CLARIN: Distributed Language Resources and Technology in a European Infrastructure

Maria Eskevich[1], Franciska de Jong[1], Alexander König[1], Darja Fišer[2], Dieter Van Uytvanck[1], Tero Aalto[3], Lars Borin[4], Olga Gerassimenko[5], Jan Hajic[6], Henk van den Heuvel[7], Neeme Kahusk[5], Krista Liin[5], Martin Matthiesen[8], Stelios Piperidis[9], and Kadri Vider[5]

[1] CLARIN ERIC, Utrecht, The Netherlands ·[2] University of Ljubljana and Jožef Stefan, Ljubljana, Slovenia ·
[3] The Language Bank of Finland·[4] Språkbanken, University of Gothenburg, Sweden·[5] University of Tartu, Estonia·
[6] Charles University, Prague, Czech Republic ·[7] CLST, Radboud University, Nijmegen, The Netherlands ·
[8] CSC - IT Center for Science, Espoo, Finland ·[9] ILSP/Athena RC, Greece
clarin@clarin.eu

Abstract

CLARIN is a European Research Infrastructure providing access to digital language resources and tools from across Europe and beyond to researchers in the humanities and social sciences. This paper focuses on CLARIN as a platform for the sharing of language resources. It zooms in on the service offer for the aggregation of language repositories and the value proposition for a number of communities that benefit from the enhanced visibility of their data and services as a result of integration in CLARIN. The enhanced findability of language resources is serving the social sciences and humanities (SSH) community at large and supports research communities that aim to collaborate based on virtual collections for a specific domain. The paper also addresses the wider landscape of service platforms based on language technologies which has the potential of becoming a powerful set of interoperable facilities to a variety of communities of use.

Keywords: CLARIN, language resources, research infrastructure, repositories, interoperability

1. Introduction

CLARIN[1] is a European Research Infrastructure providing access to language resources and tools. It focuses on the widely acknowledged role of language as cultural and social data and the increased potential for comparative research of cultural and societal phenomena across the boundaries of languages. Since its establishment as a European Research Infrastructure Consortium (ERIC) in 2012, CLARIN has grown both in terms of number of members and observers (21 and 3 respectively in Spring 2020, see Figure 1.), and in terms of the variety of specific communities served (diverse subfields within the humanities and social science, such as literary studies, oral and social history, political studies, historical linguistics, developers of analysis systems based on machine learning, etc.). A strong focus on interoperability between the wide variety of resources ensures the steady and reliable development of the infrastructure, which is also reinforced by the polices for research infrastructures that have been established in alignment with the European Strategy Forum for Research Infrastructures (ESFRI)[2].

This paper is organised as follows: Section 2. describes the general principles that are to be followed to secure the interoperability for resources, and provides motivation for the CLARIN use case; and gives an overview of repository solutions that are being used within CLARIN; Section 3. provides a number of examples of data-driven communities that are brought together through the access to language resources that can be explored using approaches and methods of diverse academic fields; and Section 4. outlines the overall landscape of technical solutions CLARIN works in.

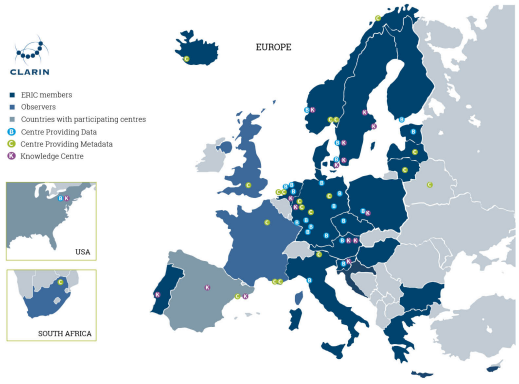

Figure 1: Map of CLARIN members, observers, and participating centres at the start of 2020.

2. CLARIN as a FAIR platform

The FAIR Guiding Principles for Data Management and Stewardship (Wilkinson et al., 2016) provide a universal framework for data management, based on the idea that research data should be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

Overall, the FAIR principles are widely being promoted as part of the Open Science paradigm and are supposed to contribute to the ease of discovery and access of research data by researchers and the general public. Reuse of data is fostered by promoting the use of widely accepted standards both for the data itself and for the metadata describing it.

