D5.4

CLARIN Value Proposition

Document information

Title	CLARIN Value Proposition
ID	CLARINPLUS-D5.4 (CE-2016-0847)
Authors	Bente Maegaard, Dieter van Uytvanck, Steven Krauwer
Responsible WP leader	Martin Wynne
Contractual Delivery Date	2016-08-31
Actual Delivery Date	
Distribution	Public
Document status in workplan	Deliverable

Project information

Project name	CLARIN-PLUS
Project number	676529
Call	H2020-INFRADEV-1-2015-1
Duration	2015-09-01 – 2017-08-31
Website	www.clarin.eu
Contact address	contact-clarinplus@clarin.eu

Table of contents

1   Executive Summary

The value proposition will serve to show to the main stakeholders (i.e. the funders, the users and the providers) what the benefits are of joining CLARIN ERIC for a country and for its research community.

In this document we first give a short formulation of the CLARIN value proposition, followed by different elaborations and examples, taking into account the different perspectives of different categories of stakeholders.

We will present the value proposition and its elaborations to interested new stakeholders. The value proposition texts will be integrated in dissemination materials such as the website, folders, poster, etc., with the goal of attracting users, funders, new consortia, centres and member countries. Furthermore, a clear understanding of the value proposition will help defining CLARIN's priorities for the investments for sustainability and new developments in the coming years.

1   Executive Summary

2   Introduction

The purpose of this document is to describe the elements that constitute the CLARIN Value Proposition, taking into account that different stakeholders may assign different values to different elements of what we offer.

In Section 3 we will first present our Mission Statement, describing our ultimate objectives, our Vision Statement, describing the state we want to achieve as opposed to the present state, and the resulting Value Proposition that explains the benefits of joining CLARIN for our stakeholders.

In the same section we will also characterize the various groups of stakeholders.

As different stakeholders may have different interests and priorities we will present different elaborations of our overall Value Proposition in Section 4, for a number of typical groups of stakeholders. In Section 5 we turn to the contributions at the European level. It should be noted that some groups of stakeholders may have overlapping interests, which may result in a repetition in some of sections.

In the appendix we are providing a one-page description with many facts that can be useful in communication contexts.

This document will serve as a basis for documents (reports, brochures, posters, websites, etc.) that will be used to attract new members, users and other contributors and it should also serve for CLARIN itself as an instrument to define priorities and initiate new developments based on the needs and expectations of the various stakeholders.

3   Value proposition and stakeholders

Value proposition: CLARIN makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analyzing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting edge data-driven research, contributing to a truly multilingual European Research Area.

Mission: Create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences*.*

Vision: All digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences.

Three comments:

Disciplines: Even if the origin of CLARIN is in the humanities and the social sciences it should be noted that the offerings of CLARIN can have relevance for researchers from many other disciplines where language plays a role, not necessarily as object of study, but also as carrier of information or instrument for communication.

Open Science: More generally CLARIN does not see itself as a stand-alone facility, but rather as a player in making the vision that is underlying the emerging European Open Science Policies a reality, interconnecting researchers across national and discipline borders by offering seamless access to data and services.

Stakeholders: The construction and operation of an infrastructure involves many different stakeholders, each with their own interests and expectations, ranging from individual researchers, research institutions, data archives, infrastructure service providers, funding bodies and governments – just to mention a few.

4   Detailed value proposition at different levels

4.1   For researchers

4.1.1   Discovering and depositing resources

Researchers can search for language resources via metadata – in the CLARIN "catalogue", the Virtual Language Observatory (VLO) -, or in the data itself (Content Search). The VLO contains references to more than 800,000 resources the majority of which are hosted at CLARIN centres, but the VLO also contains references to other relevant resource collections.

Metadata Search: www.clarin.eu/vlo
Content Search: www.clarin.eu/contentsearch

As CLARIN ERIC has now many different countries as members, the VLO covers many languages, both national and regional languages, as well as languages studied in those countries.

The advantage of the VLO is faster identification of relevant resources, allowing researchers to re-use resources that already exist, rather than having to produce their own. Additionally, the VLO allows users who create or collect their own datasets to make them better visible to others through publication of the metadata in the VLO.

Without the VLO, researchers would have to use other catalogues, or general purpose search engines, often with less precise results – the VLO has a high number of query facets which can be used to guide the search and makes the result more precise.

