

the problem

requirements



- what does he need?
 - language data
 - corpora (texts authored by the politicians but also general language corpora)
 - dictionaries / domain vocabularies / semantic lexica ...
 - tools
 - for processing and annotating
 - for analyzing and visualizing results
- where will he find them?
 - original providers
 - archives and collections
 - other universities & research centers
 - conference proceedings
 - his colleagues' drawers
- at what state?
 - raw text – ideally, in digital & processable form
 - annotated
- further requirements
 - technical know-how
 - computational power
 - legal problems



the solution: CLARIN

- CLARIN (www.clarin.eu) aims to construct an integrated interoperable research infrastructure bringing together Language Resources and Technologies
- fighting against the current fragmentation
- offering a stable, robust, user-friendly and extensible virtual workplace for accessing language data
- at the service of all researchers (focusing on SSH)

the CLARIN vision

- A researcher from her office at Corfu will be able to:
 - log in with her academic account and be authenticated only with that (Single Sign-On)
 - search, find and get the ok to use texts
 - from Oxford, Bergen and Leiden,
 - select the texts she wishes to work with and save her choice as a new collection
 - run on this collection semantic taggers from Athens and
 - statistical tools from Budapest
 - using the computational power of another computational centre, when and where required
 - save the process and results of this analysis and
 - share them with her collaborators in Paris, Vienna and Helsinki

that is...

- CLARIN aims to integrate
 - **Language Datasets:** digital content of any media type (text, sound, image, video) raw and annotated, lexica, ontologies, grammars etc.
 - **Language Technology tools:** voice recognisers, lemmatisers, taggers, term extractors etc.
- in a federation of trusted repositories which will be available to all researchers
- through **national networks of organisations within each country**, focusing on and promoting resources of/about/for the languages of each country (today: [more than 200 members from 33 countries](#))

the Greek infrastructure clarin:el

clarin in Greece

- at the construction phase (2012 – 2015)
- member of CLARIN ERIC since February 2015
- clarin:el implemented as two closely interrelated subsystems for:
 - **documenting**, depositing, **sharing** + searching, retrieving and **downloading** language resources (resources infrastructure)
 - **processing** language data by web services of language processing and producing new data (processing infrastructure)
- as a first step, aspires to offer the resources a researcher needs
 - in 5 steps
 - with an additional 6th step, when this is possible...

1. keyword search, e.g. Greek corpus in the legal domain



The screenshot shows the clarin:el website interface. At the top right, there are links for 'your profile, penny' and 'logout'. Below these are navigation links: 'browse resources', 'community', 'statistics', 'help', and 'about'. The main content area features a header with the text 'central catalogue of language resources and services' and a 'go to the athena rc repository' button. Below this is a search bar containing the text 'corpus greek law' and a 'search' button. A paragraph below the search bar reads: '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.' At the bottom of the page, there is a row of logos for various institutions and a footer with the text '©clarin:el 2015 terms of service'.

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

©clarin:el 2015 terms of service



Co-funded by Greece and the European Regional Development Fund of the European Union
(Project CLARIN Attiki, MIS 441451)

2. browsing of the results

browse resources

filter by:

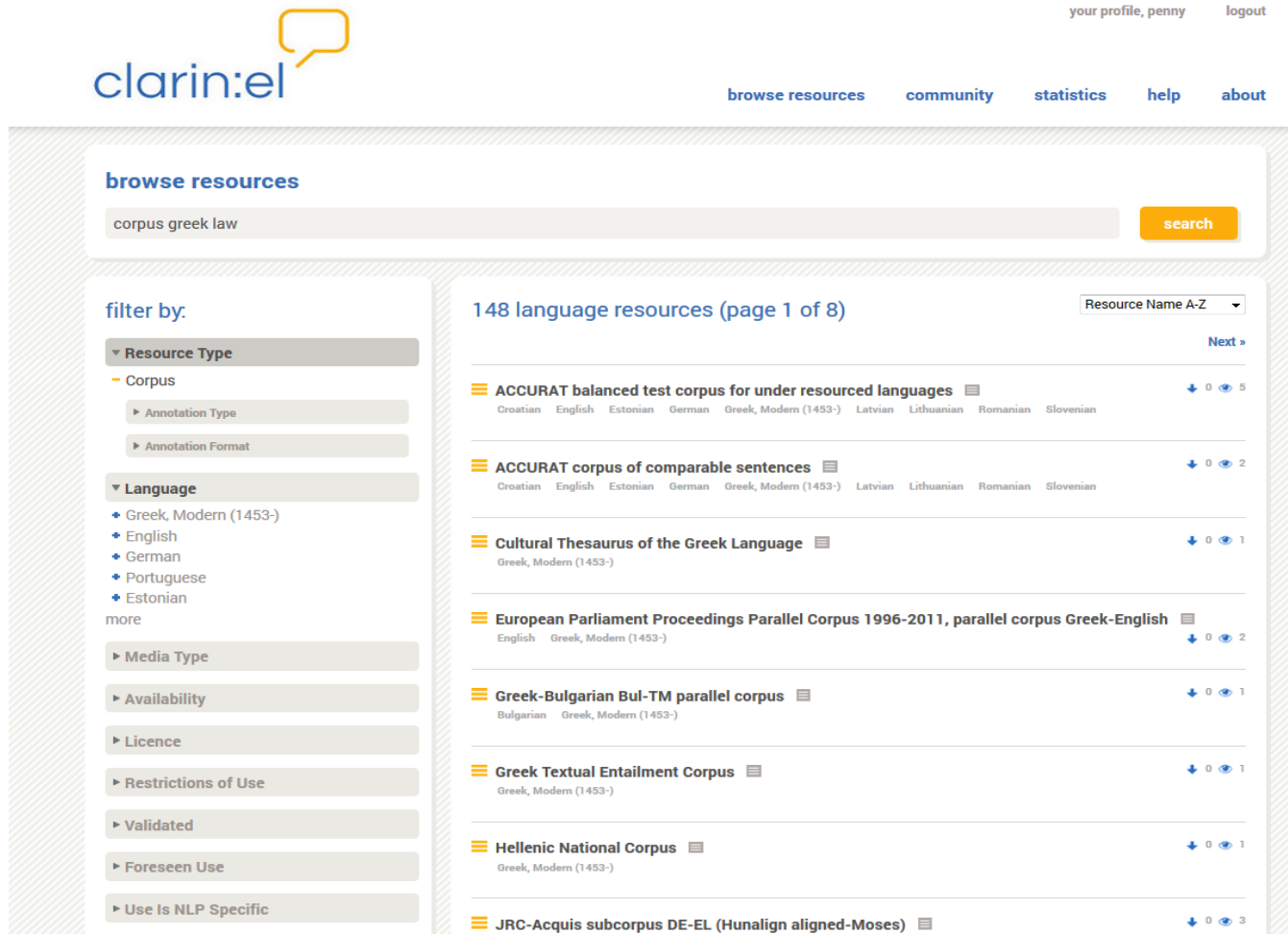
- ▶ Language
- ▶ Resource Type
- ▶ Media Type
- ▶ Availability
- ▶ Licence
- ▶ Restrictions of Use
- ▶ Validated
- ▶ Foreseen Use
- ▶ Use Is NLP Specific
- ▶ Linguality Type
- ▶ Multilinguality Type
- ▶ Modality Type
- ▶ MIME Type
- ▶ Conformance to Standards/Best Practices

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](#)

2a. filtered search



clarin:el your profile, penny logout

[browse resources](#) [community](#) [statistics](#) [help](#) [about](#)