[1] https://www.clarin.eu
[2] https://www.esfri.eu/about

Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 28–34
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

CLARIN is committed to promoting the FAIR data paradigm (de Jong et al., 2018). With the Virtual Language Observatory (VLO)[3] (see Section 2.2.) CLARIN provides a search engine that helps exploring over a million language resources from dozens of CLARIN Centres spread over all of Europe and beyond. Apart from a shared metadata paradigm that enables this kind of central discovery, technical interoperability is ensured by the technical specifications for CLARIN Centres (see Section 2.3.) and the accessibility of the data is managed with the help of a SAML-based Federated Identity[4] setup.

2.1. From FAIR to actionable

Both persistent identifiers (PIDs) and the FAIR guidelines have been existing for quite a while. Recent efforts in the context of the Research Data Alliance [5] have paved the way to enhance the already existing Handle infrastructure into an ecosystem for FAIR Digital Objects[6] (DOs) that fully supports machine-actionability: the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention. The principle behind FAIR Digital Objects is to enrich Handles with a core of directly accessible metadata descriptions (the PID kernel information, which can have community-specific extensions). These metadata elements can be unambiguously interpreted with the help of a Data Type Registry, which contains the definitions of the elements. An important difference with the more extensive metadata provided outside the Digital Objects (as described in Section 2.2.) is the speed with which the information can be retrieved and the cross-community standardization.

While FAIR DOs are not yet in a production-ready state[7] it is clearly an initiative gaining a lot of traction (Hodson et al., 2018), with the potential to bring significant progress in the field of language resource processing and beyond.

2.2. Discoverability through the VLO and CMDI metadata

Within the concept of FAIR research data, the aspect of Findability is the most important one, because data that cannot be discovered by interested parties cannot be reused, no matter how well-designed and interoperable the data itself is. CLARIN has put this aspect front and centre by making it a hard requirement for their (B and C) centres to provide metadata about their collections in a well-defined format that is shared within all of CLARIN. A CLARIN centre has to provide its metadata in the CMDI-format (Broeder et al., 2012) via the OAI-PMH protocol[8]. All of these OAI endpoints are regularly checked for updates. Any new metadata elements are harvested and fed into the VLO, a facet-based search portal, where the collected metadata can be searched by interested users.

As the next step towards interoperability, a tool has been developed (Zinn, 2016) to provide guidance on which service is recommended for which data, known as the Language Resource Switchboard [9]. It acts as a simple forwarding application that, based on the URL of an input file and a few simple parameters (language, mimetype, task), allows the user to select relevant NLP web applications that can analyze the input provided.

2.3. CLARIN landscape of repositories

This section contains an overview of repositories used throughout the CLARIN infrastructure, and internal technical solutions to support the interoperability within the network of CLARIN centres. The CLARIN infrastructure backbone is a network of CLARIN Centres that provide access to language resources in a multitude of languages from European roots and beyond, in a variety of modalities and formats. Most prominent is the role of the service providing centres, called *B Centres*, which offer services to the CLARIN community, such as access to linguistic software or language data. There are also *C Centres* which allow the harvesting of metadata for the language resources and tools by the VLO, but do not offer any additional services. The most important difference between the two types of centres is that B centres have to follow precise technical specifications[10] and are regularly evaluated and certified. The certification procedure is led by CLARIN Central Assessment Committee.[11]. One of the assessment criteria is that an application needs to be prepared for certification through the independent certification organisation CoreTrustSeal.[12]

The CLARIN network currently consists of 23 B and 22 C Centres. While C Centre status does not come with the expectation of running a research data repository, a lot of them actually do, resulting in a network of 41 centres with a repository. While the technical specifications (for B Centres, see above) have some requirements on what such a repository has to be able to do and the services it has to offer, the individual centres are free in their choice of the actual software they run and this results in a quite varied "repository landscape" within CLARIN.

Repository type	**Number of centres**
DSpace	14
Fedora	10
META-SHARE	4
Git	2
LAT	2
Dataverse	1
Custom	8
TOTAL	**41**

Table 1: Type of repositories used in CLARIN centers. This information is provided at registration stage.

[3] https://vlo.clarin.eu

[4] https://www.clarin.eu/node/3788

[5] https://www.rd-alliance.org/group/gede-group-european-data-experts-rda/wiki/gede-digital-object-topic-group

[6] https://fairdo.org/

[7] https://pti.iu.edu/centers/d2i/initiatives/rpid.html for a testbed implementation.

