The CLARIN:EL infrastructure

Maria Gavriilidou
ILSP / Athena RC, Greece
maria@athenarc.gr

Stelios Piperidis
ILSP / Athena RC, Greece
spip@athenarc.gr

Dimitrios Galanis
ILSP / Athena RC, Greece
galanisd@athenarc.gr

Juli Bakagianni
ILSP / Athena RC, Greece
julibak@athenarc.gr

Penny Labropoulou
ILSP / Athena RC, Greece
penny@athenarc.gr

Athanasia Kolovou
ILSP / Athena RC, Greece
akolovou@athenarc.gr

Dimitris Gkoumas
ILSP / Athena RC, Greece
dgkoumas@athenarc.gr

Miltos Deligiannis
ILSP / Athena RC, Greece
mdel@athenarc.gr

Kanella Pouli
ILSP / Athena RC, Greece
kanella@athenarc.gr

Iro Tsiouli
ILSP / Athena RC, Greece
tsiouli@athenarc.gr

Leon Voukoutis
ILSP / Athena RC, Greece
leon.voukoutis@athenarc.gr

Katerina Gkirtzou
ILSP / Athena RC, Greece
katerina.gkirtzou@athenarc.gr

Abstract

This paper presents the CLARIN:EL infrastructure, which comprises three pillars: the language resources and technologies Platform, the Portal and the Knowledge Centre. It serves as a comprehensive and interoperable environment that supports language-related research in the fields of language technology, language studies, digital humanities, and political and social sciences. The Platform facilitates language resources sharing by providers, and access to these resources by consumers. The Portal and the K-Centre offer complementary informative material and support services to the community, including awareness raising and training activities. This paper discusses the CLARIN:EL architecture, its design and implementation principles, the functionalities offered to the users, the support activities provided, and the network that enables its operation.

1 Introduction

CLARIN:EL is the Greek National Infrastructure for Language Resources & Technologies (LRTs), which comprises three interconnected pillars, namely, the Platform, the Portal and the NLP:EL Knowledge Centre. CLARIN:EL serves as a comprehensive and interoperable environment that supports language-related research in various fields, such as language technology (LT), linguistics, language studies, digital humanities (DH), political and social sciences. The Platform hosts the LRTs and provides the user interaction mechanisms through appropriate interfaces. The Portal and the K-Centre offer informative material and support the community as regards awareness, training, and knowledge transfer in Language Technology (LT) and Digital Humanities (DH).

The CLARIN:EL network supporting the Infrastructure consists of 14 organization members (9 Universities and 5 Research Centres). Anyone (academics, researchers, students, or the general public), whether affiliated to a network member organization or not, can have full access rights to the infrastructure. Registered users, authenticated via their academic or personal accounts, can upload their resources and/or tools, use the available resources and process them using the services offered by CLARIN:EL.

Resources provided by network members are associated, through the relevant metadata, to the specific organization, while those provided by individuals non-affiliated to a member organization are connected to the Hosted Resources Repository.

2 The CLARIN:EL Platform

CLARIN:EL is part of the Greek Roadmap for Research Infrastructures; currently, it forms part of the APOLLONIS infrastructure, together with DARIAH/DYAS. It supports the community through a certified CLARIN B-Centre and a K-Centre, it has been awarded the CoreTrustSeal, and it is listed in re3data.org.

The CLARIN:EL Platform consists of two interconnected subsystems: (a) a system for documenting (through the Metadata editor), and for storing, sharing, searching, retrieving, and downloading language resources (through the Central Inventory). It currently contains 802 resources (659 corpora, 92 lexical resources, 49 tools/services, and 2 language descriptions); and (b) the CLARIN:EL workspace: a system providing integrated services that perform core Natural Language Processing (NLP) tasks, i.e., sentence splitting, tokenization, PoS tagging, lemmatization, parsing, chunking, named entity recognition, as well as other tasks such as text classification and verbal aggression analysis. Moreover, it offers services that perform data format and character encoding conversion. The CLARIN:EL Central Inventory aims at providing a comprehensive overview of existing resources for Greek (alone or in combination with other languages). This includes listing metadata records and their CLARIN:EL hosted data or software, as well as metadata records whose corresponding data or software reside outside the CLARIN:EL platform, but also metadata records with no proper datasets (i.e., bibliographical lists, useful catalogs, etc.).

