CLARIAH-ES: Strategic network for the integration in the European research infrastructures in Social Sciences and Humanities

Francisco Javier Carreras[1], Ainara Estarrona[2], Aritz Farwell[2], Mikel Iruskieta[2], Manuel Marco[3], Maite Melero[4], Arturo Montejo-Ráez[5], Daniel Riaño[6], German Rigau[2], Dolores Romero[7], Salvador Ros[8], Elena Sánchez[9] and Xulio Sousa[10]

[1]*IATEXT, University of Las Palmas de Gran Canaria, Spain*
[2]*HiTZ Center, University of the Basque Country (UPV/EHU), Donostia-San Sebastian, Spain*
[3]*University of Alicante, Spain*
[4]*Barcelona Supercomputing Center (BSC-CNS), Spain*
[5]*CEATIC, University of Jaen, Spain*
[6]*ILC-CSIC, Madrid, Spain*
[7]*LOEP, Complutense University of Madrid, Spain*
[8]*LINDH, UNED, Madrid, Spain*
[9]*Biblioteca Nacional de España (BNE), Madrid, Spain*
[10]*Instituto da Lingua Galega, University of Santiago de Compostela, Spain*

Abstract
The CLARIAH-ES strategic research network will support and contribute to the management and national coordination of the European Research Infrastructure Consortia (ERIC) CLARIN (focused on digital data and processes related to Language) and DARIAH (focused on digital data and processes related to arts and humanities scholars). In the current context of digital transformation, study, research and development in the humanities, arts and social sciences require scientific and technological infrastructures that allow for the computational processing of textual, visual, numerical and/or audio data. These infrastructures promote multilingualism, digital methods, interoperability, maintenance and reuse of resources, open science, visibility and scientific cooperation in Europe, thus overcoming the fragmentation of research communities and increasing the impact of their research. Although both CLARIN and DARIAH are independent ERIC infrastructures, some European countries have formed joint CLARIAH consortia. The INTELE strategic network (2020-2022) articulated and agreed upon a common proposal for a national CLARIAH-ES consortium that has allowed for the official incorporation of Spain into both infrastructures as of September 2023.

Keywords
ERIC, infrastructure, Humanities, Social Sciences, Arts, CLARIN, DARIAH, CLARIAH,

1. Introduction

The CLARIAH-ES strategic research network is funded by the Spanish Ministry of Science, Innovation and Universities within the framework of the State Program to Promote Scientific-Technical Research and its Transfer (RED2022-134527-E).[1] It continues and consolidates the work carried out by the INTELE strategic network [1], also supported by the Ministry of Science, Innovation and Universities (RED2018-102797-E).[2]

CLARIAH-ES consists of Spanish researchers who are associated, by previous participation or current interest, with the two principal European research infrastructures for the humanities and social sciences CLARIN[3] [2] and DARIAH,[4] [3] each constituted as a European Research Infrastructure Consortium (ERIC). The general objective of the CLARIAH-ES network is to encourage activities that promote these infrastructures and to attain Spain's official incorporation into them, which was achieved in September 2023. Spain's participation will contribute to the advancement of Spanish research in the humanities and social sciences, as well as to its strategic positioning in international projects and programs, mainly in the context of the European Research Area.

SEPLN-CEDI2024: Seminar of the Spanish Society for Natural Language Processing at the 7th Spanish Conference on Informatics
francisco.carreras@ulpgc.es (F. J. Carreras); ainara.estarrona@ehu.eus (A. Estarrona); aritz.farwell@ehu.eus (A. Farwell); mikel.iruskieta@ehu.eus (M. Iruskieta); marco.such@gcloud.ua.es (M. Marco); melero@bsc.es (M. Melero); amontejo@ujaen.es (A. Montejo-Ráez); daniel.rianno@cchs.csic.es (D. Riaño); german.rigau@ehu.eus (G. Rigau); dromero@filol.ucm.es (D. Romero); sros@scc.uned.es (S. Ros); elena.sanchez@bne.es (E. Sánchez); xulio.sousa@usc.es (X. Sousa)

[1]https://ixa2.si.ehu.eus/red-clariah-es
[2]https://ixa2.si.ehu.eus/intele/home
[3]www.clarin.eu
[4]www.dariah.eu

With this in mind, CLARIAH-ES seeks to bring together groups that have a stake in these European research infrastructures and that wish to reduce the digital divide, promoting new multidisciplinary lines of research in the humanities, arts and social sciences (and beyond) by facilitating their digital transformation with the help of language technologies.

2. Previous and Related Work

Participants in the CLARIAH-ES network have been working with CLARIN and DARIAH for many years and their joint effort has increased the visibility of the infrastructure and helped it become operable. Their previous and current cooperation includes, for example:

The *Spanish CLARIN K-Centre*[5] [4] originated out of the University Institute of Applied Linguistics at Pompeu Fabra University (IULA-UPF).[6] Since 2015, all requests for help and/or collaboration in word processing are responded to in less than forty-eight hours, although not all requests are always adequately responded to due to a lack of resources. With this network and the creation of CLARIAH-ES it will be possible to better respond to requests, create resources and training materials and disseminate the work that will be done in the network.

The *IMPACT CLARIN K-Centre*[7] is a consortium of six European institutions and has been maintained at the University of Alicante since 2019. Its mission is to make the digitization of printed historical texts better, faster, and more economical. The center provides tools, services and facilities to further advance the fields of document digitization, language technology, and historical text processing.

The *Digital Laboratory of Digital Humanities*[8] (LINHD-UNED) has been a DARIAH Cooperating Partner since July 2022.

3. Objectives

The general objective of the strategic network is to coordinate the creation, development, and dissemination of the CLARIAH-ES research infrastructure. In so doing, we hope to contribute to the advancement of the research in the humanities, arts, and social sciences and to its strategic positioning in national and international projects, mainly in the context of the European Research Area. The more specific objectives of CLARIAH-ES are:

- Design and coordinate the CLARIAH-ES infrastructure.
- Integration within CLARIN and DARIAH research infrastructures.
- