The CLARIN:EL infrastructure: Platform, Portal, K-Centre

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

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

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

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

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

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

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

Dimitris Gkoumas
ILSP / Athena RC, Greece
dgkoumas@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 deposition, curation and sharing of digital language resources (catering for providers' needs), and access to and automatic processing of these resources (catering for consumers' needs). The Portal offers informative material about CLARIN:EL and support services to the community, including dissemination, awareness raising and training activities. The Knowledge Centre promotes digital literacy in the scientific domains served, by providing information on studies, educational and training material and publications. This paper discusses the CLARIN:EL pillars, the technical architecture, its design and implementation principles, the functionalities offered to the users, the support activities provided, usage analytics and future steps.

1 Introduction

CLARIN:EL is the Greek National Infrastructure for Language Resources & Technologies, 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 Language Resources and Technologies and provides the user interaction mechanisms through appropriate interfaces. The Portal and the K-Centre cater for dissemination, offer informative material, and support the community as regards awareness, training, and knowledge transfer in LT and DH.

At the national level, CLARIN:EL is part of the Greek Roadmap for Research Infrastructures; currently, it forms part of the APOLLONIS infrastructure, together with DARIAH/DYAS. At the European level, it is the Greek branch of the CLARIN ERIC infrastructure (Branco et al. 2023). It supports the community through a certified CLARIN B-Centre and a K-Centre[1], it has been awarded the CoreTrustSeal[2], and it is listed in re3data (the registry of research data repositories)[3]. The CLARIN:EL network supporting the Infrastructure consists of 14 organization members (9 Universities and 5 Research Centres) practically covering the whole geographical area of the country. CLARIN:EL currently (February 2024) contains 793 resources (648 corpora, 94 lexical resources, 49 tools/services, and 2 language descriptions).

The CLARIN:EL infrastructure, with its Platform, the Portal, the NLP:EL K-Centre, the network of organizations supporting it, and, additionally, with the technical and operational interconnection with CLARIN ERIC, constitutes a valuable universe supporting language technology at the national and European levels. It serves both linguists and non-linguists, academics and non-academics, students, educators and language professionals, industry, and broad public. Through the use of concrete licensing schemes for the distribution of language resources and services, it actively safeguards Open Access and Open Science and promotes the requirements for open language data and language technology.

The following sections discuss the three pillars of CLARIN:EL with their components, the design of the infrastructure and the implementation principles; the functionalities of the infrastructure (deposition, documentation, curation of resources and services, search, retrieval and processing of resources, user management, and dashboard), the technical architecture, the support activities and materials offered to the users, as well as training and dissemination activities; finally, the paper provides infrastructure usage analytics, and concludes with future steps.

2 The CLARIN:EL Platform

The CLARIN:EL Platform consists of two interconnected subsystems: (a) the Repository, a system with all functionalities related to the provision of LRs, i.e., for depositing and documenting LRs, for curating their metadata through a specially designed metadata editor, for storing, sharing, searching, retrieving, and downloading LRs, and (b) the Workbench, a system providing integrated services that perform core Natural Language Processing (NLP) tasks, such as sentence splitting, tokenization, PoS tagging, lemmatization, parsing, chunking, named entity recognition (Prokopidis & Piperidis, 2020), as well as tasks such as text classification and verbal aggression analysis (Pontiki et al., 2020). Moreover, it offers preprocessing services that perform data format and character encoding conversion.

The Platform offers access to the resources through the Central Inventory, which provides a comprehensive catalogue of the resources and tools (corpora, lexical/conceptual resources, language descriptions and tools/services), for Greek (on its own or in combination with other languages).

The Central Inventory includes metadata records for (a) CLARIN:EL hosted data or software, (b) data or software that reside outside the CLARIN:EL platform, (c) reference data (i.e., bibliographical lists, useful catalogues, etc.). The Central Inventory can be filtered according to various features, such as resource type, language, domain, depositing organization, etc.

2.1 Deposition of Language Resources

Depositors of data or language processing services must be registered CLARIN:EL users, either affiliated to network member organizations or individuals. 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 assigned to the Hosted Resources Repository. Resources deposited encompass written, spoken, or multimodal content. They can be sets of 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 corpora must be in one of the recommended text formats (plain text, XML, TMX, etc.)

