
CLARIN

Release 1

CLARIN Technical Team

Jul 26, 2023

INTRODUCTION

1	About CLARIN:EL	3
2	How to use this manual	5
3	Contact & Help	7
4	Registration/Sign in	9
4.1	Register	9
4.2	Sign in	11
5	Users	13
5.1	Registered vs Non registered	13
5.2	Types of registered users	13
6	What can I find in the CLARIN:EL central inventory?	15
7	When is the content of a resource accessible?	17
8	Central Inventory Page	19
9	Menu for registered users	21
9.1	Dashboard	22
9.2	Help	33
9.3	Your name	33
9.4	Exit	33
10	Browsing	35
11	Searching	37
12	Viewing & Downloading	41
12.1	The upper section	41
12.2	The lower section	42
13	Processing	49
13.1	1. Starting with the data	49
13.2	2. Starting with the Function	52
14	The lifecycle of a resource	59
15	Important information about all LRTs	61

16	How to create a resource by using the editor	63
16.1	Step 1: Name your resource	64
16.2	Step 2: Upload the resource data	65
16.3	Step 3: Fill in the mandatory metadata	71
16.4	Step 4: View the created record	73
17	How to create a resource by uploading an XML file	75
18	How to prepare data before uploading	79
18.1	I. Depositing data as content of a resource	79
18.2	II. Data as input of a service	83
19	Recommended File Formats	85
19.1	Guidance on selecting file formats for long-term accessibility and interoperability	85
20	What are metadata and why are they important?	87
21	FAIR principles	91
21.1	Findability	91
21.2	Accessibility	92
21.3	Interoperability	92
21.4	Reuse	93
22	Mandatory metadata	95
22.1	Common Mandatory Elements	95
22.2	Mandatory Elements per resource type	96
23	General guidelines on metadata	99
23.1	Language	99
23.2	Consistency	100
23.3	Completeness	100
23.4	Editing	100
23.5	Versioning	100
24	Specific guidelines on mandatory metadata	103
24.1	1. resourceName	103
24.2	2. description	103
24.3	3. version	104
24.4	4. keyword	105
24.5	5. additionalInformation	105
24.6	6. distribution related metadata	106
24.7	7. licenceTerms related metadata	107
24.8	8. data	108
24.9	9. personalData, sensitiveData & anonymized	108
24.10	10. Subclass related metadata	110
24.11	11. encodingLevel	112
24.12	12. function	113
24.13	13. inputContentResource	114
25	Examples of metadata	117
25.1	resourceName	117
25.2	resourceCreator	117
25.3	isPartOf	118
25.4	annotationType	119
25.5	multilingualityType	119

25.6	isDocumentedBy	120
25.7	fundingProject	120
25.8	inputContentResource	121
25.9	outputResource	122
25.10	attributionText	123
26	XML metadata descriptions	125
26.1	1. Corpora	125
26.2	2. Lexical/Conceptual resources (LCR)	144
26.3	3. Tool/Services	159
26.4	4. Language Descriptions	167
27	Full schema documentation	173
28	Actions on resources	175
28.1	I. Per resource status and user type	175
28.2	II. The actions	177
29	Information on Legal Issues	189
30	Publications	191
31	Indices and tables	193

Welcome to the user manual of CLARIN:EL!

Cite this version:

Pouli Kanella, Bakagianni Juli, Galanis Dimitris, Labropoulou Penny, Tsiouli Iro, Gavriilidou Maria. 2023. The CLARIN:EL User Manual, v.1.0.

ABOUT CLARIN:EL

CLARIN:EL is a Research Infrastructure for *Language Resources & Technologies* (LRTs); it is the greek part of the European CLARIN ERIC Infrastructure. It provides a multitude of **assets related to Language Technology (LT)** for and by **Social Sciences and Humanities** (and beyond), focusing mainly but not exclusively on **Greek LRTs**. The CLARIN:EL Research Infrastructure operates as a distributed network of repositories consisting of:

- the Institutional Repositories (created for each Organisation participating in the CLARIN:EL network) and
- the Hosted Resources Repository (HRR), maintained by ATHENA RC.

Follow the links¹ to see the contents of each repository:

- [Athena Research Centre](#),
- [Aristotle University of Thessaloniki](#),
- [Athens University of Economics and Business](#),
- [University of the Aegean](#),
- [National and Kapodistrian University of Athens](#),
- [Centre for the Greek Language](#),
- [Hosted Resources repository](#),
- [University of Crete](#),
- [Ionian University](#),
- [National Centre of Social Research \(EKKE\)](#),
- [National Centre for Scientific Research “Demokritos” \(NCSR\)](#),
- [Panteion University](#), and
- [University of West Attica](#).

In the central inventory you will find the complete list of LRTs that have been published in the aforementioned repositories.

Tip: See how *users* are connected to their repositories and the whole infrastructure.

¹ When there is no link, the respective organization has not created any resources yet.

HOW TO USE THIS MANUAL

This manual¹ aims to help you explore and/or use the CLARIN:EL infrastructure to make your resources available to the **Humanities and Social Sciences** community (and beyond). It is not meant to be read in sequence (although it can be) but to help you find specific information depending on your needs. There are chapters with general information on *basic concepts*; others describing the *process*² through which a resource comes to life and chapters which will specifically help you:

- to *browse* and *search* through the central inventory so as to find resources to *download* and *process*,
- to create resources via the *metadata editor*³ or by *uploading XML files*, and
- to perform *actions on resources* depending on your role.

If you are looking for something specific, please, use the search box on the top left side of the navigation bar, below the CLARIN:EL logo.

¹ The current version documents the third official release of the CLARIN:EL infrastructure, launched on May 31st, 2021. More functionalities are continuously added, and this manual keeps on being updated following the evolution of the CLARIN:EL platform.

² Before you start, please see *the lifecycle of a resource* to find out what each type of user needs to do throughout the procedure. To assume a role you must have *registered* first.

³ The terms **metadata editor** and **editor** are used interchangeably throughout the documentation.

CONTACT & HELP

There are three helpdesks to help you with any questions you might have: for [technical and management issues](#), [legal issues](#), and issues related to [metadata creation and documentation](#).

Links to the helpdesks (as well as FAQs and bibliography) are found at the bottom of each page in the infrastructure.

The screenshot shows the bottom section of the CLARIN:EL portal. At the top left is the 'clarin:el' logo with a 'BETA' badge. To its right are links for 'Help' and 'Sign in', and a 'CLARIN:EL portal >' button. Below this is a section titled 'Join the CLARIN:EL community!' with a brief description and a 'Learn more' button. The footer is dark grey and contains social media icons (Facebook, Twitter, YouTube), copyright information '© CLARIN:EL 2021 Terms of service | Privacy Policy', and a 'Need Help ?' section with links to 'CLARIN:EL Helpdesks', 'FAQs', and 'Bibliography'. At the very bottom is a row of logos for partner institutions and funding bodies, including 'απολλωvις', 'CLARIN', 'ΑΕΘΝΑ', 'ΑΔΜΚΠΤΟ', 'grnet', 'ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ', 'ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ', 'Ευρωπαϊκή Ένωση', 'ΕΠΛΑΝΕΚ 2014-2020', and 'ΕΣΠΑ 2014-2020'. A small upward arrow icon is on the right side of the footer.

REGISTRATION/SIGN IN

This chapter provides information to the users

1. who already have an academic account and can use it to *sign in*,
2. who do **not** have an academic account and **must first** *register* so as to create a personal account which they will later on use to *sign in*.

4.1 Register

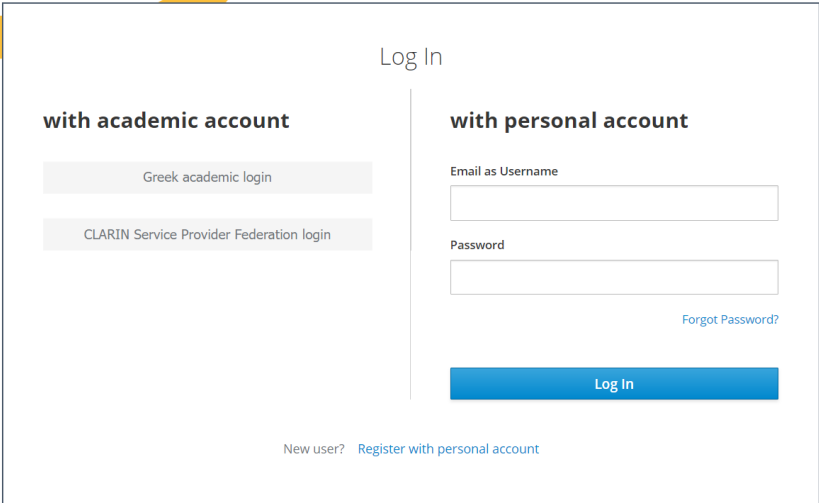
Attention: Skip this section, if you already have an accademic account which you can use to *sign in*.

To register in CLARIN:EL, follow the next steps:

- Click on the **sign in** button at the top right of the page.



- In the next window choose to **Register with personal account**.



CLARIN:EL

Log In

with academic account

Greek academic login

CLARIN Service Provider Federation login

with personal account

Email as Username

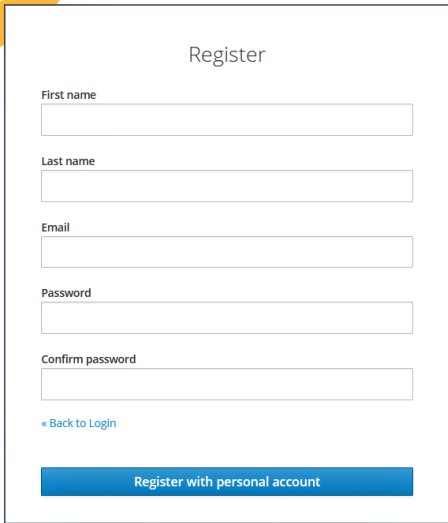
Password

[Forgot Password?](#)

Log In

New user? [Register with personal account](#)

- Then provide all the necessary information in the form that appears and click on **Register with personal account**.



CLARIN:EL

Register

First name

Last name

Email

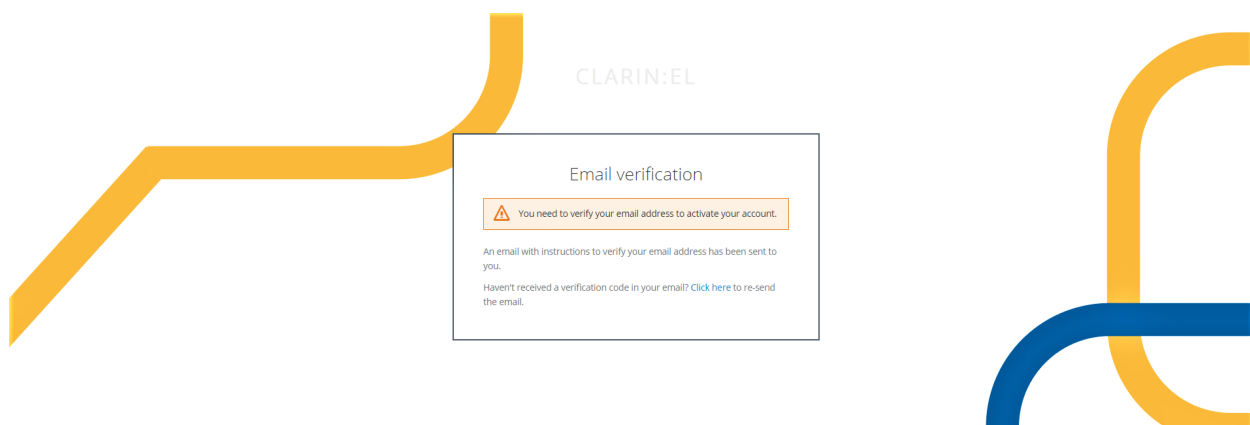
Password

Confirm password

[« Back to Login](#)

Register with personal account

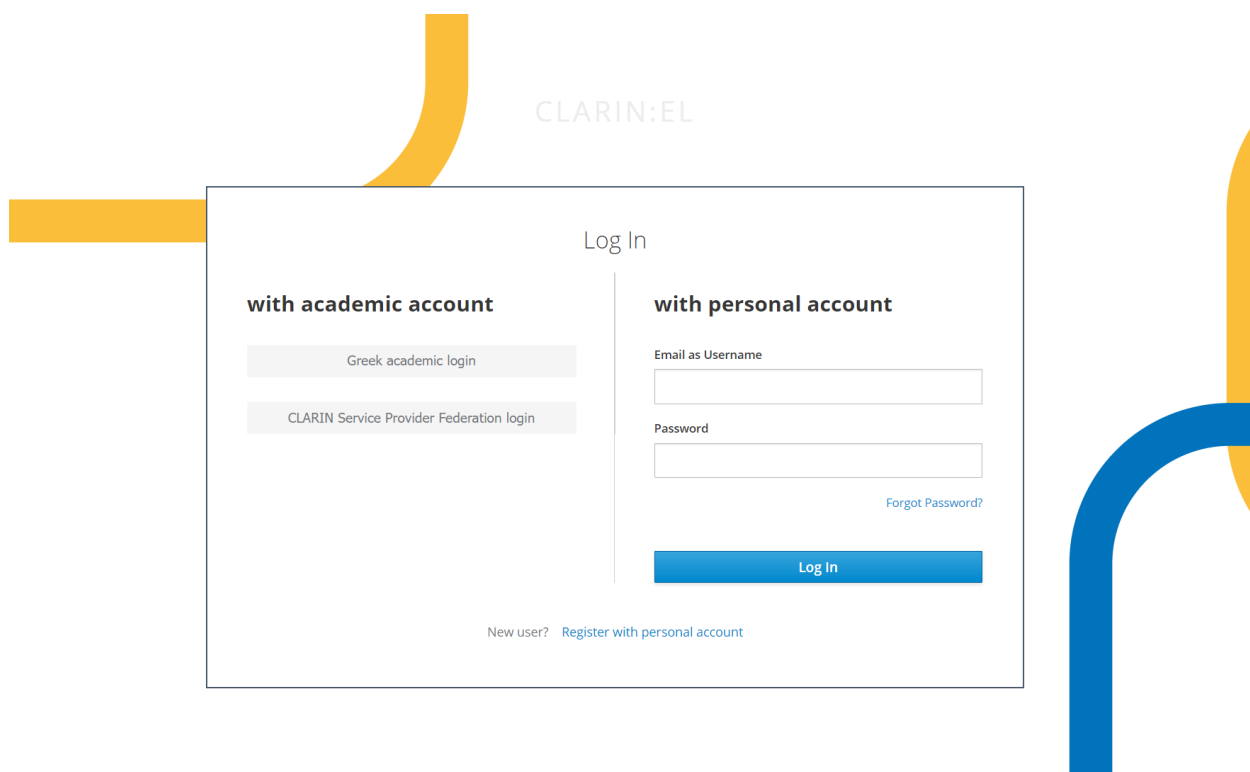
- You will receive an email with a link to confirm your address and agree to the CLARIN:EL Terms of Service & Privacy Policy.



- Once confirmed, your account is activated. After you have registered, every time you would like to use the infrastructure, simply **sign in**.


4.2 Sign in

You can sign in using either your personal or academic (Greek or Federation) account which will uniquely associate you with your organization or academic institution and consequently with the respective repository. If you are not affiliated with a specific organization/ academic institution, you will be directed to the **Hosted Resources Repository**.



To use your academic account click on **Greek academic login**. You will be redirected to a new page where you should type in the box the name of your organization/academic institution and then select **confirm**.

[GRNET AAI](#) [Select your home institution](#) [Participants](#) [Services](#) [Documentation](#) [Help](#) [Ελληνικά](#) [English](#)



GRNET AAI Federation

Authentication & Authorization Infrastructure


You were redirected to this page because you tried to access a service that participates in DELOS Federation. In order to proceed, you have to select your Home Organization from the list below. You may save your selection, in order to avoid this question during future access attempts.

Athena - Research and Innovation Center

Confirm

Save my preference: ☐

Again, in the new window that opens, use your academic credentials and finally click on **login**.



Username (user@athenarc.gr ή user@*.athena-innovation.gr)

Password

☐ Don't Remember Login

☐ Clear prior granting of permission for release of your information to this service.

Login

> Forgot your password?

> Need Help?

You are signed in!

USERS

Users have various rights depending on whether they have registered or not. Unregistered users can navigate the infrastructure with some permissions. Advanced permissions associated with the *creation*, *management* and *processing* of resources are given only to registered users according to the policy of the infrastructure.

5.1 Registered vs Non registered

The central inventory presents the published LRTs from all repositories and can be accessed with or without registration. As concerns the **consumption** of resources, there is only one extra permission the registered user has which is to use the tools and services, as shown in the table below.

Tip: Click on each of the inventory uses to find out more!

5.2 Types of registered users

Once you have registered, you must *sign in* if you wish to be directed to your repository. This will enable you to engage in the *the lifecycle of a resource* by assuming one of the following roles.

1. **Curator:** the curator is responsible for creating resources and uploading their content files, as well as managing (editing, updating, etc.) them and finally submitting them for publication.

Note: By *signing in*, you **automatically** obtain the **curator** status in your repository.

2. **Validator:** the validator is assigned resources by the supervisor in order to check whether the metadata described (and the content files uploaded) are consistent - if not, the resource returns to the curator to be edited again according to the validator's comments.

Note: See *here* how a user becomes validator.

3. **Supervisor:** the supervisor has the final word before a resource is made public and is also the only one able to unpublish it, if necessary.

Note: See *here* how a user becomes supervisor.

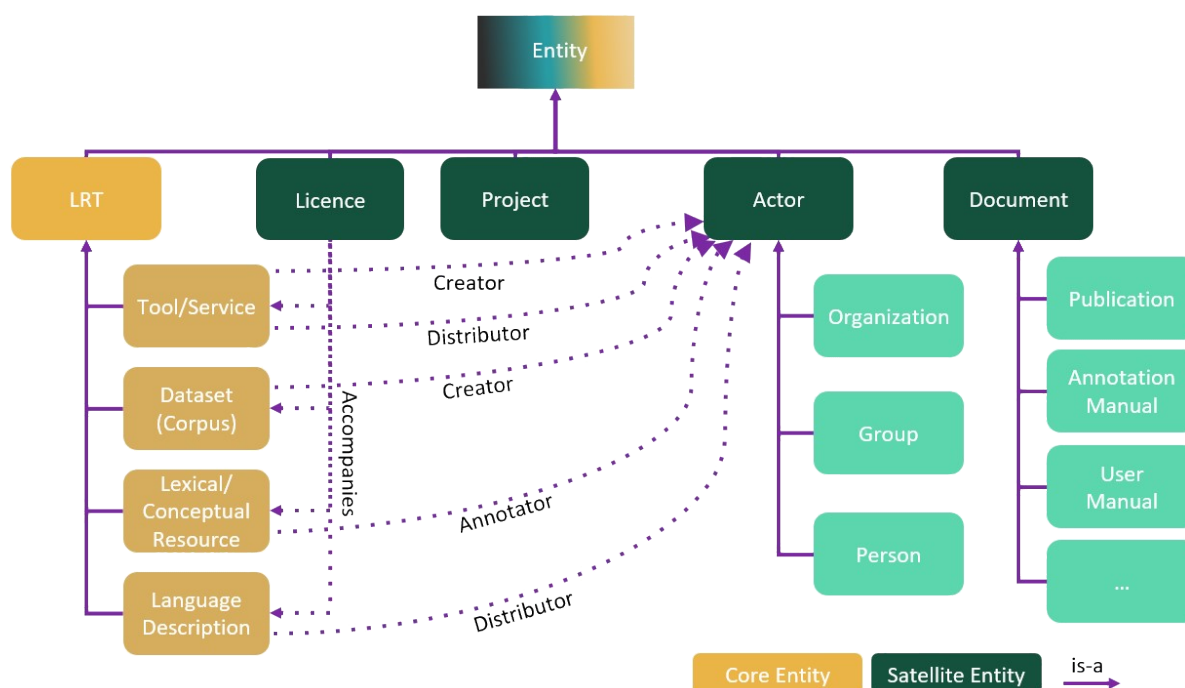
Each one controls a number of resources on which *actions* can be performed via the *dashboard* and the *resource view page*.

WHAT CAN I FIND IN THE CLARIN:EL CENTRAL INVENTORY?

The **CLARIN:EL** infrastructure includes Language Resources and Technologies (LRTs), which can be further classified according to their content into:

- **corpora** a.k.a datasets: collections of text documents, audio transcripts, audio and video recordings, etc. (for the corpora which can be used for processing see [here](#)),
- **lexical/conceptual resources**, comprising computational lexica, gazetteers, ontologies, term lists, etc.
- **tools & services**: any type of software used for LT processing (for the services integrated in the infrastructure see [here](#)), and
- models & computational grammars, collectively referred to as **language descriptions**.

The following image depicts the taxonomy of resources in relation with other entities, such as the actor -i.e. the creator, contributor or annotator- who can be a person, a group of people, or an organization.



Typically a resource consists of a description (the metadata record) and content files (e.g. the dataset for a corpus, the software for a tool etc.). A **description** is a sine qua non condition for a resource to be in the central inventory. However, a description may or may not be accompanied by content files. Therefore, the following combinations exist:

1. Resource descriptions along **with** content files,
 - 1.1 available through [CLARIN:EL](#), or

1.2 available via an external link directing to another website or via an interface needed to access the resource.

2. Resource descriptions **without** content files, which are

2.1 **for information purposes only** (the content files will be uploaded later), or

2.2 **metaresources** (there are no content files to be uploaded); these include bibliographies, conference proceedings, etc.

As concerns resources **with** content files, you can find out [here](#) the conditions under which they are accessible.

Tip: See [here](#) some important information on creating and sharing resources via the CLARIN:EL infrastructure.

WHEN IS THE CONTENT OF A RESOURCE ACCESSIBLE?

Attention: This section is for resource descriptions which **come with** content files. To see the different types of resources found in the CLARIN:EL central inventory, please see [here](#).

While navigating through the infrastructure you will see metadata records many of which, as explained earlier, come with content files. In order for the content files of a resource to be accessible two criteria must be met beforehand:

1. a resource needs to be provided under an **open access licence** (check [here](#) the *Recommended licensing scheme for Language Resources*), and
2. the resource content files must **have been uploaded** or **stored** at an access point.

This holds true both for those resources available through the infrastructure and via external links. For the latter, several other conditions must be met upon:

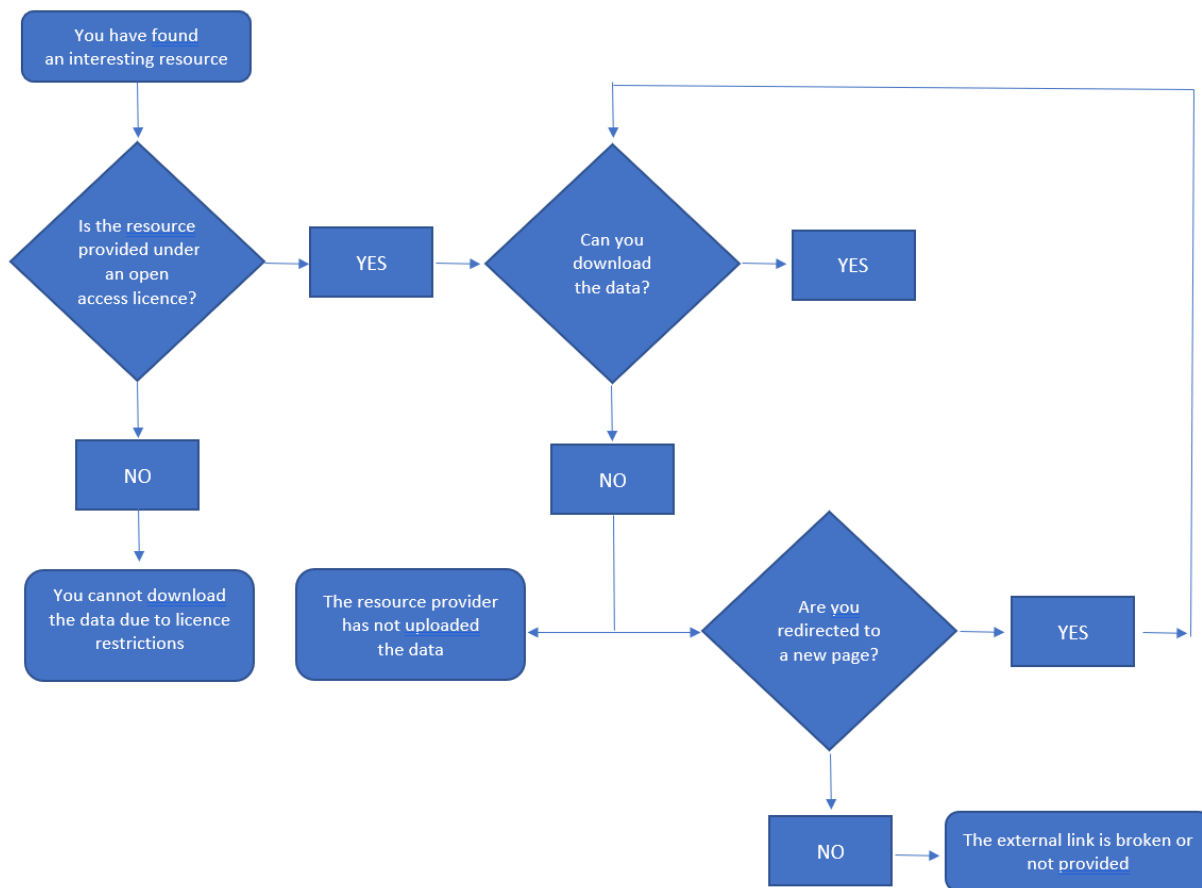
- the link must be provided in the appropriate metadata field in the metadata editor¹/xml file,
- the link must work (not be broken), and
- the content of the link must be well maintained.

Independently of the point of access to the data, directly or indirectly, the following table shows the possible combinations of actions which allow or not for downloading.

Does the licence provide Open Access?	Have the content files been uploaded (datasets, tools, lexica, etc.)?	Can the data be downloaded?
Yes	Yes	Yes
Yes	No	No
No	Yes	No
No	No	No

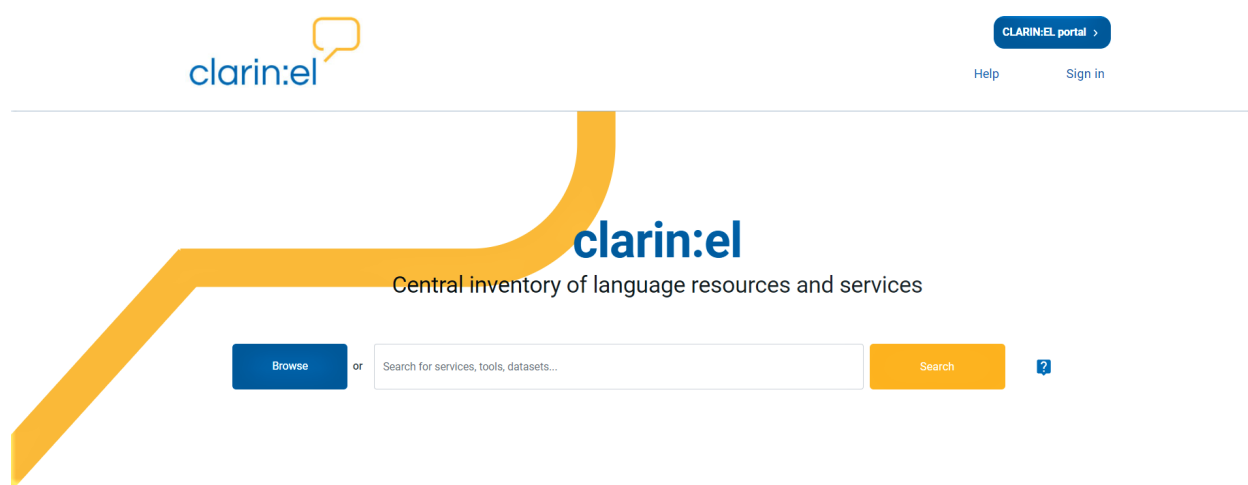
The content of the table is visualised in the following flowchart:

¹ Henceforth **editor**.



CENTRAL INVENTORY PAGE

The inventory home page of the CLARIN:EL infrastructure provides several options to the user who hasn't signed in or registered yet.



At the top right side of the page, there are three buttons:

CLARIN:EL portal: a link to the infrastructure [portal](#), where you can find information about the network, news and announcements, presentations and bibliography;

Help: a link to the infrastructure documentation and guidelines (this document), where you can learn how to navigate, create and manage resources, etc.;

Sign in: a link to a page where you can [register](#) or [sign in](#).

Following, there is a search box with an option on each side: **browse** on the left and **search** on the right. These are entry points to the central inventory. You can [browse](#) it as a whole or use [filters and keywords](#) to retrieve a subset of the LRTs that match your criteria.

In the middle of the page you can see two columns where specific groups of LRTs are presented.

Language Technology tools and services

Language processing tools, in the form of web applications or downloadable tools.

- > **CLARIN:EL Workflow registry**
Language processing web services integrated in the CLARIN:EL infrastructure, that operate at various levels of analysis.
- > **Downloadable tools**
Language processing tools which are downloaded to be run locally.
- > **Additional online services**
Language processing services available on the web or forthcoming services currently under construction.

Language resources

Corpora (i.e. structured collections of data such as text documents, audio transcripts, audio and video recordings); lexical and conceptual resources; treebanks etc.

- > **Processable corpora**
Corpora that have the appropriate format in order to be processed by the CLARIN:EL Workflows.
- > **Downloadable resources**
Resources whose license permits downloading.
- > **Metaresources**
Ancillary resources that provide information on various aspects of language and linguistics (such as bibliographical information, studies etc.).

The first column contains subsets of tools and/or services grouped according to whether:

- they are provided as [services](#) in the infrastructure,
- they can be [downloaded](#), and
- they can be [accessed online](#) through external links.

In the same way, the datasets are grouped in the next column as those which can be:

- [processed](#),
- [downloaded](#), and
- used only to [provide information](#).

At the bottom of the page there is a [link](#) for users who wish to learn more about the CLARIN:EL community and possibly join in.

Join the CLARIN:EL community!

By joining the CLARIN:EL community, you can share your own language data and language processing technologies and have access to resources and technologies from other repositories at the National and European level.

[Learn more](#)



© CLARIN:EL 2021 | [Terms of service](#) | [Privacy Policy](#)

Need Help ?

[CLARIN:EL Helpdesks](#)

[FAQs](#)

[Bibliography](#)

MENU FOR REGISTERED USERS

After you have signed in, the menu offers five options:

CLARIN:EL portal: a link to the infrastructure [portal](#), where you can find information about the network, news and announcements;

Dashboard: a link to your personalized dashboard, serving as an access point to the metadata editor¹, the resources you have created, your tasks and processing jobs.

Help: a link to the infrastructure documentation and guidelines (this document), where you can learn how to navigate, create and manage resources, etc.;

Your name: a link to your profile, which you can edit.

Exit: an icon to log out.



¹ Henceforth **editor**.

9.1 Dashboard

The Dashboard serves both as an overview page for your activities (where you can find information on your resources, tasks, processing jobs) and an entry point to *create resources* and use *workflows* to process resources. As shown below, it contains nine different sections. Sections 2 and 7-9 are slightly different depending on your role.

The dashboard is divided into several sections, each with a numbered callout:

- 1**: Welcome message and a list of actions: view and update your profile, view your resources and your tasks, access creation forms, upload resources, access the workflow registry, and browse your recent activity.
- 2**: Supervisor profile section (Supervisor H., supervisor_hosted@gmail.com) with buttons for 'View my profile' and 'Manage Repository Users'.
- 3**: 'Total resources' section showing 'Number of resources you have created' as 8, with a '+ Create resources' button.
- 4**: 'Upload resources' section with the instruction 'Upload single or multiple resources in XML format.' and a '+ Upload resources' button.
- 5**: 'Workflow registry' section with the instruction 'Select a workflow to process a corpus.' and a '+ Select workflow' button.
- 6**: 'Processing tasks' section showing a task with input/output details, submission date (19 May 2021, 11:31), a 'Finished' status, and a 'Download' button for the file 'archive.zip'. A 'View all processing tasks' button is at the bottom.
- 7**: 'Recent Activity' section with tabs for 'My Resources', 'Validation tasks' (selected), and 'My repository'.
- 8**: Table of recent activity under the 'Validation tasks' tab.
- 9**: Navigation buttons at the bottom: 'View my resources', 'View my validation tasks', and 'View my supervision tasks'.

Date created	Title
19 May 2021	draft Single DE (raw corpus)
19 May 2021	draft Single DE
19 May 2021	draft Threeparia
06 May 2021	Single text EL (txt)

Tip: See here how the dashboard looks for a curator, a validator and a supervisor.

9.1.1 1. Welcome

This section is introductory and informs you on what you can do here.

Welcome to your dashboard !

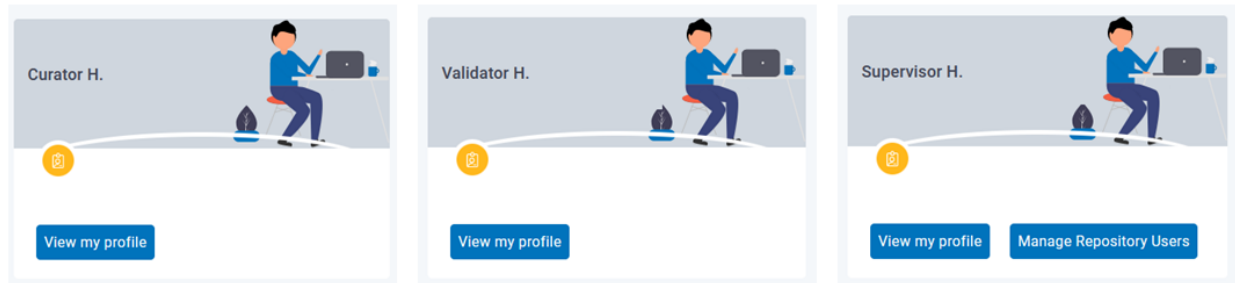
Here you can:

- view and update your profile
- view your resources and your tasks
- access creation forms
- upload resources
- access the workflow registry
- browse your recent activity

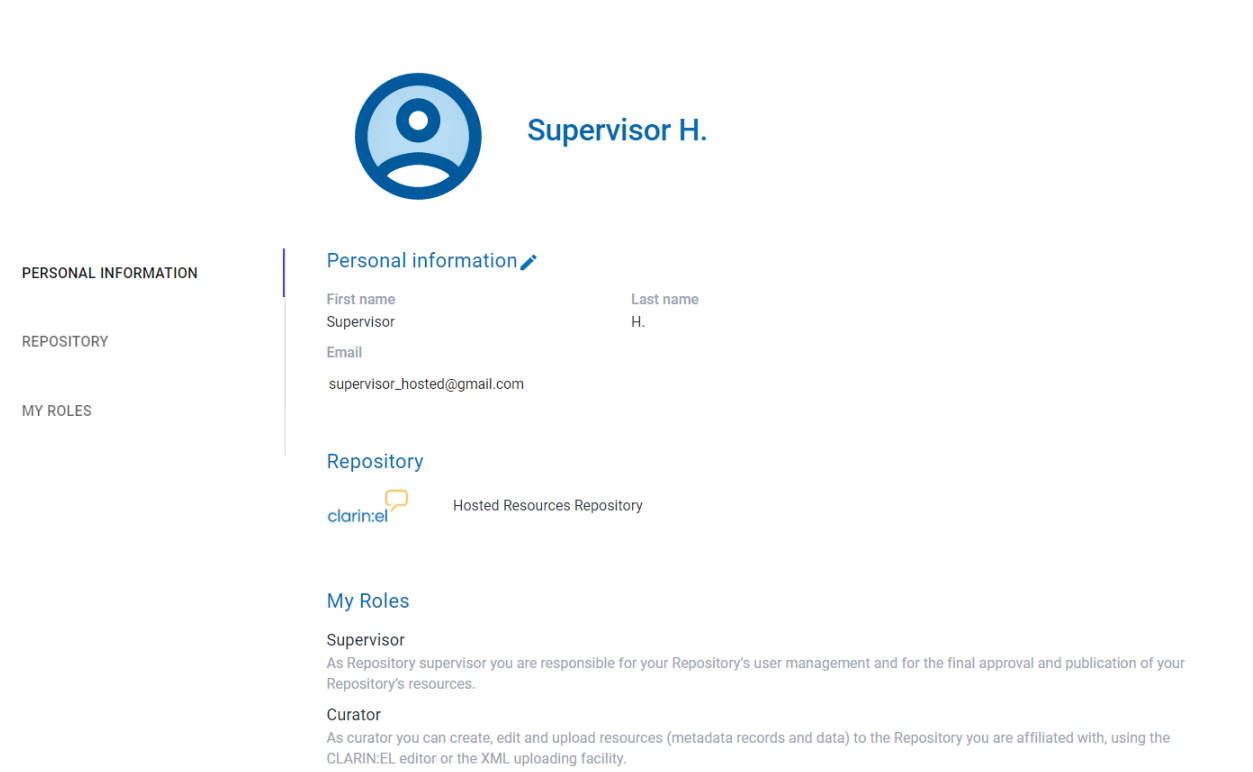


9.1.2 2. Profile

The second section is dedicated to your profile. For supervisors there is an extra functionality, namely *Manage Repository Users*, as shown below.



If you click on *View my profile*, you will be directed to a new page.



The image shows a user profile page for 'Supervisor H.'. At the top, there is a blue circular profile icon and the name 'Supervisor H.'. Below this, a sidebar on the left contains three menu items: 'PERSONAL INFORMATION', 'REPOSITORY', and 'MY ROLES'. The main content area is divided into three sections: 'Personal information' (with a pencil icon for editing), 'Repository', and 'My Roles'. The 'Personal information' section shows fields for 'First name' (Supervisor), 'Last name' (H.), and 'Email' (supervisor_hosted@gmail.com). The 'Repository' section shows the 'clarin:el' logo and the text 'Hosted Resources Repository'. The 'My Roles' section lists two roles: 'Supervisor' (As Repository supervisor you are responsible for your Repository's user management and for the final approval and publication of your Repository's resources.) and 'Curator' (As curator you can create, edit and upload resources (metadata records and data) to the Repository you are affiliated with, using the CLARIN:EL editor or the XML uploading facility.).

Some of your personal information is editable². Click on the pencil symbol next to Personal information and you will be directed to a new page where you can edit your personal data. After you have filled in the fields, save your changes.

Edit Account * Required fields

Email *	<input type="text" value="supervisor_hosted@email.com"/>
First name *	<input type="text" value="Supervisor"/>
Last name *	<input type="text" value="H."/>
Affiliation	<input type="text"/>
Position	<input type="text"/>
Personal website	<input type="text"/>

Enter an https:// URL



User management

Attention: This functionality is available **only to supervisors**.

If you click on *Manage Repository Users*, you will be directed to a page where all the users of your repository are listed. You can search for a specific user by using the search box on top. Once you have found the user, you must select the box on the left of the user's name.

² You cannot change the repository you are affiliated with, since this is done automatically during your registration or the roles you have, which have been assigned to you by the supervisor.

USERS

Filters

Action

Select an action to perform on your selected items

4 search results

	User name	Repository	Role	Actions
<input type="checkbox"/>		Hosted Resources Repository		Actions ▾
<input checked="" type="checkbox"/>	K P	Hosted Resources Repository		Actions ▾
<input type="checkbox"/>	Supervisor H.	Hosted Resources Repository	Supervisors	<input checked="" type="checkbox"/> Make metadata validator
<input type="checkbox"/>	Validator H.	Hosted Resources Repository	Metadata Legal	<input checked="" type="checkbox"/> Make legal validator

Then you are provided with two actions to choose from: you can make the user a **legal** or a **metadata** validator. The same action can be also performed from the action box on the top of the user list. Whatever you choose, a new window will open asking you to confirm your decision.

You are about to Make legal validator the following users.

K P

[Close](#) [Make legal validator](#)

By clicking on **make legal validator** you are giving the selected user permission to validate resources.

9.1.3 3. Create resources

Supervisor H.

supervisor_hosted@gmail.com

13

View my profile

Manage Repository Users

Welcome to your dashboard !

Here you can:

- view and update your profile
- view your resources and your tasks
- access creation forms
- upload resources

Total resources

Number of resources you have created.