[8] https://www.openarchives.org/pmh/

[9] https://switchboard.clarin.eu

[10] http://hdl.handle.net/11372/DOC-78

[11] https://www.clarin.eu/governance/centre-assessment-procedure

[12] https://www.coretrustseal.org/

Looking at the current install base for repositories (see Table 1) two solutions appear to be prevalent, namely DSpace[13] (14 installations) and Fedora Commons[14] (10 installations). Both are general data management solutions that need some custom adaptions to be suitable for a CLARIN Centre, but while there are currently quite a number of different adaptions of Fedora Commons within the CLARIN community, most DSpace installations are using the modifications made by the CLARIN DSpace project[15].
The CLARIN DSpace project was started by LINDAT/CLARIAH-CZ, the Czech node of the CLARIN network based at the Charles University in Prague. But in the course of the last couple of years, as the DSpace repository has been installed at various CLARIN centres across the network, developers from those centres have started contributing to the project as well. CLARIN DSpace comes with very detailed installation instructions that include the various prerequisites and different software stacks that need to be installed for DSpace to work, for example, it bundles a handle server that is used to issue persistent identifiers to each new data submission and is also responsible for resolving those identifiers later on. Additionally, CLARIN DSpace is also available as a Docker project[16] which makes it easy for a new CLARIN centre to get started with their own repository. The project is working with Overlays[17] to make adaptions to the look and feel as easy as possible without having to touch the actual codebase. This means that the project can be customized to change the branding by each centre, while still being able to quickly update to new versions should they become available.

3. Enhanced multidisciplinarity through increased resource visibility

In order to target specific communities of researchers from the domains of humanities, social sciences and human language technologies, in 2017 CLARIN started an initiative called "Resource Families"[18], the goal of which is to collect and present in a uniform way prominent data types in the network of CLARIN consortia that display a high degree of maturity, are available for most EU languages, are a rich source of social and cultural data, and are as such highly relevant for research from a wide range of disciplines and methodological approaches in SSH as well as for cross-disciplinary and trans-national comparative research. (Fišer et al., 2018)
Currently, CLARIN Resource Families feature 10 families of corpora, 5 families of lexical resources, and 3 families of natural language processing (NLP) tools. The overviews are organized according to the types of data featuring in the resources and include listings sorted by language. The listings include the most important metadata and brief descriptions, such as resource size, text sources, time periods, annotations and licences, as well as links to download pages and concordancers, whenever available. Where applicable, overviews of other existing prominent language resources which have not yet been integrated in the infrastructure have also been provided. As a side project, overviews of related materials such as thematic CLARIN workshops and tutorials along with their accompanying VideoLectures recordings[19], as well as a list of key publications on the surveyed resources have also been generated [20].
The overviews serve as an entry point to the CLARIN infrastructure for individual researchers, lecturers and students from SSH, but have also proved to be a highly valuable instrument for further improvement of the infrastructure, either by improving the identified issues with the findability or documentation of the resources, or by working towards better interoperability of the resources (e.g. by developing common corpus encoding standards).

3.1. Parliamentary data

Parliamentary data is a major source of socially relevant content. It is available in ever larger quantities, is multilingual, accompanied by rich metadata, and has the distinguishing characteristic that it is spoken language produced in controlled circumstances which has traditionally been transcribed but is now increasingly released also in audio and video formats. All these factors require solutions related to structuring, synchronization, visualization, querying and analysis of parliamentary corpora. Furthermore, approaches to the exploitation of parliamentary corpora to their full extent also have to take into account the needs of researchers from vastly different SSH fields, such as political sciences, sociology, history, and psychology.
An inspiring and highly successful series of workshops focusing on parliamentary data, such as CLARIN+[21], ParlaCLARIN[22] and ParlaFormat[23] resulted in a comprehensive overview of a multitude of existing parliamentary resources worldwide,[24] a detailed needs analysis[25] as well as tangible first steps towards better harmonization, interoperability and comparability of the resources and tools relevant for the study of parliamentary debate[26].
In the context of H2020 cluster projects PARTHENOS[27]