[1] https://www.clarin.eu/content/clarin-centres
[2] https://www.coretrustseal.org/
[3] https://www.re3data.org/

(Piperidis et al., 2016). Data providers can get guidance and assistance via the Help pages[4] and the relevant Policy documents, Data Collection Policy[5] and Deposition Documentation[6]. CLARIN:EL offers support on various issues such as data formats, metadata, and legal aspects, through the Recommended Formats guidelines[7], online documentation for metadata and data preparation, documentation and deposition, video tutorials, and helpdesks.

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

2.2 Documentation of resources with metadata

To ensure appropriate description of deposited resources, CLARIN:EL has adopted a rich metadata schema, CLARIN-SHARE[8], 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., 2020), 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 CLARIN-SHARE schema supports the objectives of the Platform. In particular, it ensures that all resources are discoverable and accessible by human users and machines (e.g., including links to URLs that offer direct access to the resource), addresses researchers' needs such as data citation (through the use of persistent and unique identifiers) and replicability of experiments (ensuring persistence and reusability of resources), and facilitates the integration of processing services with data resources (using as interoperability anchors the same attributes and values, such as data formats and annotation types, across resource types) and the documentation of processing activities and their outcomes (provenance/lineage metadata). Overall, the CLARIN-SHARE schema is an important factor contributing towards achieving FAIR data (Wilkinson et al., 2016).

The schema builds along three key concepts, each of which is associated with a distinctive set of metadata attributes:

- *resource type*, classifying resources into *corpora* (sets of text, audio, video or image files), *lexical/conceptual resources* (e.g. lexica, glossaries, ontologies, etc.), *language descriptions* (including models and computational grammars), and *tools/services*
- *media type*, which specifies the form or physical medium of the resource, i.e., *text*, *audio*, *image*, *video* and *numerical text* (referring to numerical data, such as biometrical, geospatial data, etc.). To cater for multimedia and multimodal language resources (e. g., a corpus of videos and subtitles, or a corpus of audio recordings and transcripts, a sign language corpus with videos and texts, etc.), language resources are represented as consisting of at least one media part, while multiple parts are also possible;
- *distribution*, which, following the DCAT vocabulary[9], refers to any physical form of the resource that can be distributed and deployed by end-users.

These concepts give rise to a modular structure, in which attributes are attached to the appropriate level. The level of "Language resource" includes properties common to all resource and media types, such as those used for identification purposes (title, description, etc.), recording provenance (creation, publication dates, creators, providers, etc.), contact points, etc. More technical features and classification attributes differ across resource and media types and are, thus, attached to combinations thereof; for example, a corpus may take properties specific to annotation processes, while the description of a

[4] CLARIN:EL User manual: https://clarin-platform-documentation.readthedocs.io/en/stable/
[5] Data Collection policy: https://www.clarin.gr/sites/default/files/CLARINELDataCollectionPolicy.pdf
[6] Deposition documentation: https://clarin-platform-documentation.readthedocs.io/en/stable/all/4_Data/DataPreparation.html?highlight=deposit
[7] Also in the Documentation: https://clarin-platform-documentation.readthedocs.io/en/stable/all/4_Data/FileFormats.html
[8] https://clarin-platform-documentation.readthedocs.io/en/stable/all/5_Metadata/Full.html
[9] https://www.w3.org/TR/vocab-dcat-2/

computational lexicon encodes whether it includes lemmas, examples, grammatical information, translation equivalents, etc. Technical features, such as format, size, information on licensing and mode of access are properties of the distribution.

The schema includes properties for the description of the full life cycle of language resources, from conception and creation to integration in applications and usage. All this information leads to a complex and demanding schema; to ensure flexibility and uptake by providers, only a carefully selected subset of these attributes are prescribed as *mandatory* and, thus, required to be filled in for metadata records to be approved for import in the Platform. The remaining attributes are *recommended* or *optional* i.e. providers are encouraged to fill them in and enhance the discoverability and usability of their resources.

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 mapped into the broadly used metadata schemas Dublin Core and OLAC, so that the metadata records of the resources are harvested by repositories and infrastructures that support the OAI/PMH harvesting protocol [10] (e.g., CLARIN Virtual Language Observatory/VLO[11]). In addition, CLARIN-SHARE has been implemented in the form of four profiles (one for each resource category) following the principles of the Component Metadata Infrastructure and integrated in the Component Metadata Registry[12], thus enhancing its reusability and interoperability in the CLARIN framework.