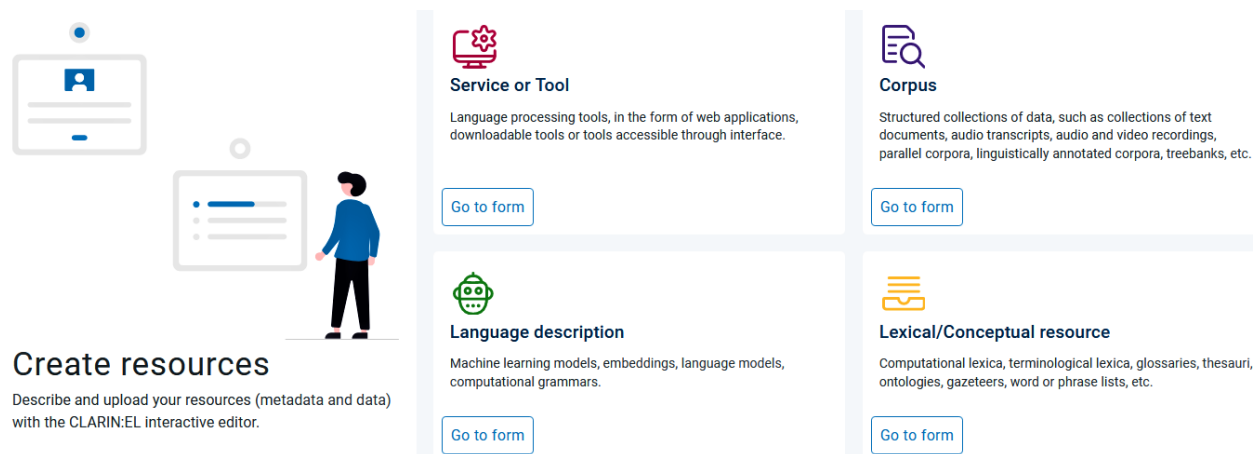
8

+ Create resources





Upload resources

Upload single or multiple resources in XML format.

In this section you see the total number of the resources you have created, **but not the resources themselves**³. By clicking on + *Create resources*, you are transferred to a new page where you have to select the type of resource you wish to create.



Create resources
Describe and upload your resources (metadata and data) with the CLARIN:EL interactive editor.

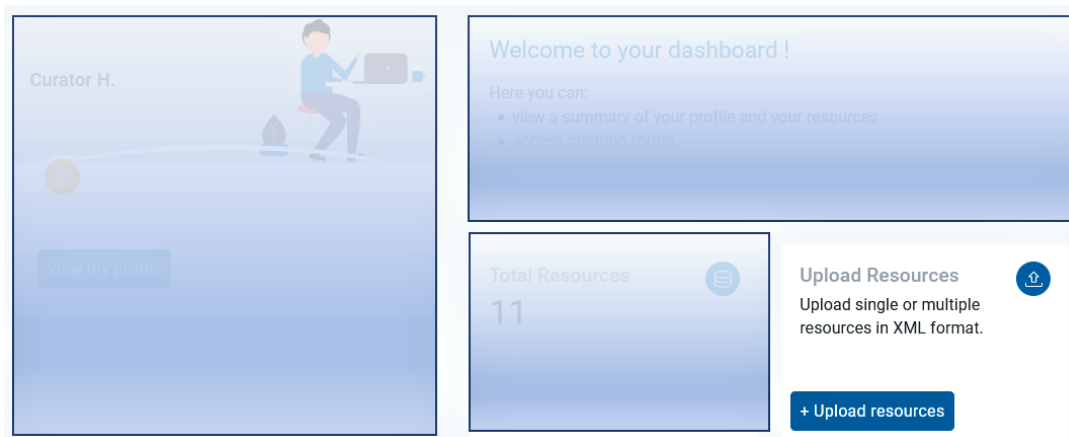
 <p>Service or Tool</p> <p>Language processing tools, in the form of web applications, downloadable tools or tools accessible through interface.</p> <p>Go to form</p>	 <p>Corpus</p> <p>Structured collections of data, such as collections of text documents, audio transcripts, audio and video recordings, parallel corpora, linguistically annotated corpora, treebanks, etc.</p> <p>Go to form</p>
 <p>Language description</p> <p>Machine learning models, embeddings, language models, computational grammars.</p> <p>Go to form</p>	 <p>Lexical/Conceptual resource</p> <p>Computational lexica, terminological lexica, glossaries, thesauri, ontologies, gazeteers, word or phrase lists, etc.</p> <p>Go to form</p>

Before you proceed to the *editor* by clicking **Go to form**, please, see the guidelines for the creation of

- a *corpus*,
- a *tool*,
- a *lexical/conceptual resource*,
- a *language description*

using the schema *mandatory* elements.

9.1.4 4. Upload resources



Welcome to your dashboard !

Here you can:

- view a summary of your profile and your resources
- access creation forms

Total Resources
11

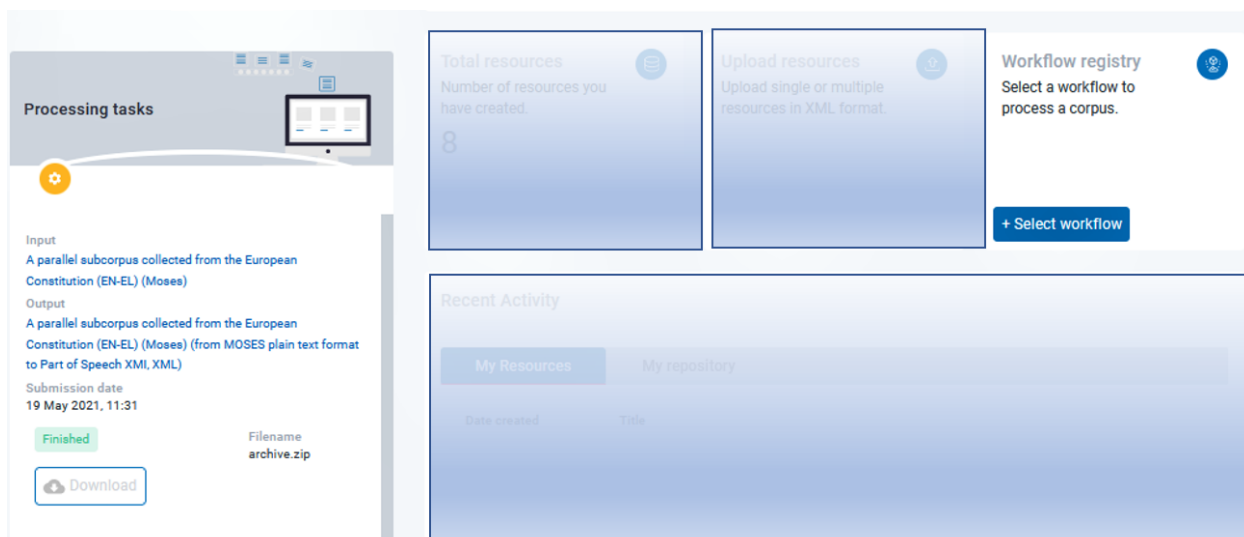
Upload Resources
Upload single or multiple resources in XML format.

[+ Upload resources](#)

You can click on + *Upload resources* if you already have one (or more) description/s in XML format. A new window will open where you are provided with several options. To find out more, see [here](#).

³ The section *My resources* has a list with what you have created.

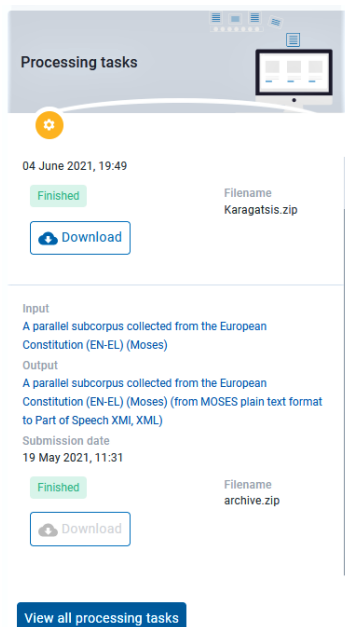
9.1.5 5. Select workflow



By clicking on + *Select workflow* you are transferred to the workflow registry. See [here](#) what the available services can do.

9.1.6 6. Processing Tasks

In this section you can see all your processing tasks and their results.



You can directly download the files that have been successfully processed from here. If you wish to have an overview, click on *View all processing tasks*.

MY SUBMITTED PROCESSING TASKS					
Input record	Filename	Output record	Submission date	Status	Download
	Karagatsis.zip		04 June 2021, 19:49	Finished	
A parallel subcorpus collected from the European Constitution (EN-EL) (Moses)	archive.zip	A parallel subcorpus collected from the European Constitution (EN-EL) (Moses) (from MOSES plain text format to Part of Speech XML, XML)	19 May 2021, 11:31	Finished	

Your submitted processing tasks are presented in a table organized in six columns:

- the *first* one contains the name of the input resource, provided you have selected it from the *processable* corpora available through CLARIN:EL; if you have uploaded your own resource, the column is empty,
- the *second* has the name of the zipped file; all the infrastructure processable corpora come in an archive.zip,
- the *third* contains the name of the output based on the name of the input; again if there was no input name, the column is empty,
- the *fourth* is the submission date,
- the *fifth* is the status of the processing, and
- the *sixth* is the download button.

9.1.7 7. My resources

This section, contains all the resources you have created via the *editor* or *XML upload*.

Recent Activity

My Resources	
Date created	Title
31 May 2021	Parla (12.1)
30 May 2021	Mandatory Grammar
30 May 2021	Mandatory LCR
29 May 2021	Tool Mandatory
29 May 2021	Mandatory Corpus
View my resources	

You can either select a resource by clicking on its name or, if you wish to have an overview of all the resources, you can click on *View my resources*.

MY RESOURCES

MY RESOURCES

Search for services, tools, datasets...

Search

Resources

+ Corpus (13)

+ Lexical/Conceptual resource (4)

+ Tool/Service (3)

+ Language description (1)

Status

+ draft (12)

+ syntactically valid (6)

+ submitted (2)

+ published (1)

Curator

Username

Search

Has data

+ no (13)

+ yes (8)

Processable

+ no (21)

21 search results

	Resource name		Actions	Status
<input type="checkbox"/>	Parla (12.1) 1.0.0 (automatically assigned) Corpus Hosted Resources Repository Created: 29 April 2021 Updated: 31 May 2021 has data	legal validator validator_hosted@gmail.com metadata validator validator_hosted@gmail.com curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Actions ▾	requested for unpublish
<input type="checkbox"/>	Mandatory Grammar 1.0.0 (automatically assigned) Language description Hosted Resources Repository Created: 30 May 2021 Updated: 30 May 2021	curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Actions ▾	syntactically valid
<input type="checkbox"/>	Mandatory LCR 1.0.0 (automatically assigned) Lexical/Conceptual resource Hosted Resources Repository Created: 30 May 2021 Updated: 30 May 2021	curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Actions ▾	syntactically valid

In this page, on the left, there are *filters* to help you sort out the resources depending on their type, status, data or processability. You can apply as many filters as you like and then clear them by clicking on the button above them.

Clear all filters

Resources

– Lexical/Conceptual resource (4)

Status

+ draft (2)

+ submitted (1)

4 search results

	Resource name		Actions	Status
<input type="checkbox"/>	Mandatory LCR 1.0.0 (automatically assigned) Lexical/Conceptual resource Hosted Resources Repository Created: 30 May 2021 Updated: 30 May 2021	curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Actions ▾	syntactically valid

Note: As a supervisor you will also see a **search box for curators**. Use the email of a curator to find only the resources created by them.

Clear all filters

Curator

Username

Search

Search using curator's email

4 search results

	Resource name		Actions	Status
<input type="checkbox"/>	Mandatory LCR 1.0.0 (automatically assigned) Lexical/Conceptual resource Hosted Resources Repository Created: 30 May 2021 Updated: 30 May 2021	curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Actions ▾	syntactically valid

9.1. Dashboard

29

As you can see, each resource occupies a row separated in four columns:

- the *first*, provides some basic information on the resource,
- the *second*, presents the name of the curator, supervisor and validator (if the resource has been validated),
- the *third*, has a button for the available actions, and
- the *fourth* is the resource status.

To learn more about the actions you can perform on a resource see [here](#).

9.1.8 8. Validation tasks

Attention: This section is visible only to **validators**.

Here you can see the list of all the resources that have been assigned to you for validation. You can select a resource by clicking on its name; you will be transferred to its view page. In case you have already validated this resource, a pop up message will inform you that you no longer have rights on it.

Recent Activity

My Resources

Validation tasks

Date created	Title
05 May 2021	Parla (12.1)
29 Apr 2021	Single text EL (txt)
28 Apr 2021	tool1_Maria

View my validation tasks

If you wish to have an overview of all the resources under validation (completed or not), you can click on *View my validation tasks*.

3 search results

Resource name	Status
Parla (12.1) 1.0.0 (automatically assigned) Corpus Hosted Resources Repository submitted: 05 May 2021 has data	legal validator validator_hosted@gmail.com published metadata validator validator_hosted@gmail.com legally valid yes curator curator_hosted@gmail.com metadata valid yes supervisor supervisor_hosted@gmail.com
Single text EL (txt) 1.0.0 (automatically assigned) Corpus Hosted Resources Repository submitted: 29 April 2021 processable has data	legal validator legalvalidator_hosted@gmail.com metadata validator validator_hosted@gmail.com syntactically valid curator supervisor_hosted@gmail.com supervisor supervisor_hosted@gmail.com
tool1 1.0	metadata validator validator_hosted@gmail.com published

Review comments
The licence chosen is inappropriate for corpora. Please check and replace it.

In this page, on the left, there are *filters* to help you sort out the resources. You can apply as many filters as you like and then clear them by clicking on the button above them.

1 search results

Clear all filters

Resource name	Status
Single text EL (txt) 1.0.0 (automatically assigned) Corpus Hosted Resources Repository submitted: 29 April 2021 processable has data	legal validator legalvalidator_hosted@gmail.com metadata validator validator_hosted@gmail.com syntactically valid curator supervisor_hosted@gmail.com supervisor supervisor_hosted@gmail.com

Review comments
The licence chosen is inappropriate for corpora. Please check and replace it.

As you can see, each resource occupies a row separated in three columns:

- the *first*, provides some basic information on the resource,
- the *second*, presents the names of the curator, supervisor and validators, and
- the *third* is the resource status. The information whether the resource is legally/metadata valid also appears here.

In addition, there is a box with the comments the validator has made, if any. See [here](#) how you can validate a resource.

9.1.9 9. My repository

Attention: This section is visible only to **supervisors**.

This section, contains all the resources of your repository, independently of who has created them. Remember, that, if you want to see the resources you have created, you must click on *my resources*.

Recent Activity

My Resources	
My repository	
Date created	Title
11 Jun 2021	corpus test upload 2
09 Jun 2021	Language Description (Mandatory elements)
09 Jun 2021	Single text EL (txt)
09 Jun 2021	Lexical/Conceptual Resource (Mandatory elements)
08 Jun 2021	draft Testrecognizer

[View my supervision tasks](#)

You can either select a resource by clicking on its name or, if you wish to have an overview of all the resources in your repository, you can click on *View my supervision tasks*. In this page, on the left, there are *filters* to help you sort out the resources depending on their type, status, data etc. You can apply as many filters as you like and then clear them by clicking on the button above them. The image below shows only the resources for which a validator must be assigned (based on the relevant filter).

Resources

+ Corpus (5)

+ Lexical/Conceptual resource (1)

Action required

+ yes (6)

Status

+ submitted (6)

Metadata valid

+ not validated (6)

Legally valid

+ yes (4)

+ not validated (2)

Has data

+ no (4)

+ yes (2)

Processable

+ no (6)

CLARIN:EL compatible service

+ no (6)

Requested for unpublish

+ no (6)

Validator assignment required

- yes (6)

☐ Created: 24 April 2021
Updated: 11 June 2021
has data

Review comments

for testing, for testing purposes2, for testing purposes

test_corpus_1

1.0.0 (automatically assigned)

Corpus

Hosted Resources Repository

Created: 05 May 2021

Updated: 06 May 2021

curator
admin@email.com

supervisor
supervisor_hosted@gmail.com

validator assignment required

yes

Actions

legally valid
not validated
metadata valid
not validated

submitted

TestK_1.1.1

1.0.0

Lexical/Conceptual resource

Hosted Resources Repository

Created: 27 April 2021

Updated: 27 April 2021

curator
curator_hosted@gmail.com

supervisor
supervisor_hosted@gmail.com

validator assignment required

yes

Actions

legally valid
yes
metadata valid
not validated

submitted

TestK_1

1.0.0

Corpus

Hosted Resources Repository

Created: 27 April 2021

Updated: 27 April 2021

curator
curator_hosted@gmail.com

supervisor
supervisor_hosted@gmail.com

validator assignment required

yes

Actions

legally valid
not validated
metadata valid
not validated

submitted

As you can see, each resource occupies a row separated in four columns:

- the *first*, provides some basic information,
- the *second*, presents the names of the curator, supervisor and validator (if the resource has been validated),
- the *third*, has a button for the available actions, and
- the *fourth* is the resource status, along with information on whether it has been validated or not.

To learn more about the actions you can perform on a resource see [here](#).

9.2 Help

This link directs you to the infrastructure [manual](#) where information is documented about resources, users, rights and processes. You can search for a specific issue of interest using the search box or go through the various chapters and sections.

The screenshot shows the CLARIN:EL manual interface. On the left is a navigation sidebar with the CLARIN logo and a search box. The sidebar menu includes sections like 'INTRODUCTION', 'BASIC CONCEPTS', and 'NAVIGATING THE INFRASTRUCTURE'. The main content area is titled 'How to use this manual' and contains introductory text about the manual's purpose, a list of actions users can perform (browse, create, perform actions), and two footnotes. At the bottom of the main area are 'Previous' and 'Next' navigation buttons and copyright information.

How to use this manual

This manual ¹ aims to help you explore and/or use the CLARIN:EL infrastructure to make your resources available to the **Humanities and Social Sciences** community. It is not meant to be read in sequence (although it can be) but to help you find specific information depending on your needs. There are chapters with general information on [basic concepts](#); others describing the [process](#) ² through which a resource comes to life and chapters which will specifically help you:

- to [browse](#) and [search](#) through the central inventory so as to find resources to [download](#) and [process](#),
- to create resources via the [editor](#) or by [uploading XML files](#), and
- to [perform actions on resources](#) depending on your responsibilities.

If you are looking for something specific, please, use the search box on the top left side of the navigation bar, below the CLARIN:EL logo.

[1] : The current version documents the third official release of the CLARIN:EL infrastructure, launched on May 31st, 2021. More functionalities are continuously added, and this manual keeps on being updated following the evolution of the CLARIN:EL platform.

[2] : Before you start, please see [the lifecycle of a resource](#) to find out what each type of user needs to do throughout the procedure. To assume a role you must have [registered](#) first.

© Copyright 2020, CLARIN Technical Team.
Built with [Sphinx](#) using a [theme](#) provided by [Read the Docs](#).

9.3 Your name

By clicking on your name you are transferred to your [profile](#) page.

9.4 Exit

If you no longer want to be signed in, you can log out by clicking on this icon.

BROWSING

From the *inventory home page* click on the **browse** (or the **search**) button. You will be directed to the central inventory. On the left side of the page you can see the available filters which are explained *here*. In the inventory each resource is provided with a snippet of information.

The screenshot displays a web interface for browsing linguistic corpora. On the left is a sidebar with filter categories: Repository, Resource type, Media type, Languages, Linguality type, Domains, Time coverage, Processable, Annotation type, Processing service, and Licences. The main area shows three resource cards:

- CGL Humanities Metalanguage Corpus: A descriptive directory of Text Corpora for Modern Greek** (Repository: CGL). Description: The resource consists of reviews of three Modern Greek-language text corpora: - The corpus of the National Thesaurus of the Greek Language (ΕΘΕΓ) (The review of the corpus is 3500 words) - The newspaper corpus by the Cen ... Language: Modern Greek (1453-) Media Type: text Keywords: monolingual text/plain Tags: processable, accessible through interface
- Hellenic National Corpus** (Repository: ANHN). Description: The Hellenic National Corpus (HNC) currently contains about 47.000.000 words, and is constantly being updated. It consists of samples of written language exclusively. Texts in the HNC represent modern Greek language use. ... Language: Modern Greek (1453-) Media Type: text Keywords: monolingual text/plain non-fiction segmentation Tag: accessible through interface
- Golden Part of Speech Tagged Corpus** (Repository: ANHN). Description: The GoldenPart-of-Speech Tagged Corpus is a subset of the Hellenic National Corpus (HNC), the size of which is 100.000 words; it consists of selected texts from a variety of sources covering various domains. These texts ... Tag: downloadable

On the left there is the logo of the resource's institutional repository. Next, there is the hyperlinked name of the resource which directs to its *view page*. Below the name, there is the resource type followed by the first lines of the description. Then there is information on the language, the keywords and the media type. On the right side of the page there are tags which provide information on the **processability**¹ and the **accessibility**² of a resource at a glance. Some of the most frequent tags are the following:

- *processable*: a corpus which is compatible with the services of the infrastructure and can be processed [to do so, you must visit the resource view page and click the *Process* button found in the *Access* tab].

¹ The following values provide information on the way a resource is accessible: downloadable, CD-ROM, DVD-R, accessible through interface, accessible through query, bluRay, hard disk, other, unspecified.

² The following values provide information on the way a tool/service is accessible: library, plugin, source code, source and executable code, web service, workflow file, docker image, other.

- *processing service*: a tool which has been integrated as a service and can be used for processing of processable resources [to do so, you must visit the resource view page and click the *Use* button found in the *Access* tab].
- *downloadable*: a resource which can be downloaded, directly through CLARIN:EL or indirectly through an external link [to do so, you must visit the resource view page and click the *Download* button found in the *Access* tab].
- *accessible through interface*: a resource which is accessed by an interface available through an external link [to do so, you must visit the resource view page and click the *Access* button found in the *Access* tab].

Just below the search box on top of the page there are two icons as indicated in the image below.

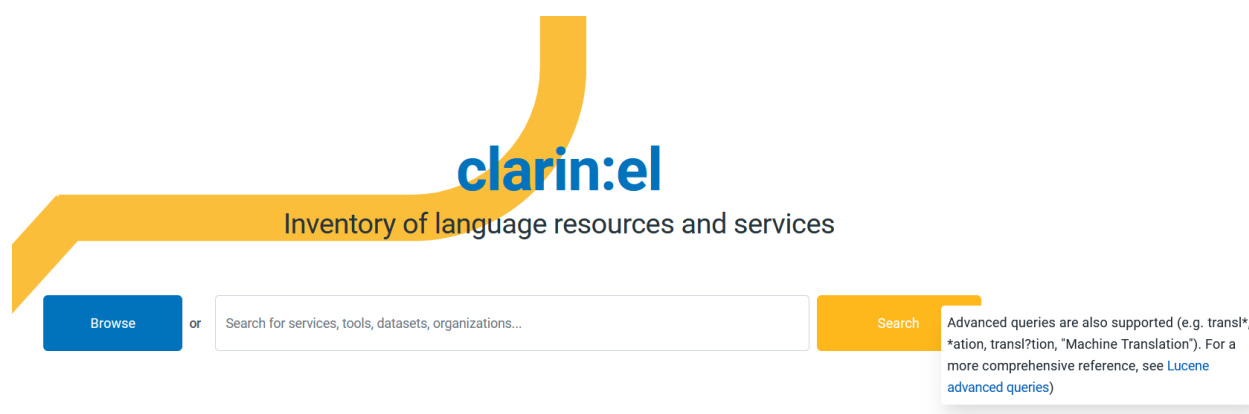


They provide two different ways of presentation for the inventory. By default, the LRTs are presented in a list but if you click on the icon on the right they are presented in a table form.

By clicking on the name of a resource, you can proceed to its *view page*.

SEARCHING

The search box is found in the inventory home page of CLARIN:EL. You can use simple keywords or, as shown in the image, special characters for [Lucene advanced queries](#)¹.



You have also the option to use **filters**. To do so, click on the browse/search button to be transferred to the central inventory. On the left side of the page, you can see all the available filters.

¹ For the time being the advanced queries can only be used for search in the central inventory, not in the search boxes found in *my resources*, *my validation tasks*, *my repository*.

clarin:el

CLARIN:EL portal >

Help Sign in

Search for services, tools, datasets...

Search ?

Repository

Resource type

Media type

Languages

Language variety

Linguality type

Multilinguality type

Domains

Time coverage

Geographic coverage

Processable

Annotation type

Processing service

Service functions

Licences

A parallel corpus collected from the European Constitution downloadable

Corpus

A parallel corpus collected from the European Constitution. 21 languages, 210 bitexts total number of files: 986 total number of tokens: 3.01M total number of sentence fragments: 0.22M

Languages: Finnish Irish Modern Greek (1453-) Slovak ...

Media Type: text

Keywords: multilingual parallel application/xml alignment

A parallel corpus of KDE4 localization files (v.2) downloadable

Corpus

92 languages, 4,101 bitexts total number of files: 75,535 total number of tokens: 60.75M total number of sentence fragments: 8.89M

Languages: Macedonian Hebrew Slovak Norwegian Nynorsk ...

Media Type: text

Keywords: multilingual parallel application/x-xml technicalTexts

A parallel subcorpus collected from the European Constitution (DE-EL) (Moses) processable downloadable

Corpus

The EUconst subcorpus (DE-EL) (Moses) is a parallel subcorpus for German and Greek, subset of the EUconst, a parallel corpus collected from the European Constitution. 21 languages, 210 bitexts total number of files: 986 ...

Languages: Modern Greek (1453-) German

Media Type: text

Keywords: bilingual parallel text/plain administrativeTexts

A parallel subcorpus collected from the European Constitution (DE-EL) (Moses) (from MOSES plain text format to Part of Speech XML, XMI) downloadable

Corpus

A parallel subcorpus collected from the European Constitution (DE-EL) (Moses) (MOSES plain text format) annotated by the ILSP Feature-based multi-tiered POS Tagger (ILSP workflow v1.0.0) and the OpenNLP Part-of-Speech Ta ...

Languages: Modern Greek (1453-) German

The filters' value lists are closed, but once you click on the arrow next to the filters they open.

Repository

- Athena RC Repository (341)
- University of Crete Repository (41)
- Centre For The Greek Language Repository (30)
- University of Thessaloniki Repository (28)
- Hosted Resources Repository (24)
- Show more

Resource type

- Corpus (412)
- Lexical/Conceptual resource (78)
- Tool/Service (41)
- Language description (2)

Media type

A parallel corpus collected from the European Constitution downloadable

Corpus

A parallel corpus collected from the European Constitution. 21 languages, 210 bitexts total number of files: 986 total number of tokens: 3.01M total number of sentence fragments: 0.22M

Languages: Finnish Irish Modern Greek (1453-) Slovak ...

Media Type: text

Keywords: multilingual parallel application/xml alignment

A parallel corpus of KDE4 localization files (v.2) downloadable

Corpus

92 languages, 4,101 bitexts total number of files: 75,535 total number of tokens: 60.75M total number of sentence fragments: 8.89M

Languages: Macedonian Hebrew Slovak Norwegian Nynorsk ...

Media Type: text

Keywords: multilingual parallel application/x-xces+xml technicalTexts

If you wish to remove the filter(s) applied, either click on the **clear all filters** button or on the **x** next to the filter name.

Search for services, tools, datasets... Search ?

Clear all filters 52 search results

English ATHENA RC Repository Corpus

In the central inventory you can use the following filters:

- *Repository*: It groups the LRTs of each repository separately.
- *Resource type*: It groups the LRTs by their type, i.e. corpora, tools/services, lexical/conceptual resources, language descriptions.
- *Media type*: It groups the LRTs depending on their medium, i.e. text, video, audio, image.
- *Languages*: It groups the LRTs depending on their language(s) or the language(s) they can process.
- *Language Variety*: It groups the LRTs depending on their language variety or the language variety they can process.
- *Linguality Type*: It groups the LRTs depending on whether they are monolingual, bilingual or multilingual.
- *Multilinguality Type*: It groups the LRTs depending on whether they are parallel or comparable (this filter applies only to bilingual and multilingual resources).
- *Domains*: It groups the LRTs according to the various fields of knowledge which have been used to classify it or its contents.
- *Time Coverage*: It groups the LRTs depending on the period of time their contents cover (for instance, [resources of the 16th century](#)).
- *Geographic Coverage*: It groups the LRTs depending on the geographic region(s) of their contents (for instance, [resources related somehow to France](#)).
- *Processable*: It groups the corpora which can be processed within [CLARIN:EL](#) according to whether they have the technical features which make them compatible with the CLARIN:EL integrated services of the workflow registry.
- *Annotation Type*: It groups the LRTs based on their annotation(s) (for example, [bilingual or multilingual resources which are aligned](#)).
- *Processing Service*: It groups only the tools which have been integrated in [CLARIN:EL](#) as services.

- *Service Function*: It groups services according to the function they perform (e.g. [services for lexicon creation](#)).
- *Licences*: It groups LRTs according to their licence(s) (for instance resources under [CC-BY 4.0 licence](#)).

<p>Attention: For a resource to be filtered, and presented as a result to the user, the respective metadata must have been filled in by the curator.</p>

VIEWING & DOWNLOADING

For each resource a metadata record is available providing descriptive and technical information as well as supporting documentation and other useful material.

Attention: Although the resource view page is accessible to all users, some of its functionalities (e.g. the *Process* button) are available only to users who are *signed in*.

The chosen example showcases a [corpus metadata record](#). The layout of the view page is the **same for all types of resources**. What is different is the **content of the sections** which depends on the resource type. Each view page is split in two sections with several subparts.

12.1 The upper section

It contains the resource **name** and **short name** (if any), **type**, **version**, and *PID* followed by a **description**. Underneath the description there is a tag indicating that this resource is *processable*. On the side there is a **language selection button**. By default, all metadata records are presented in English (**en**), but if you choose **el** you will see all the metadata which have been added in Greek as well. Following, there is the repository logo and just underneath a citation text. You can **copy** the text just by clicking on the icon as shown in the image.

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses)

 corpus

EUconst subcorpus EN-EL (Moses)

Version: 1

<http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A>

Το EUconst subcorpus EN-EL (Moses) 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). 21 γλώσσες, 210 bitexts συνολικός αριθμός αρχείων: 986 συνολικός αριθμός μονάδων (tokens): 3.01M συνολικός αριθμός αποσπασμάτων προτάσεων: 0.22M


 processable

Select Language

el


en

el


Ερευνα & Καινοτομία
Τεχνολογικά Πληροφορίες

Cite

Citation text

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses) (2015). Version 1. [Dataset (Text corpus)]. CLARIN-EL.
<http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A> 

Copy

The resource view page, provides extra functionalities when you *sign in* (provided you have assumed a *role* in your repository). The difference for the signed in user is the *actions* box shown in the image below.

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses)

 Corpus

EUconst subcorpus EN-EL (Moses)

Version: 1

<http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A>

The EUconst subcorpus EN-EL (Moses) is a parallel subcorpus for English and Greek, subset of the EUconst, a parallel corpus collected from the European Constitution. 21 languages, 210 bitexts total number of files: 986 total number of tokens: 3.01M total number of sentence fragments: 0.22M

processable

Select Language

en



Cite

Citation text

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses) (2015). Version 1. [Dataset (Text corpus)]. CLARIN-EL.
<http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A>

Actions

Actions

Tip: More about the actions you can perform on a resource [here](#).

Under the description, there are boxes providing information on some of the resource **key features**. For corpora and lexical/conceptual resources these are: the **language(s)** covered, **keywords** (describing the contents), the resource **domain** and **subclass**. Depending on other metadata filled in, the boxes could also present the resource **coverage** as concerns **time** and **space**, as shown in the following image.

Language Modern Greek (1453-)	Keywords monolingual other	Coverage Time coverage 1989-1994 1997-2018 Geographic coverage Greece
Domain KKE2516 Political Science	Corpus subclass annotated corpus	

In the metadata records for **tools/services** these boxes show their **function** and whether they are **language dependent**.

Function Discourse analysis	Language dependent no	Keywords text analysis concordances word frequencies n-grams
---------------------------------------	---------------------------------	---

12.2 The lower section

The lower section is designed to have the following tabs: *Overview*, *Technical*, *Relations*, *Access*. The combination of tabs appearing in each view page derives from the resource type and the metadata which have (or have not) been filled in. The selected tab each time changes from blue to orange.

Overview Technical Relations Access

12.2.1 1. Overview

This tab contains some of the *mandatory metadata*, i.e. the **language(s)** and **linguality** of the corpus part (here text) as well as information on the annotations (if any). Here you will also see the **genre** and **category** of the corpus. The next column gives you the option to **export** the metadata in a file¹ and below this you can find out who is the resource **provider** and **contact** details, such as an email and/or a landing page which you can access to learn more about the resource.

Information for Corpus part

TEXT

Language info

Linguality type
bilingual

Multilinguality type
parallel

Language
English, Modern Greek (1453-)

Categories

Text genre
official

Text type
Category label
administrative texts

Export

XML

Resource provider

OPUS

Website

Contact

Jörg Tiedemann

Email

Landing page

In case there are more than one corpus parts, these are placed vertically on the left most column and you can see their features by clicking on the **medium** (the chosen part is highlighted in orange, as shown below).

Information for Corpus part

TEXT

VIDEO

Language info

Linguality type
multilingual

Multilinguality type
parallel

Language
Greek Sign Language, British Sign Language, German Sign Language, French Sign Language

Export

XML

Resource creator

School of Computing Sciences



Website

Research Institute of Computer Science

Website

The Overview tab for tools/services contains information on the **input and output features**, i.e. **language(s)**, **data format(s)**, etc.

¹ For the time being the metadata record can only be exported in XML format. Soon, other formats will be supported as well.

Input	Output	Export
Input item Language Modern Greek (1453-) Processing resource type user input text Media type text	Output item Language Modern Greek (1453-) Processing resource type output text Media type text	Export XML
		Resource creator  Institute for Language and Speech Processing Website
		Contact  Athanassios Protopapas Email
Landing page		

The bottom of the page in all cases provides related **papers** and **documentation**. In the case of corpora and lexical/conceptual resources there is also information on whether **personal** and **sensitive** data are included in the resource.

Information for the resource

Documentation

Preferred citation
 Jörg Tiedemann, 2009, News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia
 Documented in
 News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces

Ethics

Personal data included
 no
 Sensitive data included
 no

12.2.2 2. Technical

Attention: The Technical tab appears **only** in the view pages of **tools/services**!

It provides information on the **intended** (as foreseen by the provider) and **actual use** (as deployed by the users) of the tool/service; additionally, it shows whether it has been **evaluated** or not.


Actual use Used in application: Editing Language Technology	Evaluated: no	Intended application Language Technology Editing
---	---------------	--

12.2.3 3. Relations

This tab contains links to the **resources which are related** to the resource being viewed (examples include part-whole relationships, aligned-non aligned versions etc.).


Relations to other resources

Part of

 A parallel corpus collected from the European Constitution (1)
<http://hdl.handle.net/11500/ATHENA-0000-0000-258B-A> (Handle)

Relations to other entities

Aligned versions

 A parallel subcorpus collected from the European Constitution (EN-EL) (TMX) 1


Attention: This tab is omitted when there are no relations to other resources.

12.2.4 4. Access

The last tab is the **Access** tab. It contains all the information about the **modes of distribution** (i.e. the forms a resource is accessible, e.g. CD-ROM, via a link from where it can be downloaded, etc.). Each form comes along with **licence terms** and if needed any other information such as an attribution text.



Modes of distribution

Download 

Sign in to process

Access info

Dataset distribution form
downloadable

Licencing Info

Licence

Creative Commons Attribution 4.0 International
<https://creativecommons.org/licenses/by/4.0/legalcode>
<https://creativecommons.org/licenses/by/4.0/>

Attribution text

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses). Αδεια: Creative Commons Attribution 4.0 International (<https://creativecommons.org/licenses/by/4.0/legalcode>, <https://creativecommons.org/licenses/by/4.0/>). Πηγή: <http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A> (CLARIN:EL)

Features



Attention: This tab is omitted in case the resource has no content files (see for instance [meta-resources](#)).


4.1 Download

If the resource is provided under a licence which allows you to download it, you will see the *Download* button. Simply click on it to get the resource content files.

Attention: Please check which are the prerequisites for a resource to be *accessible* (and therefore downloadable) in the infrastructure.

4.2 Process

If the resource has the features that make it compatible with the infrastructure workflows and is therefore *processable*, you will see a *Process* button². See [here](#) an example of processing a bilingual corpus³.

[Download](#)  [Process](#)

Access info	Licencing Info
Dataset distribution form downloadable	Licence Creative Commons Attribution 4.0 International https://creativecommons.org/licenses/by/4.0/legalcode https://creativecommons.org/licenses/by/4.0/

4.3 Use

Attention: The *Use* button appears only in metadata records for tools/services!

If the resource is provided under a licence which allows you to *Use* it, you will see a button. See [here](#) an example of using a tool/service⁴.

[Use](#)

Access info	Licencing Info
Software distribution form docker image Operating system OS-independent Private yes	Licence CLARIN-EL Terms of Service http://inventory.clarin.ilsp.gr/catalogue_backend/static/licence/clarinelTOS.pdf Attribution text Annotator of Named Entities GrNE-Tagger used under CLARIN-EL Terms of Service (http://inventory.clarin.ilsp.gr/catalogue_backend/static/licence/clarinelTOS.pdf). Source: None (CLARIN:EL)

² If you are not *signed in*, the button prompts you to do so (*Sign in to process*). After signing in, you are redirected to the resource view page where the *Process* button appears.

³ The corpus used for this example is A parallel subcorpus collected from the European Constitution (EN-EL) (Moses).

⁴ The resource used for this example is the Annotator of Named Entities GrNE-Tagger.

Features

Attention: The following section appears **only** in metadata records for **corpora and lexical/conceptual resources**!

At the bottom of the **Access** page, there is a hidden category of metadata called **Features**. Click on the arrow to reveal all the features per corpus part, i.e. the corpus part **size**, the **data format** and the **encoding** (of the text in this case).

Features ^

Text feature

size
280000 word
534 kilobyte

Data format
MOSES plain text format

Character encoding
UTF-8

PROCESSING

CLARIN:EL offers you, in total, three approaches to processing: one that starts from specific **datasets** which fulfill certain conditions¹ (you first select the *dataset* you want to process and then the *function*, i.e. the type of processing you want to apply) and two more which use **function** as the starting point (i.e. you select *what* you want to do, then the *tool/workflow* that suits your needs and then the *dataset* on which you will apply the service) . If you are interested in the function, please see how you can use the [workflow registry](#) or the [processing services](#). In these cases you will have to upload your own dataset. If, on the other hand, you are interested in a specific dataset on which you would like to apply one or more of the integrated services, check out the information provided on [processable corpora](#).

Attention: Processing is available **only to registered users** who are signed in. If you don't have an account, see [here](#) how to register.

13.1 1. Starting with the data

The corpora which have features that make them compatible with the workflows of the infrastructure are indicated as *processable*². These corpora are offered as a [preselection](#) in the inventory home page. These corpora are either *monolingual* in **Greek, English, German** or **Portuguese** or *bilingual* having **Greek** as one language and **English, German** or **Portuguese** as the other.

To process one of these corpora, follow the next **basic steps**:

13.1.1 Step 1. Select a corpus

The resource chosen for this scenario is a bilingual corpus: [A parallel subcorpus collected from the European Constitution \(EN-EL\) \(Moses\)](#).

¹ All the corpora which meet these criteria are indicated as *processable*. They are presented in the inventory home page as a [preselection](#) which directs to the central inventory.

² The tag is also found at the resource snippet in the [central inventory](#) and the resource [view page](#).

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses)

 Corpus

EUconst subcorpus EN-EL (Moses)

Version: 1

<http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A>

The EUconst subcorpus EN-EL (Moses) is a parallel subcorpus for English and Greek, subset of the EUconst, a parallel corpus collected from the European Constitution. 21 languages, 210 bitexts total number of files: 986 total number of tokens: 3.01M total number of sentence fragments: 0.22M

processable

Select Language

en



Cite


Citation text

A parallel subcorpus collected from the European Constitution (EN-EL) (Moses) (2015). Version 1. [Dataset (Text corpus)]. CLARIN:EL.

<http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A> 

First move to the *lower section* of the view page and choose the **Access** tab. Then click on *Process*³.

