα στα έργα του χώρου
- ◆ σχετικές δράσεις παραμένουν άγνωστες
- ◆ απουσία διαλειτουργικότητας
- ◆ μικρή βιωσιμότητα αποτελεσμάτων
- ◆ ύπαρξη τεχνογνωσίας και εμπειρίας αλλά όχι σε όλα τα κράτη
- ◆ εργαλεία ανεξάρτητα γλώσσας μπορούν άμεσα να χρησιμοποιηθούν για άλλους πόρους
- ◆ 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





# Η Ελληνική Υποδομή CLARIN

# Ο στόχος μας

---

- ◆ να ταιριάζουμε 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 ταιριάζουμε 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»

The screenshot shows the CLARIN-EL website interface. At the top, there is a navigation bar with a home icon and several menu items: 'Browse Resources', 'Manage Resources', 'Administration', 'Community', 'Statistics', 'Help', 'About', and 'Your Profile, Juli'. A 'Logout' button is located in the top right corner. Below the navigation bar, the CLARIN-EL logo is displayed on the left, and a large button labeled 'Go to the CLARIN EL Aggregator' is on the right. In the center, it states '66 language resources at your disposal'. Below this, there is a search input field containing the text 'Greek corpus law' and a 'Search' button. The main content area below the search field is titled 'What is it? - About the project' and contains two paragraphs of text describing the project's goals and scope.

# 2. Φυλλομέτρηση αποτελεσμάτων

The screenshot shows the CLARIN-EL search interface. At the top, there is a navigation bar with a home icon, a search bar containing 'Greek corpus law', and a 'Search' button. Below the search bar, there are several filter categories: Language, Resource Type, Media Type, Availability, Licence, Restrictions of Use, Validated, Foreseen Use, Use Is NLP Specific, Linguality Type, Multilinguality Type, Modality Type, MIME Type, Conformance to Standards/Best Practices, Domain, Geographic Coverage, and Time Coverage. The main content area displays '66 Language Resources (Page 1 of 4)' with navigation links for 'Previous' and 'Next', and an 'Order by' dropdown set to 'Resource Name A-Z'. Four resource cards are visible, each with a title, a list of languages, and icons for download, view, and delete. The first card is 'ACCURAT balanced test corpus for under resourced languages' with languages: Croatian, English, Estonian, German, Greek, Modern (1453-), Latvian, Lithuanian, Romanian, Slovenian. The second is 'ACCURAT corpus of Wikipedia texts' with languages: Croatian, English, Estonian, German, Greek, Modern (1453-), Latvian, Lithuanian, Romanian, Slovenian. The third is 'A parallel corpus collected from the European Constitution' with languages: Czech, Danish, Dutch, Flemish, English, Estonian, Finnish, French, German, Greek, Modern (1453-), Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovenian, Spanish, Castilian, Swedish. The fourth is 'A parallel corpus of KDE4 localization files (v.2)' with a long list of languages including Afrikaans, Arabic, Armenian, Assamese, Asturian, Babel, Leonese, Asturianese, Basque, Belarusian, Bengali, Bokmål, Norwegian, Norwegian Bokmål, Breton, Bulgarian, Catalan, Valencian, Central Khmer, Chinese, Crimean Tatar, Crimean Turkish, Croatian, Czech, Danish, Dutch, Flemish, English, Esperanto, Estonian, Finnish, French, Galician, Georgian, German, Greek, Modern (1453-), Gujarati, Hausa, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Kannada, Kashubian, Kazakh, Kinyarwanda, Korean, Kurdish, Latvian, Lithuanian, Low German, Low Saxon, German, Low, Saxon, Low, Luxembourgish, Letzeburgesch, Macedonian, Maltese, Malay, Malayalam, Maltese, Marathi, Nepali, Northern Sami, Norwegian Nynorsk, Nynorsk, Norwegian, Occitan (post 1500), Oriya, Panjabi, Punjabi, Pedi, Sepedi, Northern Sotho, Persian, Polish, Portuguese, Pushto, Pashto, Romanian, Russian, Serbian, and others.

## 2α. Αναζήτηση με χρήση φίλτρων

The screenshot shows the CLARIN-EL search interface. At the top, there is a navigation bar with a home icon and a 'Logout' button. Below the navigation bar are several menu items: 'Browse Resources', 'Manage Resources', 'Administration', 'Community', 'Statistics', 'Help', 'About', and 'Your Profile, Juli'. The main content area features a search bar with a 'Search' button. Below the search bar, there are two columns. The left column is titled 'Filter by:' and contains three filter sections: 'Language' (with a dropdown arrow), 'Resource Type' (with a dropdown arrow), and 'Domain' (with a dropdown arrow). Under 'Language', there is a red minus sign and 'Greek, Modern (1453-) (2)'. Under 'Resource Type', there is a red minus sign and 'Corpus (2)', with two sub-items: 'Annotation Type' and 'Annotation Format'. Under 'Domain', there is a red minus sign and 'Law (2)', a green plus sign and 'Politics (1)', and a green plus sign and 'Travel (1)'. The right column is titled '2 Language Resources' and has an 'Order by: Resource Name A-Z' dropdown menu. It displays two resource cards. The first card is for 'Greek Textual Entailment Corpus' with a document icon, a red arrow pointing down, '0', an eye icon, and '1'. Below the title is a tag 'Greek, Modern (1453-)'. The second card is for 'OROSSIMO Corpus -Law' with a document icon, a red arrow pointing down, '0', an eye icon, and '1'. Below the title is a tag 'Greek, Modern (1453-)'. The background of the interface is light gray.



# 3. Επιλογή του πόρου προτίμησης

Home
Logout

**CLARIN-EL**

Browse Resources
Manage Resources
Administration
Community
Statistics
Help
About
Your Profile, Juli

## OROSSIMO Corpus -Law 👁️ 1 ✓ Z

▶ View resource name in all available languages

PID: <http://hdl.gnnet.gr/11239/ATHENA-TEST-0000-0000-1702-4-TEST>

A corpus compiled of texts belonging to the Law domain, according to the XCES standard. There is also a terminological resource (see Orossimo terminological resource - Law)

▶ View resource description in all available languages

« Back
Download
Edit Resource

**Distribution**

**Availability**

Available - Restricted Use

**Licence**

**CC - BY**

**Distribution Access/Medium:**

Downloadable

**Attribution Details:** Orossimo Corpus by Athena R.C./ILSP used under CC-BY licence

**Contact Person**

**Penny Labropoulou**

**text**

**Monolingual text corpus**

**Languages**

Greek, Modern (1453-)

**Linguality**

Linguality type: Monolingual

**Size**

2,500,000 Words

**Character encoding**

UTF - 8

**Domains**

law

**Resource Creation**

**Resource Creator**

Institute for Language and Speech Processing / Athena R.C.

**Creation lasted:** 01/01/1996 - 12/31/1998

**Funding Project**

**OROSSIMO**

**URL:** <http://www.ilsp.gr/e...>

**Funding Type:** National Funds

**Funding Country:** Greece

**Project duration:** 04/01/1996 - 12/31/1998

**Metadata**



# 4. Όροι χρήσης - αδειοδότηση

🏠
Logout

**CLARIN-EL**

Browse Resources
Manage Resources
Administration
Community
Statistics
Help
About
Your Profile, Juli

## OROSSIMO Corpus -Law

Licence Agreement – CC-BY

Creative Commons Corporation (“Creative Commons”) is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an “as-is” basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

**Creative Commons Attribution 4.0 International Public License**

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License (“Public License”). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

**Section 1 – Definitions.**

I agree to these licence terms and would like to download the resource.

Download

« Resource View
Contact Resource Maintainer

# 5. Καταφόρτωση του πόρου

The screenshot shows the CLARIN-EL website interface. At the top, there is a navigation bar with a home icon and buttons for 'Browse Resources', 'Manage Resources', 'Administration', 'Community', 'Statistics', 'Help', 'About', and 'Your Profile. Juli'. A 'Logout' button is in the top right corner. The main content area is titled 'OROSSIMO Corpus -Law' and contains a 'Licence Agreement - CC-BY' section. The text in this section states: 'Creative Commons Corporation ("Creative Commons") is not a law firm or legal services or legal advice. Distribution of Creative Commons public licenses does not create any other relationship. Creative Commons makes its licenses and related information available on an "as is" basis. Creative Commons gives no warranties regarding its licenses, any other intellectual property rights, or any related information. Creative Commons disclaims all liability from their use to the fullest extent possible.'

Below this is the 'Creative Commons Attribution 4.0 International Public License' section, which begins with: 'By exercising the Licensed Rights (defined below), You accept and agree to the following conditions of this Creative Commons Attribution 4.0 International Public License. In that you are considering the License in light of the consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.'

Underneath is 'Section 1 - Definitions.' and a checkbox that is checked: 'I agree to these licence terms and would like to download the resource.' Below the checkbox are three buttons: 'Download', '<< Resource View', and 'Contact Resource Maintainer'.

Overlaid on the right side of the page is a Windows-style 'Save' dialog box titled 'Εισάγετε το όνομα αρχείου για αποθήκευση...'. The 'Save in' field shows 'working files'. The left sidebar lists 'My Recent Documents', 'Desktop', 'My Documents', 'My Computer', and 'My Network'. The 'File name' field contains 'archive.zip' and the 'Save as type' is set to 'WinZip File'. 'Save' and 'Cancel' buttons are at the bottom right of the dialog.

