

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

Σαράντος Καπιδάκης¹, Στέλιος Πιπερίδης², Πένυ
Λαμπροπούλου², Μαρία Γαβριηλίδου²

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



το πρόβλημα

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

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

η λύση: CLARIN

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

Ένας ερευνητής από το γραφείο του στην Κέρκυρα θα μπορεί:

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

δηλαδή...

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

πώς;

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

η ελληνική υποδομή clarin:el



το clarin στην Ελλάδα

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

χρήση της υποδομής από καταναλωτές πόρων

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



central catalogue
of language resources and services

go to the athena rc repository

167 language resources at your disposal

corpus greek law search

Here you can browse the central catalogue of language resources and services of clarin:el.
clarin:el is the Greek national network of language resources, a nation-wide Research Infrastructure devoted to the sustainable storage, sharing, dissemination and preservation of language resources.

Εθνικό Κέντρο Αθηνών gnet

©clarin:el 2015 terms of service

2. φυλλομέτρηση αποτελεσμάτων

browse resources

filter by:

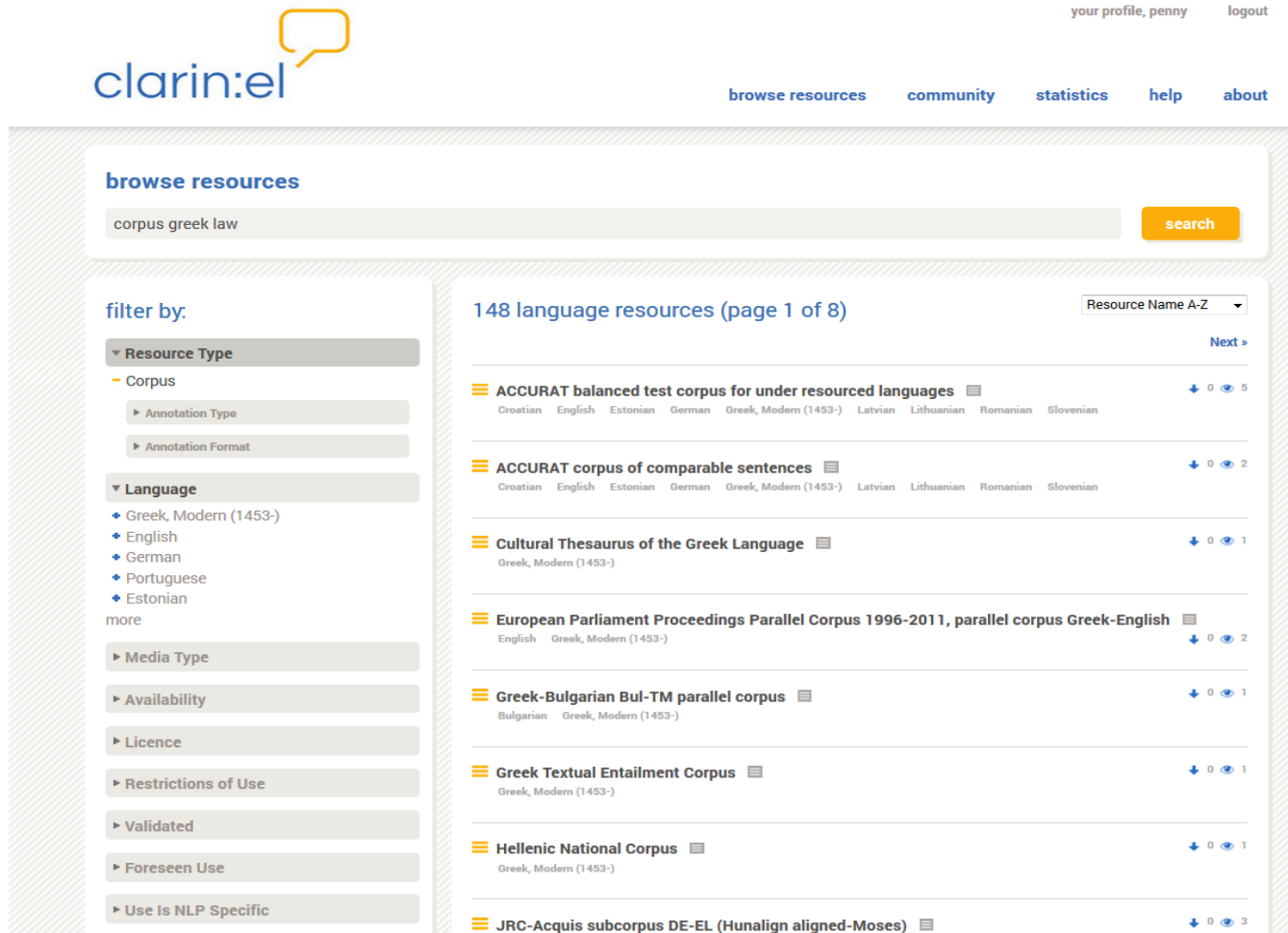
- ▶ Language
- ▶ Resource Type
- ▶ Media Type
- ▶ Availability
- ▶ Licence
- ▶ Restrictions of Use
- ▶ Validated
- ▶ Foreseen Use
- ▶ Use Is NLP Specific
- ▶ Linguality Type
- ▶ Multilinguality Type
- ▶ Modality Type
- ▶ MIME Type
- ▶ Conformance to Standards/Best Practices