Resource providers in CLARIN:EL have two options for creating metadata for their resources: to create and upload XML files that adhere to the CLARIN-SHARE metadata schema or to create metadata records using the platform's metadata editor. Users can choose to validate XML files prior to uploading them. If any inconsistencies or missing metadata are detected, error messages will be displayed, prompting the user to address the issues. Once corrections have been implemented, the XML files can then be successfully uploaded. The infrastructure supports the uploading of files either individually or in batches. The submitted XML files are once more automatically checked for completeness and well-formedness.

If users opt to create metadata for their resource through the metadata editor, they will be presented with four specific forms tailored to the four types of resources. After selecting the type of resource they want to deposit, users are redirected to the relevant form (Figure 1), where they are required to fill in at least the mandatory metadata elements.

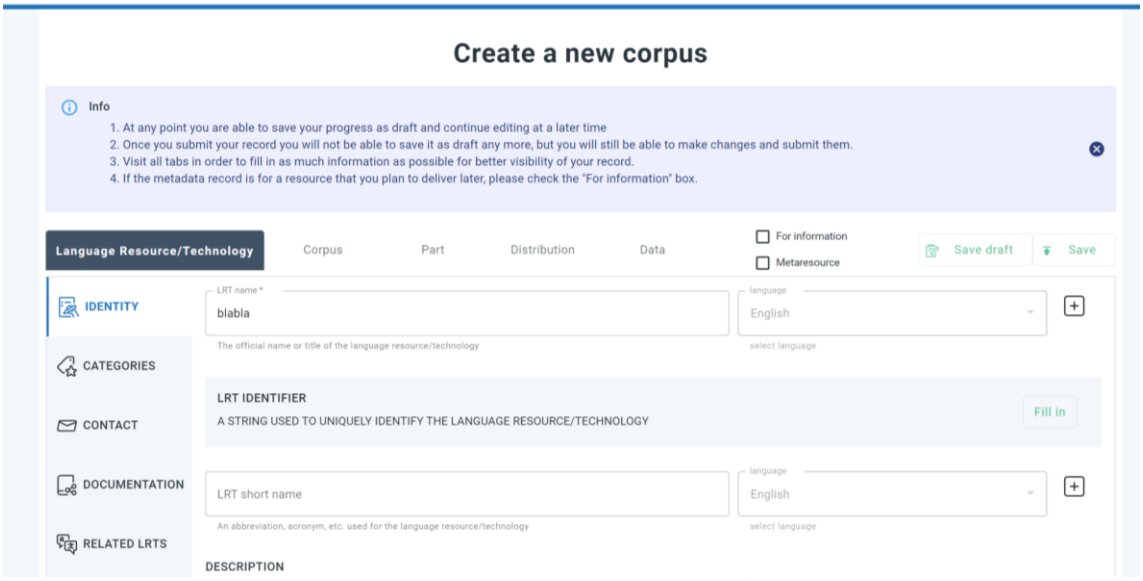

Figure 1: Creation of the resource record for a corpus

[10] https://www.openarchives.org/pmh/
[11] https://vlo.clarin.eu
[12] https://catalog.clarin.eu/ds/ComponentRegistry/#/

The metadata record cannot be saved until all required fields are filled in as indicated. Thus, the metadata editor guides the providers to the complete description and uploading of their resources, through iterative checks that make sure that all obligatory elements are filled in; it safeguards well-formedness and facilitates metadata interoperability by using controlled vocabularies (where applicable), and assists them with examples and tips.

2.3 Curation of metadata

The completion of the description by the depositor and the automatic checking by the system 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. 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. The dataset(s) associated with a metadata record are also automatically checked to assess their conformity with technical specifications as regards format and processability (interoperability with processing services). Providers are notified in cases of erroneous or sub-optimal codification of metadata, which they are requested to rectify.

Even when a resource needs to be removed from the central inventory, its findability is ensured; in these instances, users will be directed to a tombstone page, where a message indicates that the resource is (temporarily or permanently, as appropriate) unavailable.

2.4 Search and retrieval of Language Resources

CLARIN:EL presents resources in a Central Inventory (Figure 2), which users can search using keywords and facets, or browse and select a resource to view its full description; if interested, they can download it, or use the CLARIN:EL services to process it. The inventory lists metadata records containing the descriptions of the LRs (datasets or tools), which are normally (but not necessarily) accompanied by datasets or software code. 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.). The functionalities of 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 the use of the platform's 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.