Overview
Relations
Access


Modes of distribution

Download
Process

Access info

Dataset distribution form
downloadable

Licensing Info

Licence
Creative Commons Attribution 4.0 International
<https://creativecommons.org/licenses/by/4.0/legalcode>
<https://creativecommons.org/licenses/by/4.0/>

Attribution text
A parallel subcorpus collected from the European Constitution (EN-EL) (Moses) used under Creative Commons Attribution 4.0 International
(<https://creativecommons.org/licenses/by/4.0/legalcode>, <https://creativecommons.org/licenses/by/4.0/>). Source: <http://hdl.handle.net/11500/ATHENA-0000-0000-23FF-A> (CLARIN:EL)

Features

³ If you are not *signed in*, the button prompts you to do so (*Sign in to process*). After signing in, you are redirected to the resource view page where the *Process* button appears.

13.1.2 Step 2. Select a function

Once you click on *Process*, you will be directed to a selection of workflows from the *workflow registry*. These are the ones that can be used on the Greek part of the corpus you have chosen (you will also see a notification at the top of the page). Since the corpus is bilingual you will later need to select a workflow for the English part as well.

Available services and functions for Modern Greek (1453-)

BelowPosTagging
NamedEntityRecognition
Chunking

ILSP Feature-based multi-tiered POS Tagger

The ILSP Feature-based multi-tiered POS Tagger is a language processing workflow that annotates every word of a text with the corresponding part of speech (POS) tag, e.g. noun, verb, adjective, adverb, etc., based on its context. This tool performs POS tagging for texts in Modern Greek and it is the last step of this workflow, which consists of the ILSP Sentence Splitter and Tokenizer and the POS Tagger. It accepts as input plain text (txt) in UTF8 encoding and generates as output an XMI document with POS tags assigned to each token, besides labelled sentences and tokens.

Input	Output
input_language	output_language
Modern Greek (1453-)	Modern Greek (1453-)
input_character_encoding	output_character_encoding
UTF-8	UTF-8
input_data_format	output_data_format
MOSESFormat	Xmi

Use this workflow

HBrill

HBrill is used for annotating a word in Greek texts with the corresponding part of speech (POS) tag (e.g. noun, verb, adjective, adverb, etc.). It accepts as input XML documents in UTF8 encoding, and generates as output a standoff XML document in UTF8 encoding, with part of speech annotations. It belongs to the language processing workflows of NCSR Demokritos.

Input	Output
input_language	output_language
Modern Greek (1453-)	Modern Greek (1453-)
input_character_encoding	output_character_encoding
UTF-8	UTF-8
input_data_format	output_data_format
MOSESFormat	Xml

Use this workflow

cancel Select service for Modern Greek (1453-)

Click on **use this workflow** (it automatically changes from light blue to green) and then proceed by selecting this service.

Available services and functions for Modern Greek (1453-)

BelowPosTagging
NamedEntityRecognition
Chunking

ILSP Feature-based multi-tiered POS Tagger

Input	Output
input_language	output_language
Modern Greek (1453-)	Modern Greek (1453-)
input_character_encoding	output_character_encoding
UTF-8	UTF-8
input_data_format	output_data_format
MOSESFormat	Xmi

Use this workflow

HBrill

Input	Output
input_language	output_language
Modern Greek (1453-)	Modern Greek (1453-)
input_character_encoding	output_character_encoding
UTF-8	UTF-8
input_data_format	output_data_format
MOSESFormat	Xml

Use this workflow

cancel Select service for Modern Greek (1453-)

Then repeat the same procedure for the English part.

Available services for BelowPosTagging English

BelowPosTagging

OpenNLP Part-of-Speech Tagger (English)

Input	Output
input_language	output_language
English	English
input_character_encoding	output_character_encoding
UTF-8	UTF-8
input_data_format	output_data_format
MOSESFormat	Xml

Use this workflow

Back cancel Select service for English

A new window appears asking you to review the workflows you have selected before submitting them.

Review your selections and press submit to start the processing or back to choose again

BelowPosTagging - ILSP Feature-based multi-tiered POS Tagger

Input	Output
input_language Modern Greek (1453-)	output_language Modern Greek (1453-)
input_character_encoding UTF-8	output_character_encoding UTF-8
input_data_format MOSESFormat	output_data_format Xml

BelowPosTagging - OpenNLP Part-of-Speech Tagger (English)

Input	Output
input_language English	output_language English
input_character_encoding UTF-8	output_character_encoding UTF-8
input_data_format MOSESFormat	output_data_format Xml

[Back](#) [cancel](#) [Submit for process](#)

As soon as you hit the *Submit* button a message will appear, informing you that you will be notified by email when the processing is over.

The dataset is being processed. You will be notified by email when it's finished.

[close](#)

13.1.3 Step 3. Get the processed files

You will be notified by email once the processing is finished. To see the results go to your dashboard and check the *Processing tasks*.

Attention: A metadata record with the annotated data is **automatically** created and the resource is published to the inventory.

13.2 2. Starting with the Function

13.2.1 2.1 Workflow Registry

You can access the workflow registry either from the *inventory home page* or from your *dashboard*.

1 ————— 2 ————— 3
 Select function Upload data Process

SentenceSplitting
Tokenization
MorphosyntacticTagging
Lemmatization
DependencyParsing
NamedEntityRecognition
Chunking

TextCategorization
Verbal aggression analysis

ILSP Sentence splitter and tokenizer for Greek text (ILSP workflow v1.0.0)

The ILSP Sentence splitter and tokenizer for Greek texts splits the input text into smaller units, such as sentences, words, punctuation marks, dates, numbers or symbols. It is the first step of all ILSP language processing workflows for Greek texts. It accepts as input plain text (txt) in UTF8 encoding and generates as output a standoff XML document in UTF8 encoding, with labelled sentences and tokens.

Input	Output
input_language el	output_language el
input_character_encoding UTF-8	output_character_encoding UTF-8
input_data_format Text	output_data_format Xmi
input_typesystem	output_typesystem ILSP_NLP typesystem

Use this workflow

OpenNLP Sentence Detector (English) (OpenNLP workflow v1.0.0)

The OpenNLP Sentence Detector for English splits the input English text into smaller units, such as sentences, words, punctuation marks, dates, numbers or symbols. It accepts as input plain text (txt) in UTF8 encoding and generates as output XML document in UTF8 encoding. It is the first step of OpenNLP language processing workflows for English.

Input	Output
input_language en	output_language en
input_character_encoding UTF-8	output_character_encoding UTF-8
input_data_format Text	output_data_format Xml
input_typesystem	output_typesystem ILSP_OpenNLP typesystem

Use this workflow

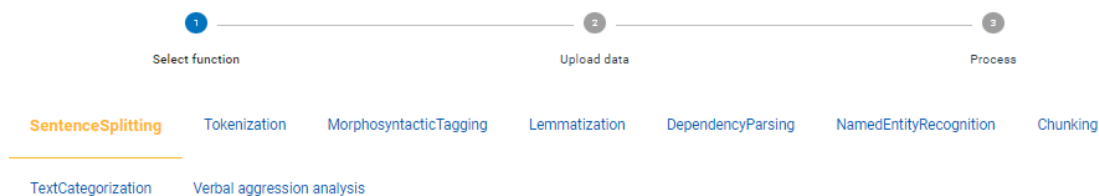
OpenNLP Sentence Detector (German) (OpenNLP workflow v1.0.0)

The OpenNLP Sentence Detector for German splits the input German text into smaller units, such as sentences, words, punctuation marks, dates, numbers or symbols. It accepts as input plain text (txt) in UTF8 encoding and generates as output XML document in UTF8 encoding. It is the first step of OpenNLP language processing workflows for German.

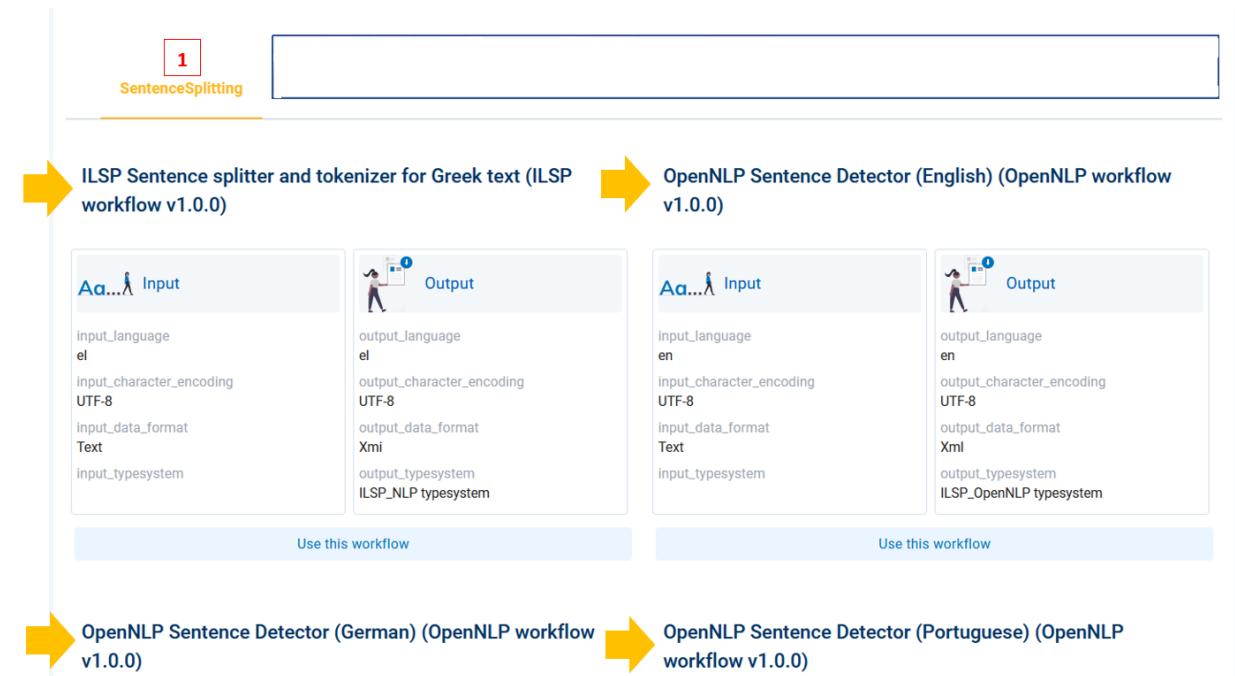
OpenNLP Sentence Detector (Portuguese) (OpenNLP workflow v1.0.0)

The OpenNLP Sentence Detector for Portuguese splits the input Portuguese text into smaller units, such as sentences, words, punctuation marks, dates, numbers or symbols. It accepts as input plain text (txt) in UTF8 encoding and generates as output XML document in UTF8 encoding. It is the first step of OpenNLP language processing workflows for Portuguese.

At the moment there are **nine** functions offered.



For each function CLARIN:EL offers a number of workflows, as shown in the image below.



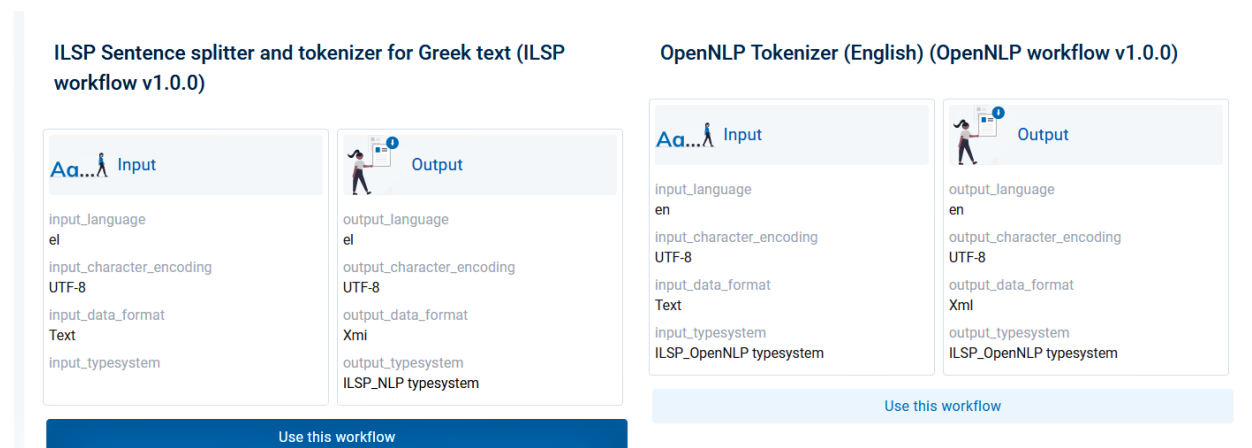
The **basic steps** to use the workflow registry are the following:

Step 1. Select a function

Select a function, according to which type of processing you want to perform, by clicking on its name. The selected function (e.g. tokenization) changes colour from blue to orange.

Step 2. Select a workflow

For tokenization there are multiple available workflows: various for Greek corpora, one for English, one for German and one for Portuguese. Select the workflow you want by clicking on *Use this workflow*.



Step 3. Upload your data

In the new window, you are informed about the prerequisites of the processing, i.e. the specifications of the dataset to be uploaded. If you wish to process your own dataset, it needs to fulfil these conditions; then, you can upload it.

Attention: You can **only** upload *monolingual corpora* in **Greek, English, German or Portuguese**. The workflows can also process the infrastructure bilingual corpora which are tagged as *processable*.

You can now upload your zip file, which should

1. have a size of max 2MB
2. have a filename in Latin characters, with no spaces in it
3. contain only plain txt files in UTF-8 encoding
4. not contain other zip files

Drag & Drop your files or Browse

ILSP Sentence splitter and tokenizer for Greek text (ILSP workflow v1.0.0)

Aa... Input

input_language
el

input_character_encoding
UTF-8

input_data_format
Text

input_typesystem

Output

output_language
el

output_character_encoding
UTF-8

output_data_format
Xmi

output_typesystem
ILSP_NLP typesystem

Reset Back Next

f t y
Need Help ?

File has been successfully uploaded.

After the dataset has been successfully uploaded, the **next** button is activated and you can click on it.

✓
 Select function

✓
 Upload data

3
 Process

The dataset is being processed. You will be notified by email when it's finished.

Reset Back Finish

Step 4. Get the processed files

You will be notified by email once the processing is finished. To see the results go to your dashboard and check the *Processing tasks*.

Attention: Both the data uploaded for processing and the data which result from the processing are **not stored** permanently in the infrastructure; the CLARIN:EL policy is to delete the annotated data 48 hours after processing has been completed. If you wish to download them, please, do so during this time frame.

13.2.2 2.2 Processing Services

Go to the central inventory and apply the processing service filter. You will be presented with all the available services in the infrastructure. To use them, follow the next **basic steps**:

The screenshot shows the CLARIN central inventory search results for processing services. The interface includes a sidebar with filters, a search bar, and a list of results.

Filters:

- Repository: ▾
- Resource type: ▾
- Languages: ▾
- Processable: ▾
- Processing service: ▴**
 - yes (19)
- Service functions: ▾
- Licences: ▾

Search results: 19 search results

yes (19)

Results:

- Annotator of Named Entities GrNE-Tagger**
 - Tool/Service
 - Ο Επισημειωτής Ονοματικών Οντοτήτων GrNE-Tagger επεξεργάζεται κείμενα αναγνωρίζοντας αυτόματα βάσει κανόνων και ταξινομώντας τις Ονοματικές Οντότητες (κύρια ονόματα) που αυτά περιέχουν στις ακόλουθες πέντε κατηγορίες: Α ...
 - Language: Modern Greek (1453-)
 - Keywords: named entities text
 - processing service
 - docker image
- HBrill**
 - Tool/Service
 - HBrill
 - Language: Modern Greek (1453-)
 - Keywords: segmentation writtenLanguage morphosyntactic analysis text
 - processing service
 - docker image
- HNPCChunker**
 - Tool/Service
 - HNPCChunker
 - processing service
 - docker image

Step 1: Select a service

Click on the name of the service you would like to use. You will be transferred to the resource view page. Move to the lower section of the page and choose the **Access** tab.

Annotator of Named Entities GrNE-Tagger



GrNE-Tagger

Version: 0.4

<http://hdl.handle.net/11500/ATHENA-0000-0000-23F2-7>

Ο Επιστημιακής Ονομαστικής Οντοτήτων GrNE-Tagger επεξεργάζεται κείμενα αναγνωρίζοντας αυτόματα βάσει κανόνων και ταξινομώντας τις Ονομαστικές Οντότητες (κύρια ονόματα) που αυτά περιέχουν στις ακόλουθες πέντε κατηγορίες: Άτομο/Πρόσωπο (PERSON), Τοποθεσία (LOCATION), Οργανισμός (ORGANIZATION), Ονόματα κτιρίων ή/και άλλων ανθρώπινων κατασκευών (FACILITY), Γεωπολιτική Οντότητα (Geopolitical entity - GPE). Ο GrNE-Tagger χρησιμοποιείται για την επεξεργασία κειμένων της νέας Ελληνικής και δέχεται ως είσοδο κείμενα σε μορφή ... [Read More](#)

Select Language

en



ATHENA
Επιστήμη & Καινοτομία
Τεχνολογία Πληροφοριών

Cite

Citation text

Annotator of Named Entities GrNE-Tagger
(2015). Version 0.4. [Software (Tool/Service)]
CLARIN-EL: None

Function

Named Entity Recognition

Language dependent

yes

Keywords

named entities text

Overview

Technical

Relations

Access



Modes of distribution

Use

Access info

Software distribution form
docker image
Operating system

Licensing Info

Licence
CLARIN-EL Terms of Service
http://inventory.clarin.llsp.gr/catalogue_backend/static/licence/clarinelTOS.pdf

Click on the *Use* button. In the next window, you will be presented with the workflow created for the service you chose. You must click on **Use this workflow**.

1

Select function

2

Upload data

3

Process

NamedEntityRecognition

Annotator of Named Entities GrNE-Tagger (ILSP workflow v1.0.0)

Input

input_language
el
input_character_encoding
UTF-8
input_data_format
Xml
input_typesystem

Output

output_language
el
output_character_encoding
UTF-8
output_data_format
Xml
output_typesystem

Use this workflow

Step 2. Upload your data

In the new window, you are informed about the prerequisites of the processing, i.e. the specifications of the dataset to be uploaded. If you wish to process your own dataset, it needs to fulfil these conditions; then, you can upload it.

1. have a size of max 2MB
2. have a filename in Latin characters, with no spaces in it
3. contain only plain txt files in UTF-8 encoding
4. not contain other zip files

Drag & Drop your files or Browse

Annotator of Named Entities GrNE-Tagger (ILSP workflow v1.0.0)

Input	Output
input_language	output_language
el	el
input_character_encoding	output_character_encoding
UTF-8	UTF-8
input_data_format	output_data_format
Xmi	Xml
input_typesystem	output_typesystem

Reset Back Next

File has been successfully uploaded.

After the dataset has been successfully uploaded, the **next** button is activated and you can click on it.

Select function Upload data Process

The dataset is being processed. You will be notified by email when it's finished.

Reset Back Finish

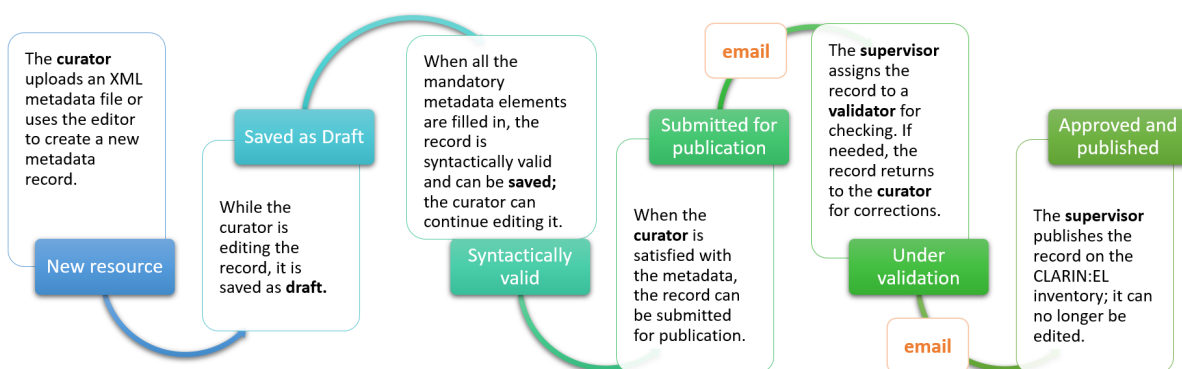
Step 3. Get the processed files

You will be notified by email once the processing is finished. To see the results go to your dashboard and check the *Processing tasks*.

Attention: Both the data uploaded for processing and the data which result from the processing are **not stored** permanently in the infrastructure; the CLARIN:EL policy is to delete the annotated data 48 hours after processing has been completed. If you wish to download them, please, do so during this time frame.

THE LIFECYCLE OF A RESOURCE

CLARIN:EL is an infrastructure for language resources and tools/services which reside in the repositories of the members of the *CLARIN:EL network*; they are harvested and presented in the central inventory. A resource (i.e., a metadata record and, optionally, content files uploaded with it) goes through a set of states in the process of being prepared for publication on the inventory, as depicted in the following figure:



The states are:

- **new resource**: A *curator* creates a resource, by creating a metadata record through the *metadata editor*¹ or by uploading an *XML metadata file* and, optionally, content files.
- **draft (internal)**: When using the interactive editor, the curator can save² the metadata record as draft, even without filling all *mandatory* elements; only compliance as to the data type of the elements is checked (e.g. elements that take URL must be filled in with the accepted pattern).
- **syntactically valid (ingested)**: The system checks that the metadata record complies with the *CLARIN:EL metadata schema* and all *mandatory* elements are filled in. The curator can continue to edit it until satisfied with the description and can then submit it for publication.

¹ Henceforth **editor**.

² See the differences in **save as draft** and **save** *here*.

- **submitted (assigned for validation):** Once the curator submits the resource for publication, the record is no longer editable. The *supervisor* receives an email to assign³ *validators*. Depending on the resource type (and the uploaded content files, if any), the resource is checked by the metadata and legal validator⁴. The validation aims to check the consistency of the description; it doesn't include any qualitative evaluation. When validators identify a problem, they contact the curator for further information and may ask the curator to edit the metadata; in such cases, the status of the resource is changed to **syntactically valid** again and the curator is notified to make the appropriate amendments.
- **published:** When the validator(s) have **approved** a resource, the supervisor is notified by email and must publish it. After being made visible in the CLARIN:EL inventory the metadata record cannot be edited any more. It can return to the syntactically valid status only if **(requested to be) unpublished**.

<p>Attention: Only the metadata records created via the editor can be saved as draft. The XML metadata files cannot be imported unless they are syntactically valid (which is the status they are automatically set to).</p>

³ When there are more than one supervisors in a repository, one should first be assigned to the resource. See [here](#) how to do this. Then, the supervisor must choose and *assign* validators.

⁴ When no content files have been uploaded, the resource is considered automatically legally valid.

IMPORTANT INFORMATION ABOUT ALL LRTS

Before you embark on creating a language resource for CLARIN:EL, please keep in mind the following:

- the resource must comply with our [terms of service](#)¹;
- the legal status of the resource must be clear and a licence must be provided;
- the resource can be created either by using the *interactive editor* or by *uploading an XML file* with at least the mandatory metadata per resource type;
- for the resource to be complete, the physical data (henceforth **content files**) must be provided as well: *directly to the infrastructure* or indirectly with the use of an external link.

Depending on the type of resource you would like to create, please refer to the respective chapters for the mandatory elements needed to create a metadata record for:

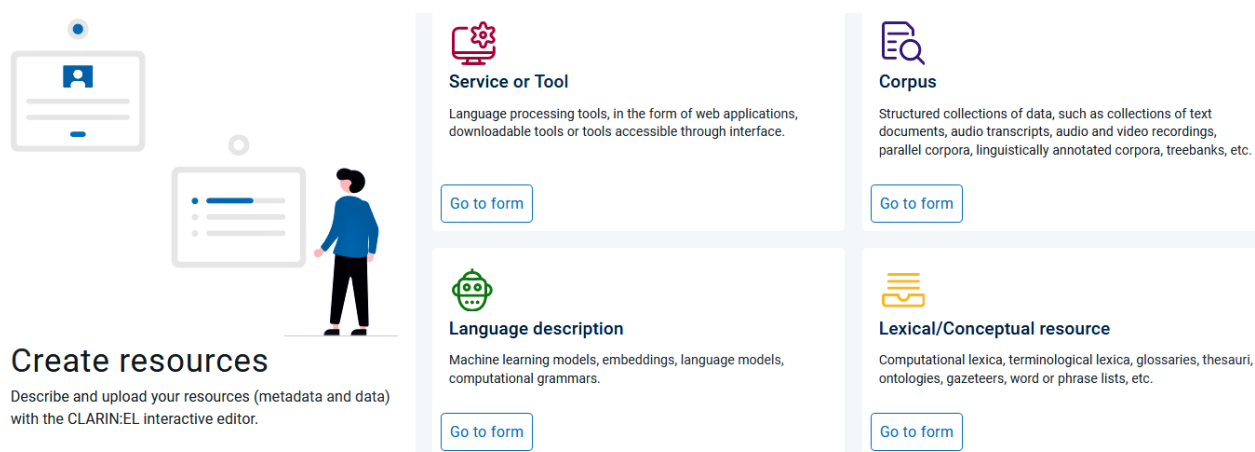
1. a *corpus*,
2. a *tool*,
3. a *lexical/conceptual resource*,
4. a *language description*

¹ See also our [privacy policy](#).

HOW TO CREATE A RESOURCE BY USING THE EDITOR





Attention: This section is dedicated to the **basic steps** you must take to create a resource using the metadata editor¹. Although the example showcases how a corpus is created, the steps (except indicated otherwise) are common for all LRTs.

In order to describe a resource through the editor you must be signed in. If so, visit your *dashboard* and click on *create resources*. In the new window, choose the resource type you wish to create by clicking on **Go to form**.



Create resources

Describe and upload your resources (metadata and data) with the CLARIN:EL interactive editor.

 Service or Tool Language processing tools, in the form of web applications, downloadable tools or tools accessible through interface. Go to form	 Corpus Structured collections of data, such as collections of text documents, audio transcripts, audio and video recordings, parallel corpora, linguistically annotated corpora, treebanks, etc. Go to form
 Language description Machine learning models, embeddings, language models, computational grammars. Go to form	 Lexical/Conceptual resource Computational lexica, terminological lexica, glossaries, thesauri, ontologies, gazeteers, word or phrase lists, etc. Go to form

Before you start, make sure you have checked which are the *mandatory* elements in order to be able to create a metadata record for

- a *corpus*,
- a *tool*,
- a *lexical/conceptual resource*,
- a *language description*

¹ Henceforth **editor**.

16.1 Step 1: Name your resource

The first step is to give your resource a name; in order to avoid confusion, the system checks if this name is already in use. If this is the case, you are presented with a list of resources which match your proposed resource name (wholly or partly). If you want to create a new record, you have to use another name².

corpus name

test corpus

Check if the corpus is already listed in the catalogue

[ACCURAT balanced test corpus for under resourced languages](#)
[new test corpus m](#)

test corpus	
Test corpus1	
Test corpus1	
Test corpus1	
Test corpus1	
test corpus for validation	
test corpus	

[Create test corpus](#)
[Cancel](#)

If the name is not found, you can proceed to the next phase by clicking the **create** button³.

corpus name

Demo resource

Check if the corpus is already listed in the catalogue

No matches

[Create Demo resource](#)
[Cancel](#)

Attention: After a tool name has been checked, you will be presented with a different screen asking you whether you would like it to be **integrated in the CLARIN:EL infrastructure as a service**.

Do you want to contribute a tool that will be integrated in CLARIN:EL as a functional service (i.e., available through the CLARIN:EL APIs)?

- ☐ Yes
- ☐ No

Depending on your answer, the respective box will/will not be checked in the editor. You can change your decision anytime.

² Otherwise, you can edit one of the resources in the list, provided you have the rights to do so.

³ The button is deactivated when matches of the name are found. As soon as you use a new name, it becomes activated.

CREATE A NEW SERVICE OR TOOL

Info

1. At any point you are able to save your progress as draft and continue editing at a later time
2. Once you submit your record you will not be able to save it as draft any more, but you will still be able to make changes and submit them.
3. Visit all tabs in order to fill in as much information as possible for better visibility of your record.
4. If the metadata record is for a resource that you plan to deliver later, please check the "work in progress" box.

LANGUAGE
RESOURCE/TECHNOLOGY

TOOL/SERVICE

DISTRIBUTION

DATA

☐ For information
☒ CLARIN:EL compatible service

Save draft

Save

IDENTITY

CATEGORIES

CONTACT

DOCUMENTATION

RELATED LRTs

LRT name *

Demo tool

The official name or title of the language resource/technology

LRT identifier

A string used to uniquely identify the language resource/technology

Fill in

language

English

select language

+

LRT short name

An abbreviation, acronym, etc. used for the language resource/technology

Description *

language

English

select language

+

16.2 Step 2: Upload the resource data

Attention: Although the functionality is available for all types of resources, it is advisable to **integrate tools as services**, in collaboration with the CLARIN:EL technical team, rather than upload the software.

The content files must be in a **compressed folder** in one of the following formats: **.zip, .tgz, .gz, .tar**⁴. When naming the folder you must use the latin alphabet and leave no spaces between the words. See [here](#) how to prepare the data before uploading.

16.2.1 Skip and upload later

When you click on **create**, a new window appears asking you to upload your data. If you want to, you can **skip** this step and upload your data later.

Please upload your datasets for this record.

Upload data

Skip

⁴ If the files are available in multiple formats, (e.g. in XML, TXT and PDF formats), you are advised to package them in different compressed files by data format.

If you decide to upload your content files at a later time, simply visit the editor and select the *Data* section where the **upload** button is found.

The screenshot shows the CLARIN editor interface. At the top, there is a navigation bar with several tabs: 'LANGUAGE RESOURCE/TECHNOLOGY', 'CORPUS', 'PART', 'DISTRIBUTION', and 'DATA'. The 'DATA' tab is currently selected and highlighted in blue. To the right of these tabs, there are two checkboxes: 'For information' and 'Metaresource', both of which are unchecked. Further right is a green 'Save' button with a downward arrow icon. Below the navigation bar, the main content area is divided into two sections. On the left, there is a vertical sidebar with the word 'DATA' in blue. The main area on the right contains a large, light gray rectangular box with the text 'Upload data' centered inside it.

16.2.2 Upload immediately

Otherwise, you can directly upload your data to the metadata record you are about to create. A series of screens (as shown in the image below) will be presented to guide you through:


1. select the dataset,
2. upload it,
3. see the details of your upload when it is completed, and
4. click on the **finish** button to go back to the editor.

Please select a .zip file in order to upload a dataset for this record

Drag & Drop your files or [Browse](#)

[cancel](#) [upload dataset](#)


Please select a .zip file in order to upload a dataset for this record



[cancel](#) [upload dataset](#)

Please upload your datasets for this record.

[Upload data](#)

Name	Upload date	Assigned to distribution	Actions
singleTXT.zip	27 May 2021	No	

*In order to delete a dataset you should first unlink it from the corresponding distribution and submit/save your record.

[4](#) [Finish](#)

16.2.3 Assign to distribution

As shown in the last image, the dataset is **not assigned** to any distribution; that means that you have to create an **association** between the dataset and the form or delivery channel through which it is distributed (e.g. a CD-ROM, a link from where the dataset can be downloaded, etc.). In addition, you must define the **licence terms** under which this form of distribution is available.

Attention: You do not have to associate the dataset with the distribution form and licence terms upon uploading the dataset; you can create the link at anytime. To do so, click on the *Distribution* section.

Dataset distribution form *

downloadable

Select the form or delivery channel through which the corpus is distributed

Private

☐ Yes

☒ No

Specifies whether the resource is private so that its access/download location remains hidden

Associate a dataset with this distribution

singleTXT .zip

First select the form of distribution from the dropdown list and then click on the name of the zip file. Save the changes you have made. The next time you check the metadata record you will see that the dataset is assigned.

LANGUAGE RESOURCE/TECHNOLOGY CORPUS PART DISTRIBUTION **DATA** ☐ For information ☐ Metaresource Save

DATA Upload data

Name	Upload date	Assigned to distribution	Actions
singleTXT.zip	27 May 2021	Yes	

*In order to delete a dataset you should first unlink it from the corresponding distribution and save your record.

You will **not be able to submit a record for publication**, unless all uploaded datasets are assigned to their respective distributions. Upon submitting a record, you will be notified and transferred to the editor.

LANGUAGE RESOURCE/TECHNOLOGY CORPUS PART DISTRIBUTION **DATA** ☐ For information ☐ Metaresource Save

IDENTITY LRT name * demo corpus language English +

The official name or title of the language resource/technology select language

LRT Identifier Fill in

A string used to uniquely identify the language resource/technology

LRT short name language English +

An abbreviation, acronym, etc. used for the language resource/technology select language

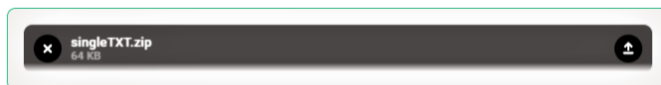
Description * language

Invalid record. The record has data not related with any Dataset Distribution

16.2.4 Replace content files

If you need to replace the content files you have uploaded, create a new compressed file **with the correct data** and the **same file name (and format)**. From the *Data* section choose to upload data. You will be presented with a warning message.

Please select a .zip file in order to upload a dataset for this record



Warning

There is already a file with the same name. By proceeding you are going to delete the already uploaded file and replace it with this one. Are you sure?

[cancel](#) [replace dataset](#)

Click on **replace dataset** and if the replacement is successful, you will be notified by a message at the bottom right side of your page. The new dataset is indicated by the **new upload date** as shown in the image below.

LANGUAGE RESOURCE/TECHNOLOGY
CORPUS
PART
DISTRIBUTION
DATA

☐ For information
☐ Metaresource

Save

DATA

Upload data

Name	Upload date	Assigned to distribution	Actions
singleTXT.zip	03 July 2021	Yes	

*In order to delete a dataset you should first unlink it from the corresponding distribution and save your record.

File has been replaced successfully.

Save the metadata record. You will be transferred to the resource view page while being informed that the resource has been updated successfully.

Single text EL (txt)

Corpus

Version: 1.0.0 (automatically assigned)

<http://hdl.handle.net/11239/CLARIN-EL-0000-0000-60FD-7-TEST>

Singletxt RbNh6VUcQs7yYdVUhsCR6A (text/plain) annotated by the ILSP Sentence splitter and tokenizer for Greek text (ILSP workflow v1.0.0).

Automatically generated metadata. Please edit them.

Cite

Citation text

Single text EL (txt) (2021). Version 1.0.0 (automatically assigned). [Dataset (Text corpus)]. CLARIN:EL.
<http://hdl.handle.net/11239/CLARIN-EL-0000-0000-60FD-7-TEST>

Actions

Language
Modern Greek (1453-)

Keywords
corpus
annotated corpus

Corpus subclass
annotated corpus

Record updated successfully.

16.2.5 Delete content files

Once you have assigned the content files to a distribution, the delete action in the *Data* section will be deactivated and you will be presented with a warning message.

LOGY CORPUS PART DISTRIBUTION **DATA** ☐ For information ☐ Metaresource Save

Upload data

Name	Upload date	Assigned to distribution	Actions
singleTXT.zip	27 May 2021	Yes	

*In order to delete a dataset you should first unlink it from the corresponding distribution and save your record.

In order to delete a dataset you should first unlink it from the corresponding distribution and submit/save your record.

To delete them, follow the next steps:

1. Go to the *Distribution* section.
2. Select the void option in the respective metadata field (this action will **disconnect** the data from the distribution).

Dataset distribution form *
downloadable

Select the form or delivery channel through which the corpus is distributed

Private
☐ Yes
☒ No

Specifies whether the resource is private so that its access/download location remains hidden

Associate a dataset with this distribution

singleTXT .zip

3. Save the metadata record.
4. Choose **again** to edit the metadata record.
5. Go to the *Data* section. The delete button is now activated.
6. Delete the data.

You will see a message that the action has been successfully completed.

LANGUAGE RESOURCE/TECHNOLOGY CORPUS PART DISTRIBUTION **DATA** ☐ For information ☐ Metaresource Save

DATA Upload data

File deleted.

7. Save the metadata record.

16.3 Step 3: Fill in the mandatory metadata

Whichever type of resource you choose, you have to provide information on some *mandatory* metadata elements (different per resource type) which are distributed in several **sections** (organized horizontally) and various **tabs** (presented vertically) in the editor.

During the creation process you can stop any time and save your record **as draft**⁵. You will not be able to **save** your record unless you have filled in all the mandatory metadata. Every time you click on **save**, the metadata you have entered are checked; each time mandatory metadata are missing or have false values, you will get a message prompting you to correct your errors.

! Correct the following errors in order to proceed

Language Resource/Technology > Identity > Description is required
 Language Resource/Technology > Identity > Description language is required
 Language Resource/Technology > Categories > Keyword is a required field
 Language Resource/Technology > Contact > Additional information is a required field
 Corpus > Technical > Corpus subclass is a required field
 Corpus > Technical > Personal data included is a required field
 Corpus > Technical > Sensitive data included is a required field
 Part > Media part > Corpus part is a required field
 Distribution > Technical > Dataset distribution is a required field

Close Proceed

After closing this message a new highlighted area appears above the sections. It contains the **path** (section > tab) to each one of the missing/incorrect metadata. When you click on it, you are **automatically transferred** to the respective section tab where the missing metadata are highlighted with a vertical red line.

⁵ You can even save as draft a record with only the resource name.

CREATE A NEW CORPUS

Correct the following errors in order to proceed

1. Language Resource/Technology > Identity > Description is required
2. Language Resource/Technology > Identity > Description language is required
3. Language Resource/Technology > Categories > Keyword is a required field
4. Language Resource/Technology > Contact > Additional information is a required field
5. Corpus > Technical > Corpus subclass is a required field
6. Corpus > Technical > Personal data included is a required field
7. Corpus > Technical > Sensitive data included is a required field
8. Part > Media part > Corpus part is a required field
9. Distribution > Technical > Dataset distribution is a required field

LANGUAGE RESOURCE/TECHNOLOGY CORPUS PART DISTRIBUTION DATA

☐ For information
☐ Metaresource

[Save draft](#) [Save](#)

IDENTITY

LRT name * Demo resource
The official name or title of the language resource/technology

language English
select language

LRT identifier
A string used to uniquely identify the language resource/technology [Fill in](#)

LRT short name
An abbreviation, acronym, etc. used for the language resource/technology

language English
select language

Description *
General information on the language resource/technology

language English
select language

Upon checking whether you can save this record after filling in the required metadata, you might be informed again with a message that more metadata are needed. This happens because some of the values you have entered have generated new requirements (these are the *mandatory upon condition* metadata). Again, the new metadata you must fill in appear at the top of the editor page.

Correct the following errors in order to proceed

- Part > Media part > Corpus text part > Language is a required field
- Distribution > Technical > Text features is a required field
- Distribution > Technical > Licence is a required field
- Distribution > Technical > Distribution location is a required field

[Close](#) [Proceed](#)

This message might appear several times before all the necessary metadata have been filled in. After the process has been completed, you will be finally allowed to save the metadata record.

☒ Save

You are about to update Demo resource Are you sure?

[Close](#) [Proceed](#)

16.4 Step 4: View the created record

After you click on **proceed**, you get a message that the metadata record has been successfully created and you are transferred to the resource view page.

The screenshot displays the CLARIN-EL portal interface. At the top, the 'clarin:el' logo is on the left, and navigation links for 'Dashboard', 'Help', and 'Curator H.' are on the right. A 'CLARIN-EL portal >' button is also present. A blue banner at the top right contains a 'Go to catalogue' link. Below this, a 'STATUS' section explains the workflow: draft, syntactically valid, submitted, approved, and published. A progress bar shows the current status as 'syntactically valid'. The main content area shows the record details for 'corpus test', including its version (1.0.0), a link to the pid url, and the text 'Hello'. Metadata fields are displayed in a grid: 'Language' (Kalaallisut, Greenlandic), 'Keyword' (Hello), and 'Corpus subclass' (raw corpus). A green notification bar at the bottom right states 'Record updated successfully.'.