browse resources

corpus greek law search

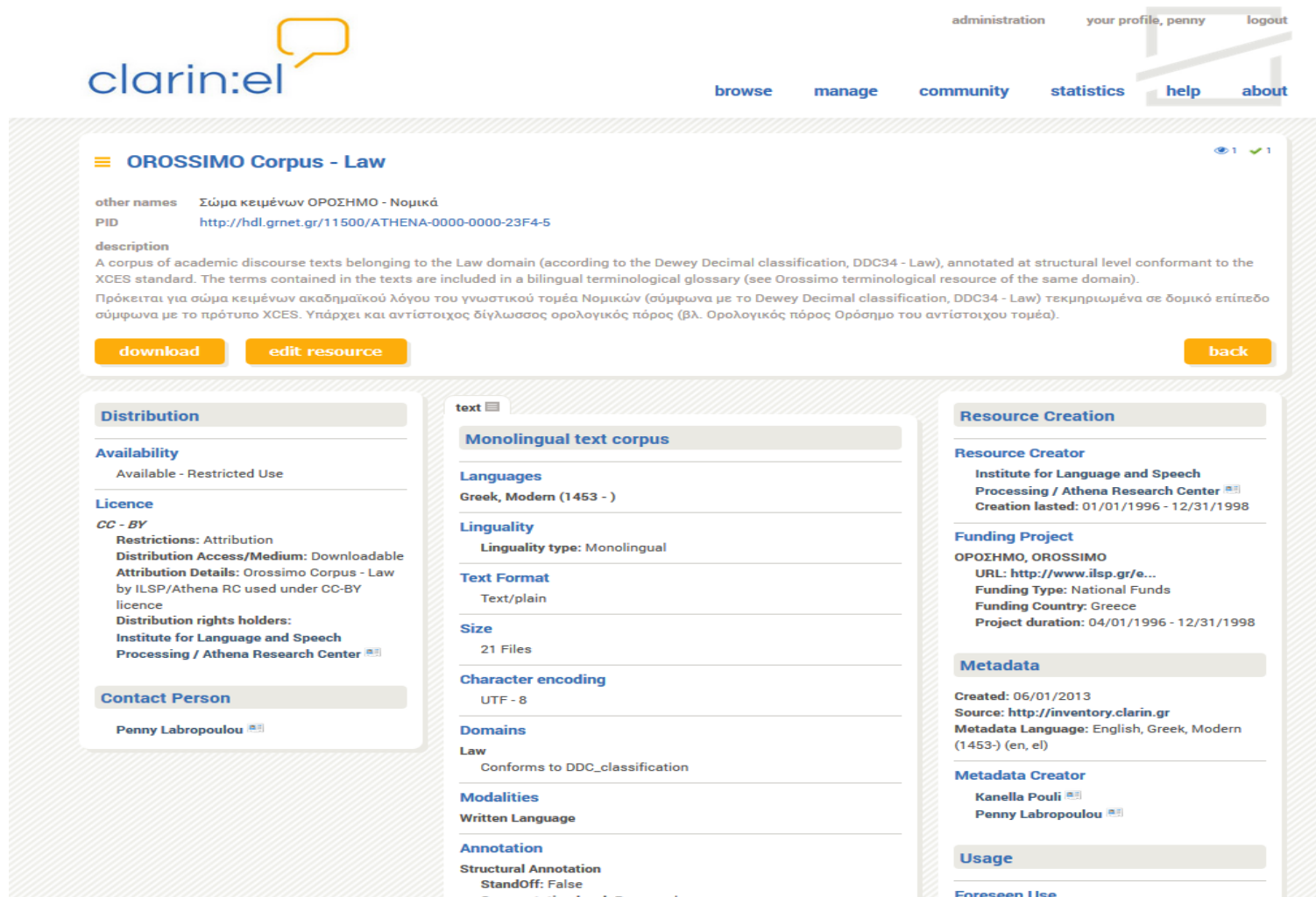
filter by:

- Resource Type**
 - Corpus
 - Annotation Type
 - Annotation Format
- Language**
 - Greek, Modern (1453-)
 - English
 - German
 - Portuguese
 - Estonian[more](#)
- Media Type
- Availability
- Licence
- Restrictions of Use
- Validated
- Foreseen Use
- Use Is NLP Specific

148 language resources (page 1 of 8) 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-)
- 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
Greek, Modern (1453-)
- JRC-Acquis subcorpus DE-EL (Hunalign aligned-Moses)** 0 3

3. viewing a selected resource



The screenshot shows the 'OROSSIMO Corpus - Law' resource page on the clarin:el website. The page includes a header with navigation links (administration, your profile, penny, logout, browse, manage, community, statistics, help, about) and a main content area with several sections:

- OROSSIMO Corpus - Law**: Includes other names (Σώμα κειμένων ΟΡΟΣΗΜΟ - Νομικά), PID (http://hdl.gnmet.gr/11500/ATHENA-0000-0000-23F4-5), and a description of the corpus.
- download** and **edit resource** buttons.
- Distribution** section: Availability (Available - Restricted Use), Licence (CC-BY), and Contact Person (Penny Labropoulou).
- text** section: Monolingual text corpus, Languages (Greek, Modern (1453 -)), Linguality (Monolingual), Text Format (Text/plain), Size (21 Files), Character encoding (UTF-8), Domains (Law), Modalities (Written Language), and Annotation (Structural Annotation).
- Resource Creation** section: Resource Creator (Institute for Language and Speech), Funding Project (OROSSHMO, OROSSIMO), and Metadata (Created: 06/01/2013, Source: http://inventory.clarin.gr).