161 language resources (page 1 of 9) Resource Name A-Z ▾

[Next »](#)

- ACCURAT balanced test corpus for under resourced languages** ▾ [↓](#) [0](#) [👁](#) [5](#)
Croatian English Estonian German Greek, Modern (1453-) Latvian Lithuanian Romanian Slovenian
- ACCURAT corpus of comparable sentences** ▾ [↓](#) [0](#) [👁](#) [2](#)
Croatian English Estonian German Greek, Modern (1453-) Latvian Lithuanian Romanian Slovenian
- Cultural Thesaurus of the Greek Language** ▾ [↓](#) [0](#) [👁](#) [1](#)
Greek, Modern (1453-)
- European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English** ▾ [↓](#) [0](#) [👁](#) [2](#)
English Greek, Modern (1453-)
- FIND - Tool for extracting words based on a specific spelling or pronunciation pattern** ▾ [↓](#) [0](#) [👁](#) [2](#)
Greek, Modern (1453-)
- Greek-Bulgarian Bul-TM parallel corpus** ▾ [↓](#) [0](#) [👁](#) [1](#)
Bulgarian Greek, Modern (1453-)
- Greek Textual Entailment Corpus** ▾ [↓](#) [0](#) [👁](#) [1](#)
Greek, Modern (1453-)
- Hellenic National Corpus** ▾ [↓](#) [0](#) [👁](#) [1](#)