Long term preservation: One of the fundamental services of the CLARIN infrastructure is making sure that language resources can be archived and made available to the community in a safe and sustainable manner. To help researchers to store their resources (e.g. corpora, lexica, audio and video recordings, annotations, grammars, etc.) in a sustainable way, at least one CLARIN data centre in each country offers a **depositing service**. These centres are willing to store the resources in their repository and assist with the technical and organisational details. This has a wide range of advantages:
- **Long-term archiving**: In every CLARIN member country there are one or more CLARIN data centres that have committed themselves to offering a storage guarantee for a longer period of time
- The resources can be cited easily and reliably as they will get a **Persistent Identifier**.
- All resources and their metadata will be equally accessible and searchable throughout the CLARIN infrastructure, irrespective of their physical location.
- All resources and their metadata can be combined, analysed and enriched with various **linguistic tools** (e.g. automated part-of-speech tagging, information extraction, phonetic alignment or audio/video analysis), irrespective of the physical location of data and tools.

Researchers as providers: Researchers are not just consumers of data and tools, but also providers in that they are encouraged to share their data and tools with others, if necessary in a protected way, so that others can build further on their results. This is supported by the availability of repositories (see above) and by facilitating data citation and licensing (see below).

Data citation: It is a major task to produce a corpus or a data collection. More and more, the scientific world recognises the value of such contributions and mechanisms have been developed for data citation that encourage creators of corpora or other data collections to publish their data. CLARIN is a perfect platform for this type of publication and subsequent citation. This is because CLARIN offers a good search tool (VLO) for better publicity and better sharing, and because of the use of Persistent Identifiers to refer to them instead of notoriously unstable URLs.

Licensing: CLARIN centres make data available through licensing and clear conditions for use. This involves CLARIN centres making deals with rights owners, signing Deposition License Agreements which include End User License Agreements, categorizing licenses in clearly marked license categories, and writing Terms of Service. Guidance is offered to creators of data who want to select the most appropriate licensing conditions when publishing their data.

4.1.2   Advanced tools and computing facilities

CLARIN offers state-of-the-art tools and on-line services for many languages. These support researchers to annotate, analyse and publish their language data. Automatic analyses such as e.g. annotation can be performed faster and more easily on large amounts of data. Examples of the functionality offered:

(1) Advanced analysis and visualizations for large data sets that may help gaining deeper insights, e.g.:
   a. DiaCollo: collocation analysis in diachronic perspective
   b. Stylo: state-of-the-art tool for stylometric analysis
(2) Faster automated analysis, leading to more time for the actual research, e.g.:
   a. WebMAUS: Automatic Segmentation and Labelling of Speech Signals over the Web
   b. AVAtech: audio and video recognizers
(3) Reproducible scientific analysis flows, leading to more data sharing and better replicability of research results, e.g.:
   a. Mind Repository: a platform for researchers to share their papers together with the data they have used and the scripts to analyse them
   b. Web service orchestration engines
(4) Access to first-class corpora through specialized query interfaces, e.g.:
   a. The Corpus of Contemporary Dutch
   b. The ICAME corpora

For large and computation-intensive tasks, CLARIN can connect scientists to highly ranked High Performance Computing (HPC) centres. Depending on the amount of computing resources needed, the researcher might need to enrol in a competitive call. In any case there are no costs for the use of the HPC facilities.

4.1.3   Federated Login: easier access to more resources

CLARIN has established a Service Provider Federation, i.e. a trusted network of identity providers which offer "single sign-on". This means that researchers can login with their institutional credentials to get access to password-protected language resources and applications in other countries.

One advantage for the researchers is that they gain time and have the benefit of having to use only one access code. Another advantage is that they also get access to protected resources in other countries. Without this single sign-on, researchers would either have no access to otherwise valuable resources, or they would have to apply for accounts for each repository.

Statistics show that about 60 unique visitors per day use the federated login to access CLARIN resources[1].

You can find some of the online resources to which CLARIN gives easy access here: https://www.clarin.eu/content/easy-access-protected-resources

4.1.4   Access to expertise for researchers

Complementary to the access to data and tools CLARIN offers researchers access to expertise through its Knowledge Sharing Infrastructure. First of all, there are help desks and knowledge centres.

All CLARIN centres offering access to data and tools operate a helpdesk (in English) where (potential) users can get information about the data and services offered, get help in using the services, and report problems.

Certified CLARIN knowledge centres (https://www.clarin.eu/content/knowledge-centres) offer access to expertise in specific areas, such as treebanking, speech analysis, audio-visual fieldwork, language learning, the Danish language, the languages of Spain, the languages of Sweden.