[13] https://duraspace.org/dspace/
[14] https://duraspace.org/fedora/
[15] https://github.com/ufal/clarin-dspace
[16] https://gitlab.inf.unibz.it/commul/docker/clarin-dspace
[17] https://github.com/ufal/clarin-dspace/wiki/Overlays
[18] https://www.clarin.eu/resource-families/
[19] http://videolectures.net/clarin/
[20] https://www.clarin.eu/resource-families/parliamentary-corporapublications-on-the-parliamentary-corpora
[21] https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records
[22] https://www.clarin.eu/ParlaCLARIN, https://www.clarin.eu/ParlaCLARIN-II
[23] https://www.clarin.eu/event/2019/parlaformat-workshop
[24] https://www.clarin.eu/resource-families/parliamentary-corpora
[25] https://office.clarin.eu/v/CE-2017-1091-Focus-group-UI-2017-03-27.pdf
[26] https://github.com/clarin-eric/parla-clarin
[27] https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/researching-parliamentary-records-in-the-digital-humanities/, https://www.clarin.eu/event/2019/parthenos-workshop-cee-countries

and SSHOC[28], representatives from the CLARIN network have started to develop training materials for the community of researchers using CLARIN resources and tools using parliamentary data. They will be integrated in the SSH Open Marketplace that is planned to result from the collaborative efforts that the European research infrastructures for the social sciences and humanities have taken up as part of the SSHOC workplan.

3.2. DELAD

DELAD[29] (meaning 'shared' in Swedish) is an initiative to establish a digital archive of disordered speech and share this with interested researchers within CLARIN. The DELAD community consists of researchers involved in collecting and analysing Corpora of Disordered Speech (CDS), research data and infrastructure specialists, and legal experts. DELAD has chosen the CLARIN infrastructure as primary space for storing and sharing CDS. More specifically, DELAD has linked up with CLARIN's Knowledge Centre for Atypical Communication Expertise (ACE)[30] (Van den Heuvel et al., 2020b) for making CDS available through The Language Archive (TLA)[31] at the Max Planck Institute for Psycholinguistics in Nijmegen (being a CLARIN Data Centre) and CMU's Talkbank[32] (Clinical Banks). DELAD has organised four workshops over the years 2015-2019, the latter two of which were held under the umbrella of CLARIN ERIC. Topics addressed in these workshops were: Guidelines for collecting and sharing CDS (in the light of the General Data Protection (GDPR)[33]), levels of anonymisation, layered access, integration of CDS in the CLARIN infrastructure, formats, and relevant metadata. More information about DELAD and the application of the GDPR on CDS can be found in (Van den Heuvel et al., 2020a). The workshops are extremely fruitful since researchers from various disciplines (clinical researchers, speech and language scientists and technologists, infrastructural specialists and legal experts) can apply their own knowledge in a new context and learn about the practical challenges that their colleagues in other domains come across (e.g. clinical researchers facing ICT and legal issues).

3.3. Europeana

Part of the materials that has been aggregated in Europeana[34], Europe's platform for digital cultural heritage, consists of language data and is therefore of potential added value for researchers studying heritage data in spoken or textual form. This premise led to a joint project between CLARIN and Europeana that has been set up with an aim to bring the visibility of Europeana data through the VLO. CLARIN and Europeana do not share a common metadata model, and therefore a semantic and structural mapping had to be defined, and a conversion implemented. CLARIN's ingestion pipeline was extended to retrieve a set of selected collections from Europeana and apply this conversion in the process.

Currently about 775 thousand Europeana records can be found in the VLO, with several times more records expected in the foreseeable future. About 10 thousand records are technically suitable for processing via the Language Resources Switchboard already. Relatively straightforward improvements to the metadata on the side of Europeana and/or its data providers could substantially increase this number.

4. CLARIN in the landscape of language technology platforms

CLARIN operates in the broader context of international initiatives that aim to support a diverse set of scenarios of use for services based on language technologies for a wide range of communities. As an initiative positioned in the wider European landscape of research infrastructures[35], CLARIN's service offer is strongly focusing on the needs of researchers. This mission comes with strong demands for both sustainability and interoperability. The Open Science agenda that by the various stakeholders is seen as a major driver for the investments, has added incentives for the support of multidisciplinary work and the integration of language data in interdisciplinary paradigms. (de Jong et al., 2018)

The value proposition put forward by CLARIN implies that an adequate level of alignment with other infrastructural initiatives is sought, and conversely: that there are several language technology platforms that reference the service offer of CLARIN and have adopted measures to ensure interoperability. In this section a number of these existing European initiatives are presented with the aim to articulate both the potential for collaboration and the complementarity of the services.