4. licensing

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.

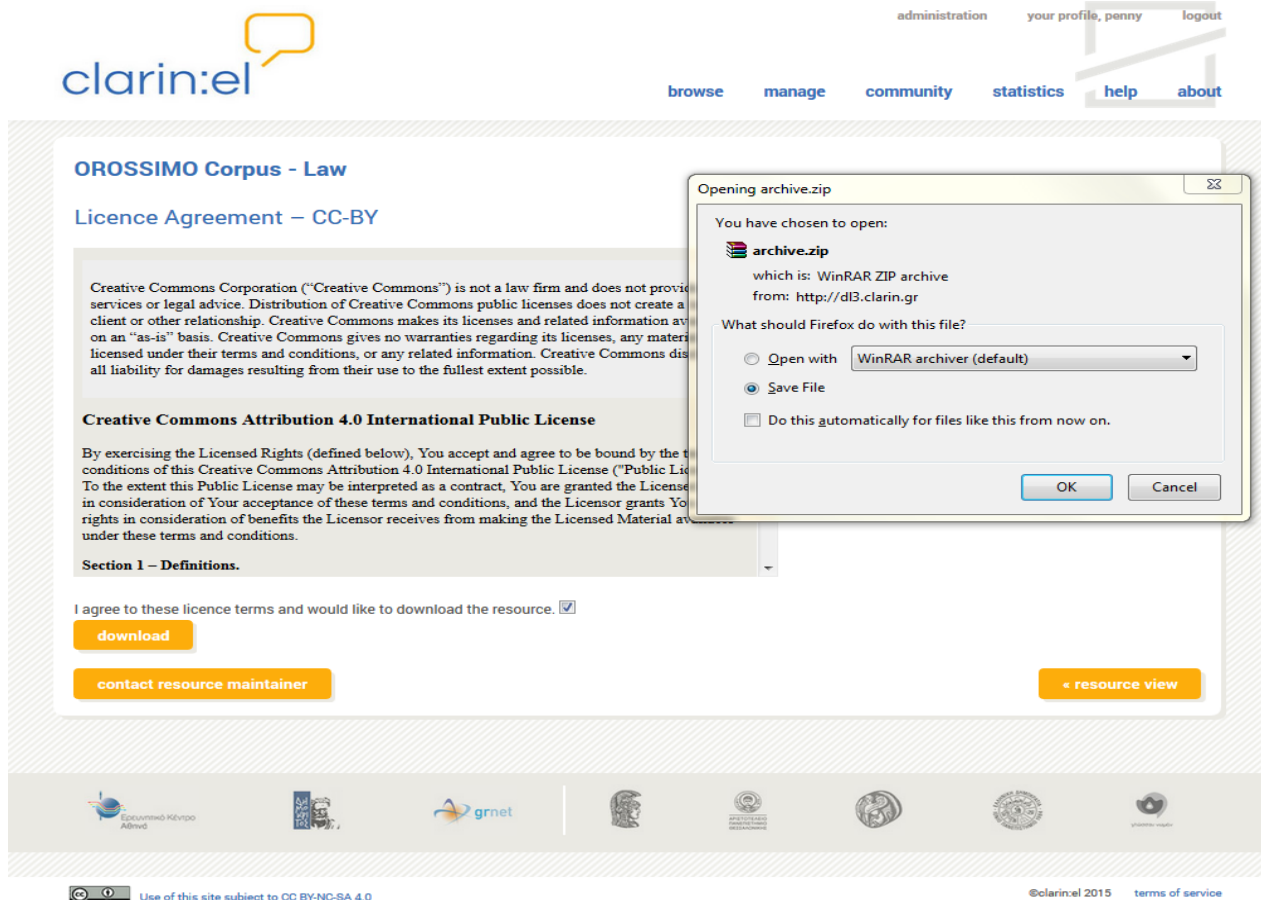
[download](#)