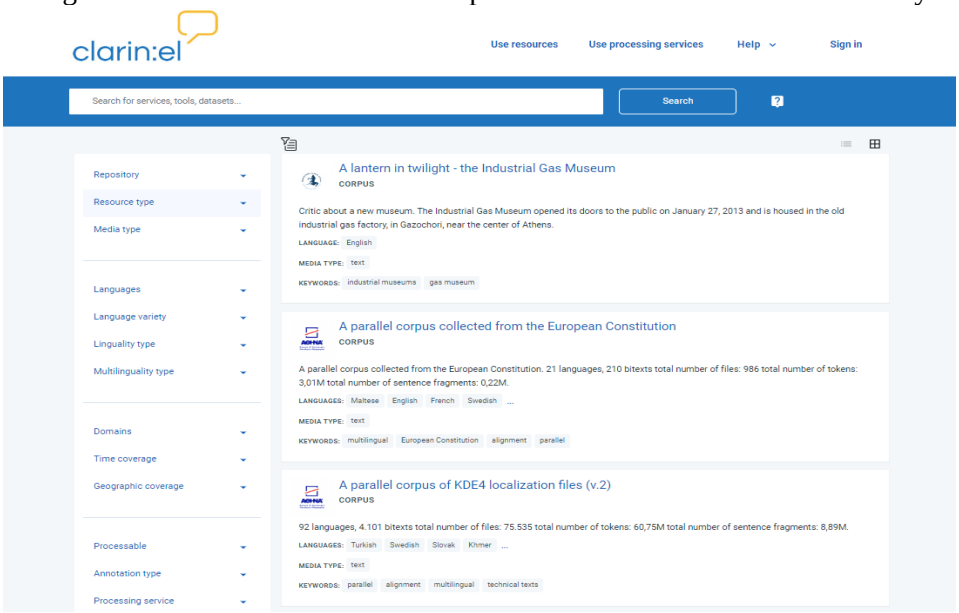

Figure 2: The CLARIN:EL Central Inventory

2.5 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 CLARIN:EL Workbench[13], 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 tokenization and sentence splitting, continuing with PoS tagging, lemmatization and concluding with named entity recognition). The Workbench is designed to support non-expert users in their data processing tasks, by providing ready-to-use pipelines (workflows) of interoperable tools at a single click, thereby relieving them from the burden of selecting and assembling tools in a workflow from scratch (Figure 3).

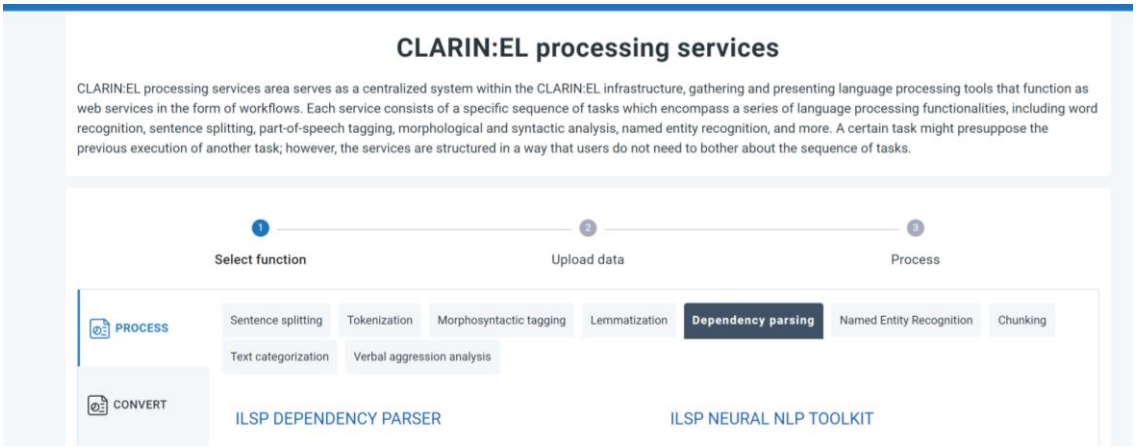

Figure 3: Selecting a processing service

Users can process datasets hosted at CLARIN:EL or upload and process their own datasets (with a size limit of 2MB, currently). In the former case, the outcome of the processing is stored in the infrastructure as a new resource, with its metadata automatically generated by combining the metadata of the dataset with those of the processing service used. The outcome of the processing is available both in the data format generated by the workflow (such as XML or XMI), and in Comma Separated Values (CSV) format. The latter is provided for reasons of user-friendliness and interoperability, given that such files can be fed to other NLP services or to visualization tools residing within CLARIN:EL and/or externally.