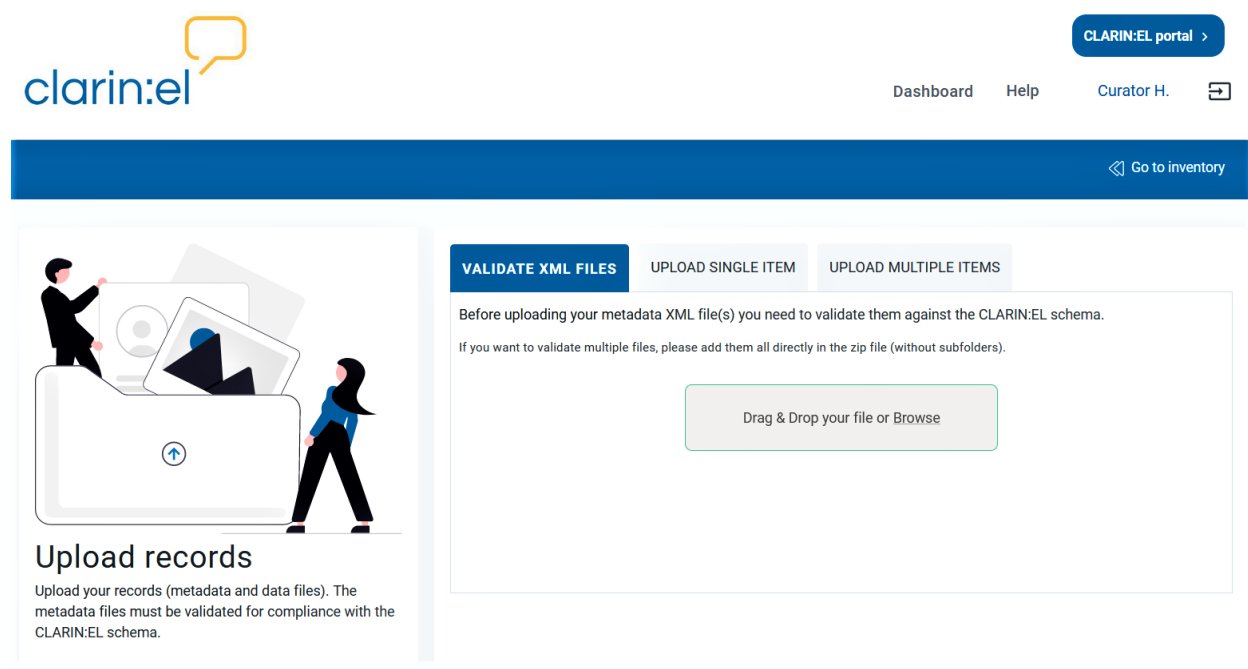
You can still *edit* the metadata you have added before you decide to submit the record for publication.

HOW TO CREATE A RESOURCE BY UPLOADING AN XML FILE

You can create any type of resource by uploading an XML file with at least the *mandatory* metadata.

Attention: This section is dedicated to the **basic steps** you must take to create a resource by XML upload. Although the example showcases how a corpus is created, the steps (except indicated otherwise) are common for all LRTs. See [here](#) specific XML files describing different types of resources.

In order to upload an XML you must be *signed in*. If so, visit your *dashboard* and click on *upload resources*. In the new window, you are presented with three tabs, as shown in the image below.



First you must validate your XML description file, i.e. you must upload it to be checked. Although you can skip that step, it's highly recommended and it will save you time; the validation shows inconsistencies or missing metadata, as in the image below.

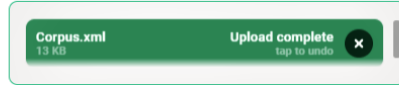
VALIDATE XML FILES

UPLOAD SINGLE ITEM

UPLOAD MULTIPLE ITEMS

Before uploading your metadata XML file(s) you need to validate them against the CLARIN:EL schema.

If you want to validate multiple files, please add them all directly in the zip file (without subfolders).



Validation report

```
{
  "info": "file 'Corpus.xml' not valid",
  "errors": [
    {
      "line": 213,
      "error": "Element 'LicenceIdentifier': This element is not expected. Expected is one of (
    }
  ]
}
```

Once you have corrected them, try to validate the file again. If everything is as it should be you will be notified that the xml file is valid.

VALIDATE XML FILES

UPLOAD SINGLE ITEM

UPLOAD MULTIPLE ITEMS

Before uploading your metadata XML file(s) you need to validate them against the CLARIN:EL schema.

If you want to validate multiple files, please add them all directly in the zip file (without subfolders).



Validation report

```
{
  "info": "file 'Corpus.xml' valid"
}
```

Then, you can proceed to the next tab to successfully upload your file. If your resource falls into one of the following categories¹ you must click the appropriate box before uploading the XML file.

¹ See [here](#) the various categories of resources.

VALIDATE XML FILES

UPLOAD SINGLE ITEM

UPLOAD MULTIPLE ITEMS

You can upload one **XML** file at a time.

It is highly recommended that you [validate](#) your XML file against the CLARIN:EL schema before you proceed.

☒ For information

If the metadata record is for a resource that you plan to upload at a later stage, please check this box.

☐ CLARIN:EL compatible service


If the metadata records are for services to be integrated in CLARIN:EL, please check the box "CLARIN:EL compatible service".

☐ Metaresource


If this is a metaresource, please check the box "metaresource".

Drag & Drop your file or [Browse](#)

You will be informed that the process has been successfully completed and you will be transferred to the resource view page.



[CLARIN:EL portal >](#)

[Dashboard](#)
[Help](#)
[Curator H.](#)



[Go to inventory](#)

STATUS

You can continue editing your metadata record while the status is draft (syntactically valid); when you are satisfied with it, you can submit it for publication; the resource will be validated and published by the supervisor of the repository, or, if required, you will be contacted for further information

[draft](#)
[syntactically valid](#)
[submitted](#)
[approved](#)
[published](#)

Demo corpus

 Corpus

short name

Version: 1.0.0

Short description of the resource for human readers with the basic features (e.g., language, size, format, domain or genre, etc); the example for this template is for an video of lectures in English, the audio recordings, their transcriptions, the subtitles in English and their translations into German

[for info](#)

Language

English German

Keywords

multimedia corpus video lectures subtitles

Corpus subclass

raw corpus

File has been successfully uploaded.

HOW TO PREPARE DATA BEFORE UPLOADING

The CLARIN:EL infrastructure allows data uploading for two reasons:

- to deposit the **content** of a resource (see [section 1](#)), or
- to use the data as **input** to a processing service (see [section 2](#)).

Attention: In each situation both the user involved and the requirements for the data are different. In both cases *signing in* is a **prerequisite**. If you wish to upload data and you are not a CLARIN:EL user you must first *register*.

18.1 I. Depositing data as content of a resource

18.1.1 1. TYPES

Not all *resources* in the CLARIN:EL infrastructure have content files (see for example *metaresources* and resources only “for info”) but they all have metadata descriptions. When they do have content, it varies depending on the *resource type*.

- **Corpora** are collections of:
 - *primary data* of various media:
 - * digital/digitised written texts (e.g. digitised books, web texts, newspapers, corpora etc.), recordings of spoken language (e.g. interviews, radio broadcasts etc.)
 - * video recordings (e.g. TV shows, facial expressions collections, gestures etc.)
 - * images (e.g. digital/digitised photographs with their captions etc.)
 - or
 - *processed data*
 - * various types of annotations of texts,
 - * sound and multimedia data, automatically or manually created (e.g. morphosyntactically annotated texts, transcriptions of spoken data, video annotations etc.)
- **Lexical/conceptual resources and language descriptions** are:
 - structured language data (e.g. word lists, lexica, thesauri, grammars etc.) used for processing and study of primary and processed data.
- **Tools** are:
 - *source code*, or

- *software*

of programs/applications performing various types of language processing (e.g. multilingual text alignment, morphological annotation, lemmatisation, parsing, knowledge extraction etc.).

18.1.2 2. USAGE

Uploading data to CLARIN:EL does not automatically guarantee they are directly accessible by the CLARIN:EL users nor that they are processable by the CLARIN:EL tools and services. Two factors are to be taken into consideration: **accessibility** and **processability**, which applies only to corpora (*processable*) and tools (*processing services*). In order to prepare the data in the most appropriate way you must have an answer to the following questions beforehand.

Important:

- Do I want my data to be **accessible** to the CLARIN:EL users?
 - If the answer is **yes**, please see the *respective section* before reading the following instructions.
 - Do I want my corpus data to be **processable** by the CLARIN:EL services?
 - If the answer is **yes**, please check the necessary metadata values along with all the other instructions.
 - Do I want my tool data to be converted into a CLARIN:EL **processing service**?
 - If the answer is **yes**, please check the *necessary metadata values* along with all the other instructions.
-

The following instructions are divided into two sections: general instructions apply to all types of resources while specific instructions apply only to corpora and tools, as indicated.

18.1.3 3. Steps to follow

3.1. General instructions

There are several legal documents which you need to consult before proceeding. Make sure you have read carefully the CLARIN:EL

- [Privacy Policy](#),
- [Terms of Service](#), and

as uploading data to the infrastructure entails that **you have agreed** to the aforementioned legal documents.

If you are **affiliated** to an organization member of CLARIN:EL make sure to contact your [Scientific Responsible](#) before depositing your data.

If you are **not affiliated** to an organization member of CLARIN:EL you need to sign the [depositor's agreement](#) before depositing your data.

Also ensure that the data you provide have **clear licence terms** and **permission received from all right-holders** involved. If the data have more than one *distributions* you will need to indicate the licence terms for each one of them. In addition, they can also be available under multiple **licence terms** depending on the user nature or the intended use (academic vs commercial).

Then, you can proceed with the three stages of data preparation: **collection**, **categorization** and **compression**.



Step 1: Collection

Collect data around a specific idea (e.g. a [glossary of feminist theory](#)). Collect **all** and **only** the necessary data. If **personal**, **sensitive** or **confidential** data are included, please anonymize them or remove them before uploading.

Step 2: Categorization

Collected data may be the result of various processing stages: video recordings which have been transcribed, PDFs which have been cleaned (images and URLs removed) and converted to TXT files. In such cases, the **raw** and **converted** data, comprise a unity involving multiple and various **formats**, **media** and **languages** which you might not want to break. To do so, and present everything in a single metadata record, you must organize your data in the most structured and easy to understand way. By grouping them in a coherent and cohesive way, you will not only facilitate other users but also make the data compatible with the infrastructure services and workflows. The following guidelines aim at helping you do so in such a way that **no information is lost** and the text part of your (corpus) data is processable.

Attention: These guidelines do not address categorization based on domain, time/geographic coverage etc.

Multiple formats

If the data are in various formats (e.g. XML, TXT, PDF, etc.), organize the files according to their format. Group all files of the same format in one folder (e.g. all XML files together). You can upload two different datasets (e.g. XML vs TXT) on the same metadata record by associating each one of them with a separate *distribution*.

Tip: See the list of *recommended file formats* for the CLARIN:EL infrastructure.

Multiple media

If the data have various media parts (e.g. text, audio, etc.), organize the files according to the medium. Group all files of the same medium in one folder (e.g. all text files together in one file and all audio files in another). You can upload two different datasets (e.g. audio recordings and transcripts) on the same metadata record by associating each one of them with a separate *distribution*.

Naming files and folders

Name both the files and the folders in a way that reflects **meaningfully and consistently** their content. Use the latin alphabet and leave no spaces between the words. If you have files in various formats, media and/or languages, label them accordingly (e.g. news1_el.txt, news1_en.txt).

Important: Any relevant documentation (e.g. manuals, questionnaires, codebooks, project reports, etc.) should be directly described and uploaded to the **respective field** in the metadata editor¹. Nevertheless, if you wish to include any documentation in the data folder, create a separate file and name it “README” (in TXT or PDF format). This file should contain all the necessary information on the methods used for collecting/generating the data and explanations about the structure, the naming of the files or any other kind of information that can help the user.

Consistency

The metadata used to describe your data should clearly reflect them. Make sure there are no inconsistencies (e.g. check that your files are indeed in PDF format and not just scanned images; if you provide information on an annotated corpus, indicate the annotation tool etc.) to avoid any problems. Check [here](#) the mandatory metadata for all resource types but also keep in mind that an LRT description is more complete if the recommended metadata are provided as well.

Step 3: Compression

The content files must be in a **compressed folder** in one of the following formats: **.zip**, **.tgz**, **.gz**, **.tar**. When naming the folder you must use the latin alphabet and leave no spaces between the words.

Attention: Do not compress the embedded files/folders since this makes it impossible for the CLARIN:EL services to handle them (i.e. do not include .zip files within a .zip file).

3.2. Specific instructions

Corpora

In order to become processable, a corpus must have the features described below:

- **multilinguality:**
 - for **monolingual** corpora, the language must be *Greek*, *English*, *German* or *Portuguese* (currently these are the language supported by the services),
 - for **bilingual** corpora, *Greek* should be the one language in a pair where *English*, *German* or *Portuguese* is the other.
- **medium:** *ext*
- **format:**
 - for **monolingual** corpora the formats are *Plain Text* and *XCES*,
 - for **bilingual** corpora the formats are *TMX* and *MOSES*.
- **encoding:** *UTF-8*
- **size:** *< 60Mb*
- **licence:** Creative Commons licences (CC, starting with Creative Commons Zero (CC-0) and all possible combinations along the CC differentiation of rights of use). See also the [Recommended licensing scheme for Language Resources](#).

Corpora with these features are compatible with the workflows of the infrastructure and are indicated as *processable*. The processable corpora are grouped together as a [subset](#) of the total list in the inventory home page.

¹ Henceforth **editor**.

Tools

If you would like to integrate a tool to the CLARIN:EL infrastructure as a compatible service, please indicate your choice upon the creation of the resource and contact the [CLARIN:EL technical team](#).

Do you want to contribute a tool that will be integrated in CLARIN:EL as a functional service (i.e., available through the CLARIN:EL APIs)?

☐ Yes

☐ No

18.1.4 4. UPLOAD

When you have finished you can upload the data.

Attention: This action is available only to *signed in curators*.

As a curator you are provided with two options for uploading:

- *upon the creation of the metadata record,*
- *at a later time.*

Once you are done with uploading you must associate the data with a **distribution**, the form or delivery channel through which the data are distributed, described [here](#).

You can repeat the procedure (data upload → association with distribution) as many times as you need to, having different sets of data associated with various distributions. This functionality serves not only the various ways through which **the same data** are distributed (e.g. a CD-ROM, a link from where a dataset can be downloaded, etc.) but also the **various data formats or media** (e.g. PDF vs TXT, Audio files vs Transcripts, etc.) which can be treated separately.

Tip: If you encounter any problem during uploading, please contact the [Technical helpdesk](#).

18.2 II. Data as input of a service

Attention: This action is available to all *signed in users*.

Both the data uploaded for processing and the data which result from the processing are **not stored** permanently in the infrastructure; the CLARIN:EL policy is to delete the annotated data 48 hours after processing has been completed.

Tip: If you want to, you can create a metadata record, where you can upload data, either by using the [editor](#) or by [uploading an XML file](#). Keep in mind that you must be *signed in*.

CLARIN:EL services accept as input small datasets with the following features:

- multilinguality: **monolingual** corpora in *Greek, English, German or Portuguese*,
- medium: *ext*

- format: *Plain Text*
- encoding: *UTF-8*
- size: *< 2Mb*

In addition, the data must be in a **compressed folder** in one of the following formats: **.zip, .tgz, .gz, .tar**. When naming the folder you must use the latin alphabet and leave no spaces between the words.

Attention: Do not compress the embedded files/folders since this makes it impossible for the CLARIN:EL services to handle them (i.e. do not include .zip files within a .zip file).

To find out more about processing, check:

1. how to access a *service*,
2. how to access a *workflow*.

RECOMMENDED FILE FORMATS

19.1 Guidance on selecting file formats for long-term accessibility and interoperability

This section¹ lists the file formats which are recommended for depositing in CLARIN:EL.

19.1.1 File Formats for Digital Preservation Policy

To ensure access and usability of your data to the broadest audience into the long term, the CLARIN:EL team has considered the following factors to determine which file formats are recommended in CLARIN:EL infrastructure:

- **Processability**
 - Suitability for the type of resource and/or type of processing.
 - In order to be processable by the [CLARIN:EL integrated NLP workflows](#), textual data have to be in one of the formats that the workflows can process (listed below).
- **Preservation**
 - Suitability for research by the designated communities.
 - How widespread the format is: broadly used formats, not deprecated, known to the designated communities.
 - Use of open source rather than proprietary format.
 - Whether the format employs lossy or lossless compression.

The policy, which is based upon the above-mentioned factors, meets the mission of CLARIN:EL to collect, preserve and distribute digital language resources and language processing services for the support of researchers, academics, students, language professionals, citizen scientists and the general public. In order to arrive at the appropriate recommendations for individual file formats, or to decide on their suitability for particular kinds of research activities/types, the purpose for which they are intended has to be considered. For example, while PDF/A has been developed for unproblematic long-term archiving and is an excellent format choice for documentation, it is undoubtedly *not suitable* for textual data intended for language processing. Therefore, based on the types of resources that are in the scope of the CLARIN:EL user communities and the processes offered/supported, the CLARIN:EL team discerns the following set, pertinent to the field of digital language resources, for which specific recommendations are provided:

- **CLARIN:EL processable data:** textual data² that can be input data for CLARIN:EL [workflows](#),
- **Textual Data:** written unstructured/plain text or originally structured text (e.g., HTML) without linguistic or other mark-up added for research purposes (*non-processable* by the CLARIN:EL workflows),

¹ The recommendations presented here have been created by the CLARIN:EL [technical team](#) to which you can address any suggested updates or questions.

² See [here](#) the guidelines on *processable* corpora.

- **Text Annotation:** annotations of textual source language data, with the original text included or as a stand-off document,
- **Language Description:** data that describe a language or some aspect(s) of a language via a systematic documentation of linguistic structures (Grammars, Machine learning (ML) models, -gram models),
- **Lexical/Conceptual Resource:** a resource organised on the basis of lexical or conceptual entries (lexical items, terms, concepts etc.) with their supplementary information (e.g., morphological, semantic, statistical information, etc.),
- **Image data:** digitized images of analogue sources of written language data for research purposes (e.g., scans of handwriting, photos of inscriptions) or two-dimensional pictures or figures that are distributed with associated textual data for NLP analysis (e.g., medical images, *image data*, accompanied with radiological reports, *textual data*),
- **Audio data:** audio recordings providing spoken language data for research purposes (e.g., audio files with the pronunciation of words for a lexicon, recorded interviews, radio broadcasts, etc.),
- **Video data:** video recordings providing multimodal or sign language data for research purposes.

19.1.2 Format Recommendations

Formats that fulfil the criteria of the Digital Preservation Policy, mentioned above, are preferred; however, additional formats are accepted, as a *first-entry level*, with the proposal for conversion to recommended formats.

Therefore, file formats are categorized into two preservation levels (recommended, acceptable) always in the context of each case. The acceptable list is not exhaustive, especially in the case of text annotation, but rather indicative, and it is proposed for an acceptable format to be converted to a recommended format.

	Recommended	Acceptable
CLARIN:EL processable data	Monolingual textual data: plain text Monolingual encoded data: XCES-ILSP variant (XML based format compliant with the XCES model for corpora) Bi-/Multilingual encoded data: TMX (XML based format for aligned data), MOSES (text-based format for parallel data)	
Textual Data	File Formats: plain text Formatted/Encoded: ODT, DOCX, PDF/A, HTML, Latex, TeX, MOSES	PDF, SGML, Rich Text Format (.rtf), Microsoft Word (.doc, .docx), PostScript
Text Annotation	File Formats: XML, XMI, CSV, TSV, RDF (all serialisation formats RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), JSON Models: XCES for corpora and structural annotation, TEI for structural and linguistic annotation, GrAF linguistic annotation, TMX for aligned, GATE linguistic annotation, CoNLL family (CoNLL-U, CoNLL-2000, CoNLL-2002, CoNLL-2003, CoNLL-2006, CoNLL-2008, CoNLL-2009, CoNLL-2012) for linguistic annotation, NIF linguistic annotation for RDF data, WARC for web crawled data	SGML, Plain Text, Microsoft Excel (.xlsx, .xls), ELLOGON
Language Description	ML Model: H5, ProtoBuf, ONNX, PMML, Pickle, MLeap, YAML, JSON N-gram model: ARPA	
Lexical/Conceptual Resource	File Formats: XML, CSV, TSV, RDF (RDF/XML, Turtle, Notation3, N-Triples, TriG, N-Quads, JSON-LD, HDT), OWL Models: LMF for lexica, OWL for ontologies, SKOS for thesauri, OntoLex-Lemon for lexica, TBX for terminological data	Microsoft Excel (.xlsx, .xls), Plain Text, SQL
Image data	All images: TIFF, SVG, JPEG 2000, PNG, GIF Scanned images: PDF/A	JPEG, BMP, Photoshop, NifTi, FlashPix, PDF
Audio data	WAV, AIFF, FLAC	MP3, MPEG, Windows Media Audio
Video data	AVI	MPEG-4, RealNetworks 'Real Video', Windows Media Video, Flash Video, QuickTime Video

WHAT ARE METADATA AND WHY ARE THEY IMPORTANT?

definition

Metadata are “*data that provide information about other data*”.

The data we wish to have information about are **language data** and **tools/services** which process them. The basic **metadata** elements used to describe the aforementioned are:

- corpora (i.e. collections of texts or other media),
- lexical/conceptual resources (i.e. collections of terms),
- language descriptions (i.e. grammars), and
- tools or services (i.e. software for natural language processing).

These metadata elements have multiple features and properties. For example the corpus element has several *children* (hierarchically dependent elements), as shown in the image, which are metadata themselves:

Element **ms:Corpus**

Namespace	http://w3id.org/meta-share/meta-share/
Annotations	<input checked="" type="checkbox"/> A structured collection of pieces of data (textual, audio, video, multimodal/multimedia, etc.) typically of considerable size and selected according to criteria external to the data (e.g. size, type of language, type of text producers or expected audience, etc.) to represent as comprehensively as possible the object of study
Diagram	<input checked="" type="checkbox"/>
Properties	<input checked="" type="checkbox"/> Content complex
Used by	<input checked="" type="checkbox"/> Element ms:LRSubclass
Model	<input checked="" type="checkbox"/> ms:lrType , ms:corpusSubclass , ms:CorpusMediaPart +, ms:DatasetDistribution +, ms:personalDataIncluded , ms:personalDataDetails *, ms:sensitiveDataIncluded , ms:sensitiveDataDetails *, ms:anonymized {0,1}, ms:anonymizationDetails *, ms:isAnalysedBy *, ms:isEditedBy *, ms:isElicitedBy *, ms:isAnnotatedVersionOf {0,1}, ms:isAlignedVersionOf {0,1}, ms:isConvertedVersionOf *, ms:timeCoverage *, ms:geographicCoverage *, ms:register *, ms:userQuery {0,1}
Children	<input checked="" type="checkbox"/> ms:CorpusMediaPart , ms:DatasetDistribution , ms:anonymizationDetails , ms:anonymized , ms:corpusSubclass , ms:geographicCoverage , ms:isAlignedVersionOf , ms:isAnalysedBy , ms:isAnnotatedVersionOf , ms:isConvertedVersionOf , ms:isEditedBy , ms:isElicitedBy , ms:lrType , ms:personalDataDetails , ms:personalDataIncluded , ms:register , ms:sensitiveDataDetails , ms:sensitiveDataIncluded , ms:timeCoverage , ms:userQuery
Instance	<input checked="" type="checkbox"/> <pre> <ms:Corpus xmlns:ms="http://w3id.org/meta-share/meta-share/"> <ms:lrType>{1,1}</ms:lrType> <ms:corpusSubclass>{1,1}</ms:corpusSubclass> <ms:CorpusMediaPart>{1,unbounded}</ms:CorpusMediaPart> <ms:DatasetDistribution>{1,unbounded}</ms:DatasetDistribution> <ms:personalDataIncluded>{1,1}</ms:personalDataIncluded> <ms:personalDataDetails xml:lang="">{0,unbounded}</ms:personalDataDetails> <ms:sensitiveDataIncluded>{1,1}</ms:sensitiveDataIncluded> <ms:sensitiveDataDetails xml:lang="">{0,unbounded}</ms:sensitiveDataDetails> <ms:anonymized>{0,1}</ms:anonymized> <ms:anonymizationDetails xml:lang="">{0,unbounded}</ms:anonymizationDetails> <ms:isAnalysedBy>{0,unbounded}</ms:isAnalysedBy> <ms:isEditedBy>{0,unbounded}</ms:isEditedBy> <ms:isElicitedBy>{0,unbounded}</ms:isElicitedBy> <ms:isAnnotatedVersionOf>{0,1}</ms:isAnnotatedVersionOf> <ms:isAlignedVersionOf>{0,1}</ms:isAlignedVersionOf> <ms:isConvertedVersionOf>{0,unbounded}</ms:isConvertedVersionOf> <ms:timeCoverage xml:lang="">{0,unbounded}</ms:timeCoverage> <ms:geographicCoverage xml:lang="">{0,unbounded}</ms:geographicCoverage> <ms:register xml:lang="">{0,unbounded}</ms:register> <ms:userQuery>{0,1}</ms:userQuery> </ms:Corpus> </pre>
Source	<input checked="" type="checkbox"/> <pre> <xs:element name="Corpus"> <xs:annotation> <xs:documentation xml:lang="en">A structured collection of pieces of data (textual, audio, video, multimodal/multimedia, etc.) typically of considerable size and selected according to criteria external to the data (e.g. size, type of language, type of text producers or expected audience, etc.) to represent as comprehensively as possible the object of study</xs:documentation> </xs:annotation> <xs:appinfo> <identifier>http://w3id.org/meta-share/meta-share/Corpus</identifier> <label xml:lang="en">Corpus</label> <subclassOf>http://w3id.org/meta-share/meta-share/DataLanguageResource</subclassOf> </xs:appinfo> </xs:element> <xs:complexType> <xs:sequence> <xs:element name="lrType" fixed="Corpus"> <xs:annotation> <xs:documentation xml:lang="en">Classifies the language resource described by a metadata record to one of the major classes</xs:documentation> </xs:annotation> <xs:appinfo> <indexed>true</indexed> </xs:appinfo> </xs:element> </xs:sequence> </xs:complexType> </pre>

What is shown in the image above is a part of the *CLARIN:EL metadata schema* dedicated to the corpus element. A **schema** is a complicated detailed *map* where all elements are located, defined, described and associated with each other hierarchically. All this information is stored in an external document called **XSD: XML Schema Documentation**.

XML stands for **eXtensible Markup Language**. It is a language designed to label data by using **tags** <>¹. The tags represent the data structure and contain the **metadata**. The XSD also expresses a set of rules to which an XML document must conform in order to be considered *valid* (according to a specific schema).

The schema is created to help different types of users to **describe**, **organize**, **retrieve** and **reuse** resources (for more information see the *Fair Principles* section). As for the resources found in *CLARIN:EL*, the schema created provides information on questions such as the following:

- **What** is the nature of the resources?
- **How** were the resources **created**?
- **Why** were they **created**?
- **When** were they **created**?
- **Who** **created** them?
- **What** were the **standards/tools/techniques** used, if any?
- What is their **size** (in various units)?
- What was their **source**?

The CLARIN:EL metadata schema has also foreseen for the various media, the different languages and other useful information on all types of resources which are expressed by the respective metadata elements.

¹ You can export the description of a resource in XML by visiting its *view page*.

Each piece of information encoded as a metadata element is *more or less necessary* for the description of a resource. This is expressed by the various degrees of **optionality** as depicted in the following table:

If a metadata element is	Then
mandatory	it must always be provided
recommended	it is still important, therefore should be provided
mandatory upon condition	it becomes mandatory after a certain value of <i>another element</i> has been filled in
recommended upon condition	it becomes recommended after a certain value of <i>another element</i> has been filled in
optional	“you should never say ‘this metadata isn’t useful’; be generous and provide it anyway!” ²

Tip: See [here](#) the mandatory metadata elements for CLARIN:EL.

Each element takes a specific value. This value is the acceptable content to be enclosed between the metadata tags and it varies from alphanumeric strings to float numbers, URLs etc. These values are instantiated in some of the following examples (*click on the arrow to reveal the example*).

<ms:keyword xml:lang="en"> alignment **</ms:keyword>**

<ms:categoryLabel xml:lang="en"> Political Science **</ms:categoryLabel>**

<ms:description xml:lang="en"> This is a collection of the raw minutes of the Greek Parliament plenary sessions of the last 30 years (more than 1.000.000 speeches). The existing corpus has all raw data in txt format. In order to make the resource more processable, we have also split it into smaller subcorpora, with a maximum compressed folder size of 40 Mb per subcorpus. The created subcorpora are thematically organized per Greek parliamentary terms. **</ms:description>**

<ms:creationStartDate> 2005-10-01 **</ms:creationStartDate>**

<ms:amount> 100000.0 **</ms:amount>**

<ms:website> <http://www.ilsp.gr/> **</ms:website>**

<ms:email> name@athenarc.gr **</ms:email>**

<ms:additionalInfo> **<ms:email>** name@athenarc.gr **</ms:email>** **</ms:additionalInfo>**

You can see more examples [here](#).

² FAIR Principles > F2: Data are described with rich metadata.

FAIR PRINCIPLES

The CLARIN:EL infrastructure and *metadata schema* support the FAIR principles : **F**indability, **A**ccessibility, **I**nteroperability and **R**euse of digital assets. This section provides an overview of the FAIR principles. For detailed information, please, visit the [GoFair website](#) where each one of the principles is further subdivided and analysed.


21.1 Findability

First data should be found. One of the ways to trace a digital object regardless of changes to its location on the internet is the **persistent identifier (PID)**. A PID is a string uniquely identifying a digital object. In CLARIN:EL each resource is given a PID upon being published and can be retrieved with it even if the resource has been removed from the central inventory. For example, the *Sentiment Analysis Tool* has been unpublished but by using its PID (<http://hdl.handle.net/11500/DEMOKRITOS-0000-0000-24A2-0>) the user is directed to the resource view page where a tombstone indicates that **this resource is temporarily unavailable**.

STATUS

this resource is temporarily unavailable

Sentiment Analysis Tool

 ToolService

Sent

Version: 1.0


<http://hdl.handle.net/11500/DEMOKRITOS-0000-0000-24A2-0>

The sentiment analysis tool is a text classification and sentiment extraction tool based on n-gram graph text representations. It may be paired with various machine learning algorithms for the generation of the language model. It can be accessed by a URL endpoint as a REST service. It has been used as is, or as a part

[Read more](#)

Select Language

en



Cite current version

Giannakopoulos, George (2015). Sentiment Analysis Tool. Version 1.0. [Software (Tool/Service)]. CLARIN:EL. <http://hdl.handle.net/11500/DEMOKRITOS-0000-0000-24A2-0>

[Actions](#)

Another way to find data is via their **metadata descriptions**. The more (and accurate) metadata provided, the merrier. On the GoFair website the importance of metadata is pointed out with a simple rule of thumb: “*you should never say ‘this metadata isn’t useful’; be generous and provide it anyway!*”

21.2 Accessibility

Once the data are found, the user should know how to access them: “*anyone with a computer and an internet connection can access at least the metadata*”¹. Accessibility in this context is the ability to retrieve data and metadata without specialised or proprietary tools or communication methods. However, accessibility is not condition free. Authentication and/or authorisation could be required where necessary. In CLARIN:EL authentication and authorisation are required when a user wants access to specific rights (as a *curator*, *validator* or *supervisor*) or when one wants access to processing services. In these cases the user has to *register/sign in* first; browsing, viewing and exporting metadata records as well as downloading resources are available to non registered users.

Accessibility is also assured when metadata are available even when the data are not. Besides the tombstone mentioned for resources which have been unpublished, CLARIN:EL has also **for info** resources either because data are under process and not ready to be published or because legal clearance is pending. These metadata records still provide all the necessary information about the upcoming data and offer contact details.

The screenshot shows the CLARIN:EL portal interface. At the top, there is a logo for 'clarin:el' and a 'CLARIN:EL portal >' button. Below the logo, there are links for 'Help' and 'Sign in'. A blue bar at the top right contains a 'Go to inventory' link. The main content area displays a resource entry for 'Greek Short Stories in Greek Literary Periodicals (1880-1930) [Bibliographic Database]'. The entry includes a 'GSSLP' corpus icon, the version '1.17 (2020-09-08)', and a URL. A 'Read more' button is highlighted with a yellow box. To the right, there is a 'Select Language' dropdown menu set to 'en', a logo for the University of Crete, and a 'Cite current version' section with detailed information about the resource and its availability.

21.3 Interoperability

Interoperability concerns both data and metadata and their perception from humans and computers. In simple words, exchange and interpretation of data should be a seamless effort between humans or machines. To allow for readability without the need for additional software (algorithms, translators, mappings) commonly accepted “controlled vocabularies, ontologies, thesauri and a good data model (*a well-defined framework to describe and structure (meta)data*) should be used”².


Towards this end, qualified references between resources (cross-references which explicitly state the connection of resources) are also needed. In CLARIN:EL the metadata schema has foreseen for such links, an example of which is presented in the image below. The **OROSSIMO Corpus - Economics** is, as indicated in the relations part of the resource view page, part of the **OROSSIMO Corpus** and has outcome the **Orossimo Terminological Resource - Economics**.

¹ FAIR Principles > A1.1: The protocol is open, free and universally implementable.

² FAIR Principles > I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Relations to other resources

Part of

 OROSSIMO Corpus (1.0.0 (automatically assigned))
<http://hdl.handle.net/11500/ATHENA-0000-0000-2410-5> (Handle)

Relations to other entities

Has outcome

 Orossimo Terminological Resource - Economics 1.0.0 (automatically assigned)

In addition, all resources are cited with their PID included.

OROSSIMO Corpus - Economics

 Corpus

Version: 1.0.0 (automatically assigned)

<http://hdl.handle.net/11500/ATHENA-0000-0000-240D-A>

A corpus of academic discourse texts belonging to the Economics domain (according to the Dewey Decimal classification, DDC33 - Economics), 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 reso

[Read more](#)


processable

Select Language



en



Cite current version

Institute for Language and Speech Processing
 - Athena Research Center (2015). OROSSIMO
 Corpus - Economics. Version 1.0.0
 (automatically assigned). [Dataset (Text
 corpus)]. CLARIN:EL.
<http://hdl.handle.net/11500/ATHENA-0000-0000-240D-A> 

Cite all versions

Institute for Language and Speech Processing
 - Athena Research Center (2015). OROSSIMO
 Corpus - Economics. [Dataset (Text corpus)].
 CLARIN:EL.
<http://hdl.handle.net/11500/CLARIN-EL-0000-0000-6948-A>  

21.4 Reuse

In order to be able to reuse data, the metadata used should richly describe all aspects related to the data generation. The term **plurality** is used “to indicate that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant”³.

For the same reason, the licensing status of data should be clear. The CLARIN:EL *metadata schema* has multiple metadata elements referring to all legal aspects (licence terms, URL, conditions of use etc.) of a resource. These appear as fields in the metadata editor⁴ as shown in the image below:

³ FAIR Principles > R1: (Meta)data are richly described with a plurality of accurate and relevant attributes

⁴ Henceforth **editor**.

<div>Licence name *</div> <div>Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International</div> <div>The full title of the licence (or terms of use, terms of service)</div>	<div>language</div> <div>English</div> <div>select language</div>	<div>Remove</div> <div>+</div>
<div>Licence/terms of use/service URL *</div> <div>The URL where the text of the licence is found</div>		<div>+</div>
<div>Condition of use *</div> <div>Select the condition(s) of use for accessing the corpus; the list includes the most frequently used conditions and aims to provide brief human readable information, while the proper exposition of all conditions and possible exceptions is found inside the licence text</div>		
<div>Licence identifier</div> <div>A string used to uniquely identify the licence</div>		<div>Fill in</div>
<div>Add</div>		
<div>Cost</div> <div>The cost for using the corpus</div>		<div>Fill in</div>
<div>Access rights</div> <div>The rights for accessing the distributable form(s) of the corpus (preferably in accordance to a formalised vocabulary)</div>		<div>Fill in</div>

MANDATORY METADATA

The *metadata schema* has elements of various degrees of optionality, i.e. how necessary they are considered for the description of a resource. Regardless of the mode you choose to create a resource (via the *metadata editor*¹ or by *XML description upload*), the **mandatory** metadata must be filled in. Otherwise, the record will not be saved or the XML description will not be imported. The following sections contain images which provide an overview of these metadata per resource type and they are briefly explained. The elements which are indicated by an **asterisk (*)** are **mandatory upon condition**; this means that their necessity depends on the values of other elements provided by the user. The images replicate the editor (with sections horizontally and tabs vertically) so that you can easily track each element. In the editor, all elements, mandatory or not, are explained by definitions and examples.

22.1 Common Mandatory Elements

Tip: Some elements are mandatory for all resources!

To **describe** a resource efficiently you need to **name** it and provide a description with a few words about it. To facilitate reading for the users and make your description more appealing, make use of the rich editor functionalities (styling, word formatting, hyperlinking, alignment etc.) indicated in the image below.

The screenshot displays the 'LANGUAGE RESOURCE/TECHNOLOGY' editor interface. At the top, there are tabs for 'TOOL/SERVICE', 'DISTRIBUTION', and 'DATA'. On the right, there are checkboxes for 'For information' and 'CLARIN:EL compatible service', along with 'Save draft' and 'Save' buttons. The left sidebar contains a menu with 'IDENTITY', 'CATEGORIES', 'CONTACT', 'DOCUMENTATION', and 'RELATED LRTS'. The main content area is divided into sections: 'LRT name *' (with a text input field containing 'X tagger for French, Y speech recognizer, Z corpus' and a language dropdown set to 'English'), 'LRT identifier' (with a text input field and a 'Fill in' button), 'LRT short name' (with a text input field and a language dropdown set to 'English'), and a rich text editor for 'Description'. The rich text editor has tabs for 'Description', 'Formatting', 'Hyperlinking', and 'Alignment', and a 'Styling' section. The 'Description' tab is active, showing a text area with a toolbar containing various formatting options like bold, italic, underline, link, unlink, list, and alignment. The language dropdown for the description is also set to 'English'.

¹ Henceforth **editor**.

Next, indicate the resource **version**; if no version number is provided, the system will automatically add number v1.0.0. In addition, one or more **keywords** are asked for the resource and an **email** or a **landing page** for anyone who wishes to have **additional information** about it.

These metadata are found in the section *Language Resource/Technology* in the *Identity* tab.

You also have to describe independently each **distributable** form of the resource (i.e. all the ways the user can obtain it, either in a compact form, as a CD-ROM or from access points such as a **distribution/ download/ access location**) and you must always select the **licence** and **licence terms** under which the resource is made available (see [here](#) the Recommended licensing scheme for Language Resources).

These metadata are found in the section *Distribution* in the *Technical* tab.

22.2 Mandatory Elements per resource type

22.2.1 A Corpus at a glance

LANGUAGE RESOURCE/ TECHNOLOGY	CORPUS	PART	DISTRIBUTION	DATA
IDENTITY <ul style="list-style-type: none"> Resource Name Description Version CATEGORIES <ul style="list-style-type: none"> Keyword CONTACT <ul style="list-style-type: none"> Additional Information DOCUMENTATION	TECHNICAL <ul style="list-style-type: none"> Corpus subclass Personal Data Sensitive Data Anonymized (*) 	MEDIA PART <ul style="list-style-type: none"> Corpus Part Linguality type (text, audio, video, image) Multilinguality type (text, audio, video) Language (text, audio, video, image) Type of content (video, image, textNumerical) 	TECHNICAL <ul style="list-style-type: none"> Dataset Distribution Dataset Distribution Form Distribution Location (*) Download Location (*) Access Location (*) Distribution Medium Features (*) Data Format Size Licence Terms 	DATA
RELATED LRTs				

Beyond the *common mandatory elements*, for a corpus to be described information is also needed on the **corpus subclass** (if it is raw or annotated, for example) and whether **personal or sensitive** data are included. If this is the case, you must say whether they have been **anonymized**. Your corpus could also have various **media parts** (namely: text, video, audio, image or textNumerical parts). Each of these parts must be described separately. For instance, if you have a text corpus of transcribed discussions along with the audio files you must provide information on both media indicating the **size** and **data format** of each one.

Tip: See [here](#) examples of XML metadata descriptions for corpora.

22.2.2 A Tool at a glance

LANGUAGE RESOURCE/ TECHNOLOGY	TOOL SERVICE	DISTRIBUTION	DATA
IDENTITY <ul style="list-style-type: none"> Resource Name Description Version CATEGORIES <ul style="list-style-type: none"> Keyword CONTACT <ul style="list-style-type: none"> Additional Information DOCUMENTATION	CATEGORIES <ul style="list-style-type: none"> Function TECHNICAL <ul style="list-style-type: none"> Language Dependent Input Content Resource Processing Resource Type Language (*) EVALUATION <ul style="list-style-type: none"> Evaluated 	TECHNICAL <ul style="list-style-type: none"> Software Distribution Software Distribution Form Web Service Type (*) (Download location, Access location, Execution location) Licence Terms 	DATA
RELATED LRTs			

Beyond the *common mandatory elements*, for a tool to be described you have to indicate its **function** (i.e. the task it performs) and also to choose if it is **language dependent** or not. If this is the case, you must also select one or more **languages** the tool can handle. The **resource type** the tool takes as **input** (e.g. corpus) must also be defined.

