CLARIN:EL Preservation Policy

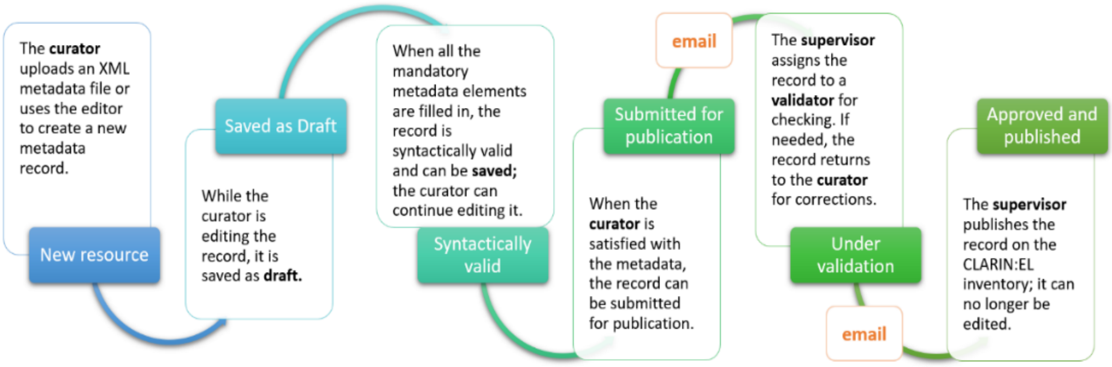
1. Purpose

CLARIN:EL repository collects Language Resources and Technologies (LRTs) with the aim to support researchers, academics, students, language professionals, citizen scientists and the general public whose activities fall into the fields of Language studies, Digital Humanities and Social Sciences, Cultural Heritage, Language Technology, Artificial Intelligence, Computer Science, Cognitive Science, etc. (https://www.clarin.gr/en/about/what-is-clarin).

This document describes how data and metadata are sustainably preserved within the CLARIN:EL platform. The document adheres to the terminology and preservation practices outlined by the Open Archival Information System (OAIS) Reference Model. An essential element of OAIS is the grouping of information into packages:

- SIP: submission information package
- AIP: archival information package
- DIP: dissemination information package

2. Archival workflow/procedure ("the LRT lifecycle")



The **curator** uploads an XML metadata file or uses the editor to create a new metadata record.

While the curator is editing the record, it is saved as **draft.**

When all the mandatory metadata elements are filled in, the record is syntactically valid and can be **saved;** the curator can continue editing it.

When the **curator** is satisfied with the metadata, the record can be submitted for publication.

The **supervisor** assigns the record to a **validator** for checking. If needed, the record returns to the **curator** for corrections.

The **supervisor** publishes the record on the CLARIN:EL inventory; it can no longer be edited.

The overall ingestion workflow is divided into the following consecutive phases (see also the diagram above):

Prepare & Ingest: Data depositors wishing to deposit resources such as datasets/tools/lexical resources/etc. are offered guidance and assistance via the repository web pages designed for this purpose (i.e., metadata editor and upload pages). Assistance and technical support about various issues such as data formats, metadata, and legal aspects is also offered through

- the Recommended Formats guidelines,
- the available online documentation about best practices on how to prepare, document and deposit their metadata and data,
- video tutorials (https://www.clarin.gr/en/support/user-manuals), and
- the respective helpdesks (https://www.clarin.gr/en/support/helpdesks).

CLARIN:EL collects resources as described in the Data Collection Policy. Once a data depositor has completed the creation of metadata for a resource and the preparation of the accompanying data (if available), then these are submitted and uploaded as Submission Information Packages (SIP). CLARIN:EL primarily acts as data repository, hosting content files (i.e., datasets and software) and their metadata, and secondarily acts as a catalog, hosting only metadata describing resources, while their datasets are not hosted within the CLARIN:EL infrastructure. The metadata are stored in a database and the data in a storage service. The resource at this stage is in "Draft mode".

Syntactic validation: To ensure completeness, accuracy and usability, all uploaded resources undergo automatic validation check for a) completeness of mandatory metadata elements b) conformity of the metadata values to the guidelines. The accompanied data are not checked automatically. Once the metadata are "syntactically valid", the respective Archival Information Packages (AIP) are created for metadata and data (if available). An AIP is the information package the system stores, preserves and sustains. At this stage the resources are still not publicly available and can be changed at any time by the depositor/curator.

Submit for publication: If the metadata are syntactically valid and the depositor is satisfied with them he/she can submit the resource for publication. Then the metadata are checked manually by validators at the technical level (inconsistencies), legal level (licence, IPR, privacy, sensitive/personal data), and metadata level (metadata quality issues). If the resource contains data, these are also manually checked. If needed, the resource (metadata and/or data) is returned to the curator/depositor for corrections or additions. When the validators give the green light, the resource is published in the catalogue and is available to all CLARIN:EL users. To ensure the integrity of the data, for every deposited file a checksum (MD5) is computed, which allows the CLARIN:EL team to identify corrupted files and restore correct versions if necessary.

Archival storage: The final checked data AIP is saved in a folder at a NAS server dedicated for CLARIN:EL. The final metadata AIP is stored in a PostgreSQL database. Items are retained indefinitely. CLARIN:EL uses Handle.net service to assign Persistent Identifiers (PIDs) to all resources, to ensure the accessibility of the data. Persistent identifiers are assigned when resources are published. When a new version of a dataset is published, a new metadata landing page is created and a new PID is generated. Thus, the already existing persistent identifier will continue to refer uniquely to the earlier version of the dataset. The new and the previous dataset are cross-referenced in their respective descriptive metadata.

Data may be removed at the request of the owner/copyright holder, in case s/he finds resources on CLARIN:EL, for which s/he has not given permission, granted a licence or which are not covered by a limitation or exception in national law[1]. Withdrawn resources' PIDs are retained indefinitely and point to tombstone pages with a note explaining the reason for withdrawal.

Access: All CLARIN:EL records are findable and accessible through the inventory pages (https://inventory.clarin.gr). The data and their associated metadata are made available to the user as two separate Dissemination Information Packages (DIP). A DIP is the information package created to distribute the digital content. In order for the content files of a resource to be accessible, two criteria must be met: (a) the resource needs to be

[1] https://www.clarin.gr/en/content/terms-service

provided under an open access licence, and (b) the resource content files must have been uploaded at CLARIN:EL or stored at an access point.[2] In contrast, the metadata DIP is always available for downloading, as metadata are available with a Creative Commons Attribution International licence (CC-BY) 4.0

3. Responsibilities

The CLARIN:EL team work together to ensure the integrity of the stored data & metadata and to guarantee their long-term accessibility. Backups are also taken on a daily basis, including an off-site copy (at GRNET). In the event of one of the partners withdraws from the CLARIN:EL Network, the respective resources will be transferred to another partner. If CLARIN:EL ceases to exist, the resources will be transferred to CLARIN ERIC.

The CLARIN:EL team cooperates with the IT administrators of ATHENA Research Center, who are responsible for providing the technical infrastructure (physical machines, networks, firewall etc.).

CLARIN:EL is responsible for the maintenance, review and revision of all its policies and documentation, including this one.

[2] https://clarin-platform-documentation.readthedocs.io/en/stable/all/1_BasicConcepts/Accessible.html#when-is-the-content-of-a-resource-accessible