There are two prerequisites for the integration of a processing tool (e.g., a PoS tagger) to the CLARIN:EL infrastructure: (a) the tool has to be wrapped and offered as a web-service (e.g. via a RESTful API), and b) the software application that offers this web-service along with its dependencies (e.g., libraries, settings, operating system etc.) has to be packaged in a Docker image. The technical team of CLARIN:EL has the responsibility to deploy the Docker images of the provided tools/services at the infrastructure's Kubernetes cluster and integrate them in a software module called Workflows Manager which acts as a processing orchestrator. The orchestrator chains the tools into workflows, provides the required readers for different types of input datasets (e.g. TXT, XCES, TMX, etc.), exports the processing results to CSV, handles failures/timeouts, etc. Obviously, a large number of processing jobs can be initiated in CLARIN:EL by different registered users simultaneously. This in many cases can lead to overload; i.e., the deployed services might not be able to serve all the required requests, for example, due to memory issues and network timeouts that might occur. To avoid overstretching the capacity of the execution system, a scheduler has been implemented that decides when a processing job will start running at the available computing resources (VMs, containers, etc.).

[13] https://inventory.clarin.gr/workflows/

2.6 User and Resource-lifecycle management

Members of the academia, the research community, or the general public, whether affiliated to a network member organization or not, have full access to the infrastructure. Registered users, authenticated via their academic, personal or CLARIN ERIC accounts, can upload their resources and/or tools, use the available resources and process them using the services offered by CLARIN:EL.

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. There are two ways to become a CLARIN:EL user: (a) create a personal (non-organization) account, and (b) use an existing account managed by the identity provider (IdP) of the affiliated organization (e.g., research institute, university etc.) that belongs either to the national Authentication and Authorization Infrastructure (AAI) Federation (GRNET) or to the CLARIN Service Provider Federation. In this way, CLARIN ERIC users from any country have access to the Greek network. In both cases, users are stored in the User Management module, which is based on Keycloak, an open source identity and access management solution.

The activities available to the users depend on their roles, which are also defined and managed within Keycloak. The User Roles schema foresees the roles of Curator (assigned by default to all registered users), Supervisor for each organization-member, and Validator (assigned by the 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 (license) and metadata quality check, while Curators have the basic set of actions for LRs deposition.

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[14]; 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 and ensuring completeness and correctness.

2.7 The User Dashboard

The CLARIN:EL User Dashboard, exclusively available to registered users, functions as a specialized section of the Platform. It is a key tool enhancing the overall user experience and optimizing the utilization of CLARIN:EL resources and services. The dashboard plays a dual role, not only facilitating the creation, management and processing of language resources, but also offering users a comprehensive overview of their activity within the CLARIN:EL ecosystem. Through this interface, users benefit from an assortment of features. Specifically:

- Customization: The Dashboard supports customization, offering different views tailored to the specific roles of registered users, allowing them to access information and functionalities relevant to their roles within the CLARIN:EL platform.
- Interactivity: The Dashboard is designed to be interactive, allowing users to actively engage with CLARIN:EL resources and services. This interactivity includes creating and uploading resources, utilizing Natural Language Processing services, or performing other tasks.
- Real-time Data Display: The Dashboard provides real-time data display, enabling users to access up-to-date information regarding their tasks, resources, and processing jobs.
- Alerts and Notifications: Users receive alerts and notifications through the Dashboard, which keep them informed about changes related to their activities on the CLARIN:EL platform.

The Dashboard provides the users with the following functionalities (Figure 4):

[14] https://clarin-platform-documentation.readthedocs.io/en/stable/all/3_Creating/publicationLifecycle.html#publicationlifecycle

- Resource Creation and Upload: Users can easily create and upload language resources directly from the dashboard.
- Processing Services: The Dashboard serves as a gateway for users to access and utilize the integrated NLP processing services.
- Activity History: Users can explore a comprehensive history of their activity within CLARIN:EL, including details on created resources, validation tasks, and processing jobs.
- Editable Profile: Users can customize their profiles and manage their personal information and preferences effortlessly.