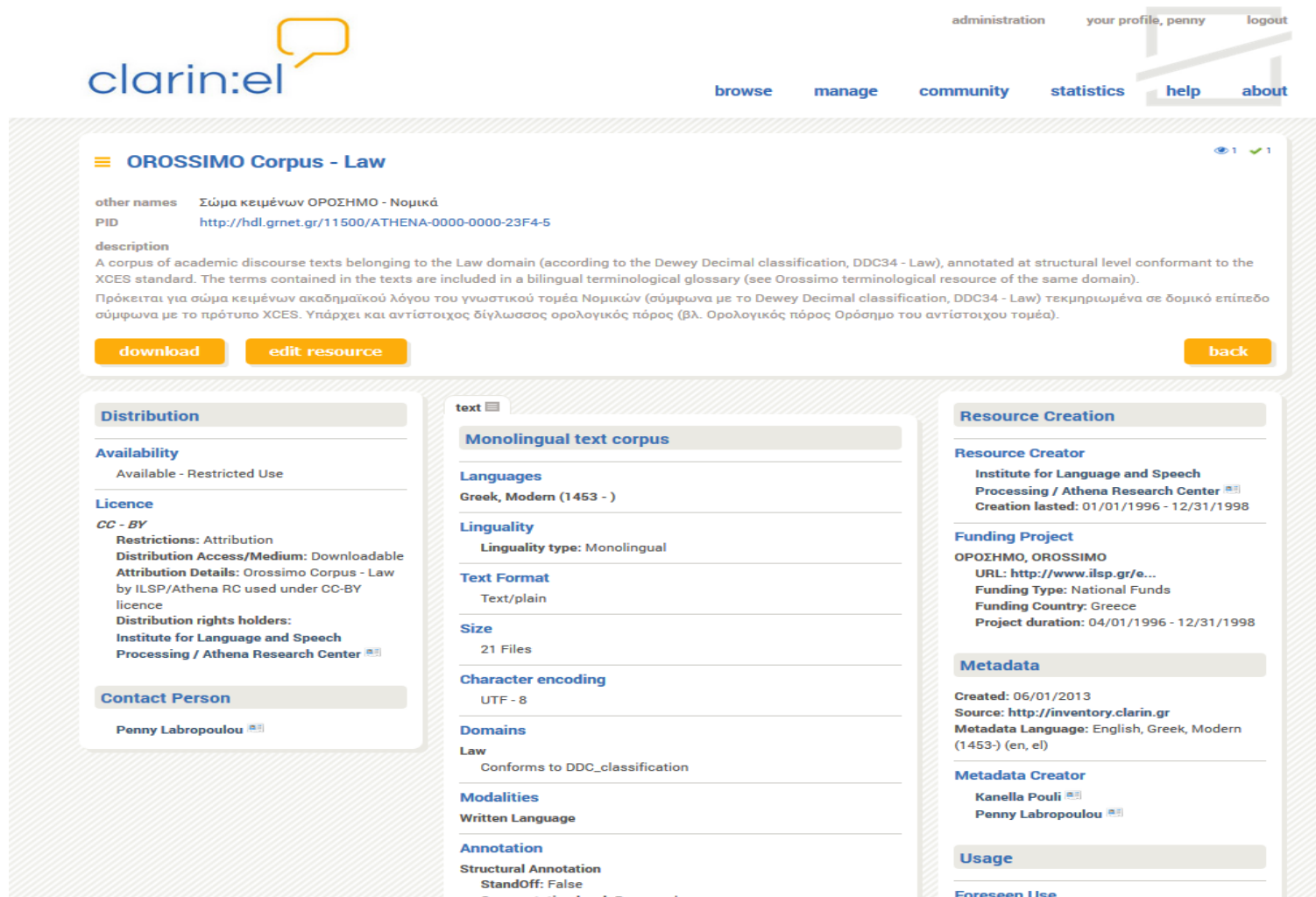
2α. αναζήτηση με χρήση φίλτρων



The screenshot shows the clarin:el website interface. At the top, there is a search bar with the text "corpus greek law" and a "search" button. Below the search bar, there are navigation links: "browse resources", "community", "statistics", "help", and "about". On the left side, there is a "filter by:" section with several filter categories: "Resource Type" (Corpus), "Language" (Greek, Modern (1453-), English, German, Portuguese, Estonian), "Media Type", "Availability", "Licence", "Restrictions of Use", "Validated", "Foreseen Use", and "Use Is NLP Specific". The main content area displays "148 language resources (page 1 of 8)" and a list of resources. Each resource entry includes a title, a list of languages, and a small icon indicating the number of resources. The resources listed are:

- ACCURAT balanced test corpus for under resourced languages (Croatian, English, Estonian, German, Greek, Modern (1453-), Latvian, Lithuanian, Romanian, Slovenian) - 5 resources
- ACCURAT corpus of comparable sentences (Croatian, English, Estonian, German, Greek, Modern (1453-), Latvian, Lithuanian, Romanian, Slovenian) - 2 resources
- Cultural Thesaurus of the Greek Language (Greek, Modern (1453-)) - 1 resource
- European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English (English, Greek, Modern (1453-)) - 2 resources
- Greek-Bulgarian Bul-TM parallel corpus (Bulgarian, Greek, Modern (1453-)) - 1 resource
- Greek Textual Entailment Corpus (Greek, Modern (1453-)) - 1 resource
- Hellenic National Corpus (Greek, Modern (1453-)) - 1 resource
- JRC-Acquis subcorpus DE-EL (Hunalign aligned-Moses) - 3 resources

3. Θέαση επιλεγμένου πόρου



clarin:el administration your profile, penny logout

browse manage community statistics help about

OROSSIMO Corpus - Law

other names Σώμα κειμένων ΟΡΟΣΗΜΟ - Νομικά
PID <http://hdl.gmet.gr/11500/ATHENA-0000-0000-23F4-5>

description
A corpus of academic discourse texts belonging to the Law domain (according to the Dewey Decimal classification, DDC34 - Law), annotated at structural level conformant to the XCES standard. The terms contained in the texts are included in a bilingual terminological glossary (see Orossimo terminological resource of the same domain).
Πρόκειται για σώμα κειμένων ακαδημαϊκού λόγου του γνωστικού τομέα Νομικών (σύμφωνα με το Dewey Decimal classification, DDC34 - Law) τεκμηριωμένα σε δομικό επίπεδο σύμφωνα με το πρότυπο XCES. Υπάρχει και αντίστοιχος δίγλωσσος ορολογικός πόρος (βλ. Ορολογικός πόρος Ορόσημο του αντίστοιχου τομέα).