**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.

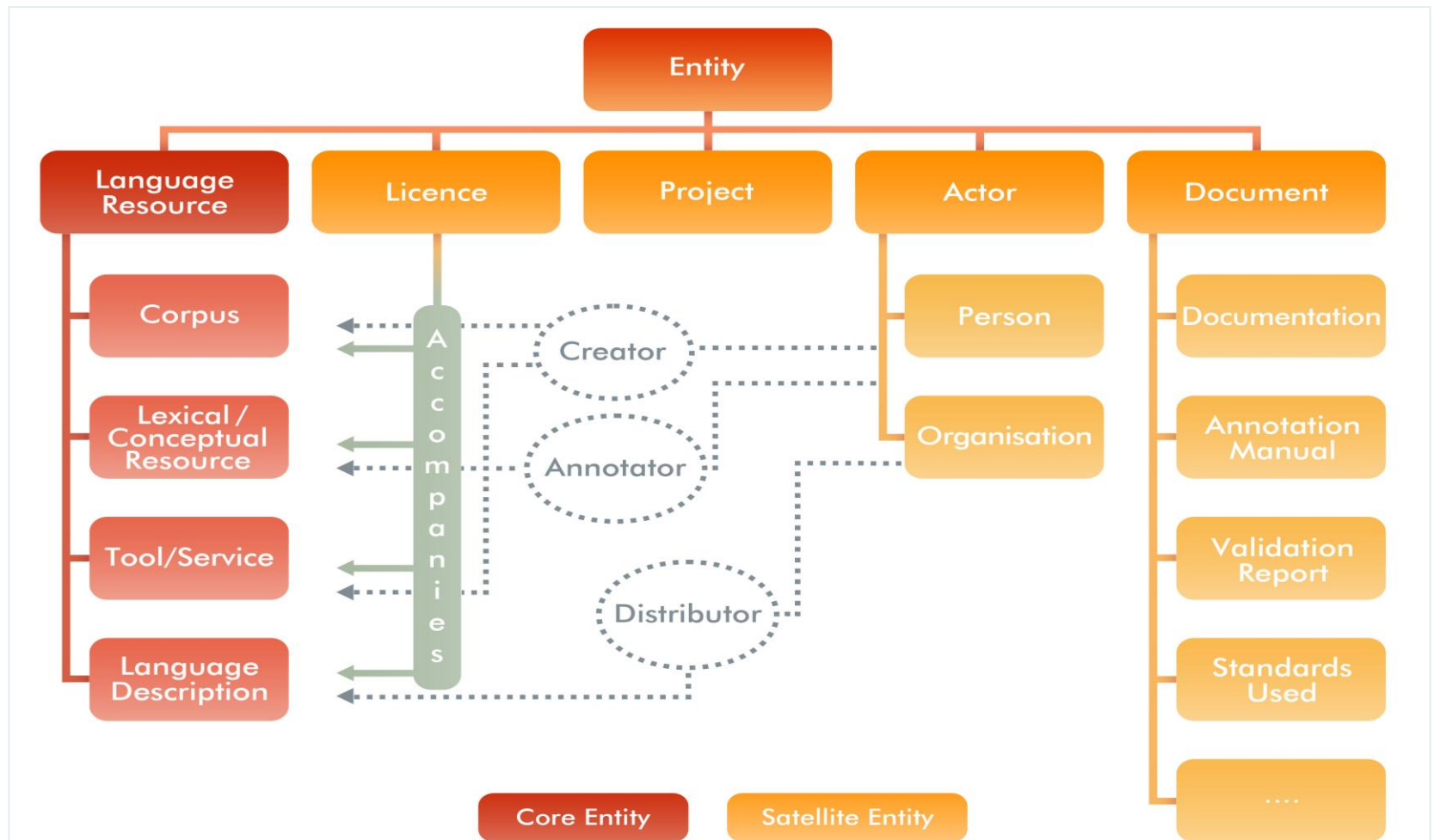
# Σχήμα μεταδεδομένων - οντολογία

---

## ◆ 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)

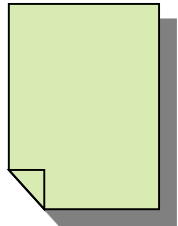
---

## ◆ Τυπολογία πόρων

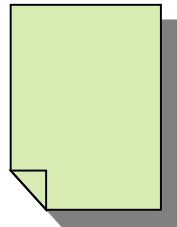
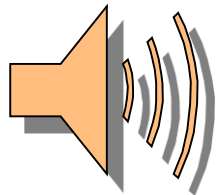
- ταξινόμηση βάσει 2 κύριων κριτηρίων: "τύπος πόρου" (resource type) & "μέσο πόρου" (media type)
- τύπος πόρου
  - **σώμα κειμένων** (γραπτού / προφορικού λόγου, πολυτροπικών/πολυμεσικών κειμένων)
  - **λεξικός / εννοιολογικός πόρος** (π.χ. λεξικό, ορολογικός πόρος, κατάλογος λέξεων, οντολογία κτλ.)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

# Τυπολογία πόρων (2)

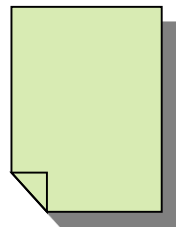
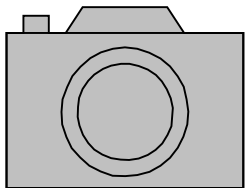
- mediaType: text, audio, image, video



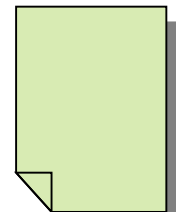
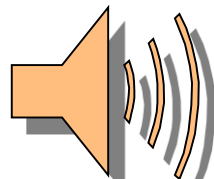
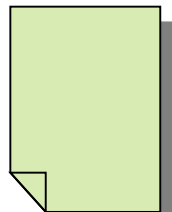
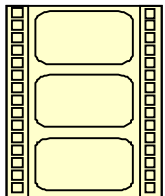
written corpora



spoken corpora



images (multimedia)

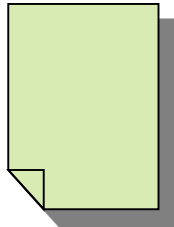


videos  
(multimedia)

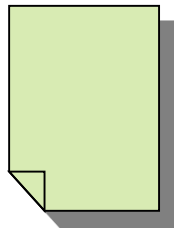


# Τυπολογία πόρων (3)

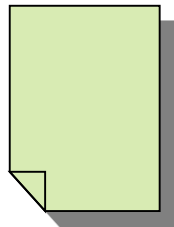
- search for text



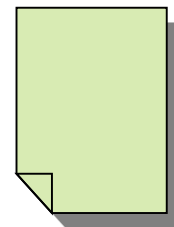
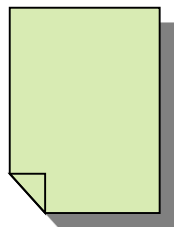
written corpora



spoken corpora

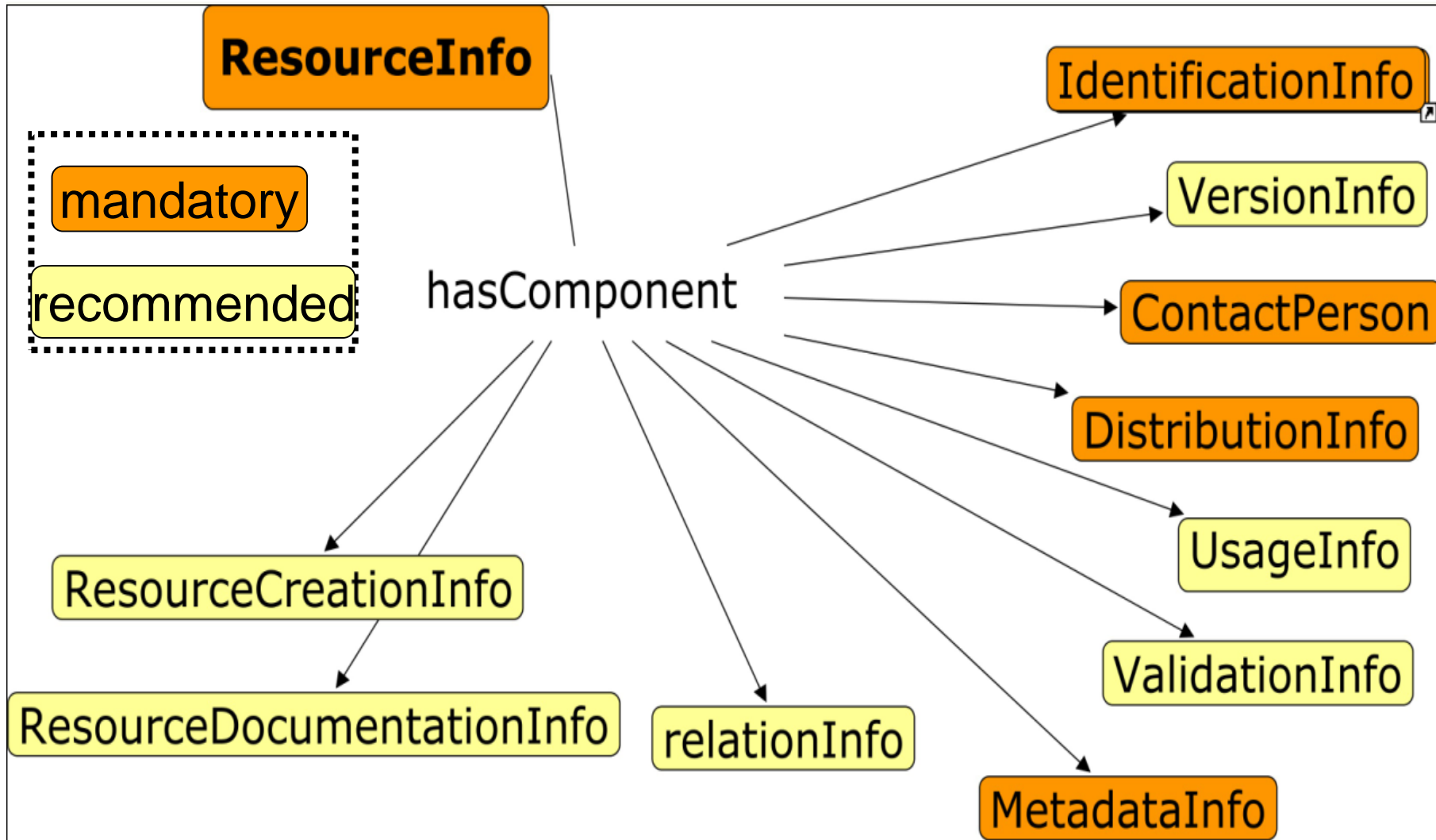


images (multimedia)

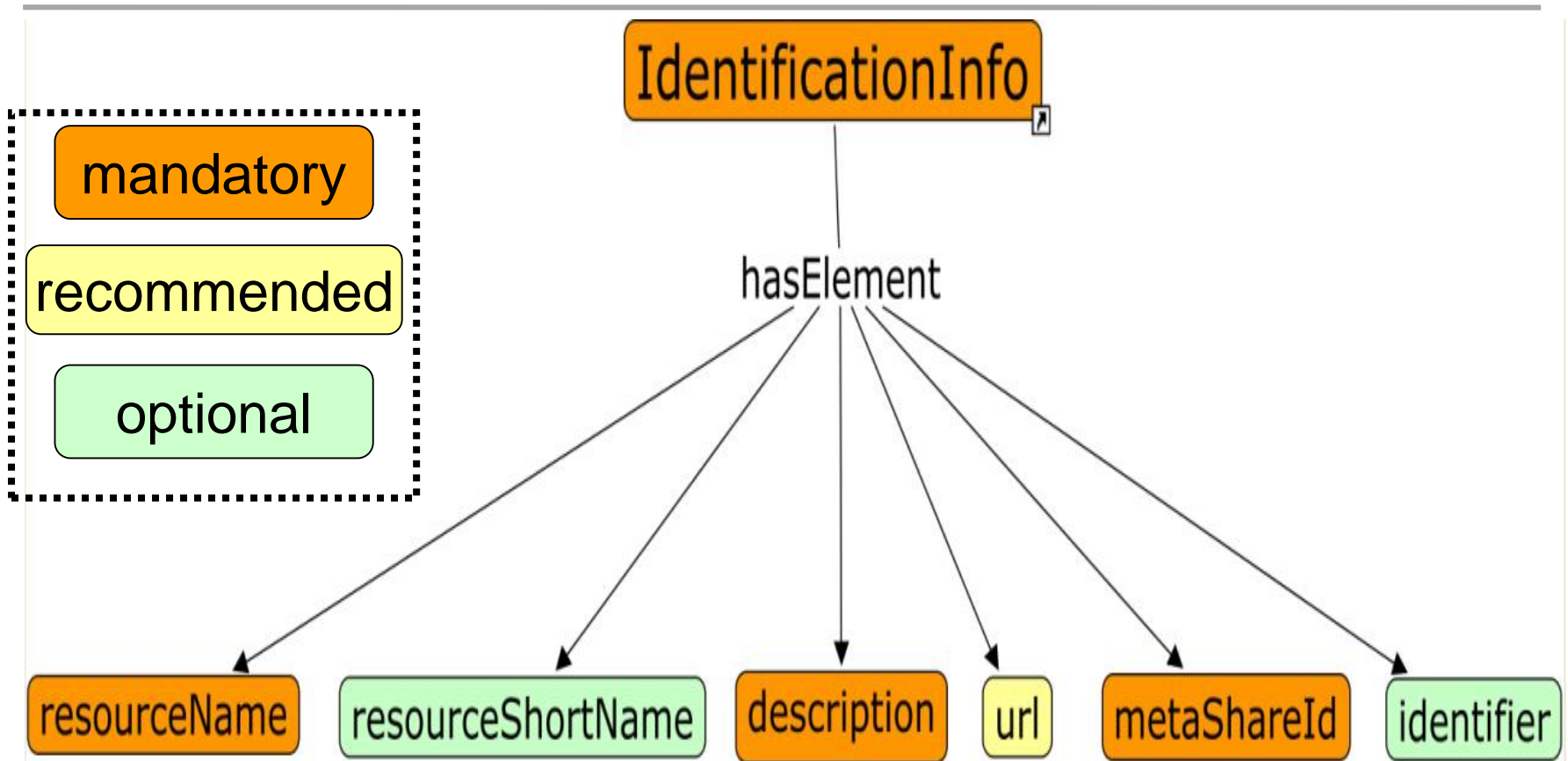


videos  
(multimedia)