This work implies the incorporation and usage of previously developed technological components; and coordination of activities and clear distinction of audiences served and regulation of access between CLARIN serving primarily the research community and other initiatives and platforms that offer access to data and tools for industry.

4.1. META

META-SHARE[36] has been developed as the infrastructural arm of META-NET[37] and has served as a component of a language technology marketplace for researchers, developers, professionals and industrial players, catering for the full development cycle of language technology, from research to innovative products and services. It has been designed as a network of repositories that store language resources (data, tools and processing services) documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access (Piperidis, 2012). Repositories can be local, set up and maintained

[28] https://www.sshopencloud.eu/news/using-corpora-implementing-validation-sshoc-masterclass

[29] http://delad.net

[30] https://ace.ruhosting.nl

[31] https://tla.mpi.nl/

[32] https://talkbank.org/

[33] https://gdpr-info.eu/

[34] https://www.europeana.eu/

[35] https://www.eric-forum.eu/the-eric-landscape/

[36] www.meta-share.eu

[37] www.meta-net.eu

by network members to store their own resources, or hosting (non-local) acting as storage and documentation facilities not only for their own resources, but also for resources developed in organisations not wishing to or not being able to set up their own repository, including donated and orphan resources. Every resource is primarily assigned to one of the network's repositories (master copy), and is formally described according to the META-SHARE metadata schema (Gavrilidou et al., 2012). The META-SHARE metadata schema has been mapped on a number of other schemas, including Dublin Core [38] and OLAC[39], the schema of the ELRA catalogue, and CLARINs CMDI. Metadata records are harvested and stored in the META-SHARE central inventory using a proprietary harvesting and synchronisation protocol, while lately an OAI-PMH bridge has been implemented as an additional harvesting protocol. While resources can be both open or with restricted access rights, free or for-a-fee, all metadata records are open, available under a Creative Commons Attribution 4.0 licence.

META-SHARE provides dedicated open-source software for setting up repositories[40], which has been used for technically setting up not only the network nodes themselves, but also for powering a number of CLARIN-related centres in a number of countries including Estonia, Finland, Greece, Portugal. While the provided solution could be readily used for setting up a language resource repository, a number of extensions were necessary to turn it into a solution satisfying the requirements set by CLARIN for establishing a CLARIN B-centre. Such extensions include: (i) assigning persistent identifiers to language resources, accommodated in and by the META-SHARE metadata schema through a dedicated metadata field, (ii) establishing an OAI-PMH bridge, implementations of which are provided, among others, by the META-SHARE nodes of Estonia and Greece, (iii) user authentication through Single Sign-On, an extension which has been implemented in variable ways by the CLARIN repositories which have opted for using the META-SHARE solution.

In the following subsections a number of META nodes for which interoperability with CLARIN has been realized are described.

4.1.1. CLARIN:EL and the Greek META-SHARE node

Prototype implementations for combining and extending data infrastructures, like META-SHARE, with linguistic processing services, have also been proposed (Piperidis et al., 2015). Such implementations aim to bring together language datasets and basic language processing services in a unified platform. The Greek META-SHARE node has been used for this prototype implementation and has been enhanced by providing a language processing mechanism for annotating content with appropriate NLP services that are documented with the appropriate metadata. Atomic services are combined into workflows modeled as an acyclic directed graph where each node corresponds to an NLP processing service (e.g. sentence splitting, part-of-speech tagging), running either locally or remotely. This implementation has been used for powering the language processing layer of the CLARIN:EL node (Piperidis et al., 2017), offering services and workflows for processing monolingual and bilingual content/resources in raw text, xces, tmx formats. From the legal framework point of view, a simple operational model has been adopted by which only openly licensed datasets can be processed by openly licensed services and workflows.

4.1.2. Language Bank of Finland

The Language Bank of Finland uses META-SHARE as its primary metadata repository. The software was deployed in 2012. Many of the Language Bank's services refer to META-SHARE directly, including the Language Bank Portal[41] and Language Bank Rights[42] the center's language resource access rights application and managing service.