download edit resource back

Distribution

Availability
Available - Restricted Use

Licence
CC - BY
Restrictions: Attribution
Distribution Access/Medium: Downloadable
Attribution Details: Orossimo Corpus - Law by ILSP/Athena RC used under CC-BY licence
Distribution rights holders:
Institute for Language and Speech
Processing / Athena Research Center

Contact Person
Penny Labropoulou

text

Monolingual text corpus

Languages
Greek, Modern (1453 -)

Linguality
Linguality type: Monolingual

Text Format
Text/plain

Size
21 Files

Character encoding
UTF - 8

Domains
Law
Conforms to DDC_classification

Modalities
Written Language

Annotation
Structural Annotation
StandOff: False

Resource Creation

Resource Creator
Institute for Language and Speech
Processing / Athena Research Center
Creation lasted: 01/01/1996 - 12/31/1998

Funding Project
ΟΡΟΣΗΜΟ, ΟΡΟΣΗΜΟ
URL: <http://www.ilsp.gr/e...>
Funding Type: National Funds
Funding Country: Greece
Project duration: 04/01/1996 - 12/31/1998

Metadata

Created: 06/01/2013
Source: <http://inventory.clarin.gr>
Metadata Language: English, Greek, Modern (1453-) (en, el)

Metadata Creator
Kanella Pouli
Penny Labropoulou

Usage

Foreseen Use

4. όροι χρήσης - αδειοδότηση

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.

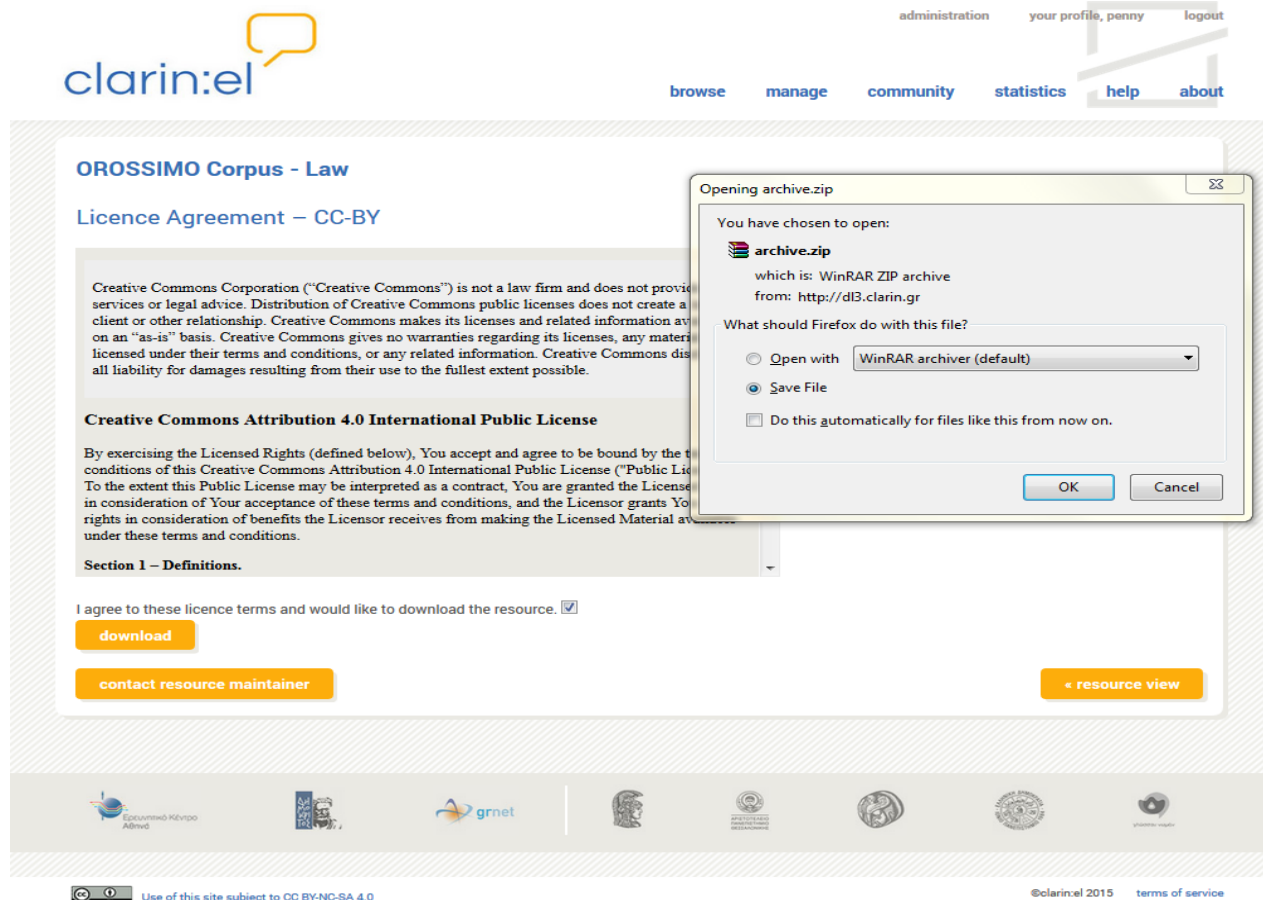
[download](#)