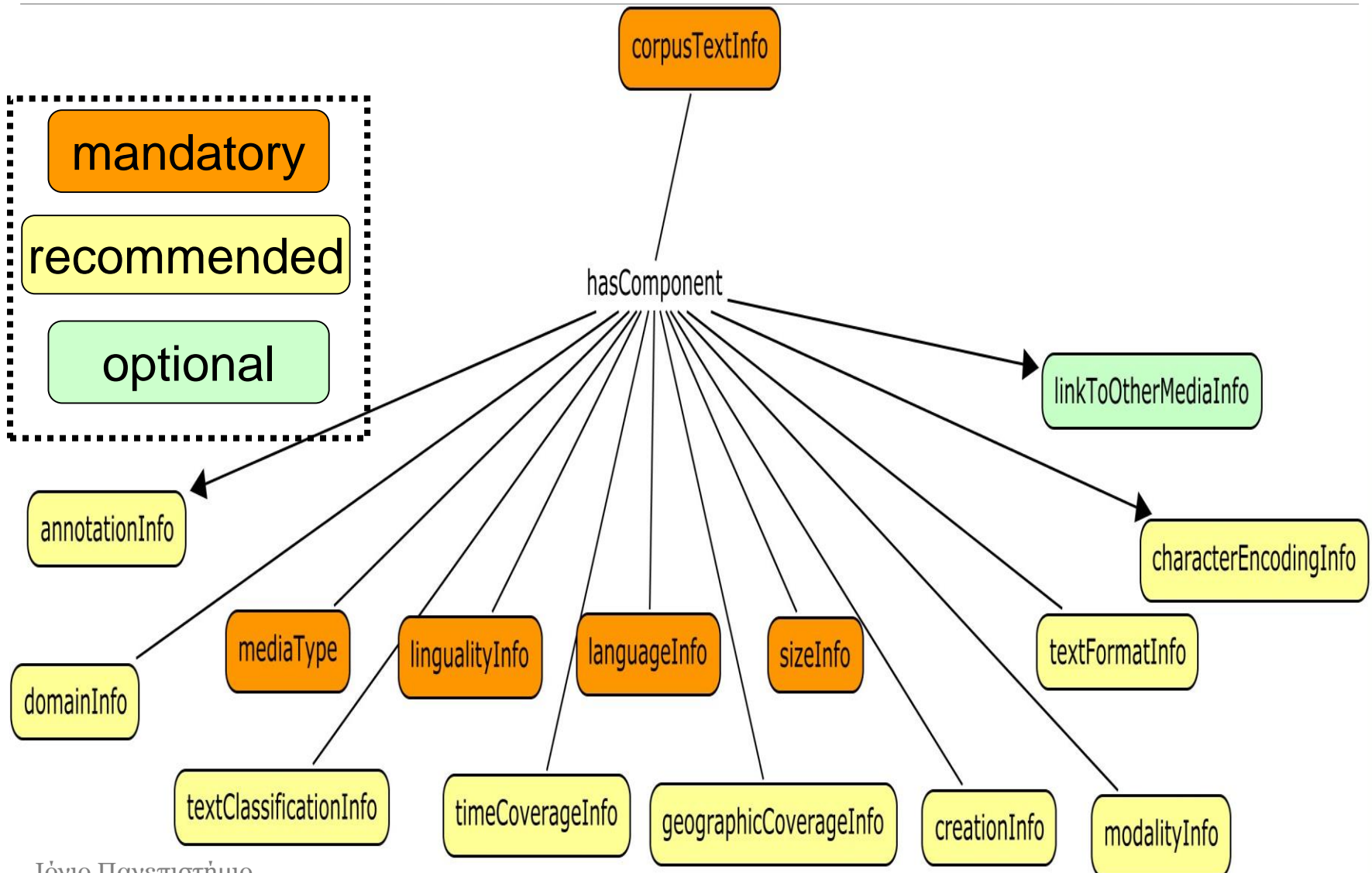
# Το βασικό συστατικό περιγραφής



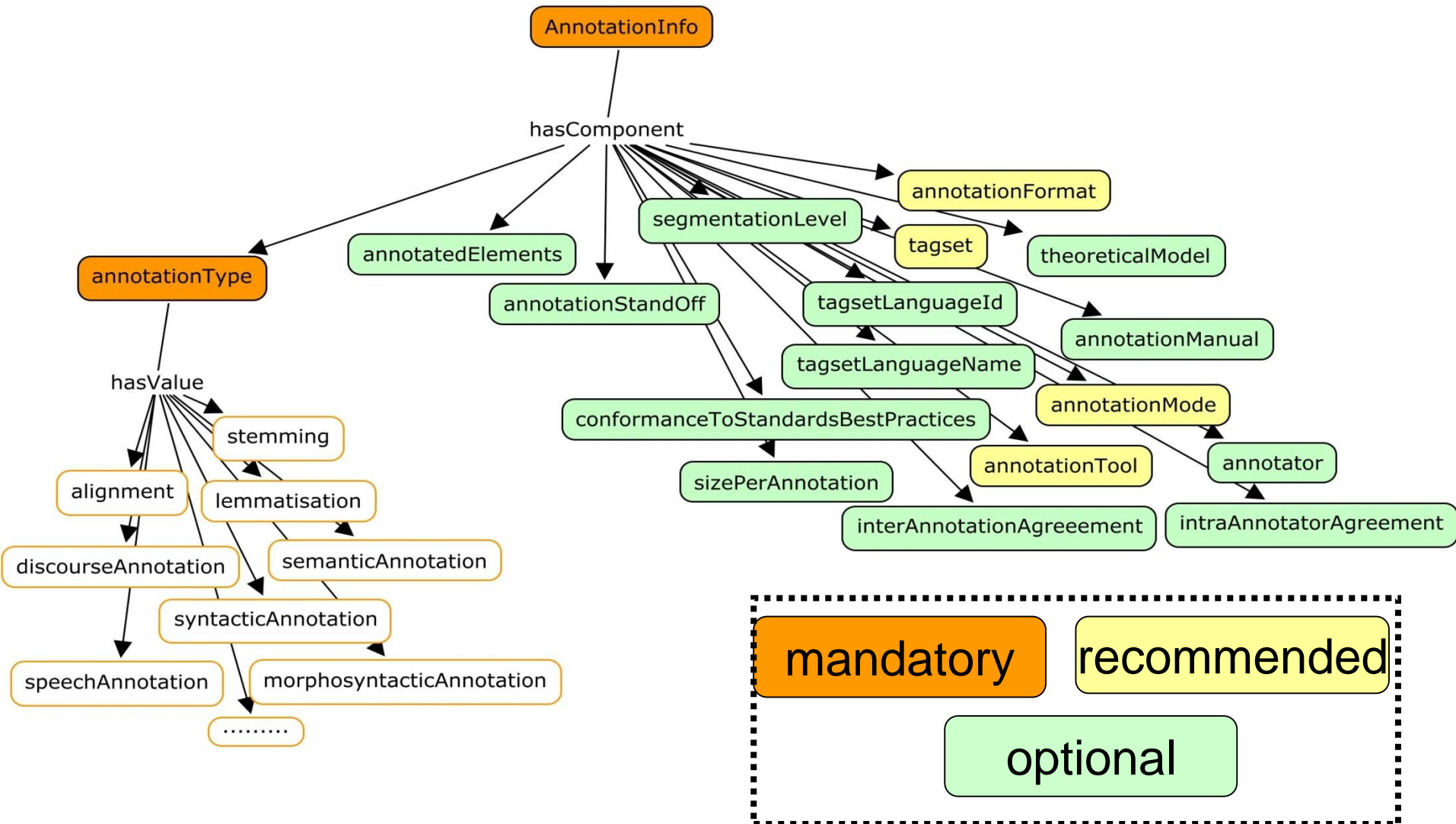
# Identification component



# corpusTextInfo



# Annotation Component



**...και σε πιο εύχρηστη μορφή**

# Τεκμηρίωση μονογλωσσικού ΣΚ

CLARIN-EL

[Logout](#)

[Browse Resources](#)
[Manage Resources](#)
[Administration](#)
[Community](#)
[Statistics](#)
[Help](#)
[About](#)
[Your Profile, Juli](#)

Home > Resources > Resources > Greek Textual Entailment Corpus

## Change Resource ⓘ

Fields marked with \* are required.

Administrative Information

Required

Recommended

Edit Corpus Text Info

Add Corpus Text Info

[Upload resource](#)
[Export Resource Description to XML](#)
[History](#)
[View on site](#)

Required administration information: Identification, Distribution, Contact person, Metadata

Identification ⌵

<b>* Resource name:</b>	* Ελληνικό Σώμα Κειμενικής Συνεπαγωγής GTEC	Language of this entry (RFC 3066 code, preferably from ISO 639-1): <input type="text" value="el"/> (Greek, Modern (1453-))
	* Greek Textual Entailment Corpus	Language of this entry (RFC 3066 code, preferably from ISO 639-1): <input type="text" value="en"/> (English)

+ Add Another Field

The full name by which the resource is known; the element can be repeated for the different language versions using the "lang" attribute to specify the language. ⓘ

\* Description:

*	Το Ελληνικό Σώμα Κειμενικής Συνεπαγωγής (Greek Textual Entailment Corpus, GTEC) αποτελείται από 600 ζεύγη T-H (κείμενο συνεπαγωγής & συνεπαγόμενη υπόθεση) τα οποία έχουν επιστημωθεί ως προς το αν η πρόταση T συνεπύγεται την πρόταση H από ανθρώπινους επιστημωτές σύμφωνα με το σχήμα επιστημωσίας που χρησιμοποιήθηκε στις δοκιμασίες RTE1 & RTE2. Το δεδομένο είναι οργανωμένο σε τρεις υποομάδες, οι οποίες αντιστοιχούν σε τρεις εφομογμές Γλωσσικής Τεχνολογίας (συστήματα Ερωματοκρίσεων, Συγκρίσιμων Αρχείων και Μηχανικής Μετάφρασης), και ανήκουν σε τρεις γνωστικούς τομείς (νομικό, πολιτική & ταξίδια). Στο σώμα περιλαμβάνονται	Language of this entry (RFC 3066 code, preferably from ISO 639-1): <input type="text" value="el"/> (Greek, Modern (1453-))
*	GTEC consists of 600 T-H pairs manually annotated for entailment (i.e. whether T entails H or not) by human annotators. The dataset which is tailored to guide training and evaluation of prospect RTE systems, is equally divided in three subsets each one representing the output of a specific HLT application: Question Answering (QA), Comparable Documents (CD) and Machine Translation (MT), and pertaining to specific subject fields (e.g. law, politics, travel). T-H examples that correspond to success and failure cases of the afore-mentioned applications have been included in the corpus. The annotations provided are conformant to the RTE1 and RTE2 challenges.	Language of this entry (RFC 3066 code, preferably from ISO 639-1): <input type="text" value="en"/> (English)

+ Add Another Field

Provides the description of the resource in prose; the element can be repeated for the different language versions using the "lang" attribute to specify the language. ⓘ

Resource short name:

*	* GTEC	Language of this entry (RFC 3066 code, preferably from ISO 639-1): <input type="text" value="en"/> (English)
---	--------	--

+ Add Another Field

The short form (abbreviation, acronym etc.) used to identify the resource; the element can be repeated for the different language versions using the "lang" attribute to specify the language. ⓘ

Url:

*	
---	--

# Τεκμηρίωση μονογλωσσικού ΣΚ (2)

## Change Corpus text

Fields marked with \* are required.