Tip: See [here](#) examples of XML metadata descriptions for tools.

22.2.3 A Lexical/Conceptual Resource (LCR) at a glance

LANGUAGE RESOURCE/ TECHNOLOGY	CORPUS	PART	DISTRIBUTION	DATA
IDENTITY <ul style="list-style-type: none"> Resource Name Description Version CATEGORIES <ul style="list-style-type: none"> Keyword CONTACT <ul style="list-style-type: none"> Additional Information DOCUMENTATION	TECHNICAL <ul style="list-style-type: none"> Encoding Level Personal Data Sensitive Data Anonymized (*) 	MEDIA PART <ul style="list-style-type: none"> Lexical/Conceptual Resource Part Linguality type (text, audio, video, image) Language (text, audio, video, image) Type of content (video, image) 	TECHNICAL <ul style="list-style-type: none"> Dataset Distribution Dataset Distribution Form Distribution Location (*) Download Location (*) Access Location (*) Distribution Medium Features (*) Data Format Size Licence Terms 	DATA
RELATED LRTs				

Beyond the *common mandatory elements*, for a Lexical/conceptual (LCR) resource to be described, information is also needed on the **encoding level** (the linguistic level of analysis of the LCR, e.g. morphology, phonology, semantics, etc.) and whether **personal or sensitive** data are included. If this is the case, you must say whether they have been **anonymized**. Your LCR could also have various **media parts** (namely: text, video, audio or image parts). Each of these parts must be described separately. For instance, if you have a digital lexicon with definitions in text and audio files for the pronunciation you must provide information on both media indicating the **size** and **data format** of each one.

Tip: See [here](#) examples of XML metadata descriptions for lexical/conceptual resources.

22.2.4 A Language Description at a glance

LANGUAGE RESOURCE/ TECHNOLOGY	CORPUS	PART	DISTRIBUTION	DATA
IDENTITY <ul style="list-style-type: none"> Resource Name Description Version CATEGORIES <ul style="list-style-type: none"> Keyword CONTACT <ul style="list-style-type: none"> Additional Information DOCUMENTATION	TECHNICAL <ul style="list-style-type: none"> Language Description Subclass ML Model (*) Grammar (*) Ngram Model (*) Encoding Level Personal Data Sensitive Data Anonymized (*) 	MEDIA PART <ul style="list-style-type: none"> Lexical/Conceptual Resource Part Linguality type (text, audio, video, image) Language (text, audio, video, image) Type of content (video, image) 	TECHNICAL <ul style="list-style-type: none"> Dataset Distribution Dataset Distribution Form Distribution Location (*) Download Location (*) Access Location (*) Distribution Medium Features (*) Data Format Size Licence Terms 	DATA
RELATED LRTs				

Beyond the *common mandatory elements*, for a Language Description to be described information is also needed on the **Language Description subclass** (whether it is a **grammar**, an **ML model** or a **n-gram model**) and whether **personal or sensitive** data are included. If this is the case, you must say whether they have been **anonymized**. Your Language Description could also have various **media parts** (namely: text, video, audio or image). Each of these parts must be described separately. For instance, if you have a text grammar along with the video files showcasing examples you must provide information on both media indicating the **size** and **data format** of each one.

Tip: See [here](#) examples of XML metadata descriptions for language descriptions.

Attention: Although only the mandatory metadata are the required in order to create a record, an LRT description is more complete if the **recommended** metadata are provided as well.

GENERAL GUIDELINES ON METADATA

23.1 Language

All metadata elements are presented and must be filled out, primarily, in *English*. Nonetheless, you can provide information in **any other language**. To do so, click on the **add symbol** **[+]** next to the metadata you wish to describe in a different language (e.g. LRT name).

The screenshot shows a form with two main sections. The first section has a text input field labeled 'LRT name *' containing 'Demo corpus', with a small text below it: 'The official name or title of the language resource/technology'. To its right is a language selection dropdown menu currently set to 'English', with a small text below it: 'select language'. To the right of the dropdown is a button with a '+' icon and the text 'Add in another language' below it. This button is highlighted with a yellow rectangle.

The metadata field is duplicated and the new language in which you can provide information is, by default, *Modern Greek*.

The screenshot shows the same form as before, but now there are two identical sections. The first section is the same as before. The second section has a text input field labeled 'LRT name *' (empty), with a small text below it: 'The official name or title of the language resource/technology'. To its right is a language selection dropdown menu currently set to 'Modern Greek (1453-)', with a small text below it: 'select language'. To the right of the dropdown is a button with an 'x' icon.

Attention: If this is not the case, **select another language** from the drop down list.

The screenshot shows the form with the first section. The language selection dropdown menu is open, showing a list of languages: English, Modern Greek (1453-), Bulgarian, Croatian, Czech, Danish, and Dutch; Flemish. The 'Bulgarian' option is highlighted. To the right of the dropdown is a button with an 'x' icon and the text 'Remove' below it. This button is highlighted with a yellow rectangle.

If you decide to **remove the added field**, click on the **remove symbol** **[x]**. This action will remove the whole field, i.e. LRT name, not just the language.

23.2 Consistency

The metadata used to describe your data should **clearly reflect them**. Make sure there are no inconsistencies (e.g. check that your files are indeed in PDF format and not just scanned images; if you provide information on an annotated corpus, indicate the annotation tool etc.). Check first which are the *mandatory* metadata per resource type and then see *how to fill them in*.

23.3 Completeness

Make sure you provide all the necessary information in all selected languages. It is highly recommended that you provide the title and the description in **English** and **Greek** (along with any other language, if needed) as the search and the resource retrieval is facilitated.

23.4 Editing

Editing metadata is necessary when there are typos, inconsistencies, missing information etc. If you are a *curator*, you can *edit* the resource metadata as many times as you want before submitting the record for publication. Once submitted, editing is permitted only to the *supervisor* (see *here* the differences in actions per role type).

Attention: Keep in mind that **editing** concerns only **minor changes** in the metadata of the resource in question. Editing takes place during the creation of a resource until it is submitted for publication. It differs from versioning in that the **changes do not constitute a new entity**. For example, if there has been a *mistake in the resource size*, **editing** is required while *if the resource size has changed* (as is the case with glossaries which are augmented) a **new version** is needed.

23.5 Versioning

If there are significant changes to the resource which differentiate it from the existing entity described, then a new version must be created. Such changes are *new editions of glossaries with more terms*, *new improved cleaned versions of corpora*, *new data (e.g. parallel sentences) with which a bilingual corpus has been enriched* or *tool/software updates*. Both the *curator* and the *supervisor* can *create a new version* after the resource has been published.

Attention: Keep in mind that only the **latest version** is found in the central inventory, while *all versions* are accessible from the **resource view page**, as shown in the image below.

Overview

Technical

Relations

Access

Information for Lexical/ Conceptual resource part

TEXT

Language info

Linguality type
bilingual

Multilinguality type
parallel

Language
Modern Greek (1453-), English

Export

XML

All versions

Terms on Fora! – English-Greek Glossary of Terms
(20.0.0)
<http://hdl.handle.net/11500/CLARIN-EL-0000-0000-6A02-7>
(Handle)

Terms on Fora! – English-Greek Glossary of Terms (16)
<http://hdl.handle.net/11500/CLARIN-0000-0000-5CF4-6>
(Handle)

SPECIFIC GUIDELINES ON MANDATORY METADATA

This section provides guidance on how to fill in specific metadata which are mandatory for a *corpus*, a *lexical/conceptual resource*, a *tool* and a *language description*. Since some of the metadata elements are *common* for all resources, they are presented first followed by metadata which are resource type specific.

Each metadata element is briefly explained and examples are provided whenever possible. The examples cover both best practices as well as common mistakes which must be avoided (marked with an asterisk *). In addition for each metadata element there is a link to the *XSD* with its full representation.

24.1 1. resourceName

~The official name or title of the language resource/technology~

The name must reflect the content (and the type) of the resource; it must present all the necessary information for the resource but it should not be too descriptive; detailed information must be provided in the description. Do not use full phrases, punctuation marks (unless necessary) or abbreviations in the resource title. Provide the full name of the resource and use the short name (if any) in the respective metadata field.

Examples

Do: Glossary of medical terms; Old and New Testament; Ellogon annotation tool

Don't: *This is a glossary of medical terms; *Old and New Testament!; *Ellogon ann. tool

-
- See how the `resourceName` element is described in detail in [XSD](#)
-

24.2 2. description

~A short presentation of the language resource/technology~

The description must contain all the important information about the resource. Don't simply repeat (or rephrase) the resource title without adding any other information. Once read and without seeing the rest of the metadata, one should be able to understand what it is about. Define the type of the resource and provide any useful information on how, when and by whom it was created, what is its language and size and what is the purpose it serves, if any. Mention any particularities or limitations about the data or the tool that users should be aware of. The description must be a free text of minimum one paragraph. You can also make use of the functionalities (formatting, hyperlinking, bullets etc.) of the metadata editor¹ to make the description easy to read.

¹ Henceforth **editor**.

Examples

Do: 1) Bilingual glossary (German / Greek) made in 2019/2020 by students of DFLTI (Ionian University) under the supervision of Mr. Olaf Immanuel Seel in the framework of the department's cooperation with the EU TermCord.

- 2) Texts corpus from the transcription of recorded children's speech focused on narration. The corpus was collected from interviews conducted by undergraduate and postgraduate students of the Department of Mediterranean Studies of the University of the Aegean with children with whom they are related either by friendship or kinship. Files with both the questions and answers are provided, where K=girl and A=boy, as well as cleaned files containing only the children's answers (clean).

Don't: *Symposium Proceedings; *Bilingual lexicon on the Greek economy

- See how the `description` element is described in detail in [XSD](#)

24.3 3. version

~A particular form of a resource differing in certain respects from an earlier form~

The recommended format for a version is: `major_version.minor_version.patch2`.

Examples

Do: 1.0.0-alpha; 2.1.1

Don't: *1.0.1-alpha; *0.0.2

The infrastructure automatically assigns the **1.0.0** version to all resources. If this is not the case with your resource, write the version number in the box (e.g. 2.0.0) and then click on the version date to reveal the calendar. Select the date when this version was released and click on OK.

The screenshot shows a metadata editor interface. The 'Version' field contains '2.0.0'. Below it, the 'Version date' field contains '2022-02-01'. A calendar pop-up is open, showing the month of February 2022. The date '1' is selected. The background shows other metadata fields like 'LRT provider', 'Actor type', and 'Source of metadata record'.

The editor also provides the possibility to automatically *create a new version* of an existing resource. See the [guidelines on versioning](#) before you proceed to do so.

² See the [semantic versioning guidelines](#) for specific instructions.

- See how the `version` element is described in detail in [XSD](#)

24.4 4. keyword

~A word or phrase characteristic of the language resource/technology that can be used at search~

Keywords are words or small phrases used to search for a resource. The more keywords used, the merrier for the resource retrieval. However, the keywords must highlight resource aspects not already covered by **mandatory** metadata. If, for example, you describe a *monolingual annotated corpus created to enhance the learning process of non native speakers*, your keywords must not be **exclusively** or **primarily** the following: “corpus”, “annotated” or “monolingual”; these are the values of the `resourceType`, `corpusSubclass` and `linguality` metadata elements respectively which are also searched and retrieved. Instead use as keywords the phrases “non native speaker” and “learning process” which emphasize the resource intended use; in addition you can add “corpus”, “annotated” and “monolingual”.

Examples

Do: non native speaker; learning process (corpus; annotated; monolingual)

Don't: *corpus; *annotated; *monolingual

After you have typed in the keyword you want, **click on the prompt** that appears under the box: **Add** “non native speaker”. Only then the value will be saved. If you omit this step, the keyword **will not be appear** when you revisit this editor section.

The screenshot shows a web form for adding keywords. At the top, it says 'Keyword' and 'A word or phrase characteristic of the language resource/technology that can be used at search (multiple values possible)'. Below this is a text input field containing 'non native speaker' and a dropdown menu set to 'English'. To the right of the dropdown is a 'Remove' button and a '+' icon. At the bottom of the form, a message says 'Missing non native speaker? Add "non native speaker"' with a yellow arrow pointing to it.

- See how the `keyword` element is described in detail in [XSD](#)

24.5 5. additionalInformation

~A URL (landing page) or email (e.g., support email) where the user can find or ask for more information~

This metadata element is either a web page with additional information on the language resource/technology (e.g., its contents, link to the access location, etc.) or the email of person responsible to provide information. Make sure to enter a valid email or URL.

Examples

Do: person@athenarc.gr; <http://www.clarin.gr>

Don't: *person@athenarc.g; <http://clarin.gr>

- See how the `additionalInformation` element is described in detail in [XSD](#)

24.6 6. distribution related metadata

~The form (or forms) in which a resource is available~

A resource might be available in more than one ways, in compact form (such as a CD-ROM, a DVD-R, a hard disk, etc.) or through an access point. If there are more than one distributions for a resource, each one must be **independently described**. A dropdown list offers a variety of forms to choose from.

Dataset distribution 1
Describe separately each distributable form of the corpus (e.g., downloadable form in CSV, XML formats, form accessible via an i/f)

Dataset distribution form *

- CD-ROM
- DVD-R
- accessible through interface
- accessible through query
- bluRay
- downloadable
- hard disk
- other
- unspecified

Once a value is selected, it generates its respective metadata elements, which must also be filled in. If, for example, a resource is *accessible through interface*, the `access location` metadata field is generated and you must fill in the URL via which the resource is accessible.

Dataset distribution 1
Describe separately each distributable form of the corpus (e.g., downloadable form in CSV, XML formats, form accessible via an i/f)

Dataset distribution form *

accessible through interface

Select the form or delivery channel through which the corpus is distributed

Private

☐ Yes

☐ No

Specifies whether the resource is private so that its access/download location remains hidden

Access location *

The URL where the corpus distribution can be accessed from (e.g. a URL with the access i/f, or a landing page where the user needs to follow some links, provide some information and check some boxes before accessing the corpus)

Attention: The CLARIN:EL infrastructure mainly hosts resources along with their data. Independently of whether the data have been *uploaded* upon the resource creation or at a later stage, they must be associated with a distribution.

The appropriate distribution has the value *downloadable* and although it generates the `download_location` metadata field, this does not need to be filled in (since the data are downloaded from the CLARIN:EL infrastructure). What must be done is to create the *association* between the distribution and the data, as shown in the image below. Click on the zip file name to provide the association. Finally, for the process to be completed, you must **save** (or **save as draft**) the metadata record.

Dataset distribution 1

Describe separately each distributable form of the corpus (e.g., downloadable form in CSV, XML formats, form accessible via an i/f)

Dataset distribution form *
downloadable

Select the form or delivery channel through which the corpus is distributed

Private
☐ Yes
☒ No

Specifies whether the resource is private so that its access/download location remains hidden

Associate a dataset with this distribution
 Parla_12_1.zip

- See how the `distribution` element is described in detail in [XSD](#)

24.7 7. licenceTerms related metadata

~The terms under which a resource is made available~

The `licenceTerms` related metadata consist of the name of the `licence`, the `licence terms` and the most frequently used `conditions of use`. The `licence` name is revealed once you start typing in the respective field. If it has already been used by another user, you will be presented with its full official name. If you click on it, the related metadata will be automatically filled in. For example, if the licence in question is the **cc-by-nc-sa**, then when you start typing, the matching options will be presented as shown in the image below.

Licence

Start typing to select the licence of the corpus distribution, or add a new value

Licence name *
cc-by-nc-sa

Missing **cc-by-nc-sa**? Add "cc-by-nc-sa"

Creative Commons Attribution Non Commercial Share Alike 3.0 Unported
<https://creativecommons.org/licenses/by-nc-sa/3.0/legalcode>
 CC-BY-NC-SA-3.0 - SPDX

Creative Commons Attribution Non Commercial Share Alike 4.0 International
<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>
 CC-BY-NC-SA-4.0 - SPDX

Click on the licence name you want to use from the suggested values. For this specific licence, the full name is **Creative Commons Attribution Non Commercial Share Alike 4.0 International**, the URL where the licence terms can be found is <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>, and the conditions of use are **attribution**, **non commercial use**, **share-alike**, all automatically provided in the respective metadata fields as shown in the image below.

Licence
Start typing to select the licence of the corpus distribution, or add a new value

Licence name *
Creative Commons Attribution Non Commercial Share Alike 4.0 International

language
English

Licence/terms of use/service URL *
<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>

Licence/terms of use/service URL *
<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Condition of use *
attribution non-commercial use share alike

Select the condition(s) of use for accessing the corpus; the list includes the most frequently used conditions and aims to provide brief human readable information, while the proper exposition of all conditions and possible exceptions is found inside the licence text

If the licence you wish to use has never been applied before, you will have to fill in manually the aforementioned metadata. See also the [Recommended licensing scheme for Language Resources](#) if you need help with which licence to choose for your resources.

- See how the `LicenceTerms` element is described in detail in [XSD](#)

24.8 8. data

~The content files of a resource~

Not all resources have content files. A metadata description may or may not be accompanied by content files (see [here](#) for more information). See also the detailed guidelines on how to [prepare data](#), the [recommended formats](#) and how to [upload](#) them.

24.9 9. personalData, sensitiveData & anonymized

~Information about whether the resource contains personal and/or sensitive data~

Attention: This metadata element is mandatory for **corpora**, **lexical/conceptual resources** and **language descriptions**.

You must specify whether the resource contains personal data (e.g. names) and/or sensitive data (e.g., medical/health-related, etc.) and thus requires special handling. If this is the case, new metadata fields are presented in which you can provide additional information on special requirements, if necessary.

Personal data included *
☒ Yes
☐ No
Specify if personal data are included

Personal data details

language
English
+
select language

Provide additional information on special requirements, if needed

Sensitive data included *
☒ Yes
☐ No
Specify whether the language resource contains sensitive data (e.g., medical/health-related, etc.) and thus requires special handling

Sensitive data details

language
English
+
select language

- See how the `personalData` element is described in detail in [XSD](#)
- See how the `sensitiveData` element is described in detail in [XSD](#)

The existence of personal and/or sensitive data generates³ another metadata element, that of **anonymization**. Here you can provide all the information on the anonymization/pseudo-anonymization, the tool used, if specific code was written, any conventions adopted, etc.

Anonymized *
☒ Yes
☐ No
Specify if the corpus has been anonymized

Anonymization details

language
English
+
select language

More information on the anonymization

- See how the `anonymized` element is described in detail in [XSD](#)

³ The `anonymized` element belongs to the **mandatory upon condition** metadata, the necessity of which depends on the values of other elements provided by the user, such as the answer “yes” to the question about the personal and/or sensitive data existence in a resource.

24.10 10. Subclass related metadata

~The classes into which a language resource can be further categorized according to its type~

Attention: This metadata element is mandatory for **corpora**, **lexical/conceptual resources** and **language descriptions**.

24.10.1 10.1 corpusSubclass

For **corpora** the corpusSubclass categories are:

- **raw**, for *non-processed* corpora,
- **annotated**, for corpora that include *both the raw corpus and the processed output*,
- **annotations**, for corpora that consist *only of the processed output*, and
- **unspecified**, for corpora that *cannot be described from one of the aforementioned categories*.

The metadata element is found in the *Corpus* section (*Technical* tab) in the editor.

- See how the corpusSubclass element is described in detail in [XSD](#)

24.10.2 10.2 lcrSubclass

A **lexical/conceptual resource** can be further categorized with the lcrSubclass element as:

- **annotation scheme**: A set of elements and values designed to annotate data. It usually consists in a formal representation. It aims to represent a specific level of information, such as morphological features of words, syntactic dependency relations between phrases, discourse level information etc. It can consist of a flat structure of elements and values (e.g. part-of-speech tags) or it can be more complex with interrelated elements (e.g. specific morphological features to be used for each part-of-speech).⁴
- **computational lexicon**: a lexicon which is intended for computational purposes and thus contains words associated with information relevant for the specific purposes.
- **dictionary**: a book or electronic resource that contains a list of words (usually in alphabetical order) and explains their meanings, or gives a word for them in another language and other information (e.g., spelling, pronunciation, etc.).
- **FrameNet**: a lexical database based on annotating examples of how words are used in actual texts in accordance to the notion of ‘semantic frame’ (schematic representation of a situation involving various participants, props and other conceptual roles); originally built for English and extended to other languages according to the same design principles.

⁴ The difference between **typesystem** and **annotation scheme** is based on whether they are used by tools or defined by users: the **annotation scheme** contains **custom types** while the typesystem is mostly used for built-in types.

- **lexicon**: (a list of) all the words used in a particular language or subject, or a dictionary.
- **Machine Readable Dictionary**: a dictionary usually meant for humans in a form that a computer can process.
- **mapping of resources**: a resource consisting of mapping values and/or rules between two resources.
- **morphological lexicon**: a lexicon with morphological information associated with its entries.
- **ontology**: a set of concepts and categories in a subject area or domain that shows their properties and the relations between them.
- **other**: value used when none of the recommended values of an element is appropriate for an item.
- **tagset**: a flat list of valid values (tags) designed to annotate data. It usually corresponds to a specific annotation type or set of annotation types.⁵
- **terminological resource**: a lexical resource that lists concepts pertaining to a specific domain.
- **thesaurus**: a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms).
- **typesystem**: a set of elements designed to annotate data. It typically contains only a list of annotation types, i.e. specific labels that are used for the annotation (e.g. part-of-speech, person, organization, etc.), and is usually inbuilt in the annotation software.⁶
- **unspecified**: value used for mandatory elements whose value is unknown or cannot be specified.
- **WordNet**: a lexical database originally created for English and extended to other languages, which groups words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.
- **wordlist**: a written collection of all words derived from a particular source, or sharing some other characteristic.

The `lcrSubclass` categories are also alphabetically presented as a dropdown list in the editor *LCR* section (*Technical* tab).

LANGUAGE RESOURCE/TECHNOLOGY

LCR PART DISTRIBUTION DATA

☐ For information ☐ Metaresource

Save draft Save

TECHNICAL

Encoding level *

Select the linguistic level of analysis the lexical/conceptual resource caters for

LCR subclass

annotation scheme

computational lexicon

dictionary

FrameNet

- See how the `lcrSubclass` element is described in detail in [XSD](#)

⁵ The difference between a **typesystem** and a **tagset** is that the **typesystem** will include only annotation types (e.g. an annotation type POS to represent part-of-speech annotations) while the **tagset** contains a list of the valid tag values (e.g. the Penn Treebank Tagset).

⁶ The difference between **typesystem** and **annotation scheme** is based on whether they are used by tools or defined by users: the **annotation scheme** contains **custom types** while the **typesystem** is mostly used for built-in types.

24.10.3 10.3 LanguageDescriptionSubclass

A **language description** has three categories from which one can choose to describe in a more fined way a resource:

- **grammar**: a set of rules governing what strings are valid or allowable in a language or text.
- **ML model**: the ML model that must be used together with the tool/service to perform the desired task.
- **n-gram model**: a language model consisting of n-grams, i.e., specific sequences of a number of words.

These categories are presented in the editor *Language Description* section (*Technical* tab) as a dropdown list.

- See how the LanguageDescriptionSubclass element is described in detail in [XSD](#)

24.11 11. encodingLevel

~Information on the contents of a resource as regards the linguistic level of analysis it caters for~

Attention: This metadata element is mandatory for **lexical/conceptual resources** and **language descriptions**.

The values for encoding refer to various linguistic levels of analysis. These levels are presented in alphabetical order below with their subject matters:

- **morphology**: word formation (such as inflection, derivation and compounding);
- **other**: value used when none of the recommended values of an element is appropriate for an item;
- **phonetics**: speech sounds;
- **phonology**: speech sounds that constitute the fundamental components of a language;
- **pragmatics**: the relationship of sentences to the environment in which they occur;
- **semantics**: the meaning of a word, phrase, etc.;
- **syntax**: the structure of linguistic units (phrases, sentences);
- **unspecified**: value used for mandatory elements whose value is unknown or cannot be specified.

The metadata field is found in the *LRC* section (*Technical* tab) for lexical/conceptual resources above the `lcrSubclass` as shown in the image below.

LANGUAGE RESOURCE/TECHNOLOGY

LCR PART DISTRIBUTION DATA

☐ For information
☐ Metaresource

Save draft Save

TECHNICAL

Encoding level *

morphology
other

For language descriptions the metadata field is found in the *Language Description* section (*Technical* tab) below the chosen *LanguageDescriptionSubclass*.

LANGUAGE RESOURCE/TECHNOLOGY

LANGUAGE DESCRIPTION PART DISTRIBUTION DATA

☐ For information
☐ Metaresource

Save draft Save

TECHNICAL

Language description type
The type of the language description (used for documentation purposes)

grammar Remove grammar

Encoding level *

morphology
other

- See how the `encodingLevel` element is described in detail in [XSD](#)

24.12 12. function

~The operation/function/task that a software object performs~

Attention: This metadata element is mandatory for **tools/services** only.

The dropdown list in the respective metadata field includes numerous values which cannot be presented all here. If you start typing, though, the list will be reduced only to the values matching your criteria. If the function of your tool/service matches one of the values suggested, **click on it** and it will be added. If the function of your tool/service **does not** match one of the values suggested, you must **click on the prompt** (*missing... ? add*). Only then the value will be saved. If you omit this step, the function **will not be appear** when you revisit this editor section.

LANGUAGE RESOURCE/TECHNOLOGY

TOOL/SERVICE DISTRIBUTION DATA

☐ For information
☐ CLARIN-EL compatible service

Save draft Save

CATEGORIES

TECHNICAL

EVALUATION

Function
Function *

event

Missing event? Add "event"

Event detection

Event labelling

Sound event annotation

The metadata element is found in the editor *Tool/Service* section (*Categories* tab).

- See how the `function` element is described in detail in [XSD](#)

24.13 13. inputContentResource

~The requirements set by a tool/service for the (content) resource that it processes~

Attention: This metadata element is mandatory for **tools/services** only.

This is a complex metadata element which requires for four other metadata fields to be described: **input resource type**, **media type**, **data format** and **annotation type**. All these elements provide the necessary information on the resource that a tool/service processes.

The screenshot shows the 'TOOL/SERVICE' tab in the CLARIN metadata form. The left sidebar has 'EVALUATION' checked. The main area is titled 'Input content resource' and contains four dropdown menus: 'Input resource type *', 'Media type', 'Data format', and 'Annotation type'. Each dropdown has a descriptive text below it. The 'Input resource type *' dropdown is highlighted with a red border. At the top right, there are checkboxes for 'For information' and 'CLARIN:EL compatible service', and buttons for 'Save draft' and 'Save'.

For the resource used as input, a dropdown list provides the values shown in the following image. To choose one, click on the value.

The screenshot shows the 'Input resource type *' dropdown menu. The dropdown is open, showing a list of values: 'corpus', 'file', 'language description', 'lexical/conceptual resource', 'other', 'output text', 'unspecified', and 'user input text'. The 'corpus' value is selected and highlighted in blue. A 'Remove Input content resource' button is visible in the top right corner of the form area.

The next field to be filled in, requires information on the medium of the resource used as input. Again, click on a value to add it.

Media type

- audio
- image
- text
- numerical text
- video

For the data format following, you must type in the box to reveal the values that match your criteria and eliminate all the others from the dropdown list. Once you have located the appropriate value, click on it.

Data format

mp

Wikipedia template filtered article

BMP

audio **mp3**

MPEG-4

audio **mpg**

Finally, if the resource provided as input is annotated, you must define the annotation type. Once more, start typing in the box to reveal the possible corresponding values. Choose one by clicking on it.

Annotation type

mov

Body **movement**

Gaze eye **movement**

Head **movement**

Lip **movement**

The `inputContentResource` element is found in the editor *Tool/Service* section (*Technical* tab).

LANGUAGE RESOURCE/TECHNOLOGY


TOOL/SERVICE


DISTRIBUTION


DATA

☐ For information

☐ CLARIN-EL compatible service


 Save draft

 Save

 CATEGORIES

Specify if the tool is language dependent *

☐ Language dependent

 **TECHNICAL**

Input content resource

Describe the requirements that a data resource must fulfill in order to be processed

- See how the `inputContentResource` element is described in detail in [XSD](#)

EXAMPLES OF METADATA

The goal of this section is to familiarize users with the use of **metadata**. To do so, resource descriptions have been exported¹ from the [CLARIN:EL infrastructure](#) and excerpts of interest have been copied verbatim. Each metadata element is presented per se and then briefly explained. There are also links to the full XML description of the resource for anyone wishing to see the metadata examined in context and the [XSD](#), for a detailed representation of the element.

25.1 resourceName

The first metadata element is the `resourceName` from the [Greek Parliament Plenary Sessions \(1989-2019\)](#), a collection of the raw minutes of the Greek Parliament plenary sessions of the last 30 years (more than 1.000.000 speeches).

XML

```
<ms:resourceName xml:lang="en">Greek Parliament Plenary Sessions (1989-2019)</  
↪ms:resourceName>  
<ms:resourceName xml:lang="el"> (1989-2019)</ms:resourceName>
```

As shown, it is possible to provide the name in more than one languages; the first language, by default, is english (xml:lang="en") while the second is free of choice. Here the language chosen is greek (xml:lang="el").

- [See the resource full XML description](#)
- [See how the element is described in detail in XSD](#)

25.2 resourceCreator

The second excerpt is taken from the [KELLY word-list](#), a monolingual lexical conceptual resource. KELLY word-lists were created to facilitate the learning of a foreign/second language. The Greek part was created by the *Institute for Language and Speech Processing* which is an **organization**.

XML

¹ These tags come in pairs; the opening and ending tags are identical except for the **forward slash**.

```
<ms:resourceCreator>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="el">    </ms:organizationName>
    <ms:organizationName xml:lang="en">Institute for Language and Speech
↳ Processing</ms:organizationName>
    <ms:website>http://www.ilsp.gr</ms:website>
  </ms:Organization>
</ms:resourceCreator>
```

The necessary information about the creator is enclosed between the `resourceCreator` tags. First, the type of the creator (`actorType`) is defined; a resource could have as creator a *person*, a *group* of people or an *organization*, as is the case for the Kelly world-list. Then the name of the organization is provided (in two languages, `xml:lang="el"` and `xml:lang="en"`) as well as its website.

- [See the resource full XML description](#)
 - [See how the element is described in detail in XSD](#)
-

25.3 isPartOf

The next example is from the [Golden Part of Speech Tagged Corpus](#), a monolingual annotated corpus in Greek with 100.000 words. This corpus is a **subset** of the [Hellenic National Corpus](#) which contains more than 97 million words from a variety of sources and various domains. The subset relationship is expressed through the `isPartOf` metadata element in the *CLARIN:EL metadata schema*.

XML

```
<ms:isPartOf>
  <ms:resourceName xml:lang="el">    </ms:resourceName>
  <ms:resourceName xml:lang="en">Hellenic National Corpus</ms:resourceName>
  <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
    >http://hdl.handle.net/11500/ATHENA-0000-0000-23E2-9</ms:LRIdentifier>
  <ms:version>3.0</ms:version>
</ms:isPartOf>
```

The `isPartOf` element includes the name of the resource (`resourceName`) from which the Golden Part has been derived, i.e. the *Hellenic National Corpus*, expressed in two languages (`xml:lang="el"` and `xml:lang="en"`) along with its identifier (`LRIdentifier`) and version (`version`).

- [See the resource full XML description](#)
 - [See how the element is described in detail in XSD](#)
-

25.4 annotationType

Alignment is the process that establishes translational equivalences between structural units (words, sentences etc.) of a text in a given language and a text with similar meaning in other language(s). The [Greek-Bulgarian Bul-TM parallel corpus](#) is a *bilingual corpus* and as the adjective *parallel* suggests has been **aligned**.

XML

```
<ms:annotation>
  <ms:annotationType>http://w3id.org/meta-share/omtd-share/Alignment1</
↪ms:annotationType>
  <ms:segmentationLevel>http://w3id.org/meta-share/meta-share/sentence</
↪ms:segmentationLevel>
  <ms:annotationStandoff>>false</ms:annotationStandoff>
  <ms:annotationMode>http://w3id.org/meta-share/meta-share/automatic</
↪ms:annotationMode>
  <ms:isAnnotatedBy>
    <ms:resourceName xml:lang="en">TrAid</ms:resourceName>
    <ms:version>unspecified</ms:version>
  </ms:isAnnotatedBy>
</ms:annotation>
```

Alignment is considered a type of annotation. The two languages have been aligned at *sentence* level (*segmentationLevel*) and there is *not* a separate (*annotationStandoff*) document with each language independently. The procedure has been *automatically* done (*annotationMode*); the tool used for the alignment (*isAnnotatedBy*) is called *TrAid* but no *version* is available (*unspecified*).

- [See the resource full XML description](#)
- [See how the element is described in detail in XSD](#)

25.5 multilingualityType

The [DICTA-SIGN corpus](#) is a **multimedia** corpus, consisting of a video part and a text part, for four sign languages (english, french, german and greek).

XML

```
<ms:multilingualityType>http://w3id.org/meta-share/meta-share/parallel</
↪ms:multilingualityType>
  <ms:language>
    <ms:languageTag>gss</ms:languageTag>
    <ms:languageId>gss</ms:languageId>
  </ms:language>
  <ms:language>
    <ms:languageTag>bfi</ms:languageTag>
    <ms:languageId>bfi</ms:languageId>
  </ms:language>
  <ms:language>
```

(continues on next page)

(continued from previous page)

```
<ms:languageTag>gsg</ms:languageTag>
  <ms:languageId>gsg</ms:languageId>
</ms:language>
<ms:language>
  <ms:languageTag>fsl</ms:languageTag>
  <ms:languageId>fsl</ms:languageId>
</ms:language>
```

Each corpus part is described separately. This excerpt describes the content of the **video part** of the resource. The languages in the video are sign languages and are aligned as indicated by the choice of the value *parallel* for the *multilingualityType* element. Then each language (*language*) is presented separately with its language tag (*languageTag*) and id (*languageId*): gss (Greek Sign Language), bfi (British Sign Language), gsg (German Sign Language) and fsl (French Sign Language).

- [See the resource full XML description](#)
 - [See how the element is described in detail in XSD](#)
-

25.6 isDocumentedBy

Sometimes there is extra information about a resource in external documents such as papers and/or conference announcements. Such is the case with [Orossimo Terminological Resource - History](#) which is documented in the *Collection of digital terminological resources: methodology and results*.

XML

```
<ms:isDocumentedBy>
  <ms:title xml:lang="el"> : </ms:title>
  <ms:title xml:lang="en">Collection of digital terminological resources:
  methodology and results</ms:title>
</ms:isDocumentedBy>
```

- [See the resource full XML description](#)
 - [See how the element is described in detail in XSD](#)
-

25.7 fundingProject

The following example is more complex as it includes various metadata elements. It is taken from the [Trilingual Terminological Dictionary](#), a lexical/conceptual resource with a threefold aim: to assist the student in learning the subject areas of the curriculum, to improve their language skills in Greek and to familiarize themselves with information technology.

XML

```

<ms:fundingProject>
  <ms:projectName xml:lang="el"> </ms:projectName>
  <ms:projectName xml:lang="en">Trilingual Terminological Dictionary</
↪ms:projectName>
  <ms:website>https://bit.ly/2V4hWLe</ms:website>
  <ms:website>https://www.ilsp.gr/projects/tol/</ms:website>
  <ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds</ms:fundingType>
  <ms:fundingType>http://w3id.org/meta-share/meta-share/nationalFunds</
↪ms:fundingType>
  <ms:funder>
    <ms:Organization>
      <ms:actorType>Organization</ms:actorType>
      <ms:organizationName xml:lang="en">Ministry of Education and
↪Religious Affairs</ms:organizationName>
    </ms:Organization>
  </ms:funder>
  <ms:funder>
    <ms:Organization>
      <ms:actorType>Organization</ms:actorType>
      <ms:organizationName xml:lang="el"> </ms:organizationName>
      <ms:organizationName xml:lang="en">European Commission</
↪ms:organizationName>
      <ms:website>https://ec.europa.eu/info/index_en</ms:website>
    </ms:Organization>
  </ms:funder>
</ms:fundingProject>

```

The resource is the result of a project (`fundingProject`) bearing the same name (`projectName`), *Trilingual Terminological Dictionary*. The information provided for the project is the `websites` available, the `fundingType` and the `funders`. The project was created with *EU and national funds* while the funders were two organizations, the *Ministry of Education and Religious Affairs* and the *European Commission*.

- [See the resource full XML description](#)
- [See how the element is described in detail in XSD](#)

25.8 inputContentResource

The following XML excerpt provides information on the **input** of *Voyant Tools*, a web-based text reading and analysis environment.

XML

```

<ms:inputContentResource>
  <ms:processingResourceType>http://w3id.org/meta-share/meta-share/corpus</
↪ms:processingResourceType>
  <ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
  <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Pdf</ms:dataFormat>
  <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Rtf</ms:dataFormat>

```

(continues on next page)

(continued from previous page)

```

<ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</ms:dataFormat>
<ms:dataFormat>http://w3id.org/meta-share/omtd-share/Conllu</ms:dataFormat>
<ms:dataFormat>http://w3id.org/meta-share/omtd-share/Html</ms:dataFormat>
</ms:inputContentResource>

```

Voyant tools can process, take as input (inputContentResource), corpora (processingResourceType) of textual data the format (dataFormat) of which is *plain text*, *PDF*, *RTF*, *XML*, *Conllu* and *HTML*.

- [See the resource full XML description](#)
- [See how the element is described in detail in XSD](#)

25.9 outputResource

The next excerpt presents the output of the [ILSP Language Identification System](#).

XML

```

<ms:outputResource>
  <ms:processingResourceType>http://w3id.org/meta-share/meta-share/corpus</
↪ms:processingResourceType>
  <ms:language>
    <ms:languageTag>el-Latn</ms:languageTag>
    <ms:languageId>el</ms:languageId>
    <ms:scriptId>Latn</ms:scriptId>
    <ms:languageVarietyName xml:lang="en">Greeklish</ms:languageVarietyName>
  </ms:language>
  <ms:language>
    <ms:languageTag>el-Grek</ms:languageTag>
    <ms:languageId>el</ms:languageId>
    <ms:scriptId>Grek</ms:scriptId>
  </ms:language>
  <ms:language>
    <ms:languageTag>fr</ms:languageTag>
    <ms:languageId>fr</ms:languageId>
  </ms:language>
  <ms:language>
    <ms:languageTag>en</ms:languageTag>
    <ms:languageId>en</ms:languageId>
  </ms:language>
  <ms:language>
    <ms:languageTag>de</ms:languageTag>
    <ms:languageId>de</ms:languageId>
  </ms:language>
  <ms:language>
    <ms:languageTag>nl</ms:languageTag>
    <ms:languageId>nl</ms:languageId>
  </ms:language>

```

(continues on next page)

(continued from previous page)

```
<ms:mediaType>http://w3id.org/meta-share/meta-share/text</ms:mediaType>
</ms:outputResource>
```

This tool performs language identification for *Greeklish, Greek, English, German, Dutch and French*. Greeklish as seen in the excerpt above is a variety (languageVarietyName) of the Greek language: the language (languageId) is defined as *Greek* (el) but the script (scriptId) is *latin* (Latn).

- [See the resource full XML description](#)
- [See how the element is described in detail in XSD](#)

25.10 attributionText

The last example showcases the attributionText of a language description resource, the [PANACEA Environment Corpus n-grams EL](#).

XML

```
<ms:attributionText xml:lang="el">PANACEA n- (n-grams) . :
- . : Creative Commons Attribution Share Alike 4.0
International (https://creativecommons.org/licenses/by-sa/4.0/legalcode,
https://creativecommons.org/licenses/by-sa/4.0/). : http://hdl.handle.net/11500/ATHENA-
0000-0000-23DA-3
(CLARIN:EL)</ms:attributionText>
<ms:attributionText xml:lang="en">PANACEA Environment Corpus n-grams EL (Greek) by
Institute for Language and Speech
Processing - Athena Research Center used under Creative Commons Attribution Share Alike
4.0 International
(https://creativecommons.org/licenses/by-sa/4.0/legalcode, https://creativecommons.org/
licenses/by-sa/4.0/). Source:
http://hdl.handle.net/11500/ATHENA-0000-0000-23DA-3 (CLARIN:EL)</ms:attributionText>
```

The licence of the resource is the *CC-BY-SA 4.0 International*. “This license lets others remix, adapt, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms. This license is often compared to “copyleft” free and open source software licenses. All new works based on yours will carry the same license, so any derivatives will also allow commercial use.”² The attribution serves this exact purpose as it provides one with text containing the information on the resource creator, the *Institute for Language and Speech Processing - Athena Research Center* and the licence under which the resource and all its derivatives are to be distributed.