[contact resource maintainer](#)

[← resource view](#)

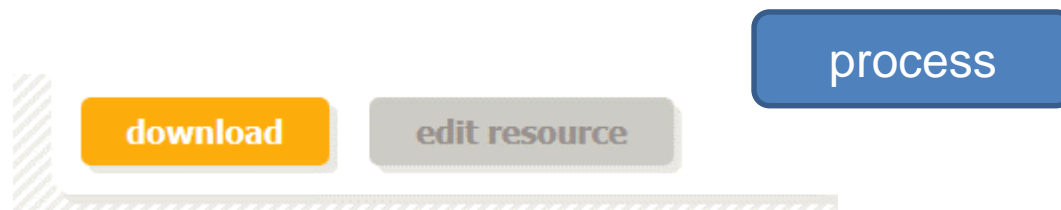


5. downloading



The screenshot shows the clarin:el website interface. At the top, there is a navigation bar with links for 'administration', 'your profile, penny', 'logout', 'browse', 'manage', 'community', 'statistics', 'help', and 'about'. The main content area is titled 'OROSSIMO Corpus - Law' and 'Licence Agreement – CC-BY'. It contains text about Creative Commons Attribution 4.0 International Public License and a 'download' button. A Firefox file dialog box is open over the page, showing 'Opening archive.zip' and options to 'Open with WinRAR archiver (default)' or 'Save File'. The 'Save File' option is selected. At the bottom of the page, there are logos for various institutions and a footer with '©clarin:el 2015 terms of service'.

6. language processing



- services for processing monolingual corpora

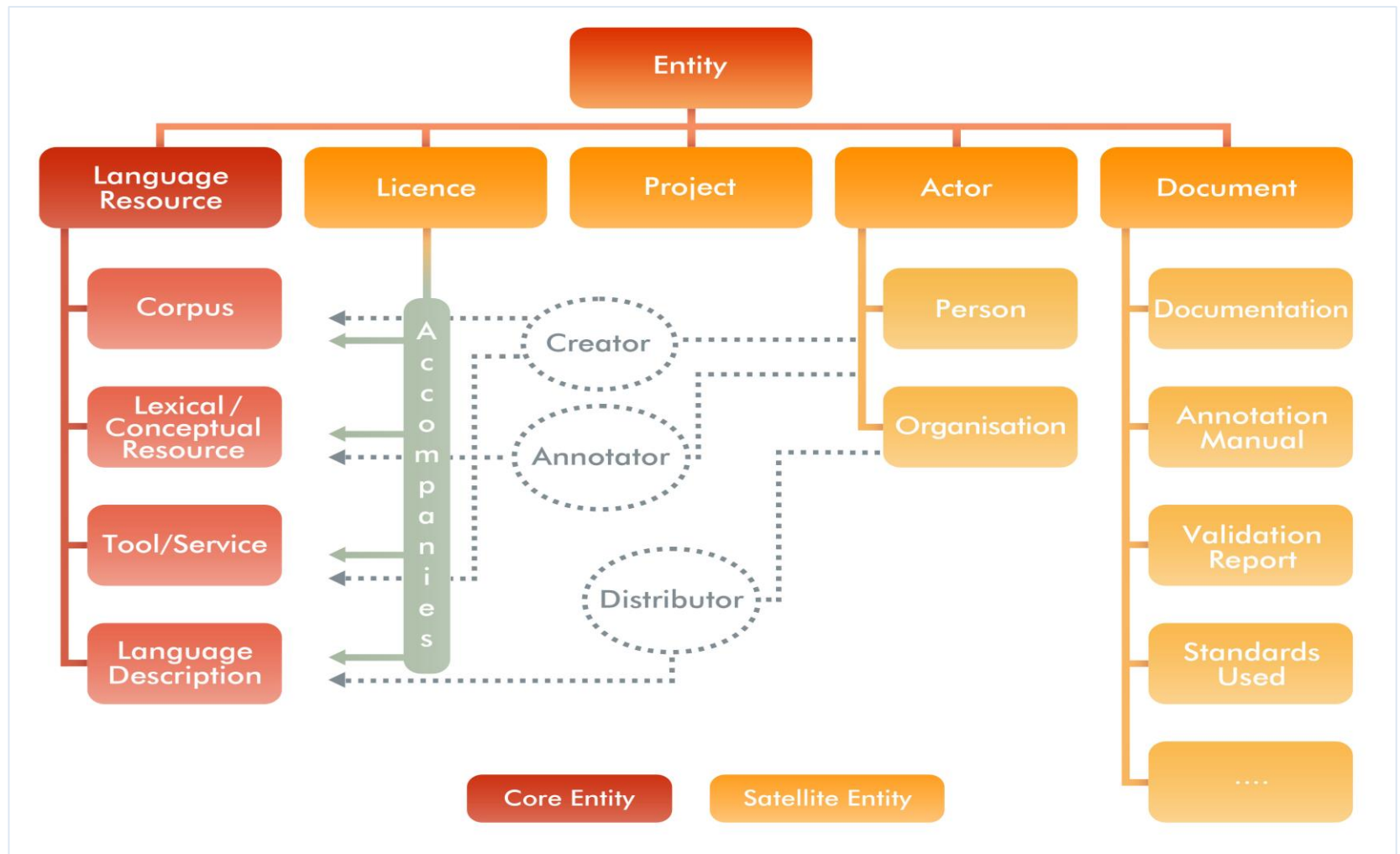
- tokenisation and sentence splitting
- part of speech tagging
- lemmatisation
- syntactic parsing
- term recognition and extraction
- named entity recognition