Specialized committees bring together expert knowledge about various topics, such as IPR and licences (CLIC – the CLARIN Legal Issues Committee) and standards (STAC - the Standards Committee).

The CLARIN ERIC office operates a central helpdesk that will reply directly to requests for help or information, or channel the requests to the best experts.

4.1.5   Online tutorials etc.

The Knowledge Sharing Infrastructure also offers online tutorials and explanations of use cases in order to facilitate the uptake of new techniques by researchers. This material is provided by experts from national CLARIN consortia, and wherever possible, offered at least in English or (in the case of movies) with English subtitles.

The use cases serve to demonstrate the successful application of digital methods to specific research questions, and to inspire researchers with similar or related questions.

4.1.6   Workshops and seminars, mobility grants

The final part of the knowledge sharing infrastructure consists of direct and physical meetings of various types.

Through an open call for workshop proposals members of the CLARIN community are invited to submit proposals for workshops on strategic priorities for CLARIN, or for the preparation of small development projects.

---

[1] Measured using Piwik during the first half of 2016 at the CLARIN discovery service. As not all

From time to time CLARIN ERIC organises workshops focused on specific topics where members of the CLARIN community get together and work on topics of common interest.

At the Annual CLARIN Conference infrastructure providers and users from all CLARIN countries exchange expertise and ideas.

A cross-border mobility scheme for short stays at CLARIN centres helps the exchange of knowledge and expertise between staff of different centres, or between staff from centres and users.

4.1.7   Support for data management plans

Since CLARIN's main purpose is to facilitate the preservation and exploitation of resources, CLARIN is in a good position to give professional support for writing a Data Management Plan (DMP), e.g. for research funding applications, thereby improving the quality of the application, and at the same time saving time and effort.

As an example, the German group CLARIN-D had a presentation at the 2015 CLARIN conference by Trippel and Zinn: DMPTY – A Wizard For Generating Data Management Plans. This is closely related to the German repositories, but may be adapted to serve in other countries.

4.2   For CLARIN institutions and institutes

4.2.1   Guaranteed access to CLARIN central services

CLARIN offers a wide range of central services from which all users and all CLARIN centres in CLARIN ERIC member countries benefit. These services are developed and offered in close collaboration between the national teams and CLARIN ERIC.

Here we describe very briefly some of those services which are offered to institutions. NB: Not only access, but also influence on the development of a wide range of tailored infrastructural services is offered.

- The unique CLARIN Service Provider Federation which makes "single sign-on" possible across countries and institutions.
- The CLARIN Identity Provider, catering for academics without an institutional Identity Provider
- The central Discovery Service[2] with high, i.e. 24/7, availability for the users. This service provides a user-friendly uniform interface to select a home organization during the federated login process.
- Robust Persistent Identifier services, including mirroring of the handle servers, via EPIC - the European Persistent Identifier Consortium[3]
- As official EUDAT community, access to advanced EUDAT services: e.g. B2DROP (workspaces) and B2SAFE (safe replication)

---

[2] See https://www.clarin.eu/content/clarin-discovery-service
[3] See http://www.pidconsortium.eu

- Services for creating, sharing and re-using rich metadata descriptions: Component Registry (creation and editing of metadata schemas), Concept Registry (registration and re-use of semantic definitions), OAI harvester (automated distribution of metadata files)
- The CLARIN centre registry, a central dashboard to register services and to connect them to automated checking and publication processes (monitoring, harvesting, content search, etc.)
- Professional monitoring services (Icinga), with high availability.
- Advanced distributed user statistics, using Piwik.
- Access to the common development environment (SVN code repository and Trac wiki).

Institutions and institutes are not only the homes of users of data and services, but they are also providers. Research institutes can share the results and spin-offs of their research through CLARIN and make them more visible, and institutions with an archiving function can use CLARIN to make their holdings, often created with public funding, more widely visible and usable.

4.2.2   Participation in H2020 projects

CLARIN ERIC is actively applying for participation in European projects when such participation supports CLARIN's general strategy and/or development. This means that joining CLARIN ERIC opens enhanced opportunities for participation in European funded projects, as staff of national CLARIN consortia may work on such projects on behalf of CLARIN ERIC.

Currently (August 2016) CLARIN ERIC is coordinating the H2020 project CLARIN-PLUS, which aims at further consolidation and expansion of the CLARIN infrastructure, both technically and organisationally.

