CLARIN:EL

Data Collection Policy

Mission

CLARIN:EL is the National Infrastructure for Language Resources and Technologies in Greece. **CLARIN:EL's Mission** is

- to **collect, document, curate** and **distribute** digital language resources, language technology tools and certified online language processing services
- for the **support** of researchers, academics, students, language professionals, citizen scientists and the general public
- whose activities fall into the fields of Language studies, Linguistics, Digital Humanities and Social Sciences, Cultural Heritage, Language Technology, Artificial Intelligence, Computer Science, Cognitive Science, etc.

CLARIN:EL is part of the National Roadmap for Research Infrastructures of Greece and is the Greek part of the CLARIN ERIC European Infrastructure. CLARIN:EL services are offered to all Greek users, as well as to all users of the CLARIN ERIC European Infrastructure.

The CLARIN:EL Research Infrastructure offers access to:

- **digital language data** of various language modalities (written, spoken, multimodal, sign, lexical/conceptual, etc.) and in various media (text, audio, video, etc.)
- **language processing tools and web services** (such as tokenizers, part of speech taggers, dependency parsers, terminology extractors, information extractors, etc.)
- **the metadata of all resources** made available through the Infrastructure.

Access to the CLARIN:EL Infrastructure is open to the entire academic and research community, industry, but also to the public, in accordance with Open Data Principles and FAIR Data Principles.

Scope

CLARIN:EL focuses on **Language Resources (LRs)** and **Language Technologies (LTs)**. By the terms Language Resources and Language Technologies we refer to language data (written, spoken or multimodal) and to the tools/services/workflows used for their processing. CLARIN:EL welcomes the deposition of digital Greek Language Resources (LRs) of or related to any media type (text, sound, video, image). We distinguish the following categories:

- **raw data:** digital/digitised resources, such as
    - written texts (e.g., web texts, digitised books, online 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, images of texts etc.)
- **annotated data**
    - various types of annotations of texts, audio and multimedia data, automatically or manually created (e.g., morphosyntactically annotated texts, video annotations etc.)
- **lexical / conceptual resources**
    - various types of structured language data (e.g. word lists, lexica, dictionaries, thesauri, ontologies etc.)
- **language technology tools/applications**

- tools and integrated applications that perform various types of language processing (e.g., multilingual text alignment, morphological annotation, lemmatisation, parsing, knowledge extraction, semantic annotation etc.)
- visualisation tools (e.g., integrated environments for the presentation of texts, mutlimedia collections, processing results etc.).

All LRs shared through the CLARIN:EL infrastructure must meet a minimum set of requirements:

- they must be formally documented with the endorsed metadata model, i.e. the CLARIN-SHARE model
- they must include a clear indication of the licensing terms under which they are provided
- it is advisable they follow the standards and technical recommendations set by CLARIN:EL.

Data formats accepted

CLARIN:EL prefers data in a readily usable format, accessible to a variety of language related computing and technological settings, CLARIN:EL integrated services included. The recommended file formats are listed in the Recommended File Formats documentation, along with the factors the CLARIN:EL team has considered to determine which formats are recommended.

The digital preservation community recommends standard preservation formats because they encode the information in a software-independent way, they facilitate interoperability between systems and applications or they encode the information with a lossless algorithm and for that reason there is no data loss when files are 'saved as' and stored in these formats.

Files deposited in CLARIN:EL are recommended to be (where possible) in open, platform-independent or non-proprietary file formats. Data in obsolete, proprietary, or hard-to-use formats may still be accepted by CLARIN:EL, although these characteristics may compromise any future use of the data.

You can read more about our preservation strategy in the Preservation Policy documentation.

Removal of resources

Once it has been published, CLARIN:EL prefers not to remove a resource with a PID, as it may have been cited by other researchers. However, if serious grounds exist (for example, if a rights holder is concerned that s/he has found Language Resources on the CLARIN:EL repository for which s/he has not given permission) CLARIN:EL can remove the resource from the infrastructure or restrict or prevent access to the dataset on a temporary or permanent basis. In case a resource is removed from public view, its landing page is still accessible from the handle PID (tombstone page), with a flag indicating the removal of the resource.

Licensing

Licensing Language Resources (i.e., selecting the appropriate licence of use that legally binds the End-User to the terms and conditions of using the resource) is a key concern of the CLARIN:EL Research Infrastructure.

With a firm orientation towards the creation of an openness culture and the relevant ecosystem for Language Resources, CLARIN:EL Research Infrastructure fosters Open Data policies. Thus, resources and services should ideally be *open* or *shared at least for research purposes*. However, legacy IPR issues connected with a resource are respected. In all cases, Language Resources must be *offered according to certain formal legal conditions and terms* clearly indicated in the *licence text*. Each Language Resource may be associated with more than one licence, if access to it is given under different conditions for different users and/or intended uses.

Access to the metadata records of language resources and processing services is open to all (i.e., registered and unregistered users) through the catalogue of the CLARIN:EL Central Inventory. CLARIN:EL registered users have access to use the language resources and

processing services and furthermore, to add their own resources or services to the institutional repository within CLARIN:EL to which they belong.