The repository is populated and curated by the Language Bank's staff at the University of Helsinki and CSC – IT Center for Science. Each item has a persistent persistent identifier. URNs are mainly used as PIDs, but Handles are also supported with a 1:1 mapping[43]. PIDs to metadata records are used as the main way of referring to the language resources in other services and publications[44]. Where applicable, the resources also have PIDs for their access locations. The metadata is exported via a custom OAI-PMH bridge[45].

4.1.3. Center of Estonian Language Resources

The Center of Estonian Language Resources (CELR) uses META-SHARE as a register of language resources where metadata is stored[46]. In addition to the standard, META-SHARE node, Simple-SAML SSO, the OAI-PMH endpoint for VLO, and DataCite DOI as persistent identifier are used.

While META-SHARE provides file storage, an external data repository, ENTU[47], is used for storing the resources themselves, as it enables a better overview of the individual files. It also enables a better integration with other services, so that for a signed-in user the access permissions are managed for both download and further processing of the resource. The djangosaml2 module is implemented and connected to the local identity provider that serves as proxy, allowing access to the users of the CLARIN service provider federation.

Currently there are four resource types in META-SHARE to select. Sometimes other types are needed, for example because a specific CMDI profile is assumed, as in the case of workflow manager Weblicht[48]. A workaround has been developed by linking an external metadata file to the META-SHARE metadata field `metadataInfo/source`.

[38] http://dublincore.org/

[39] http://www.language-archives.org/

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

[41] https://www.kielipankki.fi/language-bank/

[42] https://lbr.csc.fi/

[43] Metadata curation: http://urn.fi/urn:nbn:fi:lb-201710212

[44] Citation instructions: https://www.kielipankki.fi/corpora/

[45] http://urn.fi/urn:nbn:fi:lb-201506011

[46] https://metashare.ut.ee

[47] https://entu.keeleressursid.ee

[48] https://weblicht.sfs.uni-tuebingen.de

For DOI allocation, a custom module was made using a central Handle server for Estonian resources. All data sets registered at CELR are findable at http://datacite.org by identifier ESTDOI.KEEL[49].

4.1.4. Swedish Language Bank

Språkbanken Text at the University of Gothenburg is one of the three divisions of the Swedish National Language Bank, and also a certified B centre of SWE-CLARIN, the Swedish node of CLARIN ERIC.

The centre's META-SHARE instance[50] predates the Swedish CLARIN membership, and was the result of its participation in the META-NET collaboration (2011–2013). With CLARIN membership, a strategic decision was taken to make META-SHARE the common language resource metadata format of SWE-CLARIN. Metadata editing is done primarily by SWE-CLARIN staff; with the present low volumes of metadata addition, this turns out to be the most time-effective solution. Metadata records are persistently identified using the Handle system.

4.2. European Language Grid

The European Language Grid (ELG) is a platform under development that aims to integrate a marketplace and community meeting point for Language Technology data, tools, services and developers, users and other stakeholders, with a specific focus on non-academic use cases, both commercial and non-commercial (Rehm et al., 2020). The envisaged platform is being developed as part of a European project[51], while an alpha release is expected to be open for the public as of March 2020. Eventually the platform may offer hundreds of services, technically scalable for large projects.

The ELG platform is consisting of three layers: (i) the base infrastructure operating on a managed Kubernetes cluster, (ii) the platform back end essentially implementing a repository back end containing metadata records of language resources, tools and services, as well as meta-information about language resources and technologies stakeholders, and (iii) the platform front end consisting of interfaces for different types of ELG users, including catalogue user interfaces, trial interfaces for functional services, registering/uploading interfaces for language resources and services providers. All components of the three layers are deployed as Docker containers on the Kubernetes cluster, with functional language technology services made available also through containerization and by being wrapped with the ELG LT service API.

The ELG catalogue will point to the tools contained either locally for developers and users to be able to incorporate them in their application, or simply use them for their language technology tasks. The catalogue will also contain or point to resources available in current LT repositories, such as ELRA/ELDA, META-SHARE(Piperidis, 2012), ELRC-SHARE(Piperidis et al., 2018) and other repositories. All entities are described in compliance with the ELG-SHARE metadata schema (Labropoulou et al., 2020). The schema builds upon, consolidates and updates previous activities, especially the META-SHARE schema and its profiles (Gavrilidou et al., 2012) taking into account recent developments in the (meta)data domains (e.g., FAIR, data and software citation recommendations , Open Science movement, etc.).