[contact resource maintainer](#)

[← resource view](#)



5. καταφόρτωση του πόρου



administration your profile, penny logout

clarin:el

browse manage community statistics help about

OROSSIMO Corpus - Law

Licence Agreement – CC-BY

Creative Commons Corporation (“Creative Commons”) is not a law firm and does not provide services or legal advice. Distribution of Creative Commons public licenses does not create a 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 certain 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](#)

[contact resource maintainer](#)

[← resource view](#)

Use of this site subject to CC BY-NC-SA 4.0

©clarin:el 2015 terms of service

6. γλωσσική επεξεργασία πόρων



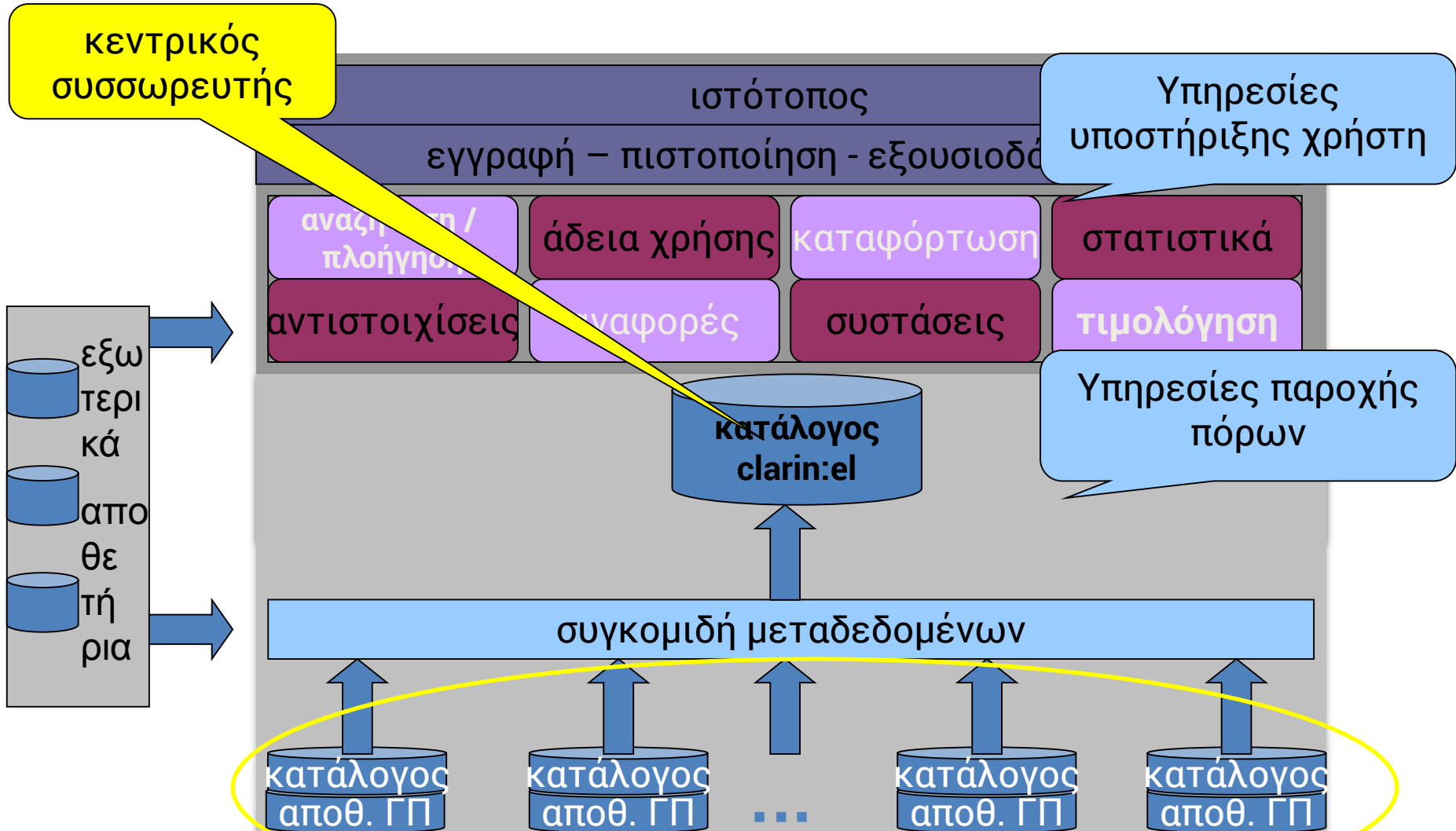
- υπηρεσίες επεξεργασίας μονογλωσσικών σωμάτων κειμένων
 - αναγνώριση λέξεων και προτάσεων (tokenisation and sentence splitting)
 - μορφοσυντακτική ανάλυση (part of speech tagging)
 - λημματοποίηση (lemmatisation)
 - συντακτική ανάλυση (syntactic parsing)
 - αναγνώριση και εξαγωγή ορολογίας (term recognition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