### Information

Required

Recommended

Optional

### Required information: Linguality, Language, Size

#### Linguality

\* **Linguality type:**  Indicates whether the resource includes one, two or more languages

**Multilinguality type:**  Indicates whether the corpus is parallel, comparable or mixed

**Multilinguality type details:**  Provides further information on multilinguality of a resource in free text

#### Languages

##### Language: Greek, Modern (1453-) ? Delete

\* **Language id:**  The identifier of the language that is included in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

\* **Language name:**  A human understandable name of the language that is used in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

**Language script:**  Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924

**Size per language:**  Provides information on the size per language component

**Language variety:**  Groups information on language varieties occurred in the resource (e.g. dialects)

**Language: #2**

\* **Language id:**  The identifier of the language that is included in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

\* **Language name:**  A human understandable name of the language that is used in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

**Language script:**  Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924



# Τεκμηρίωση πολυγλωσσικού ΣΚ

[Browse Resources](#) [Manage Resources](#) [Administration](#) [Community](#) [Statistics](#) [Help](#) [About](#) [Your Profile, Penny](#)
Logout

CLARIN-EL

Home > Resources > Resources > A parallel subcorpus collected from the European Constitution (EN-EL) (TMX)

**Change Resource**

Fields marked with \* are required.

Upload resource
Export Resource Description to XML
History
View on site

**Administrative Information**

Required
Recommended

Edit Corpus Text Info

Add Corpus Text Info

Required administration information: Identification, Distribution, Contact person, Metadata

**Identification**

<p><b>* Resource name:</b></p> <div style="border: 1px solid #d9d9d9; padding: 5px; min-height: 40px;">A parallel subcorpus collected from the European Constitution (EN-EL) (TMX)</div> <p style="text-align: center; font-size: 0.8em; margin-top: 5px;"><span style="color: green;">+</span> Add Another Field</p> <p style="font-size: 0.7em; margin-top: 5px;">The full name by which the resource is known; the element can be repeated for the different language versions using the "lang" attribute to specify the language. </p>	<p>Language of this entry (RFC 3066 code, preferably from ISO 639-1):</p> <div style="border: 1px solid #d9d9d9; padding: 2px; margin-bottom: 5px;">en</div> <p>(English)</p>
<p><b>* Description:</b></p> <div style="border: 1px solid #d9d9d9; padding: 5px; margin-bottom: 5px;"> <p><span style="color: red;">✗</span> Το EUconst subcorpus EN-EL (TMX) είναι ένα παράλληλο σώμα κειμένων για τα αγγλικά και ελληνικά που αποτελεί υποσύνολο του EUconst, a parallel corpus collected from the European Constitution (ένα παράλληλο σώμα κειμένων με υλικό από το Ευρωπαϊκό Σύνταγμα).</p> </div> <div style="border: 1px solid #d9d9d9; padding: 5px;"> <p><span style="color: red;">✗</span> The EUconst subcorpus EN-EL (TMX) is a parallel subcorpus for English and Greek, subset of the EUconst, a parallel corpus collected from the European Constitution.</p> <p style="font-size: 0.8em; margin-top: 5px;">21 languages, 210 bitexts total number of files: 986</p> </div>	<p>Language of this entry (RFC 3066 code, preferably from ISO 639-1):</p> <div style="border: 1px solid #d9d9d9; padding: 2px; margin-bottom: 5px;">el</div> <p>(Greek, Modern (1453-))</p> <p>Language of this entry (RFC 3066 code, preferably from ISO 639-1):</p> <div style="border: 1px solid #d9d9d9; padding: 2px; margin-bottom: 5px;">en</div> <p>(English)</p>

# Τεκμηρίωση πολυγλωσσικού ΣΚ (2)

Change Corpus text | Clarin EL backend - Mozilla Firefox

athena-test.clarin.gr/editor/resources/corpusinfotype\_model/41/?\_popup=1

## Change Corpus text

Fields marked with \* are required.

**Information**

Required

Recommended

Optional

Required information: Linguality, Language, Size

**Linguality**

\* **Linguality type:**  Indicates whether the resource includes one, two or more languages

**Multilinguality type:**  Indicates whether the corpus is parallel, comparable or mixed

**Multilinguality type details:**  Provides further information on multilinguality of a resource in free text


**Languages**

**Language: English ?** Delete

\* **Language id:**  The identifier of the language that is included in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

\* **Language name:**  A human understandable name of the language that is used in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

# Τεκμηρίωση πολυγλωσσικού ΣΚ (3)


**CLARIN-EL**
[Logout](#)

[Browse Resources](#)
[Manage Resources](#)
[Administration](#)
[Community](#)
[Statistics](#)
[Help](#)
[About](#)
[Your Profile, Penny](#)

---

Resource short name: ✘

[+ Add Another Field](#)

The short form (abbreviation, acronym etc.) used to identify the resource; the element can be repeated for the different language versions using the "lang" attribute to specify the language. [?](#)

Language of this entry (RFC 3066 code, preferably from ISO 639-1):

(English)

---

Url: ✘

[+ Add Another Field](#)

A URL used as homepage of an entity (e.g. of a person, organization, resource etc.) and/or where an entity (e.g.LR, document etc.) is located [?](#)

---

Identifier: ✘

[+ Add Another Field](#)

A reference to the resource like a pid or an internal identifier used by the resource provider [?](#)

# Τεκμηρίωση πολυγλωσσικού ΣΚ (4)

Change Corpus text | Clarin EL backend - Mozilla Firefox

athena-test.clarin.gr/editor/resources/corpusinfotype\_model/41/?\_popup=1

## Change Corpus text

Fields marked with \* are required.

**Information**

Required

Recommended

Optional

Required information: Linguality, Language, Size

**Linguality**

\* **Linguality type:**  Indicates whether the resource includes one, two or more languages

**Multilinguality type:**  Indicates whether the corpus is parallel, comparable or mixed

**Multilinguality type details:**  Provides further information on multilinguality of a resource in free text

**Languages**

Language: English ? Delete

\* **Language id:**  The identifier of the language that is included in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

\* **Language name:**  A human understandable name of the language that is used in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines)

# Τεκμηρίωση πολυγλωσσικού ΣΚ (5)

Change Corpus text | Clarin EL backend - Mozilla Firefox

athena-test.clarin.gr/editor/resources/corpusinfotype\_model/41/?\_popup=1

**Language: Greek, Modern (1453-) ?**  Delete

\* **Language id:**   
 The identifier of the language that is included in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines) ⓘ

\* **Language name:**   
 A human understandable name of the language that is used in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines) ⓘ

**Language script:**   
 Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924 ⓘ

**Size per language:** + Provides information on the size per language component ⓘ

**Language variety:** + Groups information on language varieties occurred in the resource (e.g. dialects) ⓘ

**Language: #3**

\* **Language id:**   
 The identifier of the language that is included in the resource or supported by the tool/service; an autocompletion mechanism with values from the ISO 639 is provided in the editor, but the values can be subsequently edited for further specification (according to the IETF BCP47 guidelines) ⓘ

\* **Language name:**   
 A human understandable name of the language that is used in the resource or supported by the

# Τεκμηρίωση πολυγλωσσικού ΣΚ (6)

Change Corpus text | Clarin EL backend - Mozilla Firefox

athena-test.clarin.gr/editor/resources/corpusinfotype\_model/41/?\_popup=1

Language script:   
Specifies the writing system used to represent the language in form of a four letter code as it is defined in ISO-15924 ⓘ

Size per language: + Provides information on the size per language component ⓘ

Language variety: + Groups information on language varieties occurred in the resource (e.g. dialects) ⓘ

[Add another Language](#)

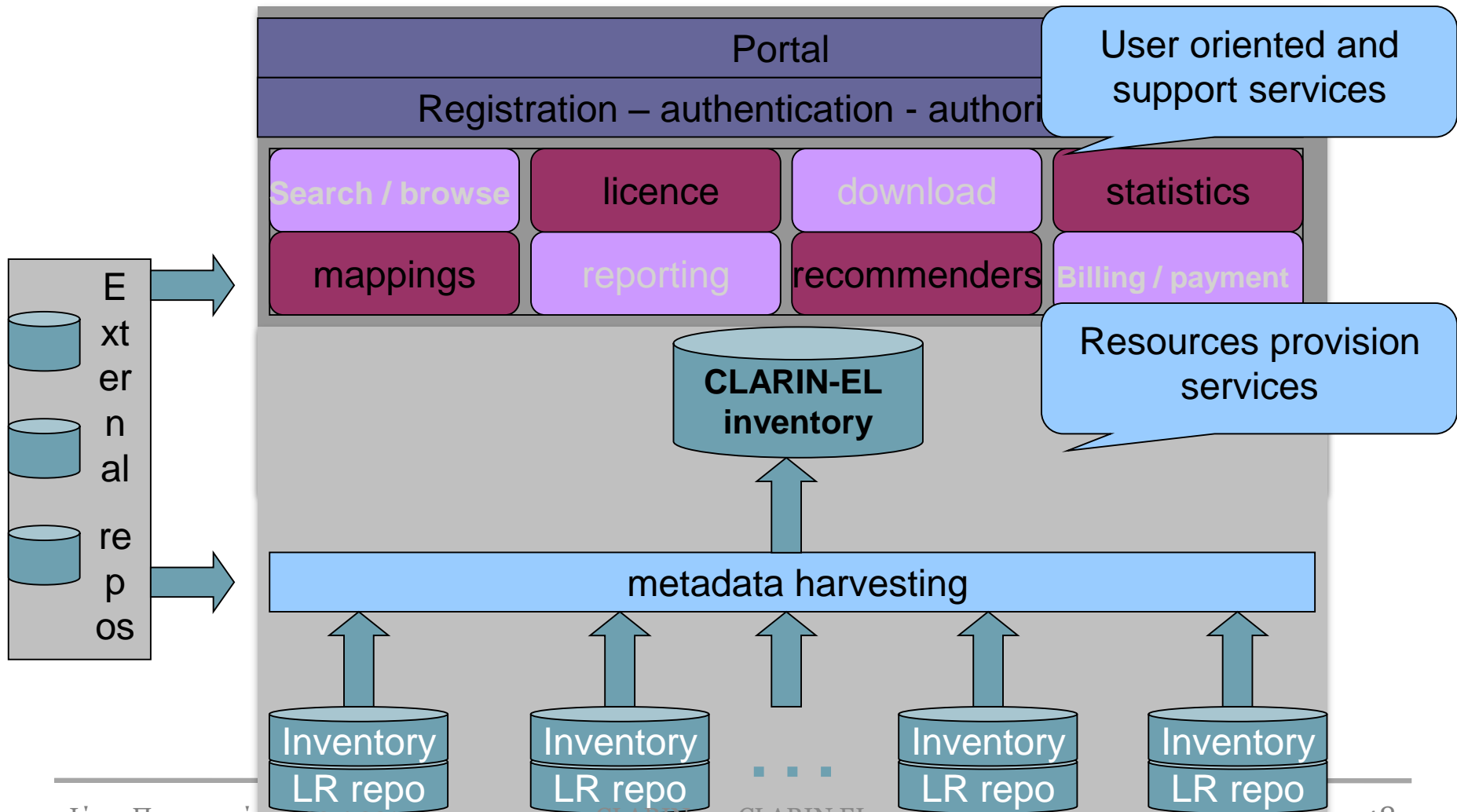
Sizes ⓘ		
* Size ⓘ	* Size unit ⓘ	Delete?
270.000 Words <input type="text" value="270.000"/>	Words ▾	<input type="checkbox"/>
611 Kb <input type="text" value="611"/>	Kb ▾	<input type="checkbox"/>
<input type="text"/>	----- ▾	