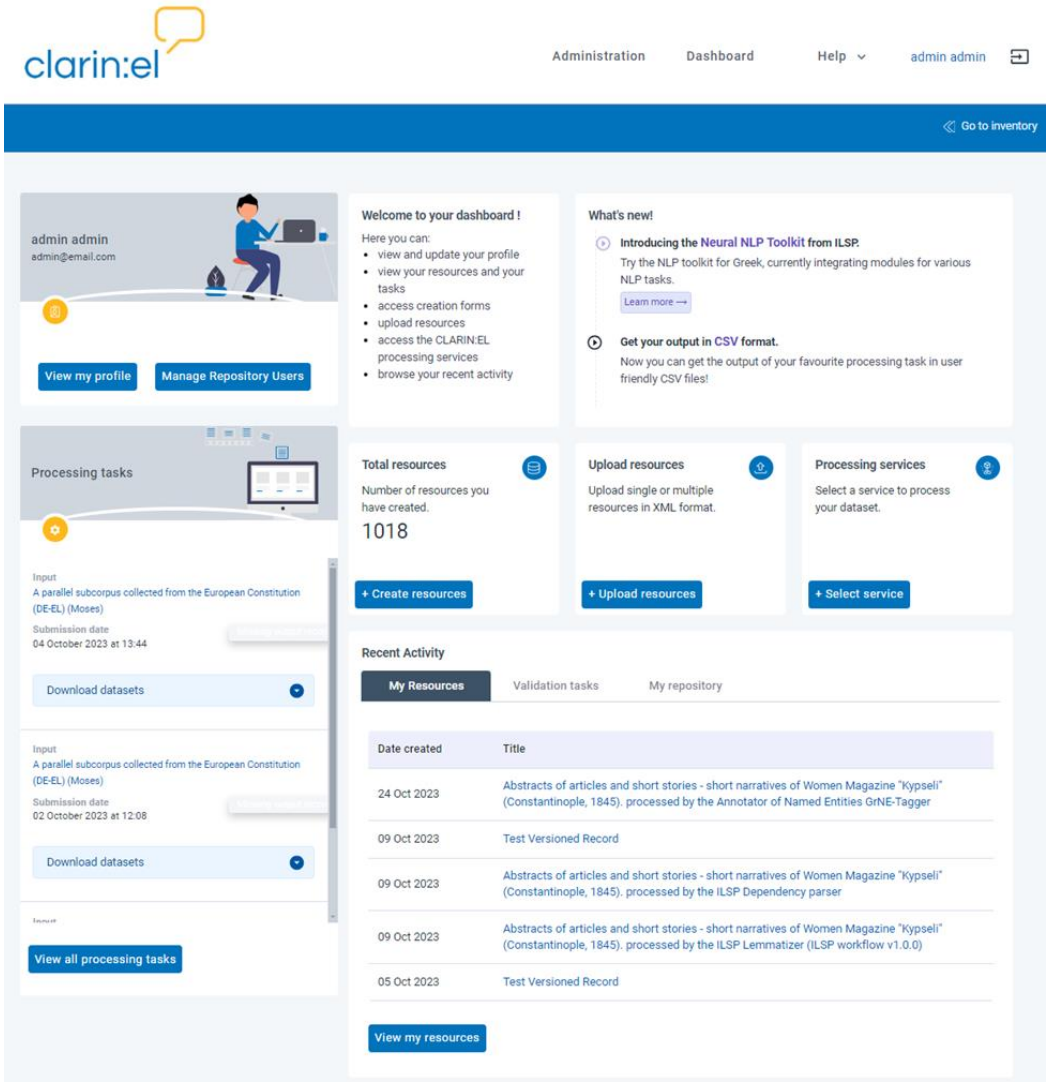

Figure 4: The User Dashboard

3 The CLARIN:EL technical architecture

The functionalities described in Section 2 are supported by the CLARIN:EL platform, designed and implemented 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[15], 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 Workflow Manager responsible for executing NLP services and a scheduler that decides where and when a user's processing request will be executed, to avoid platform overloading
- and the User Dashboard.

All the above applications run as Docker containers at a Kubernetes (k8s) cluster, maintained and supported by the CLARIN:EL development team at ILSP/Athena RC. The LRs data are saved in a dedicated Network Attached Storage (NAS), 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. For example, an automatic procedure has been set up at CLARIN:EL's k8s cluster that regularly creates backups of the LRs and metadata records to two different servers, one on-site and one off-site. Also, a recovery procedure is in place, with a set of installation scripts and procedures that are to be used when a new instantiation of CLARIN:EL must be set up, e.g., in case of disaster. The CLARIN:EL infrastructure is protected and secured using standard practices, e.g., access to VMs and containers and k8s cluster is limited only to the administrators, a firewall is appropriately configured to restrict access to (specific) ports and VMs, SSL certificates are installed and renewed as appropriate.