- [See the resource full XML description](#)
- [See how the element is described in detail in XSD](#)

² More information on the [Creative Commons website](#).

XML METADATA DESCRIPTIONS

The XML metadata descriptions presented here have been exported from the [CLARIN:EL infrastructure](#) and describe actual resources. Click on the button with the resource name to reveal its XML description. In all files the names, surnames and emails of creators, curators and contacts have been anonymized.

26.1 1. Corpora

26.1.1 Monolingual corpus #1

```
<?xml version="1.0" encoding="utf-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↪metadata-schema/CLARIN-SHARE.xsd">
<ms:metadataCreationDate>2019-12-17</ms:metadataCreationDate>
<ms:metadataLastDateUpdated>2021-11-24</ms:metadataLastDateUpdated>
<ms:metadataCurator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
<ms:metadataCreator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
  <ms:repositoryName xml:lang="el"> </ms:repositoryName>
  <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
  <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
  <ms:LanguageResource>
    <ms:entityType>LanguageResource</ms:entityType>
    <ms:resourceName xml:lang="el">
      (1989-2019)</ms:resourceName>
    <ms:resourceName xml:lang="en">Greek Parliament Plenary Sessions
      (1989-2019)</ms:resourceName>
```

(continues on next page)

(continued from previous page)

```

<ms:description xml:lang="el">
    30 (
    1.000.000 ).
    txt. ,
    40 Mb ,
    .</ms:description>
<ms:description xml:lang="en">This is a collection of the raw minutes of the
↪Greek
    Parliament plenary sessions of the last 30 years (more than 1.000.000
↪speeches). The
    existing corpus has all raw data in txt format. In order to make the
↪resource more
    processable, we have also split it into smaller subcorpora, with a maximum
    compressed folder size of 40 Mb per subcorpus. The created subcorpora are
    thematically organized per Greek parliamentary terms.</ms:description>
<ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
    >http://hdl.handle.net/11500/ATHENA-0000-0000-5D62-A</ms:LRIdentifier>
<ms:version>1.0.0 (automatically assigned)</ms:version>
<ms:additionalInfo>
    <ms:email>person@ilsp.gr</ms:email>
</ms:additionalInfo>
<ms:contact>
    <ms:Person>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:Person>
</ms:contact>
<ms:citationText xml:lang="el">
    (1989-2019) (2019). Version 1.0.0 (automatically assigned). [Dataset (Text
↪corpus)].
    CLARIN:EL. http://hdl.handle.net/11500/ATHENA-0000-0000-5D62-A</
↪ms:citationText>
    <ms:citationText xml:lang="en">Greek Parliament Plenary Sessions (1989-2019)
↪(2019).
    Version 1.0.0 (automatically assigned). [Dataset (Text corpus)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-5D62-A</ms:citationText>
<ms:keyword xml:lang="en">monolingual, political science</ms:keyword>
<ms:resourceProvider>
    <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="el">
            </ms:organizationName>
        <ms:organizationName xml:lang="en">Institute for Language and Speech
            Processing</ms:organizationName>
        <ms:website>http://www.ilsp.gr</ms:website>
    </ms:Organization>
</ms:resourceProvider>
<ms:LRSubclass>
    <ms:Corpus>
        <ms:lrType>Corpus</ms:lrType>
        <ms:corpusSubclass>http://w3id.org/meta-share/meta-share/rawCorpus</
↪ms:corpusSubclass>

```

(continues on next page)

(continued from previous page)

```

    <ms:CorpusMediaPart>
      <ms:CorpusTextPart>
        <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
        <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪monolingual</ms:lingualityType>
        <ms:language>
          <ms:languageTag>el</ms:languageTag>
          <ms:languageId>el</ms:languageId>
        </ms:language>
      </ms:CorpusTextPart>
    </ms:CorpusMediaPart>
    <ms:DatasetDistribution>
      <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪downloadable</ms:DatasetDistributionForm>
      <ms:downloadLocation>http://www.hiddenLocation.org</
↪ms:downloadLocation>
      <ms:distributionTextFeature>
        <ms:size>
          <ms:amount>5079.0</ms:amount>
          <ms:sizeUnit>http://w3id.org/meta-share/meta-share/file</
↪ms:sizeUnit>
        </ms:size>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Text</
↪ms:dataFormat>
        <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8
↪</ms:characterEncoding>
      </ms:distributionTextFeature>
      <ms:licenceTerms>
        <ms:licenceTermsName xml:lang="en">Creative Commons Attribution_
↪4.0
          International</ms:licenceTermsName>
        <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
↪legalcode</ms:licenceTermsURL>
        <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
↪</ms:licenceTermsURL>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪attribution</ms:conditionOfUse>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsDirectAccess</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsProcessing</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↪</ms:licenceCategory>
        <ms:LicenceIdentifier
          ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
↪share/SPDX"
          >CC-BY-4.0</ms:LicenceIdentifier>
      </ms:licenceTerms>
      <ms:attributionText xml:lang="el">
        (1989-2019). : Creative Commons Attribution 4.0

```

(continues on next page)

(continued from previous page)

```

International (https://creativecommons.org/licenses/by/4.0/
↪ legalcode,
    https://creativecommons.org/licenses/by/4.0/). :
    http://hdl.handle.net/11500/ATHENA-0000-0000-5D62-A
    (CLARIN:EL)</ms:attributionText>
    <ms:attributionText xml:lang="en">Greek Parliament Plenary Sessions
    (1989-2019) used under Creative Commons Attribution 4.0↪
↪ International
    (https://creativecommons.org/licenses/by/4.0/legalcode,
    https://creativecommons.org/licenses/by/4.0/). Source:
    http://hdl.handle.net/11500/ATHENA-0000-0000-5D62-A
    (CLARIN:EL)</ms:attributionText>
</ms:DatasetDistribution>
<ms:personalDataIncluded>>false</ms:personalDataIncluded>
<ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
</ms:Corpus>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

26.1.2 Monolingual corpus #2

```

<?xml version="1.0" encoding="utf-8"?>
  <ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.
  ↪ clarin.gr/metadata-schema/CLARIN-SHARE.xsd">
    <ms:metadataCreationDate>2021-02-25</ms:metadataCreationDate>
    <ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
    <ms:metadataCurator>
      <ms:actorType>Person</ms:actorType>
      <ms:surname xml:lang="en">Person_Surname</ms:surname>
      <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCurator>
    <ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</
  ↪ ms:compliesWith>
    <ms:metadataCreator>
      <ms:actorType>Person</ms:actorType>
      <ms:surname xml:lang="en">Person_Surname</ms:surname>
      <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCreator>
    <ms:sourceOfMetadataRecord>
      <ms:repositoryName xml:lang="el"> </ms:repositoryName>
      <ms:repositoryName xml:lang="en">ATHENA RC Repository</
  ↪ ms:repositoryName>
      <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
    </ms:sourceOfMetadataRecord>
    <ms:DescribedEntity>
      <ms:LanguageResource>

```

(continues on next page)

(continued from previous page)

```

<ms:entityType>LanguageResource</ms:entityType>
<ms:resourceName xml:lang="en">Golden Part of Speech Tagged Corpus</ms:resourceName>
<ms:description xml:lang="el"> Golden Part of Speech Tagged Corpus

    ,      100.000 .
      (web crawling),
:      (CC0 4.0)
    (CC BY 4.0) .

: -      (boilerplate material) -
    -
      / (ILSP Feature-based multi-tiered
POS Tagger),
    -
      , .
    Golden Corpus XML      XML,
.
      ,
      , .
      ,
      .
    XML
, .</ms:description>
<ms:description xml:lang="en">The Golden Part-of-Speech Tagged Corpus is a subset of
the
Hellenic National Corpus (HNC), the size of which is 100.000 words; it consists
of
selected texts from a variety of sources covering various domains. These texts
have
been crawled from the web and are licensed under either CC0 4.0 or CC BY 4.0. The
corpus underwent the following stages: • cleaning and removal of
boilerplate material, • manual correction of typos and spelling mistakes, •
automatic lemmatization and part-of-speech tagging for each word, using the ILSP
Feature-based multi-tiered POS Tagger, and • manual correction of the ILSP
Feature-based multi-tiered POS Tagger results. The GoldenPart of Speech Tagged
Corpus is available as a single XML file, containing all texts in the following
structure: first, some metadata about the text and then the text itself with
annotation at the level of paragraphs, sentences and words. Each word comes with
information on its lemma, POS and its boundaries (beginning and end). XML was
chosen
as the most appropriate format as it can be used in various environments,
regardless
of the operating system in use.</ms:description>
<ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
>http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C</ms:LRIdentifier>
<ms:version>1</ms:version>
<ms:additionalInfo>
  <ms:email>person@ilsp.athena-innovation.gr</ms:email>
</ms:additionalInfo>
<ms:contact>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>

```

(continues on next page)

(continued from previous page)

```

        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:Person>
</ms:contact>
<ms:citationText xml:lang="el">    -
    (2021). GoldenPart of Speech Tagged Corpus. Version 1. [Dataset (Text
    corpus)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C</ms:citationText>
<ms:citationText xml:lang="en">Institute for Language and Speech Processing - Athena
    Research Center (2021). GoldenPart of Speech Tagged Corpus. Version 1. [Dataset
    (Text corpus)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C</ms:citationText>
<ms:keyword xml:lang="en">monolingual</ms:keyword>
<ms:keyword xml:lang="en">morphosyntacticAnnotation-posTagging</ms:keyword>
<ms:resourceCreator>
    <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="el">
            </ms:organizationName>
        <ms:organizationName xml:lang="en">Institute for Language and Speech
            Processing</ms:organizationName>
        <ms:website>http://www.ilsp.gr</ms:website>
    </ms:Organization>
</ms:resourceCreator>
<ms:isPartOf>
    <ms:resourceName xml:lang="el">
        </ms:resourceName>
    <ms:resourceName xml:lang="en">Hellenic National Corpus</ms:resourceName>
    <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
        >http://hdl.handle.net/11500/ATHENA-0000-0000-23E2-9</ms:LRIdentifier>
    <ms:version>3.0</ms:version>
</ms:isPartOf>
<ms:LRSubclass>
    <ms:Corpus>
        <ms:lrType>Corpus</ms:lrType>
        <ms:corpusSubclass>http://w3id.org/meta-share/meta-share/annotatedCorpus</
    ms:corpusSubclass>
        <ms:CorpusMediaPart>
            <ms:CorpusTextPart>
                <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
                <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
    ms:mediaType>
                <ms:lingualityType>http://w3id.org/meta-share/meta-share/monolingual
    </ms:lingualityType>
                <ms:language>
                    <ms:languageTag>el</ms:languageTag>
                    <ms:languageId>el</ms:languageId>
                </ms:language>
                <ms:annotation>
                    <ms:annotationType>http://w3id.org/meta-share/omtd-share/
    ms:PartOfSpeech</ms:annotationType>
                    <ms:isAnnotatedBy>

```

(continues on next page)

(continued from previous page)

```

    <ms:resourceName xml:lang="en">ILSP Feature-based multi-
↪ tiered
        POS Tagger</ms:resourceName>
    <ms:LRIIdentifier
↪ handle"
        ms:LRIIdentifierScheme="http://purl.org/spar/datacite/
    <ms:LRIIdentifier>
        >http://hdl.handle.net/11500/ATHENA-0000-0000-23E8-3</
        <ms:version>1</ms:version>
        </ms:isAnnotatedBy>
    </ms:annotation>
    <ms:annotation>
        <ms:annotationType>http://w3id.org/meta-share/omtd-share/
↪ StructuralAnnotationType</ms:annotationType>
        <ms:segmentationLevel>http://w3id.org/meta-share/meta-share/
↪ sentence</ms:segmentationLevel>
        <ms:segmentationLevel>http://w3id.org/meta-share/meta-share/
↪ token1</ms:segmentationLevel>
        </ms:annotation>
        <ms:annotation>
            <ms:annotationType>http://w3id.org/meta-share/omtd-share/Lemma</
↪ ms:annotationType>
            </ms:annotation>
        </ms:CorpusTextPart>
    </ms:CorpusMediaPart>
    <ms:DatasetDistribution>
        <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪ downloadable</ms:DatasetDistributionForm>
        <ms:downloadLocation>http://www.hiddenLocation.org</ms:downloadLocation>
        <ms:accessLocation>http://fixme.com</ms:accessLocation>
        <ms:distributionTextFeature>
            <ms:size>
                <ms:amount>100000.0</ms:amount>
                <ms:sizeUnit>http://w3id.org/meta-share/meta-share/word3</
↪ ms:sizeUnit>
            </ms:size>
            <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</
↪ ms:dataFormat>
            <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8</
↪ ms:characterEncoding>
            </ms:distributionTextFeature>
            <ms:licenceTerms>
                <ms:licenceTermsName xml:lang="en">Creative Commons Attribution 4.0
                    International</ms:licenceTermsName>
                <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
↪ legalcode</ms:licenceTermsURL>
                <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/</
↪ ms:licenceTermsURL>
                <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/attribution
↪ </ms:conditionOfUse>
                <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪ allowsDirectAccess</ms:licenceCategory>

```

(continues on next page)

(continued from previous page)

```

        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsProcessing</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public</
↪ms:licenceCategory>
        <ms:LicenceIdentifier
            ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
↪share/SPDX"
            >CC-BY-4.0</ms:LicenceIdentifier>
    </ms:licenceTerms>
    <ms:attributionText xml:lang="el">Golden Part of Speech Tagged Corpus.
        :
        : Creative Commons Attribution 4.0 International
        (https://creativecommons.org/licenses/by/4.0/legalcode,
        https://creativecommons.org/licenses/by/4.0/). :
        http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C
        (CLARIN:EL)</ms:attributionText>
    <ms:attributionText xml:lang="en">GoldenPart of Speech Tagged Corpus by
        Institute for Language and Speech Processing - Athena Research Center
        used under Creative Commons Attribution 4.0 International
        (https://creativecommons.org/licenses/by/4.0/legalcode,
        https://creativecommons.org/licenses/by/4.0/). Source:
        http://hdl.handle.net/11500/ATHENA-0000-0000-5E7D-C
        (CLARIN:EL)</ms:attributionText>
    </ms:DatasetDistribution>
    <ms:personalDataIncluded>>false</ms:personalDataIncluded>
    <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
    </ms:Corpus>
    </ms:LRSubclass>
    </ms:LanguageResource>
    </ms:DescribedEntity>
</ms:MetadataRecord>

```

26.1.3 Bilingual corpus

```

<?xml version="1.0" encoding="utf-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↪metadata-schema/CLARIN-SHARE.xsd">
    <ms:metadataCreationDate>2015-09-11</ms:metadataCreationDate>
    <ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
    <ms:metadataCurator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCurator>
    <ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
    <ms:metadataCreator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>

```

(continues on next page)

(continued from previous page)

```

    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
    <ms:repositoryName xml:lang="el"> </ms:repositoryName>
    <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
    <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
    <ms:LanguageResource>
        <ms:entityType>LanguageResource</ms:entityType>
        <ms:resourceName xml:lang="el">- Bul-TM</ms:resourceName>
        <ms:resourceName xml:lang="en">Greek-Bulgarian Bul-TM parallel corpus</
↪ms:resourceName>
        <ms:resourceShortName xml:lang="en">Bul-TM</ms:resourceShortName>
        <ms:description xml:lang="el"> ( - )
            .
            TMX ( ).</ms:description>
        <ms:description xml:lang="en">Parallel bilingual corpus (web documents, general_
↪domain)
            aligned at sentence level; the corpus is available in TMX format.</
↪ms:description>
        <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
            >http://hdl.handle.net/11500/ATHENA-0000-0000-23E4-7</ms:LRIdentifier>
        <ms:version>1.0.0 (automatically assigned)</ms:version>
        <ms:additionalInfo>
            <ms:email>person@ilsp.gr</ms:email>
        </ms:additionalInfo>
        <ms:additionalInfo>
            <ms:email>person@ilsp.gr</ms:email>
        </ms:additionalInfo>
        <ms:contact>
            <ms:Person>
                <ms:actorType>Person</ms:actorType>
                <ms:surname xml:lang="en">Person_Surname</ms:surname>
                <ms:givenName xml:lang="en">Person_Name</ms:givenName>
            </ms:Person>
        </ms:contact>
        <ms:contact>
            <ms:Person>
                <ms:actorType>Person</ms:actorType>
                <ms:surname xml:lang="en">Person_Surname</ms:surname>
                <ms:givenName xml:lang="en">Person_Name</ms:givenName>
            </ms:Person>
        </ms:contact>
        <ms:citationText xml:lang="el"> -
            (2015). - Bul-TM. Version 1.0.0 (automatically
            assigned). [Dataset (Text corpus)]. CLARIN:EL.
            http://hdl.handle.net/11500/ATHENA-0000-0000-23E4-7</ms:citationText>
        <ms:citationText xml:lang="en">Institute for Language and Speech Processing -_
↪Athena
            Research Center (2015). Greek-Bulgarian Bul-TM parallel corpus. Version 1.0.0
            (automatically assigned). [Dataset (Text corpus)]. CLARIN:EL.

```

(continues on next page)

(continued from previous page)

```

    http://hdl.handle.net/11500/ATHENA-0000-0000-23E4-7</ms:citationText>
    <ms:keyword xml:lang="en">bilingual</ms:keyword>
    <ms:keyword xml:lang="en">parallel</ms:keyword>
    <ms:keyword xml:lang="en">alignment</ms:keyword>
    <ms:keyword xml:lang="en">writtenLanguage</ms:keyword>
    <ms:domain>
      <ms:categoryLabel xml:lang="en">Political Science</ms:categoryLabel>
      <ms:DomainIdentifier
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
        >DDC320</ms:DomainIdentifier>
    </ms:domain>
    <ms:domain>
      <ms:categoryLabel xml:lang="en">society</ms:categoryLabel>
    </ms:domain>
    <ms:resourceCreator>
      <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="el">
          </ms:organizationName>
        <ms:organizationName xml:lang="en">Institute for Language and Speech
          Processing</ms:organizationName>
        <ms:website>http://www.ilsp.gr</ms:website>
      </ms:Organization>
    </ms:resourceCreator>
    <ms:creationStartDate>2005-10-01</ms:creationStartDate>
    <ms:creationEndDate>2007-09-30</ms:creationEndDate>
    <ms:fundingProject>
      <ms:projectName xml:lang="en">Development of a Bulgarian to Greek and Greek_
↪to
      Bulgarian Translation Memory workbench</ms:projectName>
      <ms:website>https://www.ilsp.gr/projects/bultm/</ms:website>
      <ms:fundingType>http://w3id.org/meta-share/meta-share/nationalFunds</
↪ms:fundingType>
      <ms:funder>
        <ms:Organization>
          <ms:actorType>Organization</ms:actorType>
          <ms:organizationName xml:lang="en">Ministry of Economy and
            Finances</ms:organizationName>
        </ms:Organization>
      </ms:funder>
    </ms:fundingProject>
    <ms:intendedApplication>
      <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageTechnology</ms:LTClassRecommended>
    </ms:intendedApplication>
    <ms:intendedApplication>
      <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪MachineTranslation</ms:LTClassRecommended>
    </ms:intendedApplication>
    <ms:actualUse>
      <ms:usedInApplication>

```

(continues on next page)

(continued from previous page)

```

        <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪MachineTranslation</ms:LTClassRecommended>
        </ms:usedInApplication>
        <ms:usedInApplication>
        <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageTechnology</ms:LTClassRecommended>
        </ms:usedInApplication>
        </ms:actualUse>
        <ms:isDocumentedBy>
        <ms:title xml:lang="el">2.1_-_v1</ms:title>
        <ms:title xml:lang="en">Deliverable 2.1 - Text corpora-v1</ms:title>
        </ms:isDocumentedBy>
        <ms:LRSubclass>
        <ms:Corpus>
        <ms:lrType>Corpus</ms:lrType>
        <ms:corpusSubclass>http://w3id.org/meta-share/meta-share/annotatedCorpus
↪</ms:corpusSubclass>
        <ms:CorpusMediaPart>
        <ms:CorpusTextPart>
        <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
        <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪bilingual</ms:lingualityType>
        <ms:multilingualityType>http://w3id.org/meta-share/meta-share/
↪parallel</ms:multilingualityType>
        <ms:language>
        <ms:languageTag>el</ms:languageTag>
        <ms:languageId>el</ms:languageId>
        </ms:language>
        <ms:language>
        <ms:languageTag>bg</ms:languageTag>
        <ms:languageId>bg</ms:languageId>
        </ms:language>
        <ms:modalityType>http://w3id.org/meta-share/meta-share/
↪writtenLanguage</ms:modalityType>
        <ms:annotation>
        <ms:annotationType>http://w3id.org/meta-share/omtd-share/
↪Alignment1</ms:annotationType>
        <ms:segmentationLevel>http://w3id.org/meta-share/meta-share/
↪sentence</ms:segmentationLevel>
        <ms:annotationStandoff>>false</ms:annotationStandoff>
        <ms:annotationMode>http://w3id.org/meta-share/meta-share/
↪automatic</ms:annotationMode>
        <ms:isAnnotatedBy>
        <ms:resourceName xml:lang="en">TrAid</ms:resourceName>
        <ms:version>unspecified</ms:version>
        </ms:isAnnotatedBy>
        </ms:annotation>
        <ms:hasOriginalSource>
        <ms:resourceName xml:lang="en">The JRC-Aquis Corpus, version
        3.0</ms:resourceName>

```

(continues on next page)

(continued from previous page)

```

        <ms:LRIIdentifier
            ms:LRIIdentifierScheme="http://purl.org/spar/datacite/
→handle"
            >http://hdl.handle.net/11500/ATHENA-0000-0000-25C9-4</
→ms:LRIIdentifier>
            <ms:version>1.0.0 (automatically assigned)</ms:version>
        </ms:hasOriginalSource>
        <ms:hasOriginalSource>
            <ms:resourceName xml:lang="en">SETIMES - A parallel corpus.
→of the
                Balkan languages</ms:resourceName>
            <ms:LRIIdentifier
                ms:LRIIdentifierScheme="http://purl.org/spar/datacite/
→handle"
                >http://hdl.handle.net/11500/ATHENA-0000-0000-2591-2</
→ms:LRIIdentifier>
                <ms:version>1</ms:version>
            </ms:hasOriginalSource>
            <ms:creationDetails xml:lang="en">original source: EU
                texts</ms:creationDetails>
            </ms:CorpusTextPart>
        </ms:CorpusMediaPart>
        <ms:DatasetDistribution>
            <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
→downloadable</ms:DatasetDistributionForm>
            <ms:downloadLocation>http://www.hiddenLocation.org</
→ms:downloadLocation>
            <ms:accessLocation>http://fixme.com</ms:accessLocation>
            <ms:distributionTextFeature>
                <ms:size>
                    <ms:amount>100000000.0</ms:amount>
                    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/token</
→ms:sizeUnit>
                </ms:size>
                <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Tmx</
→ms:dataFormat>
                <ms:characterEncoding>http://w3id.org/meta-share/meta-share/UTF-8
→</ms:characterEncoding>
            </ms:distributionTextFeature>
            <ms:licenceTerms>
                <ms:licenceTermsName xml:lang="en">Creative Commons Attribution.
→4.0
                    International</ms:licenceTermsName>
                <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
→legalcode</ms:licenceTermsURL>
                <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
→</ms:licenceTermsURL>
                <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
→attribution</ms:conditionOfUse>
                <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
→allowsDirectAccess</ms:licenceCategory>
                <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
→allowsProcessing</ms:licenceCategory>

```

(continues on next page)

(continued from previous page)

```

        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
    ↪ </ms:licenceCategory>
        <ms:LicenceIdentifier
            ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
    ↪ share/SPDX"
            >CC-BY-4.0</ms:LicenceIdentifier>
    </ms:licenceTerms>
    <ms:attributionText xml:lang="el">-    Bul-TM.
        :
        :
        : Creative Commons Attribution 4.0 International
        (https://creativecommons.org/licenses/by/4.0/legalcode,
        https://creativecommons.org/licenses/by/4.0/). :
        http://hdl.handle.net/11500/ATHENA-0000-0000-23E4-7
        (CLARIN:EL)</ms:attributionText>
    <ms:attributionText xml:lang="en">Greek-Bulgarian Bul-TM parallel
    ↪ corpus by
        Institute for Language and Speech Processing - Athena Research
    ↪ Center
        used under Creative Commons Attribution 4.0 International
        (https://creativecommons.org/licenses/by/4.0/legalcode,
        https://creativecommons.org/licenses/by/4.0/). Source:
        http://hdl.handle.net/11500/ATHENA-0000-0000-23E4-7
        (CLARIN:EL)</ms:attributionText>
    </ms:DatasetDistribution>
    <ms:personalDataIncluded>>false</ms:personalDataIncluded>
    <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
    </ms:Corpus>
    </ms:LRSubclass>
    </ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

26.1.4 Multilingual corpus

```

<?xml version="1.0" encoding="utf-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
    ↪ metadata-schema/CLARIN-SHARE.xsd">
<ms:metadataCreationDate>2015-12-23</ms:metadataCreationDate>
<ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
<ms:metadataCurator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>
    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
<ms:metadataCreator>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>

```

(continues on next page)

(continued from previous page)

```

    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
    <ms:repositoryName xml:lang="el"> </ms:repositoryName>
    <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
    <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
    <ms:LanguageResource>
        <ms:entityType>LanguageResource</ms:entityType>
        <ms:resourceName xml:lang="el"> DICTA-SIGN</ms:resourceName>
        <ms:resourceName xml:lang="en">DICTA-SIGN corpus</ms:resourceName>
        <ms:description xml:lang="el">    (
            (, ,    ).
            14
            ,
            .

            &lt;a href="http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/index.html"
            target="_blank"&gt;&lt;/a>&gt;.</ms:description>
        <ms:description xml:lang="en">Multimedia corpus (video) for four sign languages
            (english, french, german and greek) of at least 14 informants per language.
↪and a
            session duration of approx. 2 hours using the same elicitation materials.
↪(scripts
            and tasks) across languages. The data is partially annotated. The corpus is
            available through a dedicated website, created and maintained by the.
↪University of
            Hamburg accessible &lt;a
            href="http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/index.html"
            target="_blank"&gt;here&lt;/a>&gt;.</ms:description>
        <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
            >http://hdl.handle.net/11500/ATHENA-0000-0000-28C5-5</ms:LRIdentifier>
        <ms:version>1.0.0 (automatically assigned)</ms:version>
        <ms:additionalInfo>
            <ms:landingPage>http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/index.
↪html</ms:landingPage>
        </ms:additionalInfo>
        <ms:contact>
            <ms:Person>
                <ms:actorType>Person</ms:actorType>
                <ms:surname xml:lang="en">Person_Surname</ms:surname>
                <ms:givenName xml:lang="en">Person_Name</ms:givenName>
            </ms:Person>
        </ms:contact>
        <ms:contact>
            <ms:Person>
                <ms:actorType>Person</ms:actorType>
                <ms:surname xml:lang="en">Person_Surname</ms:surname>
                <ms:givenName xml:lang="en">Person_Name</ms:givenName>
            </ms:Person>
        </ms:contact>

```

(continues on next page)

(continued from previous page)

```

<ms:citationText xml:lang="el">School of Computing Sciences - University of East
↪Anglia;
    Research Institute of Computer Science - Paul Sabatier University; Institute
↪of
    German Sign Language and Communication of the Deaf - University of Hamburg
↪(2015).
    DICTA-SIGN. Version 1.0.0 (automatically assigned). [Dataset (Text and Video
    corpus)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-00000-00000-28C5-5</ms:citationText>
<ms:citationText xml:lang="en">School of Computing Sciences - University of East
↪Anglia;
    Research Institute of Computer Science - Paul Sabatier University; Institute
↪of
    German Sign Language and Communication of the Deaf - University of Hamburg
↪(2015).
    DICTA-SIGN corpus. Version 1.0.0 (automatically assigned). [Dataset (Text
↪and Video
    corpus)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-00000-00000-28C5-5</ms:citationText>
<ms:keyword xml:lang="en">multilingual</ms:keyword>
<ms:keyword xml:lang="en">parallel</ms:keyword>
<ms:keyword xml:lang="en">scripts</ms:keyword>
<ms:keyword xml:lang="en">signLanguage</ms:keyword>
<ms:domain>
    <ms:categoryLabel xml:lang="en">Geography &amp; Travel</ms:categoryLabel>
    <ms:DomainIdentifier
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪Classification"
        >DDC910</ms:DomainIdentifier>
    </ms:domain>
<ms:resourceCreator>
    <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">School of Computing
            Sciences</ms:organizationName>
        <ms:website>https://www.uea.ac.uk/about/school-of-computing-sciences</
↪ms:website>
    </ms:Organization>
</ms:resourceCreator>
<ms:resourceCreator>
    <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">Research Institute of Computer
            Science</ms:organizationName>
        <ms:organizationName xml:lang="fr">Institut de Recherche en Informatique
↪de
            Toulouse</ms:organizationName>
        <ms:website>http://www.irit.fr</ms:website>
    </ms:Organization>
</ms:resourceCreator>
<ms:resourceCreator>
    <ms:Organization>

```

(continues on next page)

(continued from previous page)

```

        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="en">Institute of German Sign Language and
            Communication of the Deaf</ms:organizationName>
        <ms:website>https://www.idgs.uni-hamburg.de/en.html</ms:website>
    </ms:Organization>
</ms:resourceCreator>
<ms:creationStartDate>2009-02-01</ms:creationStartDate>
<ms:creationEndDate>2012-01-31</ms:creationEndDate>
<ms:fundingProject>
    <ms:projectName xml:lang="en">Sign Language Recognition, Generation and
↪Modelling
        with application in Deaf Communication</ms:projectName>
    <ms:website>http://www.dictasign.eu/</ms:website>
    <ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds</
↪ms:fundingType>
    <ms:funder>
        <ms:Organization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="el"> </ms:organizationName>
            <ms:organizationName xml:lang="en">European Commission</
↪ms:organizationName>
            <ms:website>https://ec.europa.eu/info/index_en</ms:website>
        </ms:Organization>
    </ms:funder>
</ms:fundingProject>
<ms:intendedApplication>
    <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageTechnology</ms:LTClassRecommended>
</ms:intendedApplication>
<ms:actualUse>
    <ms:usedInApplication>
        <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageTechnology</ms:LTClassRecommended>
    </ms:usedInApplication>
</ms:actualUse>
<ms:LRSubclass>
    <ms:Corpus>
        <ms:lrType>Corpus</ms:lrType>
        <ms:corpusSubclass>http://w3id.org/meta-share/meta-share/rawCorpus</
↪ms:corpusSubclass>
        <ms:CorpusMediaPart>
            <ms:CorpusTextPart>
                <ms:corpusMediaType>CorpusTextPart</ms:corpusMediaType>
                <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
                <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪multilingual</ms:lingualityType>
                <ms:multilingualityType>http://w3id.org/meta-share/meta-share/
↪unspecified</ms:multilingualityType>
                <ms:language>
                    <ms:languageTag>fr</ms:languageTag>
                    <ms:languageId>fr</ms:languageId>

```

(continues on next page)

(continued from previous page)

```

</ms:language>
<ms:language>
  <ms:languageTag>el</ms:languageTag>
  <ms:languageId>el</ms:languageId>
</ms:language>
<ms:language>
  <ms:languageTag>de</ms:languageTag>
  <ms:languageId>de</ms:languageId>
</ms:language>
<ms:language>
  <ms:languageTag>en</ms:languageTag>
  <ms:languageId>en</ms:languageId>
</ms:language>
<ms:modalityType>http://w3id.org/meta-share/meta-share/
↪signLanguage</ms:modalityType>
<ms:TextGenre>
  <ms:categoryLabel xml:lang="en">scripts</ms:categoryLabel>
</ms:TextGenre>
<ms:linkToOtherMedia>
  <ms:otherMedia>http://w3id.org/meta-share/meta-share/video</
↪ms:otherMedia>
  <ms:mediaTypeDetails xml:lang="en">scripts of the tasks for
↪the
    video</ms:mediaTypeDetails>
  <ms:synchronizedWithVideo>>false</ms:synchronizedWithVideo>
</ms:linkToOtherMedia>
</ms:CorpusTextPart>
</ms:CorpusMediaPart>
<ms:CorpusMediaPart>
  <ms:CorpusVideoPart>
    <ms:corpusMediaType>CorpusVideoPart</ms:corpusMediaType>
    <ms:mediaType>http://w3id.org/meta-share/meta-share/video</
↪ms:mediaType>
    <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪multilingual</ms:lingualityType>
    <ms:multilingualityType>http://w3id.org/meta-share/meta-share/
↪parallel</ms:multilingualityType>
    <ms:language>
      <ms:languageTag>gss</ms:languageTag>
      <ms:languageId>gss</ms:languageId>
    </ms:language>
    <ms:language>
      <ms:languageTag>bfi</ms:languageTag>
      <ms:languageId>bfi</ms:languageId>
    </ms:language>
    <ms:language>
      <ms:languageTag>gsg</ms:languageTag>
      <ms:languageId>gsg</ms:languageId>
    </ms:language>
    <ms:language>
      <ms:languageTag>fsl</ms:languageTag>
      <ms:languageId>fsl</ms:languageId>

```

(continues on next page)

(continued from previous page)

```

        </ms:language>
        <ms:modalityType>http://w3id.org/meta-share/meta-share/
↪signLanguage</ms:modalityType>
        <ms:typeOfVideoContent xml:lang="en">natural
            signers</ms:typeOfVideoContent>
        <ms:textIncludedInVideo>http://w3id.org/meta-share/meta-share/
↪none1</ms:textIncludedInVideo>
        <ms:naturality>http://w3id.org/meta-share/meta-share/elicited</
↪ms:naturality>
        <ms:conversationalType>http://w3id.org/meta-share/meta-share/
↪dialogue</ms:conversationalType>
        <ms:scenarioType>http://w3id.org/meta-share/meta-share/rolePlay</
↪ms:scenarioType>
        <ms:audience>http://w3id.org/meta-share/meta-share/none3</
↪ms:audience>
        <ms:interactivity>http://w3id.org/meta-share/meta-share/
↪interactive1</ms:interactivity>
        <ms:linkToOtherMedia>
            <ms:otherMedia>http://w3id.org/meta-share/meta-share/text</
↪ms:otherMedia>
            <ms:mediaTypeDetails xml:lang="en">scripts of the tasks for
↪the
                video</ms:mediaTypeDetails>
            <ms:synchronizedWithText>>false</ms:synchronizedWithText>
        </ms:linkToOtherMedia>
    </ms:CorpusVideoPart>
</ms:CorpusMediaPart>
<ms:DatasetDistribution>
    <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪accessibleThroughInterface</ms:DatasetDistributionForm>
    <ms:accessLocation>http://www.sign-lang.uni-hamburg.de/dicta-sign/
↪portal/index.html</ms:accessLocation>
    <ms:distributionTextFeature>
        <ms:size>
            <ms:amount>10.0</ms:amount>
            <ms:sizeUnit>http://w3id.org/meta-share/meta-share/file</
↪ms:sizeUnit>
        </ms:size>
        <ms:dataFormat>http://w3id.org/meta-share/meta-share/unspecified
↪</ms:dataFormat>
    </ms:distributionTextFeature>
    <ms:distributionVideoFeature>
        <ms:size>
            <ms:amount>25.0</ms:amount>
            <ms:sizeUnit>http://w3id.org/meta-share/meta-share/hour1</
↪ms:sizeUnit>
        </ms:size>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/mp4</
↪ms:dataFormat>
    </ms:distributionVideoFeature>
    <ms:licenceTerms>
        <ms:licenceTermsName xml:lang="en">Creative Commons Attribution
↪Non

```

(continues on next page)

(continued from previous page)

```

Commercial No Derivatives 4.0 International</
ms:licenceTermsName>
  <ms:licenceTermsURL>https://creativecommons.org/licenses/by-nc-
nd/4.0/legalcode</ms:licenceTermsURL>
  <ms:licenceTermsURL>https://creativecommons.org/licenses/by-nc-
nd/4.0/</ms:licenceTermsURL>
  <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
attribution</ms:conditionOfUse>
  <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
nonCommercialUse</ms:conditionOfUse>
  <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
noDerivatives</ms:conditionOfUse>
  <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
allowsDirectAccess</ms:licenceCategory>
  <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
</ms:licenceCategory>
  <ms:LicenceIdentifier
    ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
share/SPDX"
    >CC-BY-NC-ND-4.0</ms:LicenceIdentifier>
  </ms:licenceTerms>
  <ms:attributionText xml:lang="el"> DICTA-SIGN. : School of
    Computing Sciences - University of East Anglia, Research
Institute of
    Computer Science - Paul Sabatier University and Institute of
German Sign
    Language and Communication of the Deaf - University of Hamburg. :
    Creative Commons Attribution Non Commercial No Derivatives 4.0
    International
    (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode,
    https://creativecommons.org/licenses/by-nc-nd/4.0/). :
    http://hdl.handle.net/11500/ATHENA-0000-0000-28C5-5
    (CLARIN:EL)</ms:attributionText>
  <ms:attributionText xml:lang="en">DICTA-SIGN corpus by School of
Computing
    Sciences - University of East Anglia, Research Institute of
Computer
    Science - Paul Sabatier University and Institute of German Sign
Language
    and Communication of the Deaf - University of Hamburg used under
    Creative Commons Attribution Non Commercial No Derivatives 4.0
    International
    (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode,
    https://creativecommons.org/licenses/by-nc-nd/4.0/). Source:
    http://hdl.handle.net/11500/ATHENA-0000-0000-28C5-5
    (CLARIN:EL)</ms:attributionText>
  </ms:DatasetDistribution>
  <ms:personalDataIncluded>>false</ms:personalDataIncluded>
  <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
</ms:Corpus>
</ms:LRSubclass>
</ms:LanguageResource>

```

(continues on next page)

(continued from previous page)

```

</ms:DescribedEntity>
</ms:MetadataRecord>

```

26.2 2. Lexical/Conceptual resources (LCR)

26.2.1 Monolingual LCR

```

<?xml version="1.0" encoding="utf-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↳ metadata-schema/CLARIN-SHARE.xsd">
<ms:metadataCreationDate>2015-12-08</ms:metadataCreationDate>
<ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
<ms:metadataCurator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
<ms:metadataCreator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
  <ms:repositoryName xml:lang="el"> </ms:repositoryName>
  <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
  <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
  <ms:LanguageResource>
    <ms:entityType>LanguageResource</ms:entityType>
    <ms:resourceName xml:lang="en">KELLY word-list Greek</ms:resourceName>
    <ms:resourceShortName xml:lang="en">KELLY word-list EL</ms:resourceShortName>
    <ms:description xml:lang="el"> KELLY EL
      , 4
      ( , , ) 5
      ( , , , )
      / . ,
      36 .
      ,
      ,
      .
      (CEFR, &lt;a
href="https://www.coe.int/en/web/common-european-framework-reference-
↳ languages"

```

(continues on next page)

(continued from previous page)

```

target="_blank"&gt;Common European Framework of Reference for Languages&lt;/
→a&gt;).

(
),
(
),

,

,
CEFR,
36 . &lt;a
href="https://spraakbanken.gu.se/eng/kelly" target="_blank"&gt;
&lt;/a&gt; Kelly.

, , ,
&lt;a
href="http://kelly.sketchengine.co.uk/" target="_blank"&gt;&lt;/a&gt;

,
.</ms:description>
<ms:description xml:lang="en">The monolingual lexical conceptual resource KELLY_
→EL is
part of digital material created for educational purposes, i.e. to_
→facilitate the
learning of a foreign/second language. Nine different languages were_
→involved, four
commonly learned (English, Arabic, Russian and Chinese) and five less_
→commonly
learned (Greek, Italian, Swedish, Polish and Norwegian). More precisely,_
→KELLY EL is
the Greek part of a material which consists of monolingual and bilingual_
→word-lists
covering 36 language pairs in total. The choice of words was based for each_
→language
on digital language resources. The same standards were applied to all_
→languages for
the choice of these digital language resources in order to ensure uniformity.
→ The
material was analyzed and edited by linguists and education professionals_
→and each
word was mapped to the appropriate language level of the Common European_
→Framework
of Reference (CEFR, &lt;a
href="https://www.coe.int/en/web/common-european-framework-reference-
→languages"
target="_blank"&gt;Common European Framework of Reference for Languages&lt;/
→a&gt;).
The vocabularies produced were created by extracting knowledge from texts_
→(processes
such as lemmatization, morphological and structural annotation were used)_
→followed
by statistical metrics techniques for extracting the most frequent (and_
→therefore
necessary for language learning) words. The language technology results were