[Add another Size](#)

Save and continue editing Cancel Save

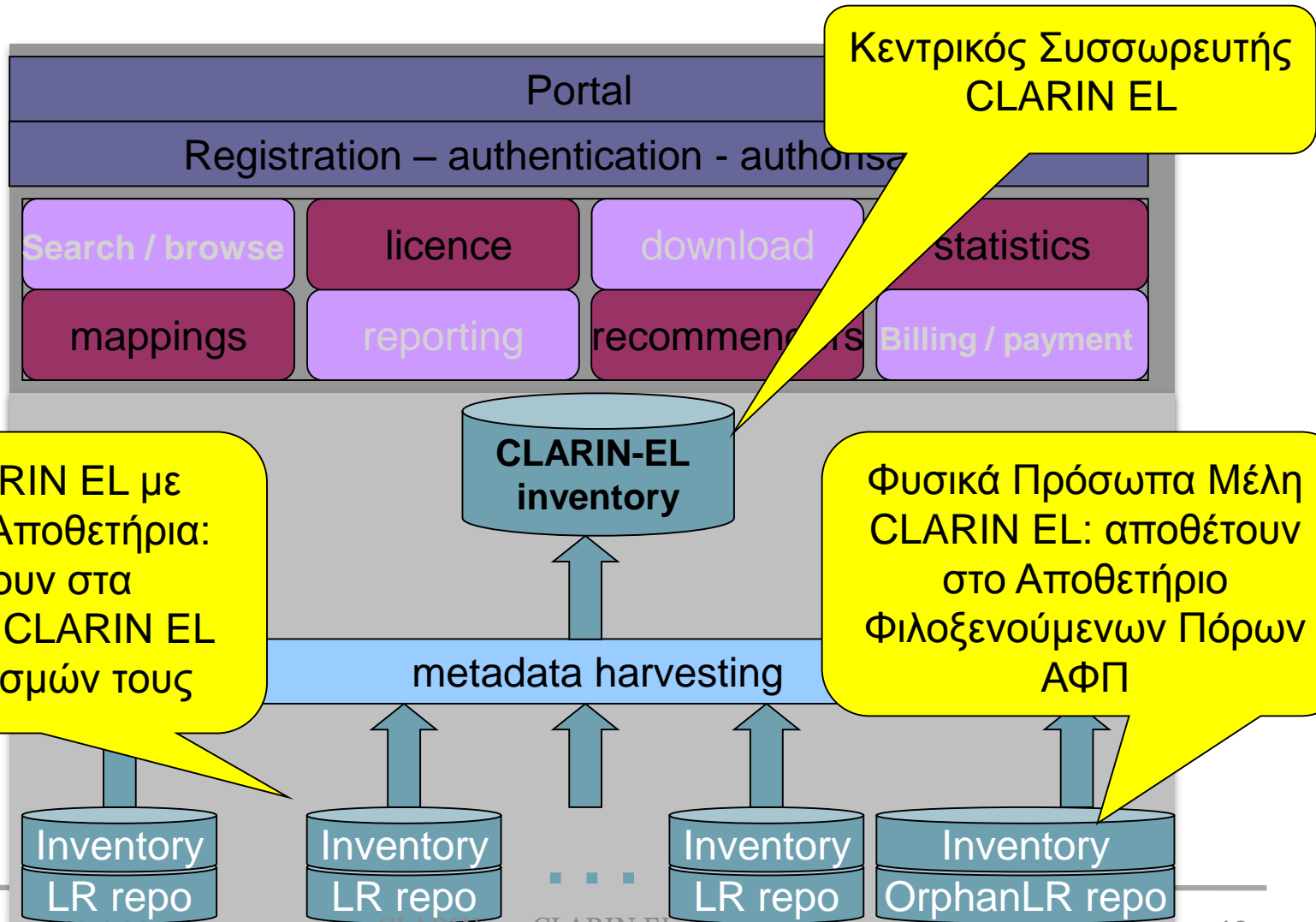
# Αρχιτεκτονική δικτύου και υποδομής

# Αρχιτεκτονική resources infrastructure





# Αρχιτεκτονική – Σύσταση Δικτύου



# Μέλη Δικτύου

---

- ◆ Κατασκευαστική Φάση (μέχρι 30/9/2015)
  - Υλοποίηση Υποδομής
    - ΕΚ «Αθηνά»
    - ΕΚΕΦΕ «Δημόκριτος»
    - Εθνικό Δίκτυο Έρευνας 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://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page)
  - <http://www.panacea-lr.eu/>
  - <http://qt21.metashare.ilsp.gr/>

# Υπολογιστική Υποδομή

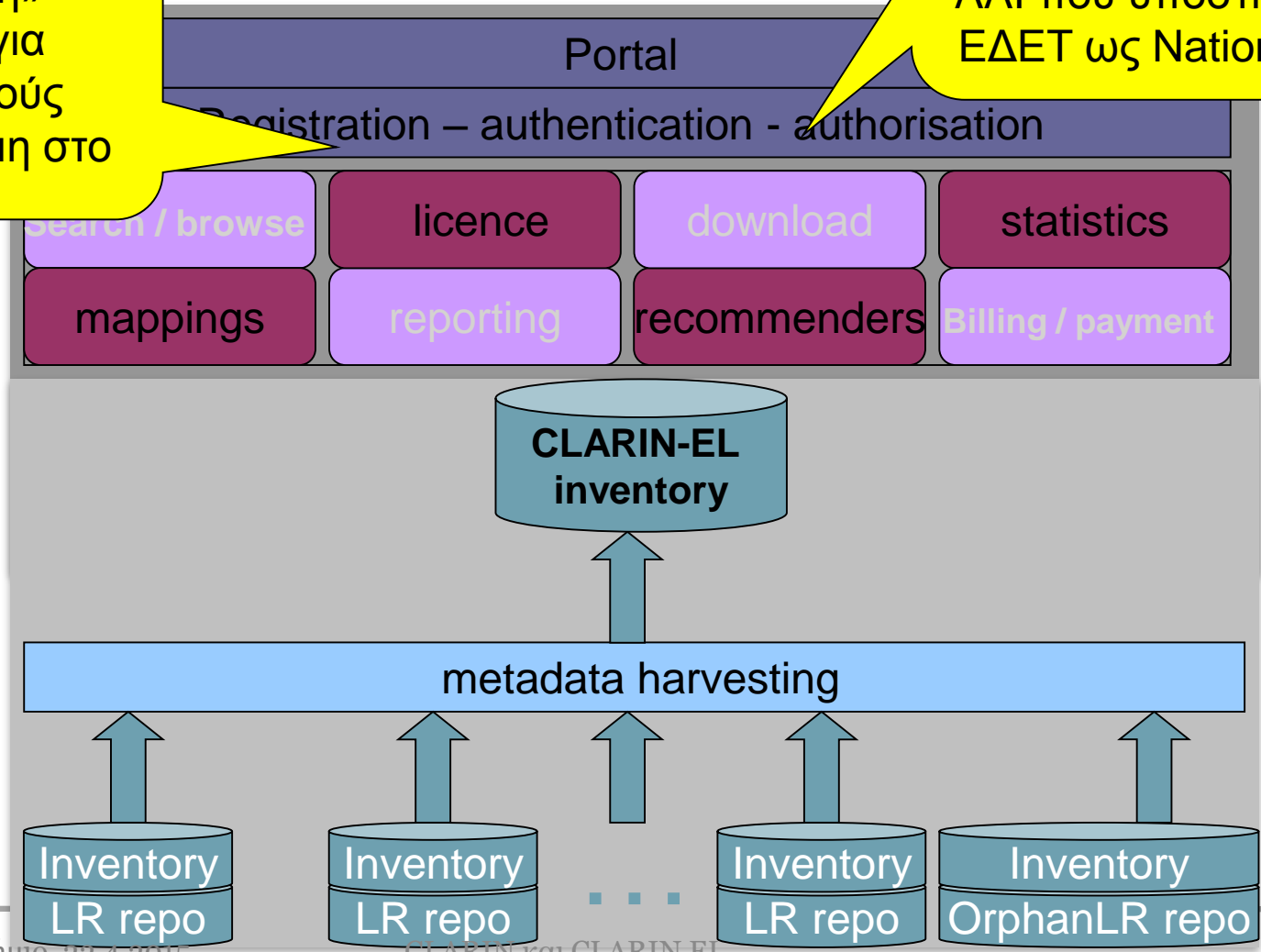
---

- ◆ Όλα τα συνιστώμενα 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 ΑΑΙ που υποστηρίζει η ΕΔΕΤ ως National IdP



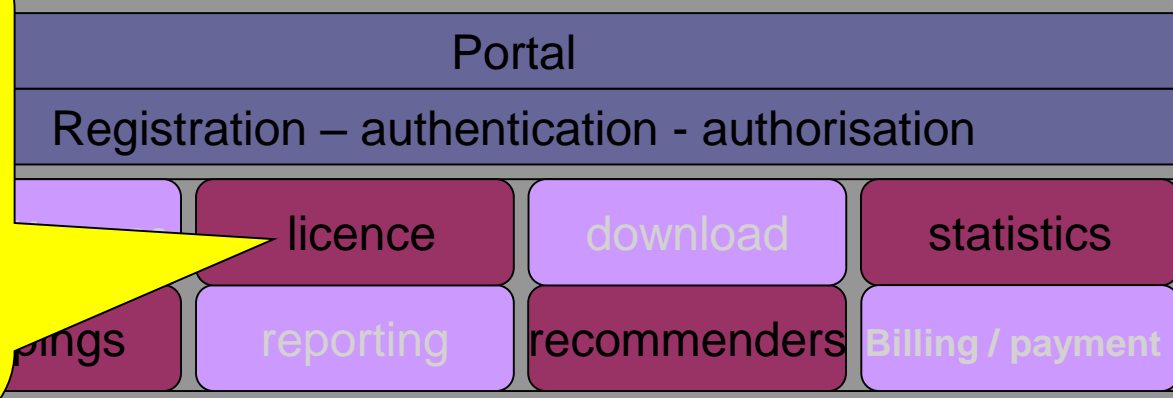
## Αρχιτεκτονική – Πρόσβαση (2)

---

- ◆ 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

Creative Commons licences 4.0, META-SHARE NoRedistribution licences 1.0, CLARIN licences (forthcoming), FOSS licences

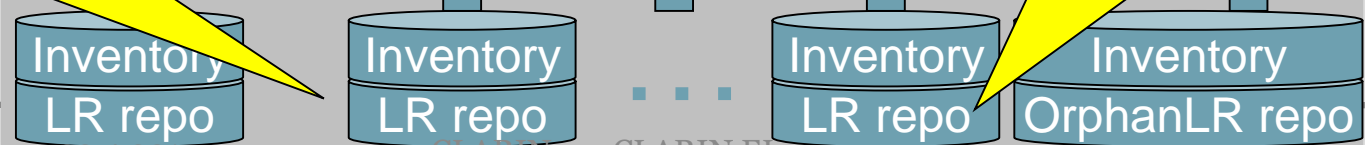


META-SHARE synchronisation protocol, OAI-PMH

META-SHARE metadata schema

metadata harvesting

Handle based PIDs



# 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](mailto:clarin-el-coordinator@clarin.gr), και [clarin-el-techadmin@clarin.gr](mailto: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](mailto:legal-helpdesk@clarin.gr)
  - Τεχνικά θέματα → [technical-helpdesk@clarin.gr](mailto:technical-helpdesk@clarin.gr)
  - Θέματα τεκμηρίωσης → [metadata-helpdesk@clarin.gr](mailto:metadata-helpdesk@clarin.gr)