4 User Support

4.1 The portal

CLARIN:EL provides several assistance mechanisms to support user needs. The Portal includes (i) information material on the infrastructure, the network and the team members working on the maintenance and operation of the Greek CLARIN, the use of the Platform and the provided services and functionalities, FAQs, as well as guidance on legal and policy issues related to data sharing, etc., (ii) dissemination material (news, events, articles and/or interviews about Language Technology), and (iii) educational and training material, namely video tutorials, scientific publications, and presentations. Publicly accessible Helpdesks enable interested parties to ask questions on technical and management issues, legal issues, or issues related to metadata creation and documentation of language data, tools and/or services. The Portal, as the main entry point of the infrastructure, besides hosting the informative material mentioned above, also provides links to direct the users to the Platform and the NLP:EL Knowledge Centre.

4.2 The NLP:EL Knowledge Centre

NLP:EL was established in March 2020 and has been since officially recognized as a CLARIN ERIC Knowledge Centre (K-Centre), while in May 2023 the Greek K-Centre has successfully received a certificate of recognition renewal. NLP:EL aims at actively supporting research and scientific advances and providing to all interested parties useful information and guidance in the fields of Natural Language Processing, Language Technology, Language Resources and Sign Language Technologies (SLT), as well as to support the digital readiness of the Greek language (Gavriilidou et al., 2023). The NLP:EL microsite (section) is organized in two main units: (a) *Knowledge*, where users can find access to an exhaustive list of tools and services for NLP and SLT, information on studies and curricula of Greek Universities related to LT and DH, educational and training material such as the *CLARIN Learning and Training Resources*[16] and the *SSH Training Discovery Toolkit*[17], a list of scientific publications (from

[15] https://github.com/metashare/META-SHARE
[16] https://www.clarin.eu/content/learning-and-training-resources
[17] https://training-toolkit.sshopencloud.eu/about

1973 to date) on NLP involving the Greek language, collected automatically from 7 databases[18], and (b) *Community*, where interested parties can be informed on NLP/LT and SLT teams active in Greece, on certified CLARIN K-Centres and on National and European LRTs Infrastructures.

4.3 Documentation

CLARIN:EL also provides detailed online documentation[19] on the Platform and all its functionalities, accessible through a "Help" button located on the home page. Through a dropdown menu, users can navigate to either the User Guide as a whole or directly to the section detailing the Recommended File Formats. This section is selectively promoted to aid users seeking information about CLARIN:EL policies regarding data and file formats during deposition. Additionally, users can utilize the search box to find particular information. The User Guide is available in both Greek and English; it is designed not for linear reading, though it can be approached that way, but rather aims to assist users in locating specific information according to their needs. It familiarizes users 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). Each chapter within the User Guide is interconnected, and external links to referenced sources and documents are also included. The User Guide is available for download in PDF, HTML, and EPub formats.

4.4 Training activities

In addition to the management and the continuous updating of the material provided through the Portal and the NLP:EL Knowledge Centre, CLARIN:EL organizes a wide range of teaching/training, scientific and user support activities and events, such as webinars, workshops, hands-on sessions, summer schools, datathons, meetings with network members and CLARIN:EL users, etc. These activities are single or recurrent, aiming to educate users on LT and DH, to introduce the functionalities of the Platform and train interested parties how to use it.

4.5 The “Me, my family and other resources” initiative

The "Me, my family & other resources" initiative draws inspiration from the CLARIN ERIC Resource Families, while the title is a paraphrase of Gerald Durrell's renowned novel "My Family and Other Animals." Similarly to CLARIN Resource Families, this initiative organizes resources into thematic families, based on shared characteristics such as domain, topic, media type, time period, etc. These resource families are virtual collections of resources from various CLARIN:EL organizations, while each one maintains their individual autonomy and distinct traits. For each family, one resource is selected and highlighted as its representative; this resource is described in a dedicated portal webpage by a person involved in its creation process. Key metadata, including type, size, medium, provider, and format are also provided alongside a preview of the content of the resource. All the members of the family are also listed as hyperlinks directing to the CLARIN:EL central inventory. Each family is presented in conjunction with significant global, European, or national observances. For instance, the inaugural family, Poetry, was introduced in March 2022 to coincide with World Poetry Day on March 21st. To date, 13 resource families have been presented, namely Poetry, Medicine & Health, Museums, Tatoeba corpora, Sign Language resources, Educational Textbooks, Human Rights, Named Entity Recognition Tools, Kypseli Women's Magazines, Elections, Parliamentary Discourse, Literary Translation and Medieval and Early Modern Greek.