```

(continues on next page)

(continued from previous page)

examined by experts according to pedagogical principles, evaluated and
 ↪ finally
 bilingual word-lists with words mapped to the different levels of CEFR for 36
 language pairs were created. More information about Kelly is available from
 ↪ the
 official
 project site, where the lists in English, Arabic, Italian, Chinese,
 Norwegian and Russian can be downloaded. There is also a database interface
 , which can be used to explore the links between words selected
 ↪ for each
 of these languages.</ms:description>
 <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
 >http://hdl.handle.net/11500/ATHENA-0000-0000-25C1-C</ms:LRIdentifier>
 <ms:version>1.0.0 (automatically assigned)</ms:version>
 <ms:additionalInfo>
 <ms:email>person@ilsp.athena-innovation.gr</ms:email>
 </ms:additionalInfo>
 <ms:contact>
 <ms:Person>
 <ms:actorType>Person</ms:actorType>
 <ms:surname xml:lang="en">Person_Surname</ms:surname>
 <ms:givenName xml:lang="en">Person_Name</ms:givenName>
 </ms:Person>
 </ms:contact>
 <ms:citationText xml:lang="el"> -
 (2015). KELLY word-list Greek. Version 1.0.0 (automatically assigned).
 [Dataset (Lexical/Conceptual Resource)]. CLARIN:EL.
 http://hdl.handle.net/11500/ATHENA-0000-0000-25C1-C</ms:citationText>
 <ms:citationText xml:lang="en">Institute for Language and Speech Processing -
 ↪ Athena
 Research Center (2015). KELLY word-list Greek. Version 1.0.0 (automatically
 assigned). [Dataset (Lexical/Conceptual Resource)]. CLARIN:EL.
 http://hdl.handle.net/11500/ATHENA-0000-0000-25C1-C</ms:citationText>
 <ms:keyword xml:lang="en">monolingual</ms:keyword>
 <ms:resourceCreator>
 <ms:Organization>
 <ms:actorType>Organization</ms:actorType>
 <ms:organizationName xml:lang="el">
 </ms:organizationName>
 <ms:organizationName xml:lang="en">Institute for Language and Speech
 Processing</ms:organizationName>
 <ms:website>http://www.ilsp.gr</ms:website>
 </ms:Organization>
 </ms:resourceCreator>
 <ms:isToBeCitedBy>
 <ms:title xml:lang="en">Kilgarriff, Adam, Frieda Charalabopoulou, Maria
 ↪ Gavrilidou,
 Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis,
 ↪ Robert Lew,
 Serge Sharoff, Ravikiran Vadlapudi and Elena Volodina (2014) Corpus-based
 vocabulary lists for language learners for nine languages. Language
 ↪ Resources

(continues on next page)

(continued from previous page)

```

        and Evaluation, 48:121-163</ms:title>
    <ms:DocumentIdentifier
        ms:DocumentIdentifierScheme="http://purl.org/spar/datacite/doi"
        >https://doi.org/10.1007/s10579-013-9251-2</ms:DocumentIdentifier>
</ms:isToBeCitedBy>
<ms:isPartOf>
    <ms:resourceName xml:lang="en">KELLY word-lists</ms:resourceName>
    <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
        >http://hdl.handle.net/11500/ATHENA-0000-0000-5862-F</ms:LRIdentifier>
    <ms:version>1.0.0 (automatically assigned)</ms:version>
</ms:isPartOf>
<ms:LRSubclass>
    <ms:LexicalConceptualResource>
        <ms:lrType>LexicalConceptualResource</ms:lrType>
        <ms:lcrSubclass>http://w3id.org/meta-share/meta-share/wordlist</
↪ms:lcrSubclass>
        <ms:encodingLevel>http://w3id.org/meta-share/meta-share/other</
↪ms:encodingLevel>
        <ms:LexicalConceptualResourceMediaPart>
            <ms:LexicalConceptualResourceTextPart>
                <ms:lcrMediaType>LexicalConceptualResourceTextPart</
↪ms:lcrMediaType>
                <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
                <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪monolingual</ms:lingualityType>
                <ms:language>
                    <ms:languageTag>el</ms:languageTag>
                    <ms:languageId>el</ms:languageId>
                </ms:language>
            </ms:LexicalConceptualResourceTextPart>
        </ms:LexicalConceptualResourceMediaPart>
        <ms:DatasetDistribution>
            <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪downloadable</ms:DatasetDistributionForm>
            <ms:downloadLocation>http://www.hiddenLocation.org</
↪ms:downloadLocation>
            <ms:distributionTextFeature>
                <ms:size>
                    <ms:amount>7385.0</ms:amount>
                    <ms:sizeUnit>http://w3id.org/meta-share/meta-share/entry</
↪ms:sizeUnit>
                </ms:size>
                <ms:dataFormat>http://w3id.org/meta-share/meta-share/unspecified
↪</ms:dataFormat>
            </ms:distributionTextFeature>
            <ms:licenceTerms>
                <ms:licenceTermsName xml:lang="en">Creative Commons Attribution_
↪Non
                    Commercial 4.0 International</ms:licenceTermsName>
                <ms:licenceTermsURL>https://creativecommons.org/licenses/by-nc/4.
↪0/legalcode</ms:licenceTermsURL>

```

(continues on next page)

(continued from previous page)

```

        <ms:licenceTermsURL>https://creativecommons.org/licenses/by-nc/4.
↪0/</ms:licenceTermsURL>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪attribution</ms:conditionOfUse>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪nonCommercialUse</ms:conditionOfUse>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsDirectAccess</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsProcessing</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↪</ms:licenceCategory>
        <ms:LicenceIdentifier
↪share/SPDX"
            ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
                >CC-BY-NC-4.0</ms:LicenceIdentifier>
        </ms:licenceTerms>
        <ms:attributionText xml:lang="el">KELLY word-list Greek. :
            - . :
            Creative Commons Attribution Non Commercial 4.0 International
            (https://creativecommons.org/licenses/by-nc/4.0/legalcode,
            https://creativecommons.org/licenses/by-nc/4.0/). :
            http://hdl.handle.net/11500/ATHENA-0000-0000-25C1-C
            (CLARIN:EL)</ms:attributionText>
        <ms:attributionText xml:lang="en">KELLY word-list Greek by Institute_
↪for
            Language and Speech Processing - Athena Research Center used_
↪under
            Creative Commons Attribution Non Commercial 4.0 International
            (https://creativecommons.org/licenses/by-nc/4.0/legalcode,
            https://creativecommons.org/licenses/by-nc/4.0/). Source:
            http://hdl.handle.net/11500/ATHENA-0000-0000-25C1-C
            (CLARIN:EL)</ms:attributionText>
        </ms:DatasetDistribution>
        <ms:personalDataIncluded>>false</ms:personalDataIncluded>
        <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
    </ms:LexicalConceptualResource>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

26.2.2 Bilingual LCR

```

<?xml version="1.0" encoding="utf-8"?>
  <ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↳ metadata-schema/CLARIN-SHARE.xsd">
<ms:metadataCreationDate>2016-12-07</ms:metadataCreationDate>
<ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
<ms:metadataCurator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
<ms:metadataCreator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
  <ms:repositoryName xml:lang="el"> </ms:repositoryName>
  <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
  <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
  <ms:LanguageResource>
    <ms:entityType>LanguageResource</ms:entityType>
    <ms:resourceName xml:lang="el"> - </ms:resourceName>
    <ms:resourceName xml:lang="en">Orossimo Terminological Resource -
      History</ms:resourceName>
    <ms:description xml:lang="el">

    .</ms:description>
    <ms:description xml:lang="en">A bilingual terminological glossary extracted from
      academic discourse texts belonging to the History domain.</ms:description>
    <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
      >http://hdl.handle.net/11500/ATHENA-0000-0000-4B4B-9</ms:LRIdentifier>
    <ms:version>1.0.0 (automatically assigned)</ms:version>
    <ms:additionalInfo>
      <ms:email>person@ilsp</ms:email>
    </ms:additionalInfo>
    <ms:contact>
      <ms:Person>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
      </ms:Person>
    </ms:contact>
    <ms:citationText xml:lang="el"> -
      (2016). - . Version 1.0.0 (automatically
      assigned). [Dataset (Lexical/Conceptual Resource)]. CLARIN:EL.
      http://hdl.handle.net/11500/ATHENA-0000-0000-4B4B-9</ms:citationText>

```

(continues on next page)

(continued from previous page)

```

<ms:citationText xml:lang="en">Institute for Language and Speech Processing -
↪Athena
    Research Center (2016). Orossimo Terminological Resource - History. Version
↪1.0.0
    (automatically assigned). [Dataset (Lexical/Conceptual Resource)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-4B4B-9</ms:citationText>
<ms:keyword xml:lang="en">bilingual</ms:keyword>
<ms:domain>
    <ms:categoryLabel xml:lang="en">History</ms:categoryLabel>
    <ms:DomainIdentifier
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪Classification"
        >DDC900</ms:DomainIdentifier>
</ms:domain>
<ms:resourceCreator>
    <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="el">
            </ms:organizationName>
        <ms:organizationName xml:lang="en">Institute for Language and Speech
            Processing</ms:organizationName>
        <ms:website>http://www.ilsp.gr</ms:website>
    </ms:Organization>
</ms:resourceCreator>
<ms:creationStartDate>1996-01-01</ms:creationStartDate>
<ms:creationEndDate>1998-12-31</ms:creationEndDate>
<ms:fundingProject>
    <ms:projectName xml:lang="el"></ms:projectName>
    <ms:projectName xml:lang="en">OROSSIMO</ms:projectName>
    <ms:website>https://www.ilsp.gr/projects/orosimo</ms:website>
    <ms:fundingType>http://w3id.org/meta-share/meta-share/nationalFunds</
↪ms:fundingType>
    <ms:funder>
        <ms:Organization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">General Secretariat for Research
↪and
                Technology</ms:organizationName>
        </ms:Organization>
    </ms:funder>
</ms:fundingProject>
<ms:hasOriginalSource>
    <ms:resourceName xml:lang="el"> - </ms:resourceName>
    <ms:resourceName xml:lang="en">OROSSIMO Corpus - History</ms:resourceName>
    <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
        >http://hdl.handle.net/11500/ATHENA-0000-0000-240F-8</ms:LRIdentifier>
    <ms:version>1.0.0 (automatically assigned)</ms:version>
</ms:hasOriginalSource>
<ms:isDocumentedBy>
    <ms:title xml:lang="el"> :
        </ms:title>
    <ms:title xml:lang="en">Collection of digital terminological resources:
↪methodology

```

(continues on next page)

(continued from previous page)

```

        and results</ms:title>
    </ms:isDocumentedBy>
    <ms:isPartOf>
        <ms:resourceName xml:lang="el"> </ms:resourceName>
        <ms:resourceName xml:lang="en">Orossimo Terminological Resource</
↪ms:resourceName>
        <ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
            >http://hdl.handle.net/11500/ATHENA-0000-0000-4B49-B</ms:LRIdentifier>
        <ms:version>1.0.0 (automatically assigned)</ms:version>
    </ms:isPartOf>
    <ms:LRSubclass>
        <ms:LexicalConceptualResource>
            <ms:lrType>LexicalConceptualResource</ms:lrType>
            <ms:lcrSubclass>http://w3id.org/meta-share/meta-share/
↪terminologicalResource</ms:lcrSubclass>
            <ms:encodingLevel>http://w3id.org/meta-share/meta-share/semantics</
↪ms:encodingLevel>
            <ms:ContentType>http://w3id.org/meta-share/meta-share/lemma</
↪ms:ContentType>
            <ms:ContentType>http://w3id.org/meta-share/meta-share/domain1</
↪ms:ContentType>
            <ms:ContentType>http://w3id.org/meta-share/meta-share/derivation</
↪ms:ContentType>
            <ms:ContentType>http://w3id.org/meta-share/meta-share/
↪translationEquivalent</ms:ContentType>
            <ms:ContentType>http://w3id.org/meta-share/meta-share/notation1</
↪ms:ContentType>
            <ms:LexicalConceptualResourceMediaPart>
                <ms:LexicalConceptualResourceTextPart>
                    <ms:lcrMediaType>LexicalConceptualResourceTextPart</
↪ms:lcrMediaType>
                    <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
                    <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪bilingual</ms:lingualityType>
                    <ms:language>
                        <ms:languageTag>en</ms:languageTag>
                        <ms:languageId>en</ms:languageId>
                    </ms:language>
                    <ms:language>
                        <ms:languageTag>el</ms:languageTag>
                        <ms:languageId>el</ms:languageId>
                    </ms:language>
                </ms:LexicalConceptualResourceTextPart>
            </ms:LexicalConceptualResourceMediaPart>
            <ms:DatasetDistribution>
                <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪downloadable</ms:DatasetDistributionForm>
                <ms:downloadLocation>http://www.hiddenLocation.org</
↪ms:downloadLocation>
                <ms:accessLocation>http://fixme.com</ms:accessLocation>
                <ms:distributionTextFeature>

```

(continues on next page)

(continued from previous page)

```

        <ms:size>
          <ms:amount>2353.0</ms:amount>
          <ms:sizeUnit>http://w3id.org/meta-share/meta-share/term</
↪ms:sizeUnit>
        </ms:size>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/MsExcel</
↪ms:dataFormat>
      </ms:distributionTextFeature>
      <ms:licenceTerms>
        <ms:licenceTermsName xml:lang="en">Creative Commons Attribution.
↪4.0
          International</ms:licenceTermsName>
        <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
↪legalcode</ms:licenceTermsURL>
        <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
↪</ms:licenceTermsURL>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪attribution</ms:conditionOfUse>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsDirectAccess</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsProcessing</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↪</ms:licenceCategory>
        <ms:LicenceIdentifier
↪share/SPDX"
          ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
          >CC-BY-4.0</ms:LicenceIdentifier>
        </ms:licenceTerms>
        <ms:attributionText xml:lang="el"> - .
          :
          : Creative Commons Attribution 4.0 International
          (https://creativecommons.org/licenses/by/4.0/legalcode,
          https://creativecommons.org/licenses/by/4.0/). :
          http://hdl.handle.net/11500/ATHENA-0000-0000-4B4B-9
          (CLARIN:EL)</ms:attributionText>
        <ms:attributionText xml:lang="en">Orossimo Terminological Resource -
↪History
          by Institute for Language and Speech Processing - Athena
↪Research Center
          used under Creative Commons Attribution 4.0 International
          (https://creativecommons.org/licenses/by/4.0/legalcode,
          https://creativecommons.org/licenses/by/4.0/). Source:
          http://hdl.handle.net/11500/ATHENA-0000-0000-4B4B-9
          (CLARIN:EL)</ms:attributionText>
      </ms:DatasetDistribution>
      <ms:personalDataIncluded>>false</ms:personalDataIncluded>
      <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
    </ms:LexicalConceptualResource>
  </ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>

```

(continues on next page)

(continued from previous page)

</ms:MetadataRecord>

26.2.3 Multilingual LCR

```

<?xml version="1.0" encoding="utf-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↳ metadata-schema/CLARIN-SHARE.xsd">
<ms:metadataCreationDate>2019-03-27</ms:metadataCreationDate>
<ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
<ms:metadataCurator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
<ms:metadataCreator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
  <ms:repositoryName xml:lang="el"> </ms:repositoryName>
  <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
  <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
  <ms:LanguageResource>
    <ms:entityType>LanguageResource</ms:entityType>
    <ms:resourceName xml:lang="el"> ()</ms:resourceName>
    <ms:resourceName xml:lang="en">Trilingual Term Dictionary (TTD)</ms:resourceName>
    <ms:description xml:lang="el"> ()
      .
      ,
      - ,
      , , , , ,
      , , , , ,
      - .
      .
    .</ms:description>
    <ms:description xml:lang="en">The Trilingual Term Dictionary (TTD) is targeted to
      foreign students of the secondary school in Thrace, Greece. The aim of the
↳ TTD is
      threefold: to assist the student in learning the subject areas of the
↳ curriculum, to
      improve their language skills in Greek and to familiarize themselves with
      information technology. TTD contains terms that are used in several subject
↳ areas

```

(continues on next page)

(continued from previous page)

```

    (e.g. biology, geography, history, social and political studies etc.) that
↪are
    taught in the secondary school. The terms of TDD (more than 5.000) have been
    collected from the schoolbooks. The terms are categorised within the subject
↪areas,
    accompanied by definitions and translated into English and Turkish.</
↪ms:description>
    <ms:LRIIdentifier ms:LRIIdentifierScheme="http://purl.org/spar/datacite/handle"
      >http://hdl.handle.net/11500/ATHENA-0000-0000-5837-0</ms:LRIIdentifier>
    <ms:version>1.0.0 (automatically assigned)</ms:version>
    <ms:additionalInfo>
      <ms:landingPage>http://www.ilsp.gr/tol/</ms:landingPage>
    </ms:additionalInfo>
    <ms:contact>
      <ms:Person>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
      </ms:Person>
    </ms:contact>
    <ms:citationText xml:lang="el"> -
      (2019). (). Version 1.0.0 (automatically
      assigned). [Dataset (Lexical/Conceptual Resource)]. CLARIN:EL.
      http://hdl.handle.net/11500/ATHENA-0000-0000-5837-0</ms:citationText>
    <ms:citationText xml:lang="en">Institute for Language and Speech Processing -
↪Athena
    Research Center (2019). Trilingual Term Dictionary (TTD). Version 1.0.0
    (automatically assigned). [Dataset (Lexical/Conceptual Resource)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-5837-0</ms:citationText>
    <ms:keyword xml:lang="en">multilingual</ms:keyword>
    <ms:domain>
      <ms:categoryLabel xml:lang="en">Mathematics</ms:categoryLabel>
      <ms:DomainIdentifier
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
        >DDC510</ms:DomainIdentifier>
      </ms:domain>
      <ms:domain>
        <ms:categoryLabel xml:lang="en">Physics</ms:categoryLabel>
        <ms:DomainIdentifier
          ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
          >DDC530</ms:DomainIdentifier>
        </ms:domain>
        <ms:domain>
          <ms:categoryLabel xml:lang="en">Biology</ms:categoryLabel>
          <ms:DomainIdentifier
            ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
            >DDC570</ms:DomainIdentifier>
          </ms:domain>
        </ms:domain>

```

(continues on next page)

(continued from previous page)

```

    <ms:categoryLabel xml:lang="en">Technology</ms:categoryLabel>
    <ms:DomainIdentifier
      ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
      >DDC600</ms:DomainIdentifier>
    </ms:domain>
    <ms:domain>
      <ms:categoryLabel xml:lang="en">Home & Family Management</
↪ms:categoryLabel>
      <ms:DomainIdentifier
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
        >DDC640</ms:DomainIdentifier>
      </ms:domain>
      <ms:domain>
        <ms:categoryLabel xml:lang="en">Chemistry</ms:categoryLabel>
        <ms:DomainIdentifier
          ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
          >DDC540</ms:DomainIdentifier>
        </ms:domain>
        <ms:domain>
          <ms:categoryLabel xml:lang="en">Language</ms:categoryLabel>
          <ms:DomainIdentifier
            ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
            >DDC400</ms:DomainIdentifier>
          </ms:domain>
          <ms:domain>
            <ms:categoryLabel xml:lang="en">Geography & Travel</ms:categoryLabel>
            <ms:DomainIdentifier
              ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
              >DDC910</ms:DomainIdentifier>
            </ms:domain>
            <ms:domain>
              <ms:categoryLabel xml:lang="en">Music</ms:categoryLabel>
              <ms:DomainIdentifier
                ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
                >DDC780</ms:DomainIdentifier>
              </ms:domain>
              <ms:domain>
                <ms:categoryLabel xml:lang="en">History</ms:categoryLabel>
                <ms:DomainIdentifier
                  ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
                  >DDC900</ms:DomainIdentifier>
                </ms:domain>
                <ms:domain>
                  <ms:categoryLabel xml:lang="en">Literature, Rhetoric &
                  Criticism</ms:categoryLabel>

```

(continues on next page)

(continued from previous page)

```

    <ms:DomainIdentifier
      ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
      >DDC800</ms:DomainIdentifier>
    </ms:domain>
    <ms:domain>
      <ms:categoryLabel xml:lang="en">Computer Science, Information & General
        Works</ms:categoryLabel>
      <ms:DomainIdentifier
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/DDC_
↪classification"
        >DDC000</ms:DomainIdentifier>
      </ms:domain>
      <ms:resourceCreator>
        <ms:Organization>
          <ms:actorType>Organization</ms:actorType>
          <ms:organizationName xml:lang="el">
            </ms:organizationName>
          <ms:organizationName xml:lang="en">Institute for Language and Speech
            Processing</ms:organizationName>
          <ms:website>http://www.ilsp.gr</ms:website>
        </ms:Organization>
      </ms:resourceCreator>
      <ms:fundingProject>
        <ms:projectName xml:lang="el"> </ms:projectName>
        <ms:projectName xml:lang="en">Trilingual Terminological Dictionary</
↪ms:projectName>
        <ms:website>https://bit.ly/2V4hWLe</ms:website>
        <ms:website>https://www.ilsp.gr/projects/tol/</ms:website>
        <ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds</
↪ms:fundingType>
        <ms:fundingType>http://w3id.org/meta-share/meta-share/nationalFunds</
↪ms:fundingType>
        <ms:funder>
          <ms:Organization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="en">Ministry of Education and
↪Religious
              Affairs</ms:organizationName>
          </ms:Organization>
        </ms:funder>
        <ms:funder>
          <ms:Organization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="el"> </ms:organizationName>
            <ms:organizationName xml:lang="en">European Commission</
↪ms:organizationName>
            <ms:website>https://ec.europa.eu/info/index_en</ms:website>
          </ms:Organization>
        </ms:funder>
      </ms:fundingProject>
    <ms:LRSubclass>

```

(continues on next page)

(continued from previous page)

```

    <ms:LexicalConceptualResource>
      <ms:lrType>LexicalConceptualResource</ms:lrType>
      <ms:lcrSubclass>http://w3id.org/meta-share/meta-share/
↪terminologicalResource</ms:lcrSubclass>
      <ms:encodingLevel>http://w3id.org/meta-share/meta-share/other</
↪ms:encodingLevel>
      <ms:LexicalConceptualResourceMediaPart>
        <ms:LexicalConceptualResourceTextPart>
          <ms:lcrMediaType>LexicalConceptualResourceTextPart</
↪ms:lcrMediaType>
          <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
          <ms:lingualityType>http://w3id.org/meta-share/meta-share/
↪multilingual</ms:lingualityType>
          <ms:language>
            <ms:languageTag>el</ms:languageTag>
            <ms:languageId>el</ms:languageId>
          </ms:language>
          <ms:language>
            <ms:languageTag>en</ms:languageTag>
            <ms:languageId>en</ms:languageId>
          </ms:language>
          <ms:language>
            <ms:languageTag>tr</ms:languageTag>
            <ms:languageId>tr</ms:languageId>
          </ms:language>
        </ms:LexicalConceptualResourceTextPart>
      </ms:LexicalConceptualResourceMediaPart>
      <ms:DatasetDistribution>
        <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
↪downloadable</ms:DatasetDistributionForm>
        <ms:downloadLocation>http://www.hiddenLocation.org</
↪ms:downloadLocation>
        <ms:accessLocation>http://fixme.com</ms:accessLocation>
        <ms:distributionTextFeature>
          <ms:size>
            <ms:amount>5224.0</ms:amount>
            <ms:sizeUnit>http://w3id.org/meta-share/meta-share/entry</
↪ms:sizeUnit>
          </ms:size>
          <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</
↪ms:dataFormat>
        </ms:distributionTextFeature>
        <ms:licenceTerms>
          <ms:licenceTermsName xml:lang="en">Creative Commons Attribution,
↪Non
          Commercial No Derivatives 4.0 International</
↪ms:licenceTermsName>
          <ms:licenceTermsURL>https://creativecommons.org/licenses/by-nc-
↪nd/4.0/legalcode</ms:licenceTermsURL>
          <ms:licenceTermsURL>https://creativecommons.org/licenses/by-nc-
↪nd/4.0/</ms:licenceTermsURL>

```

(continues on next page)

(continued from previous page)

```

        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪ attribution</ms:conditionOfUse>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪ nonCommercialUse</ms:conditionOfUse>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪ noDerivatives</ms:conditionOfUse>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪ allowsDirectAccess</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↪ </ms:licenceCategory>
        <ms:LicenceIdentifier
            ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
↪ share/SPDX"
            >CC-BY-NC-ND-4.0</ms:LicenceIdentifier>
    </ms:licenceTerms>
    <ms:attributionText xml:lang="el">    ().
        :      -      .
        : Creative Commons Attribution Non Commercial No Derivatives 4.0
        International
        (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode,
        https://creativecommons.org/licenses/by-nc-nd/4.0/). :
        http://hdl.handle.net/11500/ATHENA-0000-0000-5837-0
        (CLARIN:EL)</ms:attributionText>
    <ms:attributionText xml:lang="en">Trilingual Term Dictionary (TTD) by
        Institute for Language and Speech Processing - Athena Research.↵
↪ Center
        used under Creative Commons Attribution Non Commercial No.↵
↪ Derivatives
        4.0 International
        (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode,
        https://creativecommons.org/licenses/by-nc-nd/4.0/). Source:
        http://hdl.handle.net/11500/ATHENA-0000-0000-5837-0
        (CLARIN:EL)</ms:attributionText>
    </ms:DatasetDistribution>
    <ms:personalDataIncluded>>false</ms:personalDataIncluded>
    <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
</ms:LexicalConceptualResource>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```


26.3 3. Tool/Services

26.3.1 Single tool

```

<?xml version="1.0" encoding="utf-8"?>
  <ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↳ metadata-schema/CLARIN-SHARE.xsd">
<ms:metadataCreationDate>2015-09-14</ms:metadataCreationDate>
<ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
<ms:metadataCurator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCurator>
<ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
<ms:metadataCreator>
  <ms:actorType>Person</ms:actorType>
  <ms:surname xml:lang="en">Person_Surname</ms:surname>
  <ms:givenName xml:lang="en">Person_Name</ms:givenName>
</ms:metadataCreator>
<ms:sourceOfMetadataRecord>
  <ms:repositoryName xml:lang="el"> </ms:repositoryName>
  <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
  <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
</ms:sourceOfMetadataRecord>
<ms:DescribedEntity>
  <ms:LanguageResource>
    <ms:entityType>LanguageResource</ms:entityType>
    <ms:resourceName xml:lang="el">
      </ms:resourceName>
    <ms:resourceName xml:lang="en">ILSP Language Identification System</
↳ ms:resourceName>
    <ms:resourceShortName xml:lang="en">ILSP LangId</ms:resourceShortName>
    <ms:description xml:lang="el">
      .
      , , , ,
      greeklish,
      .</ms:description>
    <ms:description xml:lang="en">The ILSP Language Identification System is a tool_
↳ used for
    language identification in digital texts. The tool performs language_
↳ identification
    for Greek, Greeklish, English, German, Dutch and French; it can also be used_
↳ for
    other languages upon provision of specific supplementary external files for_
↳ each new
    language.</ms:description>
  <ms:LRIIdentifier ms:LRIIdentifierScheme="http://purl.org/spar/datacite/handle"
>http://hdl.handle.net/11500/ATHENA-0000-0000-23E7-4</ms:LRIIdentifier>
  <ms:version>1.0.0 (automatically assigned)</ms:version>

```

(continues on next page)

(continued from previous page)

```

<ms:additionalInfo>
  <ms:email>person@phs.uoa.gr</ms:email>
</ms:additionalInfo>
<ms:contact>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>
    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
  </ms:Person>
</ms:contact>
<ms:citationText xml:lang="el"> -
  (2015). . Version 1.0.0
  (automatically assigned). [Software (Tool/Service)]. CLARIN:EL.
  http://hdl.handle.net/11500/ATHENA-0000-0000-23E7-4</ms:citationText>
<ms:citationText xml:lang="en">Institute for Language and Speech Processing -
↪Athena
  Research Center (2015). ILSP Language Identification System. Version 1.0.0
  (automatically assigned). [Software (Tool/Service)]. CLARIN:EL.
  http://hdl.handle.net/11500/ATHENA-0000-0000-23E7-4</ms:citationText>
<ms:keyword xml:lang="en">text</ms:keyword>
<ms:resourceCreator>
  <ms:Organization>
    <ms:actorType>Organization</ms:actorType>
    <ms:organizationName xml:lang="el">
      </ms:organizationName>
    <ms:organizationName xml:lang="en">Institute for Language and Speech
      Processing</ms:organizationName>
    <ms:website>http://www.ilsp.gr</ms:website>
  </ms:Organization>
</ms:resourceCreator>
<ms:intendedApplication>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageTechnology</ms:LTClassRecommended>
</ms:intendedApplication>
<ms:intendedApplication>
  <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageIdentification</ms:LTClassRecommended>
</ms:intendedApplication>
<ms:isDocumentedBy>
  <ms:title xml:lang="el"> </ms:title>
  <ms:title xml:lang="en">Language identification in digital texts</ms:title>
  <ms:DocumentIdentifier
    ms:DocumentIdentifierScheme="http://purl.org/spar/datacite/url"
    >http://www.ilsp.gr/homepages/protopapas/pdf/Protopapas_2004_LangIDsubm.
↪pdf</ms:DocumentIdentifier>
  </ms:isDocumentedBy>
  <ms:LRSubclass>
    <ms:ToolService>
      <ms:lrType>ToolService</ms:lrType>
      <ms:function>
        <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↪LanguageIdentification</ms:LTClassRecommended>

```

(continues on next page)

(continued from previous page)

```

        </ms:function>
        <ms:SoftwareDistribution>
          <ms:SoftwareDistributionForm>http://w3id.org/meta-share/meta-share/
↪ sourceCode</ms:SoftwareDistributionForm>
          <ms:downloadLocation>http://www.hiddenLocation.org</
↪ ms:downloadLocation>
          <ms:licenceTerms>
            <ms:licenceTermsName xml:lang="en">BSD 2-Clause "Simplified"
              License</ms:licenceTermsName>
            <ms:licenceTermsURL>https://opensource.org/licenses/BSD-2-Clause
↪ </ms:licenceTermsURL>
            <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪ unspecified</ms:conditionOfUse>
            <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪ allowsDirectAccess</ms:licenceCategory>
            <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↪ </ms:licenceCategory>
            <ms:LicenceIdentifier
              ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
↪ share/SPDX"
              >BSD-2-Clause</ms:LicenceIdentifier>
          </ms:licenceTerms>
          <ms:attributionText xml:lang="el">
            . : -
            . : BSD 2-Clause "Simplified" License
              (https://opensource.org/licenses/BSD-2-Clause). :
              http://hdl.handle.net/11500/ATHENA-0000-0000-23E7-4
              (CLARIN:EL)</ms:attributionText>
          <ms:attributionText xml:lang="en">ILSP Language Identification↵
↪ System by
              Institute for Language and Speech Processing - Athena Research↵
↪ Center
              used under BSD 2-Clause "Simplified" License
              (https://opensource.org/licenses/BSD-2-Clause). Source:
              http://hdl.handle.net/11500/ATHENA-0000-0000-23E7-4
              (CLARIN:EL)</ms:attributionText>
        </ms:SoftwareDistribution>
        <ms:languageDependent>true</ms:languageDependent>
        <ms:inputContentResource>
          <ms:processingResourceType>http://w3id.org/meta-share/meta-share/
↪ corpus</ms:processingResourceType>
          <ms:language>
            <ms:languageTag>el-Latn</ms:languageTag>
            <ms:languageId>el</ms:languageId>
            <ms:scriptId>Latn</ms:scriptId>
            <ms:languageVarietyName xml:lang="en">Greeklish</
↪ ms:languageVarietyName>
          </ms:language>
          <ms:language>
            <ms:languageTag>el-Grek</ms:languageTag>
            <ms:languageId>el</ms:languageId>
            <ms:scriptId>Grek</ms:scriptId>

```

(continues on next page)

(continued from previous page)

```

        </ms:language>
        <ms:language>
            <ms:languageTag>fr</ms:languageTag>
            <ms:languageId>fr</ms:languageId>
        </ms:language>
        <ms:language>
            <ms:languageTag>en</ms:languageTag>
            <ms:languageId>en</ms:languageId>
        </ms:language>
        <ms:language>
            <ms:languageTag>de</ms:languageTag>
            <ms:languageId>de</ms:languageId>
        </ms:language>
        <ms:language>
            <ms:languageTag>nl</ms:languageTag>
            <ms:languageId>nl</ms:languageId>
        </ms:language>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
    <ms:mediaType>
        </ms:inputContentResource>
        <ms:outputResource>
            <ms:processingResourceType>http://w3id.org/meta-share/meta-share/
    <corpus></ms:processingResourceType>
        <ms:language>
            <ms:languageTag>el-Latn</ms:languageTag>
            <ms:languageId>el</ms:languageId>
            <ms:scriptId>Latn</ms:scriptId>
            <ms:languageVarietyName xml:lang="en">Greeklish</
    <ms:languageVarietyName>
        </ms:language>
        <ms:language>
            <ms:languageTag>el-Grek</ms:languageTag>
            <ms:languageId>el</ms:languageId>
            <ms:scriptId>Grek</ms:scriptId>
        </ms:language>
        <ms:language>
            <ms:languageTag>fr</ms:languageTag>
            <ms:languageId>fr</ms:languageId>
        </ms:language>
        <ms:language>
            <ms:languageTag>en</ms:languageTag>
            <ms:languageId>en</ms:languageId>
        </ms:language>
        <ms:language>
            <ms:languageTag>de</ms:languageTag>
            <ms:languageId>de</ms:languageId>
        </ms:language>
        <ms:language>
            <ms:languageTag>nl</ms:languageTag>
            <ms:languageId>nl</ms:languageId>
        </ms:language>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
    <ms:mediaType>

```

(continues on next page)

(continued from previous page)

```

        </ms:outputResource>
        <ms:evaluated>false</ms:evaluated>
    </ms:ToolService>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

26.3.2 Combined tools

```

<?xml version="1.0" encoding="utf-8"?>
<ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.
gr/metadata-schema/CLARIN-SHARE.xsd">
    <ms:metadataCreationDate>2019-04-02</ms:metadataCreationDate>
    <ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
    <ms:metadataCurator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCurator>
    <ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
    <ms:metadataCreator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCreator>
    <ms:sourceOfMetadataRecord>
        <ms:repositoryName xml:lang="el"> </ms:repositoryName>
        <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
        <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
    </ms:sourceOfMetadataRecord>
    <ms:DescribedEntity>
        <ms:LanguageResource>
            <ms:entityType>LanguageResource</ms:entityType>
            <ms:resourceName xml:lang="en">Voyant Tools</ms:resourceName>
            <ms:description xml:lang="el"> Voyant Tools
                /
                .    Voyant Tools    Stéfan
                Sinclair Geoffrey Rockwell 2016

                . ,
                Voyant Tools : -
                , -
                /
                , - ,
                , ,
                , -

```

(continues on next page)

(continued from previous page)

```

(      ),
    , -
    Voyant Tools.</ms:description>
<ms:description xml:lang="en">Voyant Tools is a web-based text reading and
↪analysis
    environment. It is a scholarly project that is designed to facilitate
↪reading and
    interpretive practices for digital humanities students and scholars as well
↪as for
    the general public. What you can do with Voyant: --Use it to learn how
    computers-assisted analysis works. Check out our examples that show you how
↪to do
    real academic tasks with Voyant. --Use it to study texts that you find on
↪the web or
    texts that you have carefully edited and have on your computer. --Use it to
↪add
    functionality to your online collections, journals, blogs or web sites so
↪others can
    see through your texts with analytical tools. --Use it to add interactive
↪evidence
    to your essays that you publish online. Add interactive panels right into
↪your
    research essays (if they can be published online) so your readers can
↪recapitulate
    your results. --Use it to develop your own tools using our functionality and
    code.</ms:description>
<ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
    >http://hdl.handle.net/11500/ATHENA-0000-0000-5827-2</ms:LRIdentifier>
<ms:version>1.0.0 (automatically assigned)</ms:version>
<ms:additionalInfo>
    <ms:landingPage>https://voyant-tools.org/</ms:landingPage>
</ms:additionalInfo>
<ms:contact>
    <ms:Person>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:Person>
</ms:contact>
<ms:contact>
    <ms:Person>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:Person>
</ms:contact>
<ms:citationText xml:lang="en">Rockwell, Geoffrey; Sinclair, Stéfan (2019).
↪Voyant
    Tools. Version 1.0.0 (automatically assigned). [Software (Tool/Service)].
↪CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-5827-2</ms:citationText>

```

(continues on next page)

(continued from previous page)

```

<ms:keyword xml:lang="en">text</ms:keyword>
<ms:resourceCreator>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>
    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
  </ms:Person>
</ms:resourceCreator>
<ms:resourceCreator>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>
    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
  </ms:Person>
</ms:resourceCreator>
<ms:LRSubclass>
  <ms:ToolService>
    <ms:lrType>ToolService</ms:lrType>
    <ms:function>
      <ms:LTClassRecommended>http://w3id.org/meta-share/omtd-share/
↳LinguisticAnalysis</ms:LTClassRecommended>
    </ms:function>
    <ms:SoftwareDistribution>
      <ms:SoftwareDistributionForm>http://w3id.org/meta-share/meta-share/
↳webService</ms:SoftwareDistributionForm>
      <ms:executionLocation>https://voyant-tools.org/</
↳ms:executionLocation>
      <ms:webServiceType>http://w3id.org/meta-share/meta-share/unspecified
↳</ms:webServiceType>
      <ms:licenceTerms>
        <ms:licenceTermsName xml:lang="en">GNU General Public License v3.
↳0 or
          later</ms:licenceTermsName>
        <ms:licenceTermsURL>https://www.gnu.org/licenses/gpl-3.0-
↳standalone.html</ms:licenceTermsURL>
        <ms:licenceTermsURL>https://opensource.org/licenses/GPL-3.0</
↳ms:licenceTermsURL>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↳unspecified</ms:conditionOfUse>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↳allowsDirectAccess</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↳</ms:licenceCategory>
        <ms:LicenceIdentifier
          ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
↳share/SPDX"
          >GPL-3.0-or-later</ms:LicenceIdentifier>
        </ms:licenceTerms>
        <ms:licenceTerms>
          <ms:licenceTermsName xml:lang="en">Creative Commons Attribution_
↳4.0
            International</ms:licenceTermsName>

```

(continues on next page)

(continued from previous page)

```

        <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
    ↪ legalcode</ms:licenceTermsURL>
        <ms:licenceTermsURL>https://creativecommons.org/licenses/by/4.0/
    ↪ </ms:licenceTermsURL>
        <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
    ↪ attribution</ms:conditionOfUse>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
    ↪ allowsDirectAccess</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
    ↪ allowsProcessing</ms:licenceCategory>
        <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
    ↪ </ms:licenceCategory>
        <ms:LicenceIdentifier
            ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
    ↪ share/SPDX"
            >CC-BY-4.0</ms:LicenceIdentifier>
    </ms:licenceTerms>
    <ms:attributionText xml:lang="el">Voyant Tools. : Geoffrey
    ↪ Rockwell and Stéfan Sinclair. : GNU General Public License v3.0
    ↪ or
        later (https://www.gnu.org/licenses/gpl-3.0-standalone.html,
        https://opensource.org/licenses/GPL-3.0) and Creative Commons
        Attribution 4.0 International
        (https://creativecommons.org/licenses/by/4.0/legalcode,
        https://creativecommons.org/licenses/by/4.0/). :
        http://hdl.handle.net/11500/ATHENA-0000-0000-5827-2
        (CLARIN:EL)</ms:attributionText>
    <ms:attributionText xml:lang="en">Voyant Tools by Geoffrey Rockwell
    ↪ and
        Stéfan Sinclair used under GNU General Public License v3.0 or
    ↪ later
        (https://www.gnu.org/licenses/gpl-3.0-standalone.html,
        https://opensource.org/licenses/GPL-3.0) and Creative Commons
        Attribution 4.0 International
        (https://creativecommons.org/licenses/by/4.0/legalcode,
        https://creativecommons.org/licenses/by/4.0/). Source:
        http://hdl.handle.net/11500/ATHENA-0000-0000-5827-2
        (CLARIN:EL)</ms:attributionText>
    </ms:SoftwareDistribution>
    <ms:languageDependent>>false</ms:languageDependent>
    <ms:inputContentResource>
        <ms:processingResourceType>http://w3id.org/meta-share/meta-share/
    ↪ corpus</ms:processingResourceType>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
    ↪ ms:mediaType>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Pdf</
    ↪ ms:dataFormat>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Rtf</
    ↪ ms:dataFormat>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Xml</
    ↪ ms:dataFormat>
        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/ConllU</
    ↪ ms:dataFormat>

```

(continues on next page)

(continued from previous page)