2.1 Documentation of resources with metadata

To ensure appropriate description of deposited resources, CLARIN:EL provides a rich metadata schema, CLARIN-SHARE, which allows coherent documentation to be added to each resource. The CLARIN-SHARE metadata model builds upon the META-SHARE metadata model (Gavriilidou et al., 2012), and its application profiles, ELG-SHARE (Labropoulou et al., 2022), ELRC-SHARE (Piperidis et al., 2018), and the MS-OWL ontology (Khan et al., 2022; McCrae et al., 2015), RDF/OWL representation of the model. The schema is intertwined with and supports the full lifecycle of language resources, from creation to annotation and usage. To foster the visibility and reusability of data, CLARIN:EL exposes metadata for harvesting, thus extending their discovery. The CLARIN-SHARE metadata schema has been converted into broadly used metadata schemas, such as CMDI, DC and OLAC, and the metadata records of the resources are harvested by repositories and infrastructures that support such metadata schemas (e.g., CLARIN Virtual Language Observatory/VLO).

Resource providers in CLARIN:EL have two options for creating metadata for their resources: create and upload XML files that adhere to the CLARIN-SHARE metadata schema or create metadata records using the platform's metadata editor. The submitted XML files are automatically checked for completeness and well-formedness. The metadata editor guides the providers to the complete description and uploading of their resources, it safeguards interoperability by using controlled vocabularies (where applicable) and assists them with examples and tips. The completion of the description and the automatic checking are followed by two rounds of manual assessment. The first round involves metadata and legal validation performed by human validators, followed by the final approval by the supervisor of each organization member, which triggers the resource's publication in the repository. Additionally, frequent quality checks, aiming at the completeness and correctness of metadata records and related datasets, are conducted centrally by the dedicated CLARIN:EL technical and metadata team.

2.2 Deposition of Language Resources

Data providers can get guidance and assistance via the Help pages[1] and the relevant Policy documents[2],[3]. The repository offers support on various issues such as data formats, metadata, and legal aspects, through

[1] CLARIN:EL User manual https://clarin-platform-documentation.readthedocs.io/en/stable/

[2] Data Collection policy: https://www.clarin.gr/sites/default/files/CLARINELDataCollectionPolicy.pdf

[3] Deposition documentation: https://clarin-platform-documentation.readthedocs.io/en/stable/all/4_Data/DataPreparation.html?highlight=deposit

the Recommended Formats guidelines, online documentation for metadata and data preparation, documentation and deposition, video tutorials, and helpdesks.

Resources deposited encompass written, spoken, or multimodal content. They can be texts, lexical resources, language models or processing tools, and they may pertain to modern Greek language, to earlier forms of Greek, or to other languages. To be processable by the integrated services of CLARIN:EL, the resources must be in one of the recommended text formats (plain text, XML, TMX, etc.).

CLARIN:EL favors and promotes Open Licenses; however, distribution and/or use restrictions on data are respected. Metadata are freely accessible to all with a CC-BY 4.0 license. The responsibility of clearing IPR and selecting the appropriate license lies with the resource provider. CLARIN:EL offers a variety of standard licenses for the provider to select from, and assistance through the Legal Helpdesk.

2.3 Searching and retrieval of Language Resources

Through the platform, users can search for resources using keywords and facets, or browse the resource catalogue and select a resource to view its full description; if interested, they can download it, or use the NLP services of CLARIN:EL to process it. CLARIN:EL presents resources in one central inventory. The catalogue lists metadata records (with or without data). Resources with no data fall into two categories: (a) metadata records in anticipation of the data that is not yet ready to be published and (b) meta-resources, i.e., ancillary resources (e.g., bibliographical lists, literature reviews, etc.). Browsing, viewing, and exporting metadata records, as well as downloading open-access resources are available to all users (registered or not), while user authentication and authorization are required for using the CLARIN:EL processing services or accessing restricted resources. The downloadability of a resource depends on the license defined by the provider. Legal and technical restrictions on resources are specified by the provider via the relevant metadata elements, based on which CLARIN:EL implements the resource's access policy. For the content files of a resource to be accessible, two criteria must be met: an open access license, and storage of the content files at an access point within CLARIN:EL or externally.