οι θεμέλιοι λίθοι της υποδομής

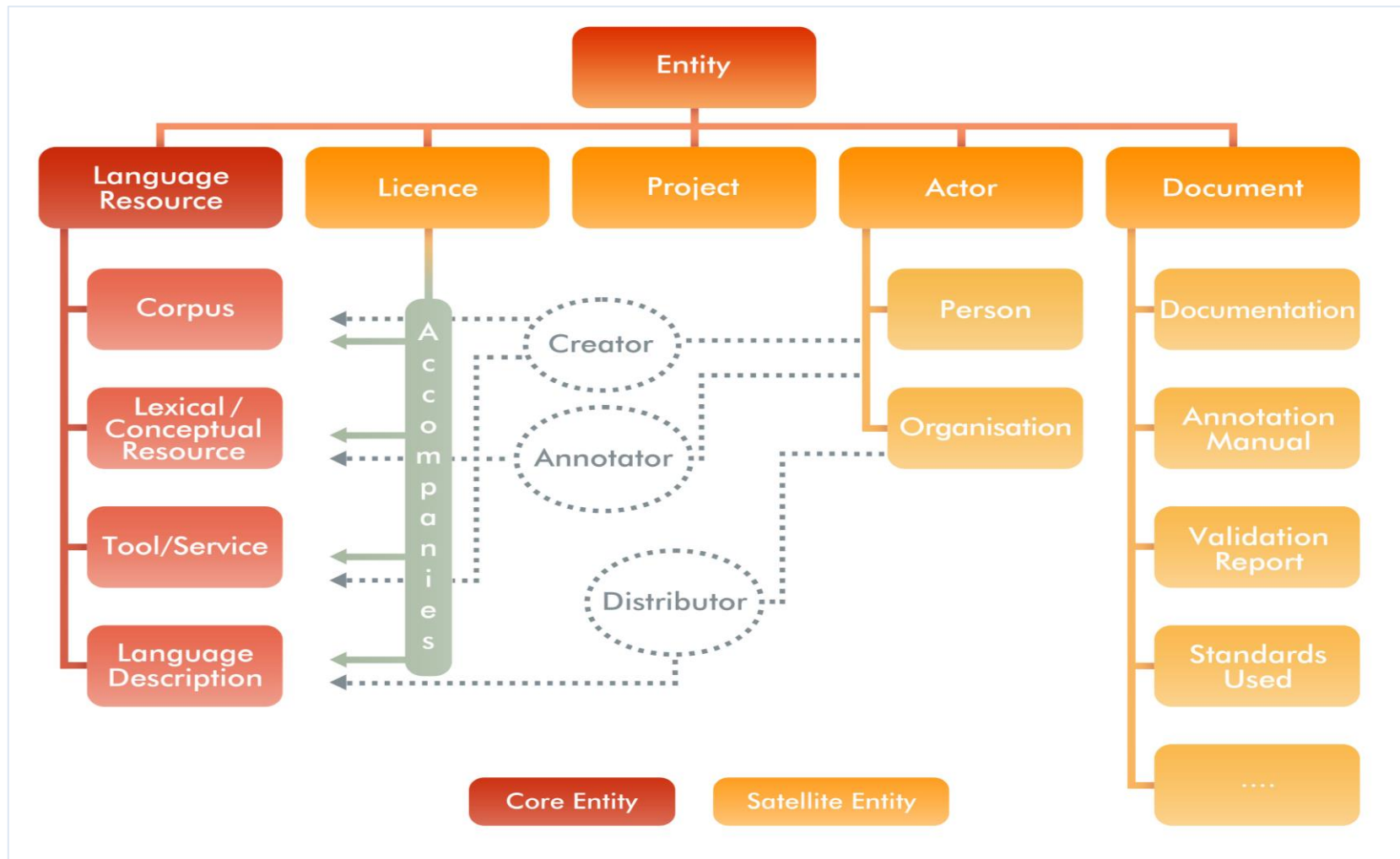
αρχιτεκτονική υποδομής & σχήμα τεκμηρίωσης