Additionally, CLARIN ERIC is participating in four other European projects: PARTHENOS, EUDAT2020, LT-Observatory, and EUROPEANA-DSI. At this moment staff from more than a dozen CLARIN ERIC Member or Observer countries is participating in EU projects on behalf of CLARIN ERIC for a broad variety of tasks in terms of topic and size. More information about ongoing EU projects with CLARIN ERIC involvement can be found on our website on https://www.clarin.eu/content/clarin-eu-projects .

4.2.3   Access to cloud services
For members CLARIN ERIC provides free access to various commercially available cloud services:
- Communication tools: Adobe Connect for video conferences, Slack for instant messaging
- Github for code sharing, including private repositories
- Basecamp for project management.

4.2.4   Access to expertise for CLARIN institutions and institutes
To institutions planning to establish themselves as CLARIN technical centres or aiming to keep themselves up-to-date with recent developments we offer various types of support, including funding for travel costs involved.

- Best practice and other documentation on the creation and maintenance of infrastructure services is available, and will be continuously updated.
- Thematic workshops or tutorial sessions on specific infrastructure topics are organized to bring together centre staff from new and established centres
- A mobility scheme for short (typically up to 1 week) exchanges between new and established centres.

4.3   For CLARIN ERIC Members (countries, ministries)

In addition to the above-mentioned benefits for the country's researchers and institutions, the following national strategic advantages can be identified:

CLARIN membership may contribute to a better position of the national language(s) in the European and international context by making data more visible and more widely accessible.

Visibility of cultural content increases: As much of a country's cultural content is of linguistic nature or described by means of language, disclosing it through CLARIN will make it more visible, both for the research community and for the public at large. Conversely membership of CLARIN ERIC will give researchers in the country full access to the cultural heritage in other CLARIN countries.

As generic infrastructure services can be used across borders, CLARIN members can benefit from the fact that the costs of construction and operation of such services can be shared between members. Software of a generic nature can be shared between members, and in many cases the development cost of more specific software can be reduced by porting it between languages rather than doing it from scratch.

CLARIN members have direct influence on the decision-making about all aspects of the infrastructure, ranging from construction, operation, longer term evolution and strategic priorities, as voting member in the General Assembly, the National Coordinators' Forum, etc., as well as through participation in the various committees. More details about the governance of CLARIN ERIC can be found in the Statutes, see http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-75.

Access to expertise for new CLARIN countries:

For emerging national consortia in countries that have just joined or are preparing themselves for joining CLARIN, CLARIN offers workshops on how to set up the CLARIN infrastructure at the national level. Four workshops for new and future members are being held 2015-2017, and the information materials produced will be available on the CLARIN website.

Additionally, a series of best practice papers is being prepared on the following themes
- Building a national consortium
- Living up to the criteria for CLARIN centres
- Building the Knowledge Sharing Infrastructure at the national level
- Cost estimates for the construction and operation
- Professional profiles for typical CLARIN infrastructure functions
- IPR and ethical issues
- Interoperability and standards

5   Values for Europe

Excellence: CLARIN's role is not just to offer infrastructure services per se, but also to contribute to the further development of the European Research Area. In this context CLARIN has a number of things to offer to support cutting edge research and excellence in European research.

Cross-border, cross-language research: CLARIN offers the same access to data and services to all users. This is a necessity for joint research and for the replication of research results, which has become a very important factor in many disciplines. At the same time (and maybe even more importantly) CLARIN is also offering access to *similar data collections[4] and services* to the European humanities and social sciences research community at large. This creates an excellent starting point for joint cross-border and cross-language research, and exchange of expertise. This will stimulate the collaboration for communities with a research agenda that requires the exploration of language data from a comparative perspective. As language is a data type that captures linguistic, cultural and social phenomena, the potential for multidisciplinary research will also increase.

Open Science: CLARIN is not a stand-alone facility, but is well-embedded in the European research infrastructure landscape at large, and as such fully committed to the European Open Science Policies, including re-use of data. As mentioned in 4.1.7 a focus on data management is inherent in CLARIN's data preservation activities. Through its focus on language it offers opportunities for cross-discipline collaboration between areas where language plays a role, both within the humanities and social sciences and in other disciplines.

Various instruments will be used to support this, including:
•   Workshops on typical cross-border, cross-language or cross-discipline topics in order to ensure a continuous flow of knowledge and expertise between different communities
•   Initiatives for or participation in EU projects with a strong language dimension that bring together researchers from national CLARIN consortia, thereby providing excellent input and expertise for these projects.