CLARIN:EL metadata are licenced under the Creative Commons Attribution International licence (CC-BY) version 4.0 or higher. All CLARIN:EL providers are obliged to allow the harvesting of their metadata.

Recommended licencing scheme for Language Resources

To limit the complexity of licensing, a range of recommended licenses are provided by CLARIN:EL in the form of templates for the Language Resource providers to choose from. The CLARIN:EL model licensing scheme, with a firm orientation towards the creation of an openness culture and the relevant ecosystem for Language Resources, is organised on the following axes:

- **Creative Commons** licences (CC, starting with Creative Commons Zero (CC-0) and all possible combinations along the CC differentiation of rights of use) for datasets and **Free Open Source Software (FOSS)** licences for s/w are the first level of legal machinery applied.
- The second legal layer is a set of licenses that allow use and exploitation of the Resources while permitting the Language Resource Owner to have full control over the Resource distribution. These **META-SHARE No Redistribution** licences will effectively help get "closed" resources safely out to the community.

The sets of licensing templates that are proposed can be found in https://www.clarin.gr/en/support/legal > Tab Recommended licences.

Versioning

A new version of a Language Resource should be created when the data of an existing Language Resource are updated, i.e., (re)processed, corrected or appended with additional data. The same applies for Language Technologies when a new unique state of the software is released.

Moreover, a new version of a Language Resource or a Language Technology should be created when the metadata of the resource are significantly modified. Correction of typos or rephrasing of description should not be a reason for a new version, whereas changes in the resource name, addition of creators, licences, size, etc. require the creation of a new version.

A new version is welcomed in CLARIN:EL by copying the metadata of the previous release and altering the version number and date on the newly created record, as well as adding the relationship between the versions in the metadata. Upon that new version record, the depositor can make changes through the editor forms and upload the new data. The new version gets its own handle PID, while the concept handle PID, that is, the handle PID that represents all versions of the LR -the concept of the data and the ensemble of the versions-, resolves on the latest versioned record of the LR.

Confidentiality and Privacy

CLARIN:EL prefers data that are open or shared at least for research purposes. Datasets including personal or sensitive data have to be pseudonymised or anonymised, so that the individuals cannot be identified. Consent forms allowing publication of the data supplied by the individuals mentioned should also accompany the datasets (where possible). The pseudonymised or anonymised data can be shared with an Open Licence.

Reuse

Since 2015 CLARIN:EL has been managed as an infrastructure for researchers, academics, students, language professionals, citizen scientists and the general public, to deposit and share their language data, as well as to download, use data to their research and analyse and process the data with online integrated processing tools and services.

We encourage users to deposit to the data repository by providing them with a friendly metadata editor for describing and uploading their resources. Anyone using data from the CLARIN:EL repository is expected to cite or reference this work as they would any other scientific research, even if the licence does not explicitly require users to do so. To facilitate

proper citation of CLARIN:EL Language Resources, we provide a ready-to-go citation text for each LR, that users can copy.

To foster the reusability of data, CLARIN:EL exposes the metadata for harvesting, thus extending their discovery through other catalogues. To this end, it has developed converters into various metadata schemas, favoured by the community (e.g. Component MetaData Infrastructure (CMDI)) and generic ones (e.g. Dublin Core (DC), Open Language Archives Community (OLAC)). In this way, the resources can be harvested by CLARIN ERIC Virtual Language Observatory (VLO) and other repositories and infrastructures.

Responsibilities

CLARIN:EL provides storage, computational resources, cataloguing, curation, maintenance and distribution of resources, for language resources and processing services; therefore, CLARIN:EL is responsible for adhering to disciplinary and ethical norms for the specific activities.

The responsibility of adherence to disciplinary and ethical norms for collection/ creation and description of resources belongs to the resource provider (Member organizations, their members and individual researchers/academics); there are procedures and measures in place, catering for compliance with disciplinary and ethical norms (CLARIN:EL Terms of Service).

Additionally, CLARIN:EL provides legal aid to depositors uncertain how to treat, describe and distribute resources including personal/sensitive data.

Each organization joining CLARIN:EL nominates a Scientific Responsible who signs the Network Statutes, by which s/he declares that s/he undertakes the responsibility for all resources and/or services deposited at CLARIN:EL repository; this entails: clearance of IPR issues, clearance of the legal base for processing of any personal data contained in the resources (including the provision of relevant consent forms), provision of the resources with a clear license (preferably one of the open licenses proposed by CLARIN:EL), description of the resources using the common metadata schema of the infrastructure, keeping the descriptions always up-to-date, and adoption of the proposed standards and

best practices. The signature of the Statutes explicitly binds not just the signatory, but also all persons affiliated to the organization; the Scientific Responsible informs them about their responsibilities and the correct use of CLARIN:EL.

Individual providers are not members of the Network; thus, they are not bound by the Statutes and therefore need to sign a Depositor's Agreement.