# Εθνικό δίκτυο CLARIN-EL



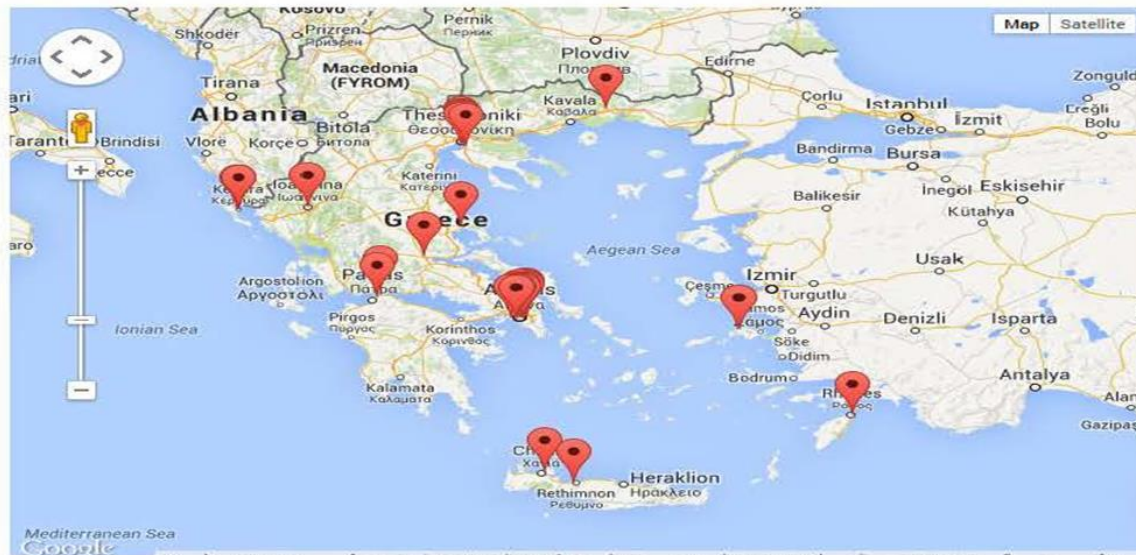
## Επιλογές εμφάνισης

Μπορείτε να περιορίσετε τη θέαση των αποτελεσμάτων σε ορισμένες κατηγορίες γλωσσικών πόρων και εφαρμογών, επιλέγοντας τις αντίστοιχες κατηγορίες και επιβεβαιώνοντας την επιλογή σας με το πλήκτρο "Φίλτρο".

- Πόροι πρωτογενούς υλικού
- Πόροι επεξεργασμένου υλικού
- Πόροι αναφοράς
- Εργαλεία/Εφαρμογές Γλωσσικής Τεχνολογίας

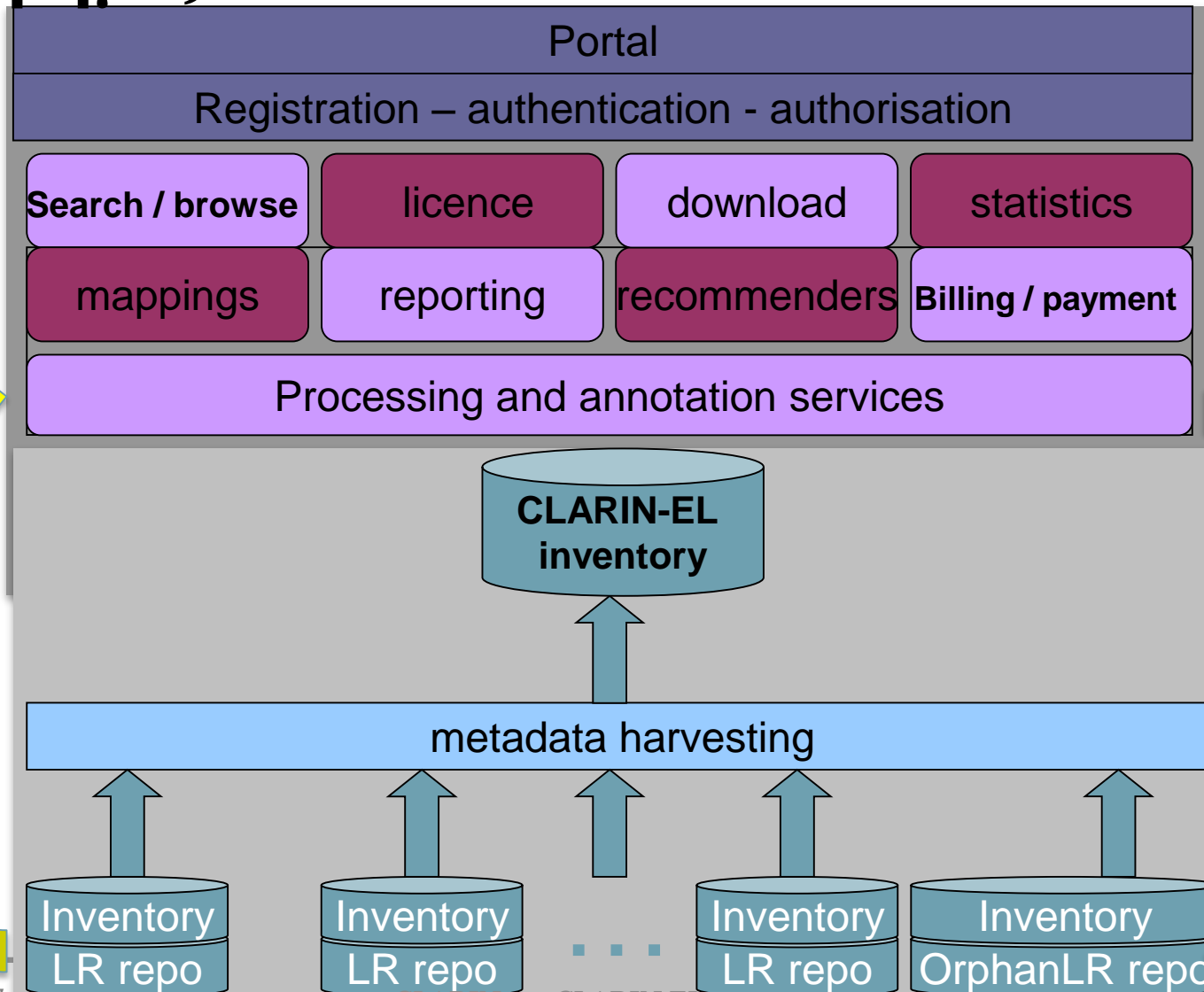
Φίλτρο

Κατάργηση φίλτρου



**τι δουλεύουμε τώρα**

# Αρχιτεκτονική – Υπηρεσίες επεξεργασίας (το 6<sup>ο</sup> 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-X (=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)

## Επεξεργασία 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...**

## eBooks

οδηγίες χρήσης και παραδειγματα  
Παρακαλούμε στείλτε τις παρατηρήσεις σας, το σχολιό σας ή/και τυχόν προβλήματα στα: [prapasas\\_AT\\_esp.gr](mailto:prapasas_AT_esp.gr)

Evaluation user: srip

ψάρωμα

(x) ψάρωμα

**Βαθμίδα**

δημοτικό (35) αρχαία (8) λύκειο (27)  
 γυμνάσιο (47) ημερήσιο-λύκειο (1)

**Μάθημα**

αγγλικά (0) ανθολόγια (1) αρχαία (5) αστρονομία (0)  
 βιολογία (1) βιομ. παραγωγή & ενέργεια (2) βιοχημεία (0)  
 Εκπαίδευση φυσικών πόρων (4) διαίτηση απορρίψεων (0)  
 εικαστικά (0) έγκριση-έκθεση (2) φιλοσοφία (2) φυσική (1)  
 φυσική αγωγή (4) γαλλικά (0) γεωμετρικά (0)  
**γεωγραφία (25)** επιστήμη (2) γλώσσα (9)  
 γραμματική (0) ηλεκτρολογία (0) ιστορία (8)  
 ιστορία σπουσμάτων και τεχνολογίας (0) ιστορία τέχνης (1)  
 κοινωνική και πολιτική αγωγή (1) κοινωνιολογία (1)  
 λατινικά (0) λεξικό (31) λογική (0) λογιστική (1)  
 μαθηματικά (1) μελέτη περιβάλλοντος (8) μουσική (0) νέα (2)  
 οικολογία (1) οικονομία (2) πληροφορική (0) ΣΠΟ (0)  
 στατιστική (1) σχέδια (0) συντακτικά (0) τεχνολογία (1)  
 τεχνολογία επικοινωνιών (0) τεχνολογία υπ. συστημάτων (0)  
 θρησκευτικά (2) χημεία (1)

**Τάξη**

A (17) A-B (0) **A-B-Γ (32)** B (25) B-Γ (0)  
 Γ (12) Γ-Δ (0) Δ (6) Δ-Ε-Στ (6) Ε (5) Ε-Στ (2) Στ (13)

< 1 2 3 ... 11 12 >

**Did you mean**  
αληθία ;

displaying 1 to 10 of 118

**9. Θέλουμε καθαρές θάλασσες και λιγότερη περιβάλλοντος** δημοτικό Δ

... πινακίδα: «Προσέχη! Μολυσμένα νερά.» Τι μπορεί να συνέβη μέσα σ' ένα χρόνο; Τα οικουσιτήματα της ακτής και της θάλασσας κινδυνεύουν από την υπερβολική **αλιεία** τη ρύπανση Τα εντατικά **ψάρωμα** Τα... μας έφρενα στο παδί. Η Ελλάδα έχει πολλές παραλίες και ακρογιάλια, που αποτελούν οικουσιτήματα. Πολλά από αυτά κινδυνεύουν από τη ρύπανση, την υπερβολική **αλιεία** αλλά και από άλλες ανθρώπινες...

**Γ' Ενότητα :Το ανθρωπογενές πε...** γεωγραφία δημοτικό E

... χώρα μας η **αλιεία**. Οι περισσότεροι κάτοικοι των παράλιων περιοχών και των νησιών ασχολούνται με το **ψάρωμα**. Η **αλιεία** - παράλια, παραμοσώγια ή υπερπόντια - και η αθυσοκαλλιέργεια συνιστούν αρκετά στην... Γεωγραφία (E', δημοτικού)-Θέλια Μαθητή Κεφάλαιο 37ο Η κτηνοτροφική παραγωγή και η **αλιεία** στην Ελλάδα για τα κτηνοτροφικά προϊόντα και τα αλιεύματα που παράγονται στη χώρα μας Κτηνοτροφία Εικόνα...

**A** λεξικό γυμνάσιο A-B-Γ

... Παράτησε σπίτι και δουλειά και το έριξε στην ~, 2 (συνεκδ.) αλήτας: Σ' αυτό το πάρκο μαζεύεται όλη η ~, αλητεύω (αμβ.), αλειώω -ομαι: [απία.] (μτβ.) **ψαρεύω**, **αλιεία** η: [επία.] **ψάρωμα**, αλειυτικός -ή -ά... αλειυτικά το: σκάφος για **αλιεία**. Από το ΑΕ άλις «θάλασσα», αλιμνο (σπαρ.): (συνήθ. + αδύν. τ. γεν. της προσωπ. αντων.) 1 για να εκφράσουμε μεγάλη λύπη, απελπισία, απόγνωση: ~, τι θα σπαγίω η δύστυχη! 2 για...