- services for processing multilingual corpora

- for the Greek-English pair
 - all of the above
- for the Greek-X (=EU language) pairs
 - sentence alignment

**behind all these,
a documentation model for describing
language resources and services**

ontology – description units





resource typology (1)

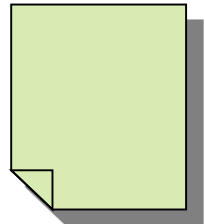
- classification on the basis of two main criteria: resource type & media type
- resource type
 - **corpus** (written / spoken / multimedia / multimodal corpora)
 - **lexical / conceptual resource** (e.g. dictionary, terminological resource, word list, ontology etc.)
 - **language description** (e.g. language model, computational grammar etc.)
 - **tool / service** (e.g. lemmatiser, annotator, machine translation tool etc.)



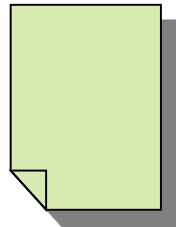
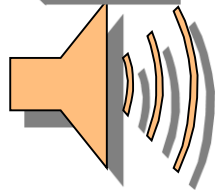
resource typology (2)

clarin:el

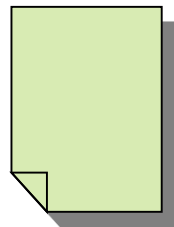
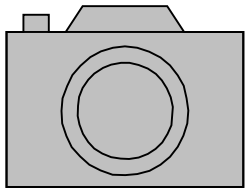
- media type: text, sound, image, video



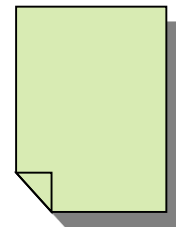
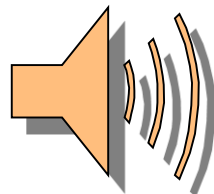
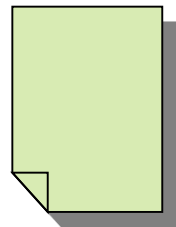
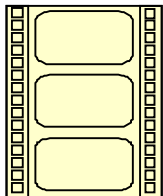
written corpora



spoken corpora



pictures

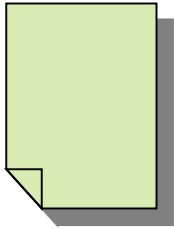


videos

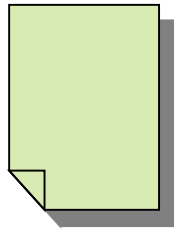


resource typology (3)

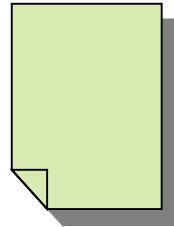
search for "text"