ELG has established a network of National Competence Centers led by country representatives who in many cases are also involved in national CLARIN consortia. It is to be expected that this will help facilitating the alignment of activities and the potential for interoperability between the platforms.

5. Concluding remarks

This paper presents the CLARIN research infrastructure as a platform for the sharing of distributed language resources in the context of the dynamics of the Open Science agenda and the inherent objective of giving sustainable access to FAIR data on the one hand, and on the other hand it positions CLARIN in the wider landscape of service platforms based on language technologies. The interoperability across platforms can be considered to bring added value for the further emergence of a seamless service offer to a variety of communities of use, both within and beyond academia.

Given that several complementary infrastructural initiatives have recently acquired public funding for the development of new services and/or deeper integration of language resources and technology into the ecosystem of digital infrastructures (e.g., EHRI[52], ELRC[53], ELEXIS[54] and Prêt-à-LLOD[55]) it is to be expected that further steps towards platform harmonization will be undertaken in the near future, and addressed in discussion fora such as the 1st International Workshop on Language Technology Platforms (IWLTP) workshop and other conversations organized in the context of networking and project events.

Bibliographical References

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012). Cmdi: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. (2018). Clarin: towards fair and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259–3264.

Fišer, D., Lenardič, J., and Erjavec, T. (2018). CLARIN's Key Resource Families. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

[49] https://search.datacite.org/works?query=estdoi.keel

[50] https://spraakbanken.gu.se/metashare/

[51] H2020 ICT Call 29a; https://european-language-grid.eu

[52] https://www.ehri-project.eu

[53] http://www.lr-coordination.eu

[54] https://elex.is

[55] https://www.pret-a-llod.eu

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE metadata schema for the description of language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1090–1097, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., and Wittenburg, P. (2018). Turning FAIR into reality. Final report and action plan from the EC expert group on FAIR data. DOI:10.2777/1524.

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Arranz, V., Choukri, K., Backfried, G., Pérez, J. M. G., and Silva, A. G. (2020). Making metadata fit for next generation language technology platforms: The metadata schema of the european language grid. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resource Association (ELRA).

Piperidis, S., Galanis, D., Bakagianni, J., and Sofianopoulos, S. (2015). Combining and extending data infrastructures with linguistic annotation services. In *International Workshop on Worldwide Language Service Infrastructure*, pages 3–17. Springer.

Piperidis, S., Labropoulou, P., and Gavrilidou, M. (2017). clarin:el: a language resources documentation, sharing and processing infrastructure [in Greek]. In Thanasis Georgakopoulos, et al., editors, *Proceedings of the 12th International Conference on Greek Linguistics*, volume 2, page 851–869, Berlin, October. Edition Romiosini/CeMoG.

Piperidis, S., Labropoulou, P., Deligiannis, M., and Giagkou, M. (2018). Managing public sector data for multilingual applications development. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Piperidis, S. (2012). The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 36–42, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajič, J., Choukri, K., Vasiļjevs, A., Backfried, G., Prinz, C., Pérez, J. M. G., Meerten, L., Lukowicz, P., van Genabith, J., Lösch, A., Slusallek, P., Irgens, M., Gatellier, P., Köhler, J., Bars, L. L., Auksoriūtė, A., Bel, N., Branco, A., Budin, G., Daelemans, W., Smedt, K. D., Garabik, R., Gavriilidou, M., Gromann, D., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Odijk, J., Ogrodniczuk, M., Ras, E., Rögnvaldsson, E., Rosner, M., Pedersen, B. S., Skadiņa, I., Tadić, M., Tufiş, D., Váradi, T., Vider, K., Way, A., and Yvon, F. (2020). The european language technology landscape in 2020: Language-centric and human-centric ai for cross-cultural communication in multilingual europe. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resource Association (ELRA).

Van den Heuvel, H., Kelli, A., Klessa, K., and Salaasti, S. (2020a). Corpora of disordered speech in the light of the gdp: Two use cases from the delad initiative. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resource Association (ELRA).

Van den Heuvel, H., Oostdijk, N., Rowland, C., and Trilsbeek, P. (2020b). The clarin knowledge centre for atypical communication expertise. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resource Association (ELRA).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Zinn, C. (2016). The CLARIN Language Resource Switchboard. In *Abstracts of the CLARIN Annual Conference 2016*, Aix-en-Provence, France.