**A** λεξικό γυμνάσιο A-B-Γ

... Παράτησε σπίτι και δουλειά και το έριξε στην ~, 2 (συνεκδ.) αλήτας: Σ' αυτό το πάρκο μαζεύεται όλη η ~, αλητεύω (αμβ.), αλειώω -ομαι: [απία.] (μτβ.) **ψαρεύω**, **αλιεία** η: [επία.] **ψάρωμα**, αλειυτικός -ή -ά... αλειυτικά το: σκάφος για **αλιεία**. Από το

Ίονιο Πανεπιστήμιο,  
22.4.2015

CLARIN και CLARIN EL

68

# Αναγνωρίζουμε οντότητες και τα ονόματά τους

Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Ήταν η Βασίλισσα μιας ολόκληρης αυτοκρατορίας όμως η καρδιά της ανήκε σε έναν μόνο άντρα **Αγγλία** 1837: η 18χρονη **Βικτώρια**, βρίσκεται στο επίκεντρο της βασιλικής διαμάχης για την εξουσία. Ο θεός της, **Βασιλιάς Γουίλιαμ** σύντομα θα πεθάνει και η **Βικτώρια** είναι η διάδοχος του θρόνου. Ενώ όλοι προσπαθούν να κερδίσουν την εύνοιά της, η μητέρα της, **Δούκισσα του Κεντ**, την κρατάει σε απόσταση από την αυλή. Παράλληλα, ο **Βασιλιάς Λεοπόλδος του Βελγίου** και θεός της **Βικτώρια**, πατρωνάει τον γοητευτικό ανηψιό του **Άλμπερτ**, για να κερδίσει την καρδιά της. Όμως η **Βικτώρια** και ο **Άλμπερτ** έχουν κουραστεί να δέχονται εντολές από τους άλλους. Και αφού η **Βικτώρια** γίνεται Βασίλισσα της **Αγγλίας**, αναπτύσσεται μεταξύ τους ένας ισχυρός δεσμός που θα οδηγήσει όχι μόνο στο γάμο και στην οικογένεια, αλλά και σε μια ουσιαστική σχέση ζωής.



Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Ήταν η Βασίλισσα μιας ολόκληρης αυτοκρατορίας όμως η καρδιά της ανήκε σε έναν μόνο άντρα **Αγγλία** 1837: η 18χρονη **Βικτώρια**, βρίσκεται στο επίκεντρο της βασιλικής διαμάχης για την εξουσία. Ο θεός της, **Βασιλιάς Γουίλιαμ** σύντομα θα πεθάνει και η **Βικτώρια** είναι η διάδοχος του θρόνου. Ενώ όλοι προσπαθούν να κερδίσουν την εύνοιά της, η μητέρα της, **Δούκισσα του Κεντ**, την κρατάει σε απόσταση από την αυλή. Παράλληλα, ο **Βασιλιάς Λεοπόλδος του Βελγίου** και θεός της **Βικτώρια**, πατρωνάει τον γοητευτικό ανηψιό του **Άλμπερτ**, για να κερδίσει την καρδιά της. Όμως η **Βικτώρια** και ο **Άλμπερτ** έχουν κουραστεί να δέχονται εντολές από τους άλλους. Και αφού η **Βικτώρια** γίνεται Βασίλισσα της **Αγγλίας**, αναπτύσσεται μεταξύ τους ένας ισχυρός δεσμός που θα οδηγήσει όχι μόνο στο γάμο και στην οικογένεια, αλλά και σε μια ουσιαστική σχέση ζωής.



Annotation Sets Annotations List Annotations Stack Co-reference Editor Text

Ήταν η Βασίλισσα μιας ολόκληρης αυτοκρατορίας όμως η καρδιά της ανήκε σε έναν μόνο άντρα **Αγγλία** 1837: η 18χρονη **Βικτώρια**, βρίσκεται στο επίκεντρο της βασιλικής διαμάχης για την εξουσία. Ο θεός της, **Βασιλιάς Γουίλιαμ** σύντομα θα πεθάνει και η **Βικτώρια** είναι η διάδοχος του θρόνου. Ενώ όλοι προσπαθούν να κερδίσουν την εύνοιά της, η μητέρα της, **Δούκισσα του Κεντ**, την κρατάει σε απόσταση από την αυλή. Παράλληλα, ο **Βασιλιάς Λεοπόλδος του Βελγίου** και θεός της **Βικτώρια**, πατρωνάει τον γοητευτικό ανηψιό του **Άλμπερτ**, για να κερδίσει την καρδιά της. Όμως η **Βικτώρια** και ο **Άλμπερτ** έχουν κουραστεί να δέχονται εντολές από τους άλλους. Και αφού η **Βικτώρια** γίνεται Βασίλισσα της **Αγγλίας**, αναπτύσσεται μεταξύ τους ένας ισχυρός δεσμός που θα οδηγήσει όχι μόνο στο γάμο και στην οικογένεια, αλλά και σε μια ουσιαστική σχέση ζωής.

- GrTriggerLexica
- Header
- LOCATION
- Lookup
- PERSON
- Paragraph
- Sentence
- Token
- Original markups



# Συνδέουμε οντότητες και έννοιες με την wikipedia



ENTER SOME TEXT AND PRESS ANNOTATE TO WIKIFY IT.  
FOR NOW, GREEK ONLY.

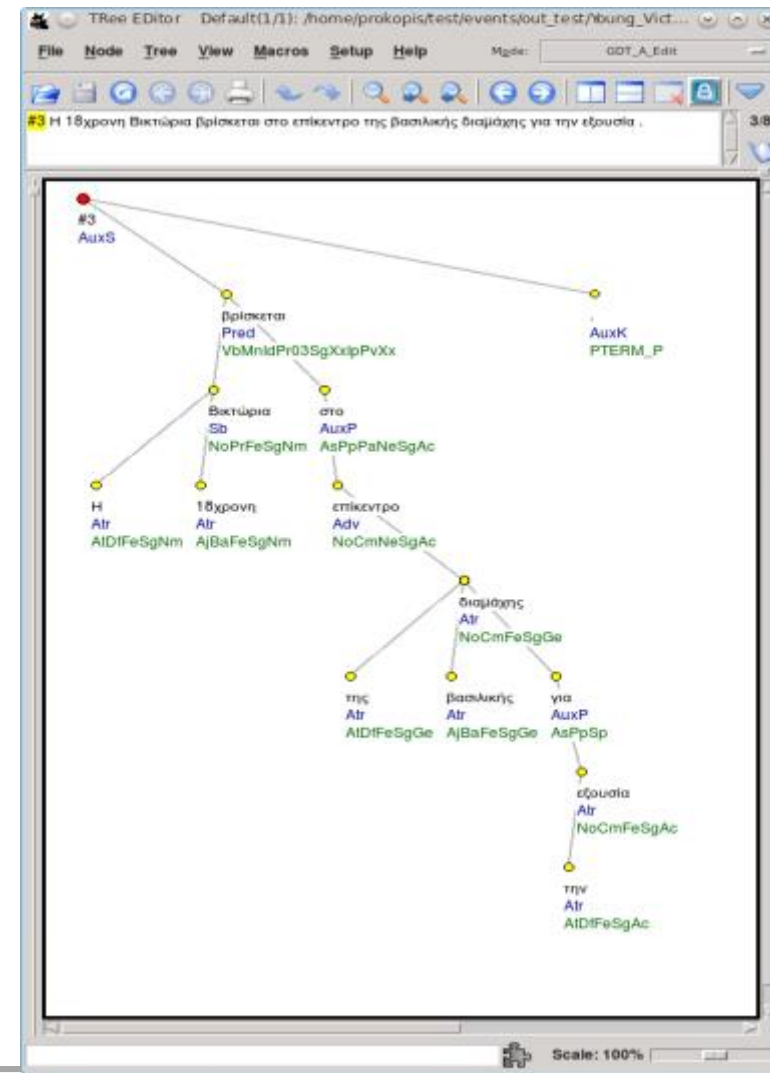
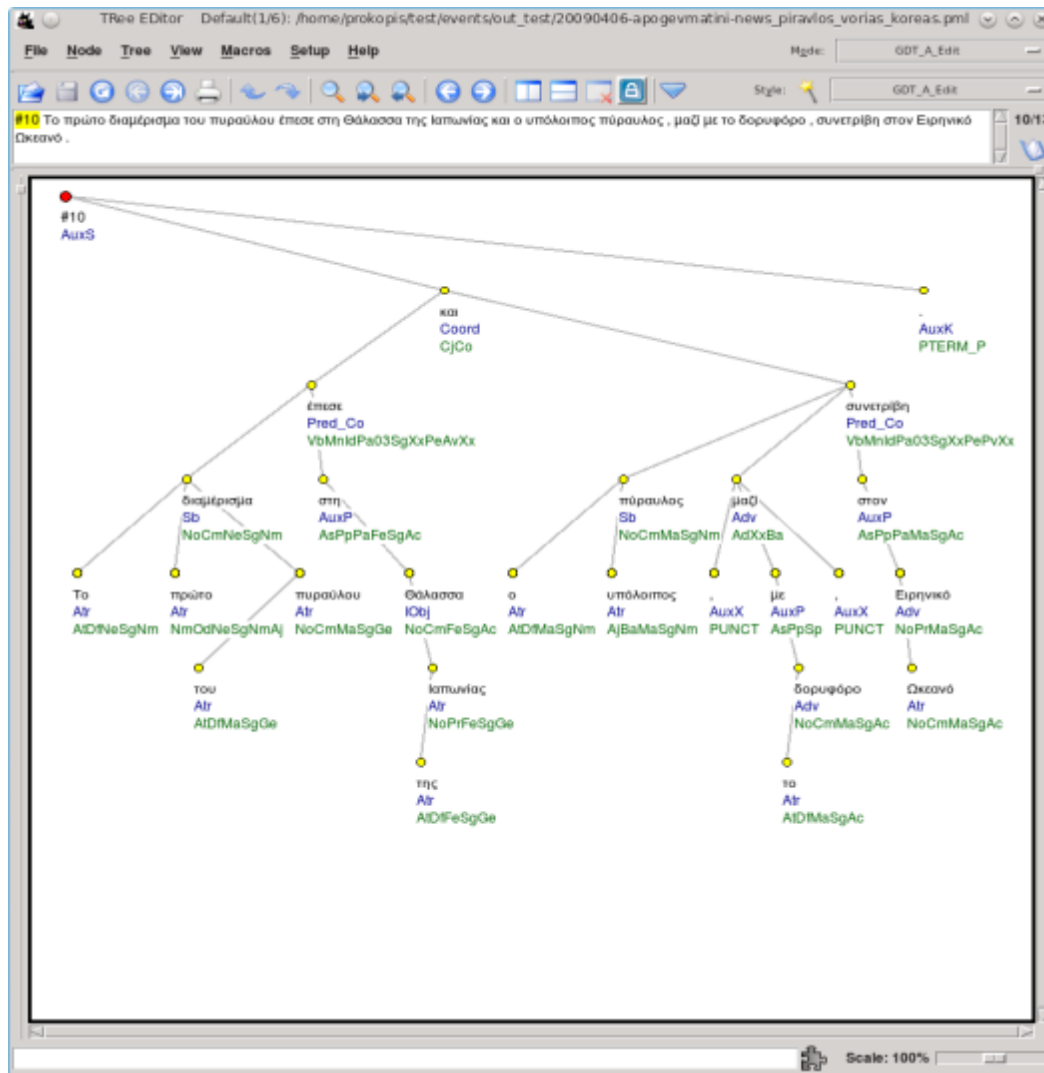
Οι Καρχηδόνιοι λένε ακόμη και το εξής: ότι υπάρχει χώρα της Λιβύης και άνθρωποι που κατοικούν σ' αυτήν έξω από τις Ηράκλειες στήλες. Λένε ακόμη ότι, όταν φθάνουν στη χώρα των ανθρώπων αυτών, βγάζουν έξω τα προϊόντα τους, τα βάζουν στη σειρά στην παραλία και μπαίνουν πάλι στα πλοία και κάνουν καπνό. Οι ιθαγενείς, όταν δουν τον καπνό, κατεβαίνουν στην παραλία, αφήνουν ποσότητα χρυσού, ανάλογης αξίας προς τα εμπορεύματα, και γυρίζουν πίσω. Οι Καρχηδόνιοι σπεύδουν στην ξηρά και εξετάζουν τον χρυσό. Αν καταλάβουν ότι ο χρυσός ισοφαρίζει την αξία του εμπορεύματος, τον παίρνουν και φεύγουν αν όχι, μπαίνουν πάλι στα καράβια και περιμένουν. Οι ιθαγενείς πλησιάζουν και προσθέτουν χρυσό παραπάνω, μέχρι να τους ικανοποιήσουν. Κανείς, όπως λένε οι Καρχηδόνιοι, δεν αδικεί.