αρχιτεκτονική clarin:el

κεντρικός
συσσωρευτής



οντολογία – οντότητες περιγραφής clarin:el





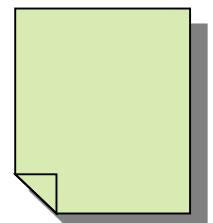
τυπολογία πόρων (1)

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

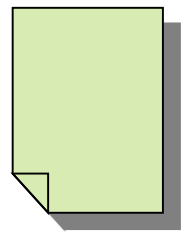
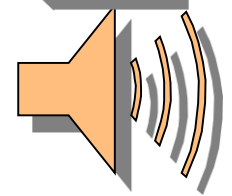


τυπολογία πόρων (2)

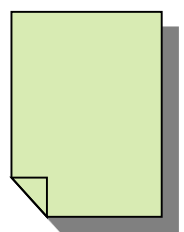
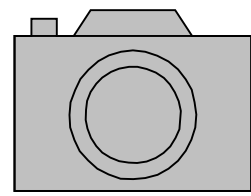
- τύπος μέσου: κείμενο, ήχος, εικόνα, βίντεο



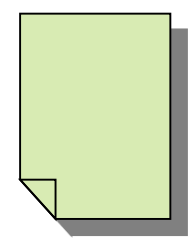
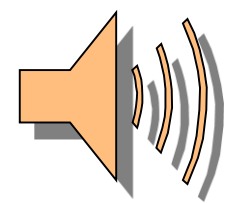
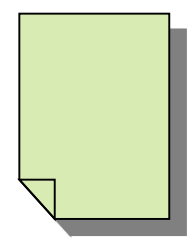
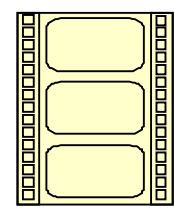
σώματα γραπτού λόγου



σώματα προφορικού λόγου



εικόνες

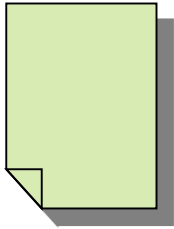


βίντεο

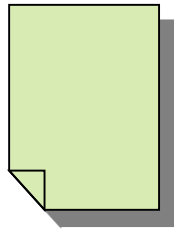


τυπολογία πόρων (3)

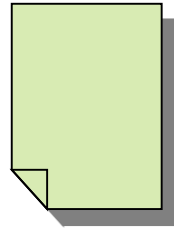
αναζήτηση "κειμένου"



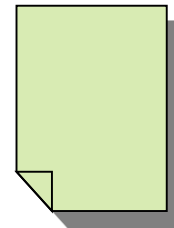
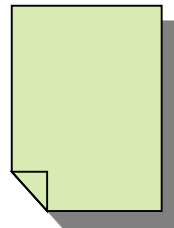
σώματα γραπτού λόγου



