Setting up the national infrastructure clarin:el

Stelios Piperidis  
Institute for Language and Speech Processing,  
Athena Research Centre,  
Greece  
spip@ilsp.gr

Maria Gavriilidou  
Institute for Language and Speech Processing,  
Athena Research Centre,  
Greece  
maria@ilsp.gr

Abstract

This paper presents the Greek national infrastructure for language resources, clarin:el, member of CLARIN ERIC since 2015. It describes the design principles for the creation of the network and lists the current members; it describes the infrastructure and its architecture and briefly elaborates on the resources and services offered.

1 Introduction

The Greek network clarin:el is the network of organizations that co-operate for the construction, operation and maintenance of the national research infrastructure (www.clarin.gr). Greece joined CLARIN ERIC in 2015 and the national infrastructure was built as part of the National Research Infrastructures Roadmap in the framework of a project co-funded by Greece and the EU.

The design and development of clarin:el extensively drew upon previous experience gained through similar infrastructural initiatives, like Clarin (Varadi et al., 2008), (Wittenburg et al., 2008), META-SHARE (Piperidis, 2012), Flarenet (Soria et al., 2012) and QTLaunchPad (Piperidis et al., 2016).

2 The Greek national network

At the core of the network lies the project consortium; that is, the organizations that implemented the infrastructure, which are the founding members of the network; namely, Athena RC with two institutes, ILSP (www.ilsp.gr) as coordinator, and IMIS (www.imis.gr), NCSR DEMOKRITOS (https://www.iit.demokritos.gr/skel/) and the Greek Research and Technology Network (www.grnet.gr).

Besides the founding members, five more members were selected according to both geographical and scientific principles. Geographic criteria ensured broad coverage of the country’s regions, while scientific criteria guaranteed a broad coverage of fields related to language research, linguistics and language technology. Thus, next to the three founding members, the following institutions were selected: the University of Athens, Dept. of Linguistics (http://en.phil.uoa.gr/the-faculty/department-of-linguistics.html) based in Central Greece, two from Northern Greece, namely the Aristotle University of Thessaloniki, Dept. of French Language and Literature (http://www.frl.auth.gr), with expertise on Linguistics, Translation and Language Resources, and the Centre for the Greek Language (www.greeklanguage.gr), an institution of the Ministry of Education responsible for language policy and language teaching; the Ionian University (www.ionio.gr) at Corfu, Western Greece, specializing on Translation studies and Information science; the University of the Aegean at Rhodes (http://dms.aegean.gr/en) with expertise in language resources collection and documentation.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/
These organisations constitute the network’s nodes; they have set up their own repositories (Institutional Repositories), in which they document, store and manage the resources they share through the infrastructure. An alternative type of repository is the Hosted Resources Repository, where organisations not wishing to set up a repository can store, document and manage their resources. Besides organisations, individuals can also share resources, technologies and services using the Hosted Resources Repository for this purpose. Providers (legal entities or individuals) are responsible for the curation of their language resources, technologies and services and for the clearance of the related intellectual property rights.

3 The infrastructure

3.1 The architecture

The architectural design of the clarin:el network and infrastructure has been inspired and, to a large extent, based on the design principles of the META-SHARE infrastructure. As such, clarin:el is a network of distributed repositories, each being set up for and maintained by each member organization of the clarin:el network. Unlike META-SHARE, clarin:el:

a. features a single central repository that harvests metadata and hosts a central inventory for the whole network, and

b. adopts a multi-tenant approach implementing a single database server for all repositories.

It is the central inventory that will assume all functionalities and responsibilities for connecting clarin:el to CLARIN ERIC (e.g. providing OAI-PMH endpoints for metadata harvesting, ensuring secure access to the CLARIN ERIC federation).

These architectural choices centralize and facilitate the creation, management and administration of repositories, thus taking the burden of local hosting from member-organizations wishing to set up an institutional LRs repository. Repositories are automatically set up on okeanos (https://okeanos.grnet.gr/home/), the research cloud infrastructure of the Greek Research and Technology Network and actual datasets are stored in its distributed permanent storage pithos+ (https://okeanos.grnet.gr/services/pithos/). In this way, maintenance operations are simplified to a large extent and the respective costs are substantially reduced.

clarin:el has adopted the META-SHARE metadata schema (Gavriilidou et al, 2011), which has been mapped to CMDI 1.0 (Broeder et al, 2008), (Broeder et al, 2012). Resources are persistently identified using handles (www.handle.net) issued upon being published in the repositories. Metadata support simple and faceted search, as shown in Figure 1.

![clarin:el interface](image)

Figure 1: Browsing and searching the clarin:el inventory
Figure 2 shows the record of the chosen resource with the full metadata description.

![Figure 2: Full metadata record](image)

The legal component draws on the META-SHARE toolkit and fosters the use of Creative Commons 4.0 licences, the recently upgraded META-SHARE NoRedistribution 2.0 licenses and existing standard open source software (FOSS) licenses for tools and services. The rights and restrictions of use of the resource are under the control of the resource owners, while metadata elements documenting the rights of use are obligatory (no resource can be deposited in the infrastructure without license metadata).

### 3.2 Language processing services

The primary repository functionalities have been enhanced by an additional language processing mechanism for the processing of language datasets with appropriate services. Inspired by previous initiatives (Piperidis et al., 2016), language processing services are documented with the appropriate metadata in the repository and are provided as web services through the language processing layer. In the typical usage scenario, when a clarin:el user selects to process a dataset, a list of all available relevant annotation/processing levels and respective services is automatically generated based on the properties of the dataset (Figure 3), i.e. its resource type, language(s), domain, format, annotation level, and licensing conditions. Currently, services for processing Greek and English datasets are offered at the following levels: tokenization & sentence splitting, POS tagging, lemmatization, dependency parsing, named entity recognition (all for Greek and English), opinion mining and summarization (for Greek only).

![Figure 3: Presenting the relevant annotation levels for a resource and selecting the desired service](image)
In view of integrating language processing services, the metadata model was extended to facilitate the documentation of language processing services as well as processing-related properties of datasets. As soon as the user selects a service, the server dispatches the dataset to the specific web service(s) for processing. When the processing has been completed, the new (annotated) dataset is automatically documented according to the metadata schema, stored and indexed in the native repository (Figure 4). Automatic creation of metadata records for newly-created datasets combines the metadata descriptions of the original dataset and the language processing tool/service used; it also creates relations between the original dataset and the annotated one.

![Figure 4: Automatic creation of metadata record for a derivative resource](image)

4 Conclusion

The national clarin:el infrastructure has been set up involving 7 institutional repositories and 1 cloud computing and storage infrastructure provider. It has been in operation since November 2015 offering language data and processing services to all academic users in Greece. Users can login with their academic credentials, if they have such credentials and their academic organisation has joined the Greek Authentication Authorisation Infrastructure (AAI), managed by GRNET, or register and create an account with clarin:el, if they are not affiliated with a Greek academic institution.

![Figure 5: The clarin:el portal](image)

clarin:el is accessible through its portal ([www.clarin.gr](http://www.clarin.gr)) which provides connection of the national to the European infrastructure CLARIN ERIC, informative material and a range of support facilities to all its users (Figure 5).
5 References