```

        <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Html</
↪ms:dataFormat>
        </ms:inputContentResource>
        <ms:outputResource>
        <ms:processingResourceType>http://w3id.org/meta-share/meta-share/
↪corpus</ms:processingResourceType>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/image</
↪ms:mediaType>
        </ms:outputResource>
        <ms:outputResource>
        <ms:processingResourceType>http://w3id.org/meta-share/meta-share/
↪corpus</ms:processingResourceType>
        <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
↪ms:mediaType>
        </ms:outputResource>
        <ms:evaluated>false</ms:evaluated>
    </ms:ToolService>
</ms:LRSubclass>
</ms:LanguageResource>
    </ms:DescribedEntity>
</ms:MetadataRecord>

```

26.4 4. Language Descriptions

```

<?xml version="1.0" encoding="utf-8"?>
    <ms:MetadataRecord xmlns:ms="http://w3id.org/meta-share/meta-share/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://w3id.org/meta-share/meta-share/ https://inventory.clarin.gr/
↪metadata-schema/CLARIN-SHARE.xsd">
    <ms:metadataCreationDate>2015-09-10</ms:metadataCreationDate>
    <ms:metadataLastDateUpdated>2021-05-28</ms:metadataLastDateUpdated>
    <ms:metadataCurator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCurator>
    <ms:compliesWith>http://w3id.org/meta-share/meta-share/CLARIN-SHARE</ms:compliesWith>
    <ms:metadataCreator>
        <ms:actorType>Person</ms:actorType>
        <ms:surname xml:lang="en">Person_Surname</ms:surname>
        <ms:givenName xml:lang="en">Person_Name</ms:givenName>
    </ms:metadataCreator>
    <ms:sourceOfMetadataRecord>
        <ms:repositoryName xml:lang="el"> </ms:repositoryName>
        <ms:repositoryName xml:lang="en">ATHENA RC Repository</ms:repositoryName>
        <ms:repositoryURL>http://inventory.clarin.gr</ms:repositoryURL>
    </ms:sourceOfMetadataRecord>
    <ms:DescribedEntity>
        <ms:LanguageResource>
            <ms:entityType>LanguageResource</ms:entityType>

```

(continues on next page)

(continued from previous page)

```

<ms:resourceName xml:lang="el">PANACEA    n- (n-grams)
  </ms:resourceName>
<ms:resourceName xml:lang="en">PANACEA Environment Corpus n-grams EL
  (Greek)</ms:resourceName>
<ms:description xml:lang="el">    n-
  ( n = 1 - 5)    n- / /
    ,
    .            PANACEA
  (http://www.panacea-lr.eu),    7 .
    FPC,
    .            31,71 .
    .            2011.</ms:description>
<ms:description xml:lang="en">PANACEA Environment Corpus n-grams EL (Greek) 1.0
↳ contains      Greek word n-grams and Greek word/tag/lemma n-grams in the "Environment"
↳ (ENV)         domain. N-grams are accompanied by their observed frequency counts. The
↳ length of    the n-grams ranges from unigrams (single words) to five-grams. The data were
               collected in the context of PANACEA (http://www.panacea-lr.eu), an EU-FP7
↳ Funded       Project under Grant Agreement 248064. The n-gram counts were generated from
↳ crawled      Web pages that were automatically detected to be in the Greek language and
↳ were         automatically classified as relevant to the ENV domain. The collection
↳ consisted of approximately 31.71 million tokens. Data collection took place in the summer
↳ of           2011.</ms:description>
<ms:LRIdentifier ms:LRIdentifierScheme="http://purl.org/spar/datacite/handle"
  >http://hdl.handle.net/11500/ATHENA-0000-0000-23DA-3</ms:LRIdentifier>
<ms:version>1.0</ms:version>
<ms:additionalInfo>
  <ms:landingPage>http://nlp.ilsp.gr/panacea/D4.3/data/201209/gms/env_el/
↳ README.txt</ms:landingPage>
</ms:additionalInfo>
<ms:contact>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>
    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
  </ms:Person>
</ms:contact>
<ms:contact>
  <ms:Person>
    <ms:actorType>Person</ms:actorType>
    <ms:surname xml:lang="en">Person_Surname</ms:surname>
    <ms:givenName xml:lang="en">Person_Name</ms:givenName>
  </ms:Person>
</ms:contact>

```

(continues on next page)

(continued from previous page)

```

<ms:citationText xml:lang="el"> -
    (2015). PANACEA n- (n-grams) .
    Version 1.0. [Model (n-gram model)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-23DA-3</ms:citationText>
<ms:citationText xml:lang="en">Institute for Language and Speech Processing -
↳ Athena
    Research Center (2015). PANACEA Environment Corpus n-grams EL (Greek).
↳ Version 1.0.
    [Model (n-gram model)]. CLARIN:EL.
    http://hdl.handle.net/11500/ATHENA-0000-0000-23DA-3</ms:citationText>
<ms:keyword xml:lang="en">monolingual</ms:keyword>
<ms:domain>
    <ms:categoryLabel xml:lang="en">environment</ms:categoryLabel>
    <ms:DomainIdentifier
↳ ClarinEL_domainClassification"
        ms:DomainClassificationScheme="http://w3id.org/meta-share/meta-share/
        >Clarin_Domain002</ms:DomainIdentifier>
    </ms:domain>
<ms:resourceCreator>
    <ms:Organization>
        <ms:actorType>Organization</ms:actorType>
        <ms:organizationName xml:lang="el">
            </ms:organizationName>
        <ms:organizationName xml:lang="en">Institute for Language and Speech
            Processing</ms:organizationName>
        <ms:website>http://www.ilsp.gr</ms:website>
    </ms:Organization>
</ms:resourceCreator>
<ms:creationStartDate>2011-06-01</ms:creationStartDate>
<ms:creationEndDate>2011-08-31</ms:creationEndDate>
<ms:fundingProject>
    <ms:projectName xml:lang="en">Platform for Automatic, Normalized Annotation.
↳ and
        Cost-Effective Acquisition of Language Resources for Human
        Language</ms:projectName>
    <ms:website>http://www.panacea-lr.eu</ms:website>
    <ms:website>http://panacea-lr.eu</ms:website>
    <ms:grantNumber>ICT-248064</ms:grantNumber>
    <ms:fundingType>http://w3id.org/meta-share/meta-share/euFunds</
↳ ms:fundingType>
    <ms:funder>
        <ms:Organization>
            <ms:actorType>Organization</ms:actorType>
            <ms:organizationName xml:lang="el"> </ms:organizationName>
            <ms:organizationName xml:lang="en">European Commission</
↳ ms:organizationName>
            <ms:website>https://ec.europa.eu/info/index_en</ms:website>
        </ms:Organization>
    </ms:funder>
</ms:fundingProject>
    <ms:creationDetails xml:lang="en">automatic web crawling, automatic language
↳ detection,

```

(continues on next page)

(continued from previous page)

```

data preprocessing (boilerpipe filtering, lemmatization &
tagging)</ms:creationDetails>
<ms:isCreatedBy>
  <ms:resourceName xml:lang="en">boilerpipe library</ms:resourceName>
  <ms:LRIIdentifier ms:LRIIdentifierScheme="http://purl.org/spar/datacite/url"
    >https://code.google.com/archive/p/boilerpipe/</ms:LRIIdentifier>
  <ms:version>unspecified</ms:version>
</ms:isCreatedBy>
<ms:isCreatedBy>
  <ms:resourceName xml:lang="en">ILSP Lemmatizer</ms:resourceName>
  <ms:version>unspecified</ms:version>
</ms:isCreatedBy>
<ms:isCreatedBy>
  <ms:resourceName xml:lang="en">ILSP Feature-based multi-tiered POS
    Tagger</ms:resourceName>
  <ms:version>unspecified</ms:version>
</ms:isCreatedBy>
<ms:isDocumentedBy>
  <ms:title xml:lang="en">PANACEA Environment Corpus n-grams EL 1.0 README</
ms:title>
  <ms:DocumentIdentifier
    ms:DocumentIdentifierScheme="http://purl.org/spar/datacite/url"
    >http://nlp.ilsp.gr/panacea/D4.3/data/201209/gms/env_el/README.txt</
ms:DocumentIdentifier>
  </ms:isDocumentedBy>
  <ms:LRSubclass>
    <ms:LanguageDescription>
      <ms:lrType>LanguageDescription</ms:lrType>
      <ms:LanguageDescriptionSubclass>
        <ms:NGramModel>
          <ms:ldSubclassType>NGramModel</ms:ldSubclassType>
          <ms:baseItem>http://w3id.org/meta-share/meta-share/word</
ms:baseItem>
          <ms:order>5</ms:order>
        </ms:NGramModel>
      </ms:LanguageDescriptionSubclass>
      <ms:LanguageDescriptionMediaPart>
        <ms:LanguageDescriptionTextPart>
          <ms:ldMediaType>LanguageDescriptionTextPart</ms:ldMediaType>
          <ms:mediaType>http://w3id.org/meta-share/meta-share/text</
ms:mediaType>
          <ms:lingualityType>http://w3id.org/meta-share/meta-share/
monolingual</ms:lingualityType>
          <ms:language>
            <ms:languageTag>el</ms:languageTag>
            <ms:languageId>el</ms:languageId>
          </ms:language>
        </ms:LanguageDescriptionTextPart>
      </ms:LanguageDescriptionMediaPart>
      <ms:DatasetDistribution>
        <ms:DatasetDistributionForm>http://w3id.org/meta-share/meta-share/
downloadable</ms:DatasetDistributionForm>

```

(continues on next page)

(continued from previous page)

```

        <ms:downloadLocation>http://www.hiddenLocation.org</
↪ms:downloadLocation>
        <ms:samplesLocation>http://nlp.ilsp.gr/panacea/D4.3/data/201209/gms/
↪env_el/ENV_EL_1000.3gms.sample</ms:samplesLocation>
        <ms:samplesLocation>http://nlp.ilsp.gr/panacea/D4.3/data/201209/gms/
↪env_el/ENV_EL_wpl_1000.3gms.sample</ms:samplesLocation>
        <ms:distributionTextFeature>
            <ms:size>
                <ms:amount>14954020.0</ms:amount>
                <ms:sizeUnit>http://w3id.org/meta-share/meta-share/five-gram
↪</ms:sizeUnit>
            </ms:size>
            <ms:size>
                <ms:amount>13683940.0</ms:amount>
                <ms:sizeUnit>http://w3id.org/meta-share/meta-share/four-gram
↪</ms:sizeUnit>
            </ms:size>
            <ms:size>
                <ms:amount>3860716.0</ms:amount>
                <ms:sizeUnit>http://w3id.org/meta-share/meta-share/bigram</
↪ms:sizeUnit>
            </ms:size>
            <ms:size>
                <ms:amount>9767383.0</ms:amount>
                <ms:sizeUnit>http://w3id.org/meta-share/meta-share/trigram</
↪ms:sizeUnit>
            </ms:size>
            <ms:size>
                <ms:amount>435189.0</ms:amount>
                <ms:sizeUnit>http://w3id.org/meta-share/meta-share/unigram</
↪ms:sizeUnit>
            </ms:size>
            <ms:dataFormat>http://w3id.org/meta-share/omtd-share/Text</
↪ms:dataFormat>
        </ms:distributionTextFeature>
        <ms:licenceTerms>
            <ms:licenceTermsName xml:lang="en">Creative Commons Attribution,
↪Share
                Alike 4.0 International</ms:licenceTermsName>
            <ms:licenceTermsURL>https://creativecommons.org/licenses/by-sa/4.
↪0/legalcode</ms:licenceTermsURL>
            <ms:licenceTermsURL>https://creativecommons.org/licenses/by-sa/4.
↪0/</ms:licenceTermsURL>
            <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪attribution</ms:conditionOfUse>
            <ms:conditionOfUse>http://w3id.org/meta-share/meta-share/
↪shareAlike</ms:conditionOfUse>
            <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsDirectAccess</ms:licenceCategory>
            <ms:licenceCategory>http://w3id.org/meta-share/meta-share/
↪allowsProcessing</ms:licenceCategory>
            <ms:licenceCategory>http://w3id.org/meta-share/meta-share/public
↪</ms:licenceCategory>

```

(continues on next page)

(continued from previous page)

```

    <ms:LicenceIdentifier
      ms:LicenceIdentifierScheme="http://w3id.org/meta-share/meta-
↪share/SPDX"
      >CC-BY-SA-4.0</ms:LicenceIdentifier>
    </ms:licenceTerms>
    <ms:attributionText xml:lang="el">PANACEA    n-
      (n-grams) . :
      - . : Creative Commons Attribution
      Share Alike 4.0 International
      (https://creativecommons.org/licenses/by-sa/4.0/legalcode,
      https://creativecommons.org/licenses/by-sa/4.0/). :
      http://hdl.handle.net/11500/ATHENA-0000-0000-23DA-3
      (CLARIN:EL)</ms:attributionText>
    <ms:attributionText xml:lang="en">PANACEA Environment Corpus n-grams↵
↪EL
      (Greek) by Institute for Language and Speech Processing - Athena
      Research Center used under Creative Commons Attribution Share↵
↪Alike 4.0
      International (https://creativecommons.org/licenses/by-sa/4.0/
↪legalcode,
      https://creativecommons.org/licenses/by-sa/4.0/). Source:
      http://hdl.handle.net/11500/ATHENA-0000-0000-23DA-3
      (CLARIN:EL)</ms:attributionText>
    </ms:DatasetDistribution>
    <ms:personalDataIncluded>>false</ms:personalDataIncluded>
    <ms:sensitiveDataIncluded>>false</ms:sensitiveDataIncluded>
  </ms:LanguageDescription>
</ms:LRSubclass>
</ms:LanguageResource>
</ms:DescribedEntity>
</ms:MetadataRecord>

```

FULL SCHEMA DOCUMENTATION

You can browse the full schema documentation here:

- [Metadata record \(Base item\)](#)
- [Language Resource](#)
 - [Tool/Service](#)
 - [Corpus](#)
 - [Language description](#)
 - [Lexical/Conceptual resource](#)

The metadata building blocks (either attributes or elements) are all listed alphabetically.

CLARIN:el XML Schema Documentation

Table of Contents

- [Schema Document Properties](#)
- [Global Declarations](#)
 - [Attribute: AccessRightsStatementScheme](#)
 - [Attribute: AudioGenreClassificationScheme](#)
 - [Attribute: ClassificationScheme](#)
 - [Attribute: DocumentIdentifierScheme](#)
 - [Attribute: DomainClassificationScheme](#)
 - [Attribute: FunderIdentifierScheme](#)
 - [Attribute: IdentifierScheme](#)
 - [Attribute: ImageGenreClassificationScheme](#)
 - [Attribute: LicenceIdentifierScheme](#)
 - [Attribute: LRIdentifierScheme](#)
 - [Attribute: MetadataRecordIdentifierScheme](#)
 - [Attribute: OrganizationIdentifierScheme](#)
 - [Attribute: PersonalIdentifierScheme](#)
 - [Attribute: ProjectIdentifierScheme](#)
 - [Attribute: RepositoryIdentifierScheme](#)
 - [Attribute: ResourceIdentifierScheme](#)
 - [Attribute: socialMediaOccupationalAccountType](#)
 - [Attribute: SpeechGenreClassificationScheme](#)
 - [Attribute: SubjectClassificationScheme](#)
 - [Attribute: TextGenreClassificationScheme](#)
 - [Attribute: TextTypeClassificationScheme](#)
 - [Attribute: VideoGenreClassificationScheme](#)
 - [Element: abstract](#)
 - [Element: accesses](#)
 - [Element: accessLocation](#)
 - [Element: accessRights](#)
 - [Element: AccessRightsStatementIdentifier](#)
 - [Element: Actor](#)
 - [Element: actualUse](#)
 - [Element: actualUseDetails](#)
 - [Element: additionalHWRequirements](#)
 - [Element: additionalInfo](#)
 - [Element: address](#)
 - [Element: addressSet](#)
 - [Element: affiliatedOrganization](#)
 - [Element: affiliation](#)
 - [Element: age](#)
 - [Element: ageGroup](#)
 - [Element: ageGroupOfParticipants](#)
 - [Element: ageRangeEndOfParticipants](#)
 - [Element: ageRangeStartOfParticipants](#)
 - [Element: algorithm](#)
 - [Element: algorithmDetails](#)
 - [Element: alias](#)
 - [Element: alternativeTitle](#)
 - [Element: amount](#)
 - [Element: annotatedElement](#)
 - [Element: annotation](#)
 - [Element: annotationEndDate](#)
 - [Element: annotationMode](#)
 - [Element: annotationModeDetails](#)
 - [Element: annotationReport](#)

If you want to find out more about an element, simply click on it and you will be transferred to its full description. In the following example two elements are presented: **download location** and **duration of audio**.

Element: downloadLocation

Name	downloadLocation
Type	ms: httpURI
Nilable	no
Abstract	no
Documentation	A URL where the language resource (mainly data but also downloadable software programmes or forms) can be downloaded from
Application Data	<div><identifier> http://w3id.org/meta-share/meta-share/downloadLocation </identifier> <label xml:lang="en"> download location </label> <note xml:lang="en"> rule: must be filled in when SoftwareDistributionForm = downloadable; must not be filled in for web services; optional in all other cases </note></div>

XML Instance Representation

<ms:downloadLocation> ms:httpUri </ms:downloadLocation>

Schema Component Representation

<xs:element name="downloadLocation" type="ms:httpURI"/>

Element: durationOfAudio

Name	durationOfAudio
Type	ms: Duration
Nilable	no
Abstract	no
Documentation	Specifies the duration of the audio recording including silences, music, pauses, etc.
Application Data	<div><identifier> http://w3id.org/meta-share/meta-share/durationOfAudio </identifier> <label xml:lang="en"> duration of audio </label></div>

XML Instance Representation

<ms:durationOfAudio>
 <ms:amount> ... </ms:amount> [1]
 <ms:durationUnit> ... </ms:durationUnit> [1]
</ms:durationOfAudio>

Schema Component Representation

<xs:element name="durationOfAudio" type="ms:Duration"/>

ACTIONS ON RESOURCES

28.1 I. Per resource status and user type

You can manage your resources by performing actions on them. All actions are available from the list of resources you have **permissions** on. To reach this list go to your *dashboard* and choose:

- View my resources from *My resources* if you are a *curator* (C),
- View my supervision tasks from *My repository* if you are a *supervisor* (S),
- View my validation tasks from my *Validation tasks* if you are a *validator* (V).

You will be presented with a list of resources. Each row is dedicated to a single resource and has an **Actions** button as shown in the image. When you click on it, you will see a dropdown list.

The screenshot displays a table of resources. The first resource is 'Lexical/Conceptual Resource (Mandatory elements)' with version '1.0.0 (automatically assigned)' and user 'curator'. It has a checkbox, an 'Actions' dropdown button, and a 'syntactically valid' status button. The second resource is 'Testrecognizer' with version '1.0.0 (automatically assigned)' and user 'curator'. It has a checkbox, an 'Edit Metadata' button, a 'Copy record' button, a 'Submit for publication' button, and a 'draft' status button. The third resource is 'democorpus' with version '1.0.0 (automatically assigned)' and user 'curator'. It has a checkbox, a 'Delete Metadata' button, and a 'syntactically valid' status button. The table also includes columns for 'Hosted Resources Repository', 'Created', and 'Updated' dates.

Resource Name	Version	User	Hosted Resources Repository	Created	Updated	Actions	Status
Lexical/Conceptual Resource (Mandatory elements)	1.0.0 (automatically assigned)	curator	Lexical/Conceptual resource	30 May 2021	09 June 2021	Actions	syntactically valid
Testrecognizer	1.0.0 (automatically assigned)	curator	Tool/Service	08 June 2021	08 June 2021	Edit Metadata, Copy record, Submit for publication	draft
democorpus	1.0.0 (automatically assigned)	curator	Corpus	08 June 2021	08 June 2021	Delete Metadata	syntactically valid

Alternatively, you can choose a resource by clicking on its name and see the available actions from its *view page*¹.

¹ You will not be able to see the resource view page until you have provided all the correct *mandatory* elements. Only then the metadata record, which is syntactically valid, can be **saved** and presented as a view page.

draft

syntactically valid

submitted

approved

published

Lexical/Conceptual Resource (Mandatory elements)

LexicalConceptualResource

Version: 1.0.0 (automatically assigned)

This is a resource with only the mandatory (and mandatory upon condition) elements filled in.

Language

English Modern Greek (1453-)

Keyword

mandatory, testing

Overview

Access

clarin:el

Actions

Edit Metadata

Copy record

Submit for publication

Delete Metadata

The actions you see are the actions **you have the right to perform** depending on your *role* and the resource status. The table below shows the actions each type of user has permission to do at every stage of the *publication lifecycle*. Click on the action and you will be transferred to the respective section to learn more about it.

Action/Status	draft	synt.valid	submit- ted	ap- proved	pub- lished	re- quested ²	unpub- lished
<i>edit metadata</i>	C	C, S	S	S	n/a	n/a	S
<i>copy record</i>	n/a	C, S	C, S	C, S	C, S	C, S	C, S
<i>submit for publica- tion</i>	n/a	C, S	n/a	n/a	n/a	n/a	n/a
<i>assign supervisor</i>	n/a	n/a	S	n/a	n/a	n/a	n/a
<i>assign validator</i>	n/a	n/a	S	n/a	n/a	n/a	n/a
<i>validate</i>	n/a	n/a	V	n/a	n/a	n/a	n/a
<i>publish</i>	n/a	n/a	n/a	S	n/a	n/a	S
<i>unpublish</i>	n/a	n/a	n/a	n/a	S	S	n/a
<i>request to unpublish</i>	n/a	n/a	n/a	n/a	C	n/a	n/a
<i>create new version</i>	n/a	n/a	n/a	n/a	C, S	C, S	n/a
<i>delete metadata</i>	C	C, S	S	S	n/a	n/a	S

- n/a = not applicable

Actions can be performed on a **single** resource (from the resource *view page*) or on **multiple** resources (via the list of resources you have **permissions** on). To select more than one resources click on the box on the left of their name and choose an action from the **action box** at the top of the page. If you select resources **the status of which is different**, you will see **only the available actions** for all resources (e.g. export metadata as shown in the image below).

² Resources which have been requested to unpublish.

Curator

Username

Resources

- + Corpus (18)
- + Lexical/Conceptual resource (5)
- + Tool/Service (4)
- + Language description (1)

Status

- + draft (12)
- + submitted (7)
- + syntactically valid (7)
- + published (1)
- + unpublished (1)

Has data

- + no (18)

28 search results

Resource name	Actions	Status
Language Description (Mandatory elements)	1.0.0 (automatically assigned) Language description Hosted Resources Repository Created: 30 May 2021 Updated: 22 June 2021	metadata validator validator_hosted@gmail.com curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com
testing	1.0.0 (automatically assigned) Corpus Hosted Resources Repository Created: 08 June 2021 Updated: 22 June 2021	metadata validator validator_hosted@gmail.com curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com

Agatha Christie (NE tagged)

In your list of resources, there are also filters to facilitate your tasks. The following table shows the available filters per user type. Click on the filter and you will be transferred to the respective section to learn more about it.

Filter	Curator	Supervisor	Validator
<i>resource type</i>	X	X	X
<i>status</i>	X	X	X
<i>has data</i>	X	X	X
<i>processable</i>	X	X	X
<i>action required</i>		X	
<i>metadata valid</i>		X	X
<i>legally valid</i>		X	X
CLARIN:EL compatible service		X	X
<i>requested for unpublish</i>		X	
<i>validator assignment required</i>		X	
<i>for information</i>		X	
<i>metaresources</i>		X	

28.2 II. The actions

28.2.1 Edit

By clicking on **edit metadata** you will be transferred to the resource metadata record in the *metadata editor*³. You can edit it as many times as you like. After editing you can save the resource **as draft** or if you have filled in all the *mandatory* metadata and you are satisfied with the resource description you can finally **save** it. By saving it, it acquires the **syntactically valid** status. It remains editable and can also be edited by the repository supervisor (something not possible while it was at the draft stage where it is accessible only by the curator).

- If you are a **curator** you will be able to edit your resource from the draft to the syntactically valid stage.
- If you are a **supervisor** you will be able to edit a resource from the syntactically valid stage up to when it is approved and then again when it is *unpublished*.

³ Henceforth **editor**.

A metadata record will have to go through editing **again** if the metadata validator has rejected it. This will result in the resource returning to the **syntactically valid** status again. As a curator you will receive an email with the validator's comments and you will have to edit the resource and submit it for publication when you are over.

28.2.2 Copy record

You can copy a metadata record after you have saved it and it is **syntactically valid**.

The screenshot shows a table of metadata records. The first record is 'Lexical/Conceptual Resource (Mandatory elements)' with version '1.0.0 (automatically assigned)'. It is assigned to a 'curator' and is in the 'Hosted Resources Repository'. The status is 'syntactically valid'. An 'Actions' dropdown menu is open for this record, showing options: 'Edit Metadata', 'Copy record', 'Submit for publication', and 'Delete Metadata'. The 'Copy record' option is highlighted with a blue box. The status 'syntactically valid' is shown in a pink box next to the record.

Click on the action and a new window will open, in which you must name the resource and define its version⁴.

Please fill in the following fields in order to create a copy of Lexical/Conceptual Resource (Mandatory elements)

resource name *
LCR (mandatory elements) - COPY

The official name or title of the language resource/technology

Version
2.0.0

The new version of the record that will be created. Recommended format: major_version.minor_version.patch (see semantic versioning guidelines at <http://semver.org>)

cancel Create copy

When you are done, click on **Create copy**. You will be informed that the copy of the metadata record was successfully created while transferred to its view page.

⁴ Version 1.0.0 is automatically given to resources in which the *mandatory* element has not been filled in.

STATUS

You can continue editing your metadata record while the status is draft (syntactically valid); when you are satisfied with it, you can submit it for publication; the resource will be validated and published by the supervisor of the repository, or, if required, you will be contacted for further information

draft **syntactically valid** submitted approved published

LCR (mandatory elements) - COPY

LexicalConceptualResource

Version: 1.1.0

This is a resource with only the mandatory (and mandatory upon condition) elements filled in.

clarin:el

Actions

Language: English, Modern Greek (1453)

Keyword: mandatory, testing

Success

The copied metadata record is **syntactically valid** and must go through the stages described in the *publication lifecycle*.

28.2.3 Submit for publication

This action takes place on **syntactically valid** resources. When you are satisfied with the metadata description of a resource (either as a curator or supervisor), you can **submit it for publication**.

Resource name	Actions	Status
LCR (mandatory elements) - COPY 1.1.0 Lexical/Conceptual resource Hosted Resources Repository Created: 20 June 2021 Updated: 20 June 2021 curator curator_hosted@gmail.com	Actions Edit Metadata Copy record Submit for publication	syntactically valid
TestK_1.1 1.0.0 Tool/Service Hosted Resources Repository Created: 27 April 2021 Updated: 12 June 2021 curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Submit for publication	syntactically valid
Language Description (Mandatory elements) 1.0.0 (automatically assigned) Language description Hosted Resources Repository Created: 30 May 2021 Updated: 09 June 2021 metadata validator validator_hosted@gmail.c curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com	Delete Metadata Actions	

Record submitted for publication successfully.

- If you are a **supervisor** you will receive an email asking you to “assign validators to the following record, as it is ready for legal and metadata validation”.

28.2.4 Assign supervisor

Attention: This action is available only to **supervisors** and is required only when there are **more than one supervisors** in a repository. If there is only one supervisor in a repository, the system automatically assigns all resources to him/her.

Once a resource has been submitted for publication you will receive an email asking you to assign yourself as supervisor to the resource: *Please assign yourself as supervisor to the following record and then assign legal and metadata validators, as it is ready for validation.*

The screenshot shows a table of resources with two entries:

Grammar (Mandatory elements)	test
1.0.0 (automatically assigned) Language description Hosted Resources Repository Created: 22 June 2021 Updated: 22 June 2021	Corpus Hosted Resources Repository Created: 16 June 2021 Updated: 16 June 2021

Below the table, there are two rows of information:

curator user1@gmail.com validator assignment required yes	legally valid yes metadata valid validated
curator user2@gmail.com validator assignment required no	draft

An 'Actions' dropdown menu is open over the second row, showing two options: 'Assign supervisor' (with a person icon) and 'Copy record' (with a plus icon).

By clicking on the action, a new window will open asking you to select the user you wish to make supervisor of the resource.

The dialog box is titled 'Please select a supervisor for Grammar (Mandatory elements)'. It contains a label 'Assign supervisor *' and a dropdown menu. The dropdown menu is open, showing two options: 'supervisor1@gmail.com' and 'supervisor2@gmail.com'. At the bottom right of the dialog, there are 'Submit' and 'Cancel' buttons.

After you have selected the supervisor, click on submit and you will see a success message at the bottom right side of your page.



28.2.5 Assign validator

Attention: This action is available only to **supervisors**. In order to be able to assign a resource to the legal and metadata validators, you must first have assigned these roles to users in your repository. See [here](#) how to do this.

When a resource is submitted for publication, you will receive an email informing you that you must assign it to validators. If the submitted resource does **not** have content files, it is automatically considered **legally valid**⁵ and you only have to assign a metadata validator.

The screenshot shows a list of resources in the CLARIN interface. The first resource, 'TestK_1.1.1', is a Lexical/Conceptual resource with version 1.0.0, created and updated on 27 April 2021. It is assigned to a curator (curator_hosted@gmail.com) and a supervisor (supervisor_hosted@gmail.com). The 'legally valid' status is 'yes'. The second resource, 'TestK_1', is a Corpus resource with version 1.0.0, also created and updated on 27 April 2021. It is assigned to the same curator and supervisor. The 'legally valid' status is 'yes', and it has a 'has data' tag. The 'TestK_1' resource is highlighted with a yellow box, and the 'Assign metadata validator' action is visible in the dropdown menu.

If the submitted resource **has** content files, you have to assign a legal and a metadata validator.

The screenshot shows a list of resources in the CLARIN interface. The first resource, 'TestK_1.1.1', is a Lexical/Conceptual resource with version 1.0.0, created and updated on 27 April 2021. It is assigned to a curator (curator_hosted@gmail.com) and a supervisor (supervisor_hosted@gmail.com). The 'legally valid' status is 'yes', and the 'metadata valid' status is 'not validated'. The second resource, 'TestK_1', is a Corpus resource with version 1.0.0, also created and updated on 27 April 2021. It is assigned to the same curator and supervisor. The 'legally valid' status is 'not validated', and the 'metadata valid' status is 'not validated'. The 'TestK_1' resource is highlighted with a yellow box, and the 'Assign legal validator' action is visible in the dropdown menu. The third resource, 'Maria's corpus text multilingual', is a Corpus resource with version 1.0, created and updated on 27 April 2021. It is assigned to a curator (curator_hosted@gmail.com) and a supervisor (supervisor_hosted@gmail.com). The 'legally valid' status is 'not validated', and the 'metadata valid' status is 'not validated'. The 'test ingest' resource is also visible at the bottom of the list.

The procedure must be repeated for each validator type separately. You must click on the action (assign legal or metadata validator) and then you will be presented with a new window.

⁵ Since there are no content files, there is no need for licence policy; hence the resource is considered legally valid.

Please select a validator in order to validate TestK_1

Assign validator *

legalValidator_1@email.com
 legalValidator_2@email.com

Depending on the number of validators existing in your repository you will see a dropdown list. Choose the validator you want and then click on **submit**. A success message will appear at the right down side of your window.

When you return to the resources in your *supervision tasks*, you will be able to see the validators you have assigned to each resource, or the need to assign validators, as well as any comments they might have made during validation.

testing

1.0.0 (automatically assigned)
Corpus
Hosted Resources Repository
Created: 08 June 2021
Updated: 22 June 2021

metadata validator
validator_hosted@gmail.com
curator
curator_hosted@gmail.com
supervisor
supervisor_hosted@gmail.com
validator assignment required
no

Actions

syntactically valid

Review comments
[21/06/2021]Metadata validation review: The resource is bilingual but the same language has been used twice. Please check the metadata on linguality and languages.

TestK_1

1.0.0
Corpus
Hosted Resources Repository
Created: 27 April 2021
Updated: 21 June 2021
has data

legal validator
validator_hosted@gmail.com
curator
curator_hosted@gmail.com
supervisor
supervisor_hosted@gmail.com
validator assignment required
yes

Actions

submitted
legally valid
not validated
metadata valid
not validated

28.2.6 Publish

Attention: This action is available only to **supervisors**.

Once a resource is both legally and metadata valid, its status changes from **submitted** to **approved**. You will be notified by email that “it has been approved by the validators, therefore it is ready for publication”.

Language Description (Mandatory elements)

1.0.0 (automatically assigned)
 Language description
 Hosted Resources Repository
 Created: 30 May 2021
 Updated: 09 June 2021
 metadata validation date: 09 June 2021

metadata validator
 validator_hosted@gmail.com
 curator
 curator_hosted@gmail.com
 supervisor
 supervisor_hosted@gmail.com
 validator assignment required
 no

test_upload

1.0.0 (automatically assigned)
 Corpus
 Hosted Resources Repository
 Created: 20 April 2021
 Updated: 26 April 2021

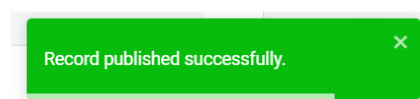
legal validator
 user1@athenarc.gr
 metadata validator
 user1@athenarc.gr
 curator
 admin@email.com

Actions

- Edit Metadata
- Publish
- Copy record
- Delete Metadata

approved
 legally valid
 yes
 metadata valid
 approved
 gally valid
 metadata valid

A message at the bottom right side of the page will inform you upon the successful publication of the resource and the resource curator will also be notified by email.



28.2.7 Request to unpublish

If you are a **curator** and you think that a published resource shouldn't be in the central inventory, you can ask for it to be unpublished.

Grammar (Mandatory elements)

1.0.0 (automatically assigned)
 Language description
 Hosted Resources Repository
 Created: 22 June 2021
 Updated: 22 June 2021

metadata validator
 user1@gmail.com
 curator
 user1@gmail.com
 supervisor
 user1@gmail.com

Actions

- Create new version
- Copy record
- Request to unpublish
- Unpublish

published

When you click on the action, a new window opens asking you to declare the reasons for your request.

Request to unpublish

Unpublication request reason *

The resource contains personal data and must be anonymized.

Please specify the reason for your unpublication request.

cancel

Request to unpublish

Once you indicate the reasons, press **Request to unpublish**; you will see a message, at the bottom right side of the page, that your request has been successfully submitted and in the list of resources, the resource status has changed.

Grammar (Mandatory elements)	
1.0.0 (automatically assigned)	metadata validator
Language description	user1@gmail.com
Hosted Resources Repository	curator
Created: 22 June 2021	user1@gmail.com
Updated: 22 June 2021	supervisor
	user1@gmail.com

Actions

requested for unpub...

28.2.8 Unpublish

Attention: This action is available only to **supervisors**.

Only **published** or **requested for unpubl**ish records can be unpublished.

[illegible]

When the action is successfully completed the resource status is set to **unpublished**.

Grammar (Mandatory elements)

<input type="checkbox"/> 1.0.0 (automatically assigned) Language description Hosted Resources Repository Created: 22 June 2021 Updated: 22 June 2021 metadata validation date: 22 June 2021	metadata validator user1@gmail.com curator user1@gmail.com supervisor user1@gmail.com validator assignment required no	Actions ▾	unpublished legally valid yes metadata valid yes
--	---	-----------	--

test maria

Corpus Hosted Resources Repository Created: 16 June 2021 Updated: 16 June 2021	curator user2@gmail.com validator assignment required no	draft
---	---	-------

Record Unpublished successfully. ✕

28.2.9 Create New Version

Once a resource is published, you can create a new version of it, if needed, e.g. due to updates.

tool1

<input type="checkbox"/> 1.0 Tool/Service Hosted Resources Repository Created: 28 April 2021 Updated: 15 June 2021	metadata validator validator_hosted@gmail.com curator curator_hosted@gmail.com supervisor supervisor_hosted@gmail.com validator assignment required no	Actions ▾	published legally valid yes metadata valid yes
--	---	-----------	--

Click on the action and a new window will open. Fill in the new version number and the date and then click on **create new version**.

Please fill in the following fields in order to create a new version for the record tool1

Version *
2.1.1

The new version of the record that will be created. Recommended format: major_version.minor_version.patch (see semantic versioning guidelines at <http://semver.org>)

version date
2021-06-01

The date of the LRT version (latest update of the particular version if possible)

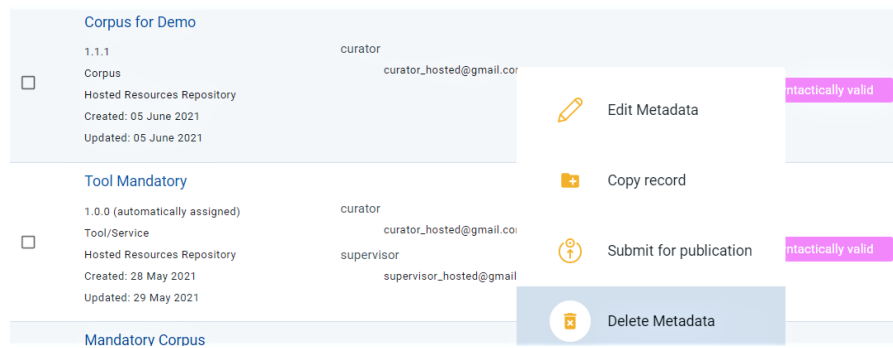
cancel Create new version

You will see a success message at the bottom right side of your page.


 Success

28.2.10 Delete metadata

If you are not satisfied with the description of a resource, you can delete its metadata record.



When the action has been completed successfully, you will see a success message at the bottom right side of your page. The metadata record you deleted does no longer exist.


 Record Deleted successfully.

28.2.11 Validate

Attention: This action is available only to **validators**.

When a resource has been assigned to you, you will receive an email informing you that you must validate it. From your *validation tasks* find the resource and click on its name. You **cannot** perform any actions from the resource list but you **can** from the resource view page. Click on the actions dropdown list as shown in the image below. You will be presented with the appropriate option⁶ of validation. The process described is the same both for metadata and legal validation.

⁶ Resources that do not have content files are automatically considered **legally valid** and only go through metadata validation.

STATUS

Your metadata record has been submitted for technical validation by the CLARIN-EL team and can no longer be edited; you will be notified when it is published or, if needed, for further information

draft syntactically valid **submitted** approved published

testing
Corpus
Version: 1.0.0 (automatically assigned)
Testing resource

clarin:el

Actions ▾

Metadata validate

Language
Aragonese

Keyword
testing, and testing again

Corpus subclass
raw corpus

When you click on the action, a new window will open asking you to accept or reject the metadata record.

Please fill in the following fields in order to validate testing

☐ Approve

☐ Reject

Submit Cancel

- If you choose **accept** and submit, the window closes and the resource is **metadata valid**.
- If you choose **reject**, a new window opens where you must write the reasons for rejecting the resource.

Please fill in the following fields in order to validate testing

☐ Approve

☒ Reject

Reasons *

The resource is bilingual but the same language has been used twice. Please check the metadata on linguality and languages.

Please state the reasons for rejecting this record

Submit Cancel

When you return to the resources found in your validation tasks, you will see your comments.

testing

1.0.0 (automatically assigned)

Corpus

Hosted Resources Repository

submitted: 08 June 2021

metadata validator

validator_hosted@gmail.com

curator

curator_hosted@gmail.com

supervisor

supervisor_hosted@gmail.com

syntactically valid

Review comments

[21/06/2021]Metadata validation review: The resource is bilingual but the same language has been used twice. Please check the metadata on linguality and languages.

Your comments will be communicated to the resource creator by email. The resource status will then return to **syntactically valid** so that it can be edited again according to your remarks.

INFORMATION ON LEGAL ISSUES

Click on the links to read about the [Privacy Policy](#) and the [Terms of Service](#) (these are found at the bottom of each page in the infrastructure, as shown in the image below).



Tip: Please, also check the [Recommended licensing scheme for Language Resources](#) which has been created to help you limit the complexity of licensing.

PUBLICATIONS

A full list of publications and presentations is available [here](#).

INDICES AND TABLES

- genindex
- modindex
- *Searching*