σώματα προφορικού λόγου

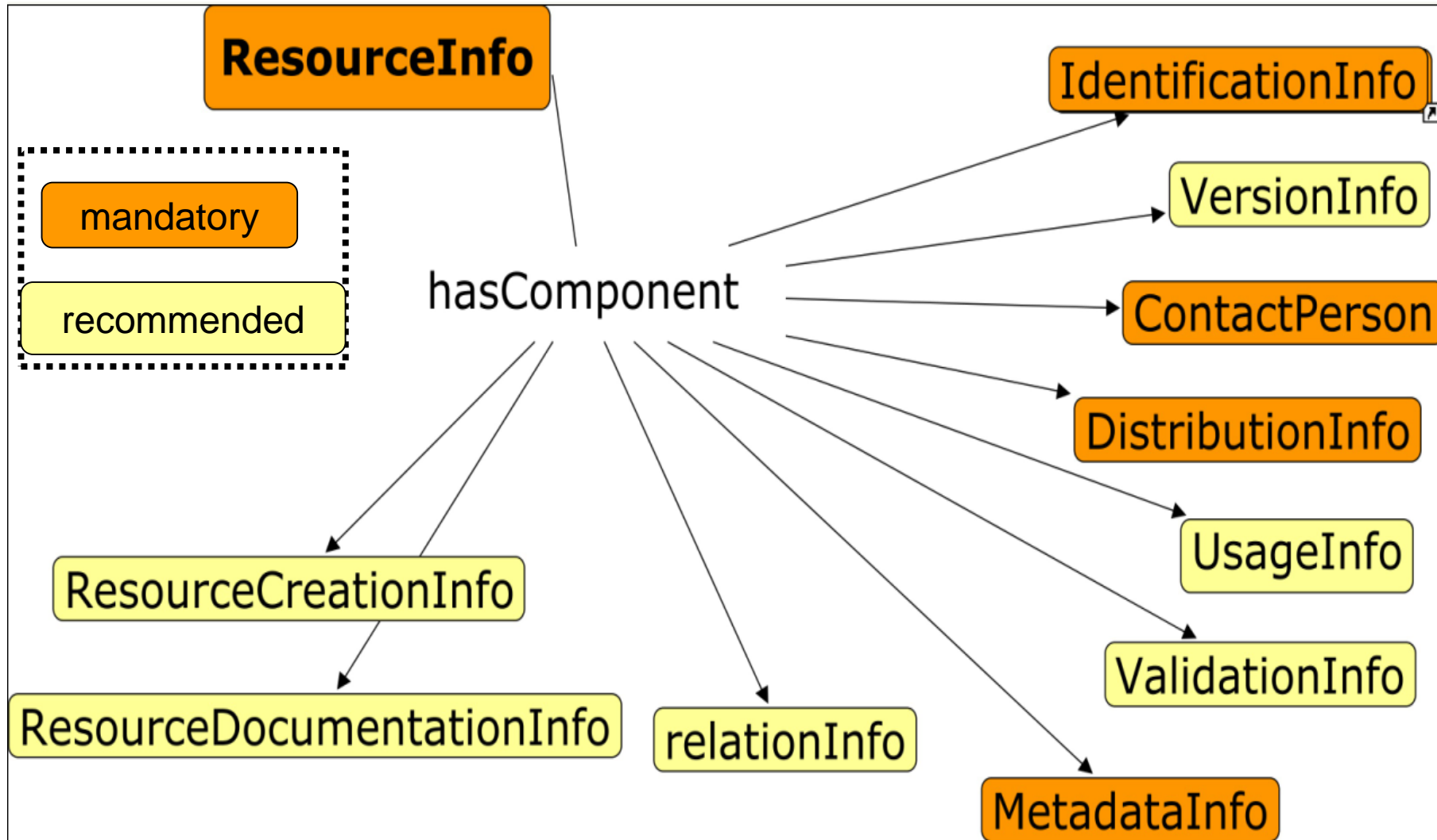


εικόνες



βίντεο

σχήμα τεκμηρίωσης: υποχρεωτικά και προαιρετικά συστατικά πληροφορίας



δίκτυο clarin:el

χρήστες υποδομής

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



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



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

περισσότερες πληροφορίες



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



info@clarin.gr



clarin.gr



@CLARIN_el



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

Ευχαριστώ πολύ!