2.4 Processing of Language Resources

CLARIN:EL offers two types of tools/services for processing data: (a) single-task tools (e.g., lemmatizers, tokenizers, etc.) available as web services or as downloadable tools, accessible either from within CLARIN:EL or through an external link, and (b) the Workspace, which includes NLP web services integrated in the CLARIN:EL infrastructure. Each single-task web service can also be part of a workflow, i.e., of a pipeline of tools that operate at multiple levels of analysis (e.g., a workflow starting from sentence splitting, continuing with tokenization, POS tagging, lemmatization and concluding with named entity recognition). The Workspace is designed to support non-expert users in their data processing tasks, by providing ready-to-use pipelines of interoperable tools at a single click, thereby relieving them from the burden of selecting, downloading, and assembling tools from scratch. Users can process datasets hosted at CLARIN:EL or upload and process their own datasets (with a size limit of 2MB). In the former case, the processing results are stored in the infrastructure as a new resource, with its metadata automatically generated combining the metadata of the dataset and of the processing service used. The outcome of the processing is available both in the data format generated by the last service of the workflow (such as XML or XMI), and also in Comma Separated Values (CSV) format. The latter is provided for reasons of user-friendliness and interoperability (such files can be fed to other NLP services or to visualization tools).

2.5 User and Resource-lifecycle management

Registered users have full access to all CLARIN:EL functionalities and are considered potential resource providers, either as individuals or as members of their organization. The activities available to the users depend on their roles, which are defined in the User Management module (based on Keycloak). The User Roles schema comprises the roles of Curator (assigned to all registered users), Validator (assigned by the Supervisor), and Supervisor. These roles are involved in the creation and publishing of a resource, with varying rights: Supervisors have the full list of permitted actions, Validators are responsible for the legal and metadata quality check, while Curators have the basic set of actions.

The set of states of a resource in the process of being prepared for publication in the Central Inventory is depicted in the Resource Lifecycle; these states include the creation of a new resource by a curator (resource status: Draft), the automatic checking of its syntactic validity and conformity with the specifications (status: Ingested/Syntactically valid), the submission of the resource by the curator and the assignment of the resource to validators by the supervisor (status: Assigned for Validation); after the approval of the resource by the validators (status: Approved), the supervisor publishes the resource, making it visible on the CLARIN:EL inventory (status: Published).
Each member organization is responsible for its internal User Role Management, i.e. assigning roles (Curator, Validator, Supervisor), and for ensuring efficient creation, description, and publication of their own resources. Above this User Role Management at the level of member organizations, additional Validator and Supervisor roles exist at the central level of the CLARIN:EL Platform, with rights on all resources, facilitating quality assessment to ensure completeness and correctness.

3 The CLARIN:EL technical architecture

All the above functionalities are supported by the CLARIN:EL architecture, designed with state-of-the art technologies. Its subsystems are built with robust, open-source, scalable technologies, and consist of several applications: the PostgreSQL database (DB) used for storing several types of data, such as user data, the metadata records of the LRs, etc.; ElasticSearch for indexing; the repository backend, built using the Django web framework, offers REST services for managing metadata (import, create, update, delete), authorizes access to the resources etc.; the repository, based on the META-SHARE software[4], with many improved architectural choices, new functionalities and features; the User Interface that consists of web pages for searching/browsing the catalogue, the metadata editor for creating/updating metadata, admin pages for validating resources etc.; Keycloak, an identity and access management solution used for securing the applications; the integrated NLP services; a manager responsible for executing NLP services and a scheduler that decides where/when a user's processing request will be executed, to avoid platform overloading; and finally the User Dashboard.