ANNOTATE

ANNOTATED TEXT

Οι **[[Καρχηδόνα|Καρχηδόνιοι]]** λένε ακόμη και το εξής: ότι υπάρχει χώρα της **[[Λιβύη|Λιβύης]]** και άνθρωποι που κατοικούν σ' αυτήν έξω από τις **[[Ηράκλειες στήλες|Ηράκλειες στήλες]]**. Λένε ακόμη ότι, όταν φθάνουν στη χώρα των ανθρώπων αυτών, βγάζουν έξω τα προϊόντα τους, τα βάζουν στη σειρά στην παραλία και μπαίνουν πάλι στα πλοία και κάνουν **[[Καπνός (φυτό)|καπνό]]**. Οι ιθαγενείς, όταν δουν τον **[[Καπνός (φυτό)|καπνό]]**, κατεβαίνουν στην παραλία, αφήνουν ποσότητα **[[Χρυσός|χρυσού]]**, ανάλογης αξίας προς τα εμπορεύματα, και γυρίζουν πίσω. Οι **[[Καρχηδόνα|Καρχηδόνιοι]]** σπεύδουν στην ξηρά και εξετάζουν τον **[[Χρυσός|χρυσό]]**. Αν καταλάβουν ότι ο **[[Χρυσός|χρυσός]]** ισοφαρίζει την αξία του εμπορεύματος, τον παίρνουν και φεύγουν αν όχι, μπαίνουν πάλι στα καράβια και περιμένουν. Οι ιθαγενείς πλησιάζουν και προσθέτουν **[[Χρυσός|χρυσό]]** παραπάνω, μέχρι να τους ικανοποιήσουν. Κανείς, όπως λένε οι **[[Καρχηδόνα|Καρχηδόνιοι]]**, δεν αδικεί.

# 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;

The screenshot shows a text analysis interface. On the left, a text document is displayed with several words highlighted in different colors: 'φιλικά' (cyan), 'λογική τιμή' (purple), 'νόστιμο' (cyan), 'άψογη εμφάνιση' (purple), 'δρροσερή' (cyan), 'ικανοποιητική μερίδα' (purple), and 'ικανοποιητική' (cyan). On the right, a vertical list of sentiment labels is shown, each with a checkbox and a colored background:

- Attitude
- AttitudeLexicon
- Modifiers
- PRAISE
- Paragraph
- Sentence
- TARGET
- Token

At the bottom right of the label list is a 'New' button. The background text is a restaurant review in Greek, discussing the service, food quality, and overall experience.





# ή πόσο καλή ήταν μια ταινία

1 <EvalLex>Αυστηρά</EvalLex> και μόνο για οσους μπορούν να εκτιμήσουν ένα ποίημα. Οι υπολοίποι να δουν το Island (κι εγω το ειδα αλλωστε).

2 Από τις πολύ <EvalLex>καλές</EvalLex> και επιδραστικές ταινίες του Γκοντάρ, τότε που κάθε δουλειά του ήταν κάτι που άξιζε να συζητιέται για καιρό.

3 Η ταινία , ειδικά αν την εξετάσουμε στο πλαίσιο της εποχής που δημιουργήθηκε , ήταν <EvalLex>ιδιαίτερα</EvalLex> <EvalLex>ευρηματική</EvalLex>.

4 Αυτός ο χαρακτηρισμός αφορά κυρίως τον τρόπο σκηνοθεσίας και την <EvalLex>εξαιρετική</EvalLex> μέθοδο αφήγησης της ιστορίας της Νανά.

5 Ωστόσο βρήκα το τέλος πολύ <EvalLex>απλοϊκό</EvalLex> - αυτή είναι και η <EvalLex>μοναδική</EvalLex> μου ένσταση σε μια κατά τα άλλα <EvalLex>όμοια</EvalLex> ταινία.

6 Διόρθωση: Όταν δεν μιλάει κανείς δεν σκέφτεται. Απίστευτη η αμεσότητα του διαλόγου στο καφέ.

1 Αυστηρά και μόνο για οσους μπορούν να εκτιμήσουν ένα ποίημα. Οι υπολοίποι να δουν το Island (κι εγω το ειδα αλλωστε).

2 Από τις <Attitude>πολύ καλές</Attitude> και <Attitude>επιδραστικές</Attitude> ταινίες του Γκοντάρ, τότε που κάθε δουλειά του ήταν κάτι που άξιζε να συζητιέται για καιρό.

3 Η ταινία , ειδικά αν την εξετάσουμε στο πλαίσιο της εποχής που δημιουργήθηκε , ήταν <Attitude>ιδιαίτερα ευρηματική</Attitude>.

4 Αυτός ο χαρακτηρισμός αφορά κυρίως τον τρόπο σκηνοθεσίας και την <Attitude>εξαιρετική</Attitude> μέθοδο αφήγησης της ιστορίας της Νανά.

5 Ωστόσο βρήκα το τέλος <Attitude>πολύ απλοϊκό</Attitude> ταινία.

6 Διόρθωση: Όταν δεν μιλάει κανείς δεν σκέφτεται. Απίστευτη η αμεσότητα του διαλόγου στο καφέ.

The screenshot shows a text analysis interface. The main window displays a document with several segments highlighted in different colors: cyan for 'καλές', purple for 'επιδραστικές ταινίες', cyan for 'εξαιρετική', cyan for 'ευρηματική', cyan for 'απλοϊκό', cyan for 'μοναδική', cyan for 'όμοια', cyan for 'αμεσότητα', and purple for 'διαλόγου'. The sidebar on the right contains a list of analysis categories with checkboxes: Attitude (unchecked), AttitudeLexicon (unchecked), CRITICISM (checked), Modifiers (unchecked), PRAISE (checked), Paragraph (unchecked), Sentence (unchecked), TARGET (checked), and Token (unchecked). A 'New' button is located at the bottom right of the sidebar.

**Κι αν πάμε να μεταφράσουμε...**

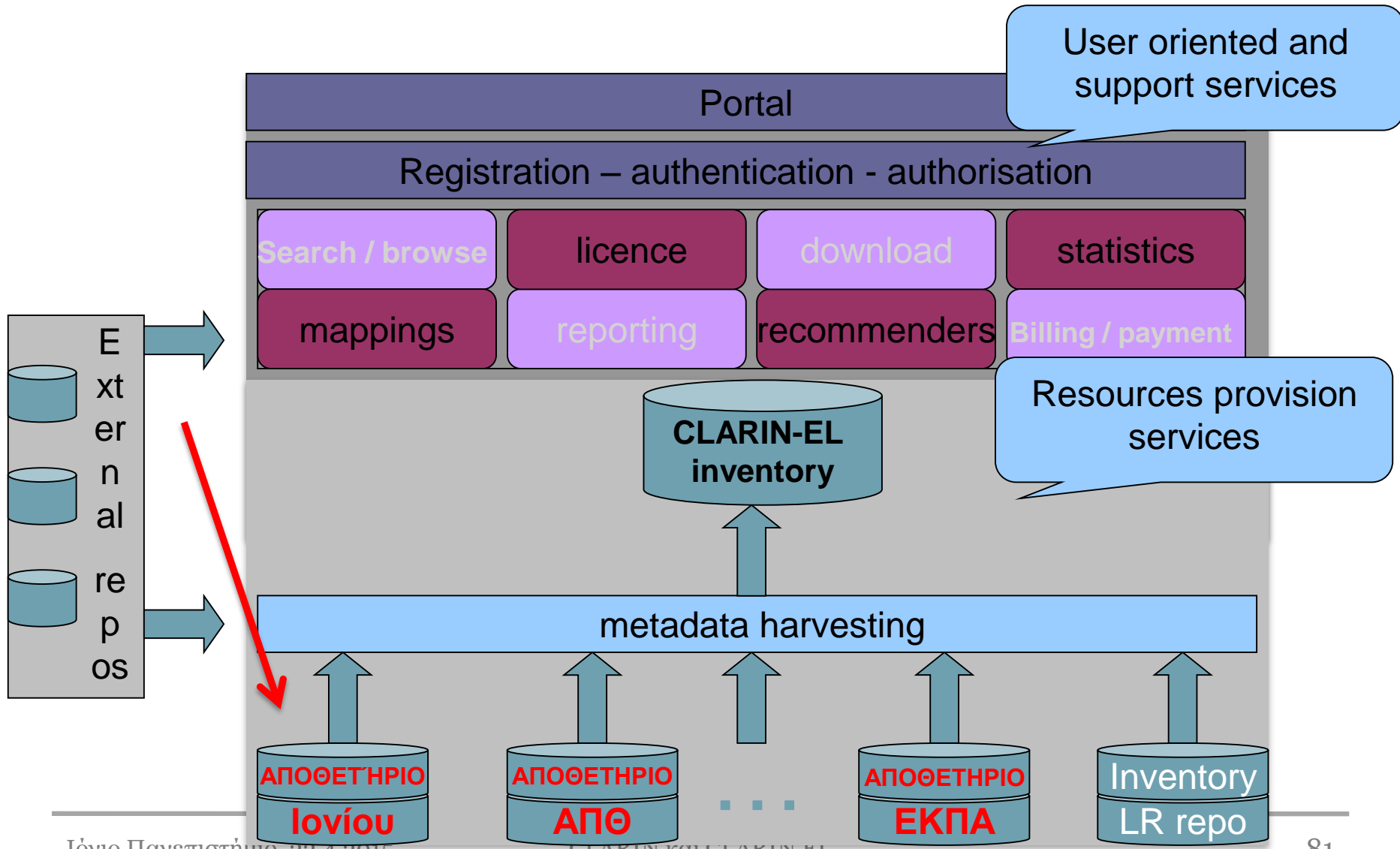
## ...προς τα ελληνικά

Αγγλικά	Ελληνικά	OK ?
<p>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.</p>	<p>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.</p>	<p>✓</p>
<p>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.</p>	<p>υπό την προϋπόθεση ότι οι κανόνες για την προστασία των δεδομένων προσωπικού χαρακτήρα, κατά περίπτωση, <u>να τηρούνται, τα δεδομένα, μια φορά,</u> να μπορούν να επαναχρησιμοποιηθούν πολλές φορές χωρίς απώλεια της πιστότητας.</p>	<p>✗</p>

## ...προς τα αγγλικά

Ελληνικά	Αγγλικά	OK?
<p>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.</p>	<p>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.</p>	<p>✓</p>
<p>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.</p>	<p>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.</p>	<p>✗</p>

# **Το Ιόνιο Πανεπιστήμιο στο CLARIN EL**



# 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	
μορφώματος αρχείων (txt, doc, xml, tnx, 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 ΔΑΡΙΑΗ-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 εποχή
- ◆ Την υποστήριξη του πολιτισμού και της πολιτιστικής κληρονομιάς

# Το CLARIN EL στο οικοσύστημα EY

---

- ◆ 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!