[18] ACL Anthology, ACM Digital Library, ArxiV, IEEE Xplore, ResearchGate, Semantic Scholar, and Springer. For ACL Anthology, papers were found on the official GitHub repository in XML format, whereas for the rest the software JabRef was used through its search interface.

[19] https://clarin-platform-documentation.readthedocs.io/en/stable/index.html

5 CLARIN:EL analytics

Matomo, formerly known as Piwik, is an open-source web analytics platform[20], which has been integrated on-premise into both the portal website (https://www.clarin.gr/) and inventory (https://inventory.clarin.gr/), providing detailed insights into visitor behaviour and site performance. With Matomo, website administrators gain valuable data to optimize content, improve user experience, and enhance overall website effectiveness. CLARIN:EL leverages Matomo Analytics to gather anonymized information about how visitors interact with the infrastructure. Using Matomo Analytics, CLARIN:EL tracks visitor interactions in real-time and over-time, offering a detailed overview of user engagement, including page views, visit duration, and bounce rates. This data helps understand how visitors navigate the website and which content attracts the most interest. Other valuable information include statistics on location and device used to access the infrastructure, as well as basic infrastructure performance statistics. Location statistics are evaluated based on the anonymized IP (e.g., 192.168.xxx.xxx). Anonymization in Matomo refers to the process of concealing or obscuring personally identifiable information (PII) from the data collected by the platform. This is crucial for privacy compliance, especially with regulations like GDPR in Europe.

In December 2023, the total number of CLARIN:EL registered users was 1,560 (+10,4% compared to December 2022). Based on Matomo Analytics, in 2023 the three pillars of the infrastructure (Portal, Platform, K-Centre) gathered a total of approximately 6,300 visits, 20,600 pageviews, 8,931 resource downloads and 3,843 processing tasks.

6 Conclusion

This paper presented the CLARIN:EL infrastructure, the design and implementation principles as reflected in its architecture, the functionalities available to the users, the support activities provided to the community, and finally, usage analytics. Future steps include the maintenance and upgrading of the infrastructure's modules, and the population of the repository with new resources (datasets and workflows). As regards dissemination and training, future objectives include the continuous support and training of CLARIN:EL users, and the roll-out of outreach activities aiming to raise awareness about LT in the research community. At the technical level, CLARIN:EL aims to ensure and increase interoperability with other infrastructures and repositories, at the national, European and international levels. At the strategy level, concrete aims constitute the enlargement of the network with new organization members and end-users, and ensuring its connectivity with EOSC, the SSHOC Marketplace, the Language Data Space and similar emerging initiatives.

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

Branco, A., Eskevich, M., Frontini, F., et al. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. Lang Resources & Evaluation. https://doi.org/10.1007/s10579-023-09658-z

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

[20] https://matomo.org/

Gavriilidou, M., Giagkou, M., Loizidou, D., & Piperidis, S. (2023). Language Report Greek. In: Rehm, G., Way, A. (eds) European Language Equality. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-28819-7_19

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., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Rigault, M., Arranz, V., Choukri, K., Backfried, G. Gómez Pérez, J.M., & Garcia Silva, A. (2020). 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., Galanis, D., Bakagianni, J., & Sofianopoulos, S. (2016). Combining and Extending Data Infrastructures with Linguistic Annotation Services. In: Murakami, Y., Lin, D. (eds) Worldwide Language Service Infrastructure. WLSI 2015. Lecture Notes in Computer Science(), vol 9442. Springer, Cham. https://doi.org/10.1007/978-3-319-31468-6_1

Piperidis, S., Labropoulou, P., Deligiannis, & M., Giagkou, M. (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

Pontiki, M., Gavriilidou, M., Gkoumas, D., & Piperidis, S. 2020. Verbal Aggression as an Indicator of Xenophobic Attitudes in Greek Twitter during and after the Financial Crisis. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 19–26, Marseille, France. European Language Resources Association.

Prokopidis, P., & Piperidis, S. (2020). A Neural NLP toolkit for Greek. In *11th Hellenic Conference on Artificial Intelligence* (pp. 125-128).

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18