The CLARIN:EL User Dashboard, available only to registered users, is a Single Page Application (SPA), built using React, providing users with a quick and easy way to monitor and track the performance of their tasks while interacting with various CLARIN:EL resources and services. The main features of the dashboard include customization (different dashboard for each user role), interactivity, real-time data display, alerts, and notifications. The User Dashboard serves both as an entry point to create and upload resources and use NLP processing services, and also as an overview page presenting the users' activity history (information on the resources created, tasks and processing jobs), as well as their editable profile.

All the above applications run as Docker containers at a Kubernetes (k8s) cluster, maintained and supported by the CLARIN:EL development team in ILSP/Athena RC. The checked LRs data are saved in a dedicated Network Attached Storage, while metadata are stored in PostgreSQL. CLARIN:EL uses Handle.net service to assign PIDs to resources, to ensure data accessibility. Procedures are in place for ensuring that hardware, software, and storage media containing archival copies of digital content are managed in accordance with security control, data protection and recovery standards.

4 User Support

CLARIN:EL provides several assistance mechanisms to support user needs. The Portal includes (i) information material on the infrastructure, the use of the Platform and the provided services, FAQs, etc., (ii) dissemination material (news, events, etc.), and (iii) educational material (video tutorials, scientific publications, and presentations). Publicly accessible Helpdesks enable interested parties to ask questions on technical, documentation and legal issues. The Portal, besides hosting the informative material mentioned above, also provides links to redirect the users to the Platform and the NLP:EL Knowledge Centre.

The K-Centre, which aims at actively supporting research and scientific advances in the relevant fields, is organized in two main units; *Knowledge*, where users can find LT tools and services, information on studies and curricula, educational and training material regarding NLP, and *Community*,

[4] https://github.com/metashare/META-SHARE

where they can be informed on NLP/LT teams in Greece, certified CLARIN K-Centres and National and European LRTs Infrastructures.

CLARIN:EL also provides detailed online documentation on the Platform and all its functionalities. The User Manual familiarizes the users (provider, curator or consumer) with the basic concepts of the Infrastructure, guides them through its main functionalities (browsing, searching, viewing, downloading, and processing Language Resources), instructs them how to create and manage their resources, and explains the role and the significance of the metadata schema used for this purpose. Finally, it provides crucial information on legal issues connected to the publication, distribution, and use of language resources (licensing), as well as those connected to the use of the infrastructure itself (Privacy policy and Terms of Use).

In addition to the management and the continuous updating of the material provided through the Portal and the NLP:EL Knowledge Centre, the CLARIN:EL team organizes training activities, such as webinars, workshops, summer schools, datathons, etc., (single or recurrent) in order to educate users on Language Technology and Digital Humanities, to raise awareness or to introduce new functionalities of the Platform.

5 Conclusion

We have presented the CLARIN:EL infrastructure, the functionalities available to the users, the design and implementation principles as reflected in its architecture, and the support activities provided to the community. Future steps include the maintenance and upgrading of the infrastructure's modules, the population of the repository with new resources, including workflows, the continuous support of its users, the enlargement of the network with new organization members and end-users, the interoperability with other infrastructures and repositories, and, finally, the hosting of outreach activities aiming to raise awareness about LT in the research community.

Acknowledgements

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.), under the Emblematic Action "The emerging landscape of digital work in Humanities in the context of the European infrastructures DARIAH and CLARIN" (Project Number: 7982), https://digital-landscape.gr/.

References

Gavriilidou, M., et al. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* pp. 1090-1097, http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf

Khan, A.F., et al. (2022). When linguistics meets web technologies. Recent advances in modelling linguistic linked data. Semantic Web Preprint (2022): 1-64. https://www.semantic-web-journal.net/content/when-linguistics-meets-web-technologies-recent-advances-modelling-linguistic-linked-open

Labropoulou, P., et al. (2022). Making metadata fit for next generation language technology platforms: The metadata schema of the European Language Grid. arXiv preprint arXiv:2003.13236

McCrae, J.P., et al (2015). One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. *The Semantic Web: ESWC 2015 Satellite Events*. ESWC 2015. Lecture Notes in Computer Science, vol 9341. Springer, Cham. https://doi.org/10.1007/978-3-319-25639-9_42

Piperidis, S., et al. (2018). Managing public sector data for multilingual applications development. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018*), Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1205.pdf