---

[4] newspaper collections, collections of parliamentary debates etc

6   Concluding remarks

This value proposition will serve to show to the main stakeholders (i.e. the funders, the users and the providers) what the benefits of joining CLARIN ERIC are for a country and for its research community.

We present the value proposition and its elaborations to interested new stakeholders in the form of e.g. information packages or presentations at events where relevant potential stakeholders are present. It is also used to formulate texts on dissemination material such as the website, folders, poster, etc., with the goal of attracting users, funders, new consortia, centres and member countries.

Furthermore, it will also be used as a reference point for CLARIN itself, as a clear understanding of the value proposition will help defining CLARIN's priorities for the investments for sustainability and new developments in the coming years.

A last important comment: the value proposition should not be seen as set in concrete: the whole research landscape is changing fast in many respects, and CLARIN will adapt itself to new developments if this is necessary to achieve its goals as reflected by its mission and vision statements.

References

CLARIN ERIC Statutes

ESFRI 2016 Strategy Report of Research Infrastructures. (The appendix on next page is page 82 of this report, part of the Landmark section. Note however, that the list of countries has been updated with the two countries that joined in the meantime: Latvia and Hungary).

Apart from this, throughout the document many links are made to the CLARIN ERIC website, www.clarin.eu , and other websites where documentation is found. These links are not repeated here.

Appendix

CLARIN ERIC, Common Language Resources and Technology Infrastructure, cf. ESFRI 2016

Single sign-on access to language data repositories and tools to explore, exploit, analyse, enrich or combine them

TYPE: distributed, virtual
COORDINATING COUNTRY
NL
MEMBER COUNTRIES
AT
BG
CZ
DE
DK
EE
EL
FI
HU
IT
LT
LV
NL
NO
PL
PT
SE
SI
PARTICIPATING COUNTRIES
UK, ES, USA
OTHER MEMBERS
Dutch Language Union

TIMELINE
- ESFRI Roadmap entry: 2006
- Preparation phase: 2008-2011
- Construction phase: 2011-2015
- Operation phase: 2012
- Legal entity establishment: ERIC, 2012

ESTIMATED COSTS
- Capital value: not applicable
- Operation: 12 M€/year

HEADQUARTERS
CLARIN ERIC
Utrecht University
The Netherlands

WEB SITE
http://www.clarin.eu

DESCRIPTION

The Common Language Resources and Technology Infrastructure (CLARIN) provides easy and sustainable access for scholars in the humanities and social sciences to digital language data – in written, spoken or multimodal form – and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of their location. To this end CLARIN is building a networked federation of language data repositories, service centres and centres of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres are interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. CLARIN integrates existing data and service centres, without major capital investments.

Currently the CLARIN infrastructure is under construction, but a growing number of participating centres is already offering services to data, tools and expertise. Since 2012, CLARIN is a European Research Infrastructure Consortium (CLARIN ERIC).

ACTIVITY

The operation, maintenance and continuous expansion of the infrastructure is carried out in the participating countries at the individual CLARIN centres (currently over 30), which have to meet clearly defined technical and organizational criteria, ensuring the coherence of the whole infrastructure. CLARIN also works closely together with research communities in creating and expanding a knowledge infrastructure that can support developers of language resources and tools, as well as the end-users of the available data and services. In total over 100 institutes across Europe are involved. CLARIN has been offering operational access and services since 2012, and the offer will grow as centres expand and new countries join. To stimulate the uptake and increase the insight in the usability of the services, CLARIN participates in the development of courseware and organises workshops and data camps focussing on the processing and analysis of specific data types, such as parliamentary data, social media data, and historical texts. The collaboration with other RIs that support scholarly communities in Social Sciences and Humanities (SSH) takes places via the collaboration in H2020 consortia. CLARIN promotes open access to research data within the present legislation and ethical standards and will develop outreach activities to stimulate the reuse and repurposing of the data resources integrated.

IMPACT

CLARIN stimulates the reuse and repurposing of available research data, thereby enabling scholars in SSH (including digital humanities) to increase their productivity and, more importantly, open new research avenues in and across disciplines that address one or more of the multiple societal roles of language: as a carrier of cultural content and information, both synchronically and diachronically, as a reflection of scientific and societal knowledge, as an instrument for human communication, as one of the central components of the identity of individuals, groups, cultures or nations, as an instrument for human expression, or as an object of study or preservation. Through the access and discovery services, CLARIN increases the potential impact of data and tools produced with publicly funded projects. Working with CLARIN data and tools will increase the skill levels for data analysis tasks among the new generations of SSH students, which is likely to be welcomed by the data science sector.