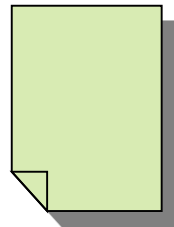
written corpora



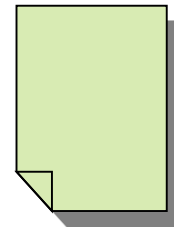
spoken corpora

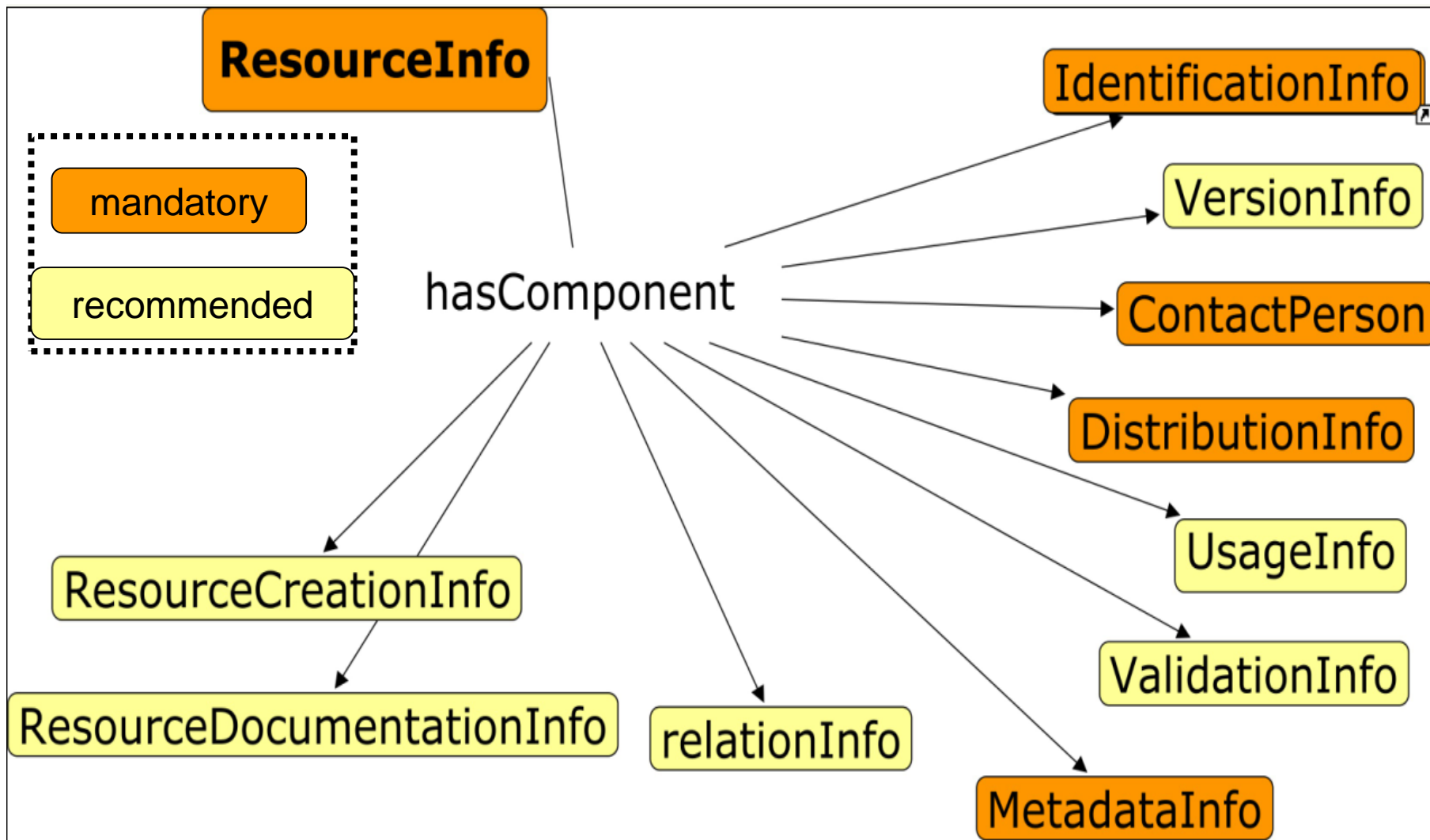


pictures



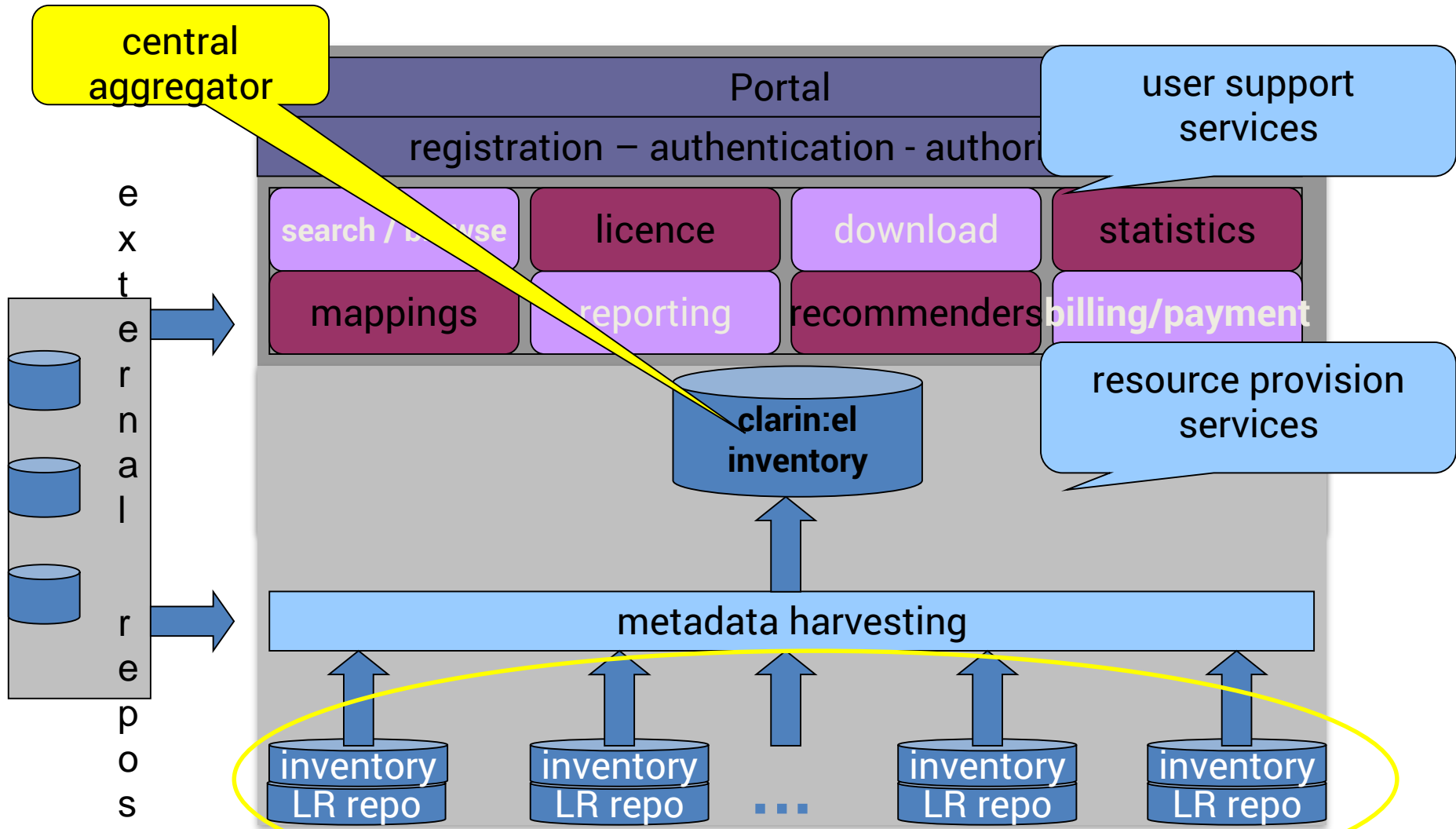
videos





architecture of the infrastructure and network

clarin:el architecture



clarin:el users

- behind the infrastructure, the content!
- users: both **consumers** and **providers**
- why should one share their resources;
- sharing, enrichment, exploitation, proliferation



clarin:el network

- only LR providers
- network members can be
 - institutions (with or without their own institutional repos)
 - collaborating researchers (at the Hosting Repository)
- current members – Construction Phase (until 31/12/2015)



- Pilot Operation Phase (from 1/1/2016)
 - All academic and research institutions

more information



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



info@clarin.gr



clarin.gr



@CLARIN_el



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

Thank you!

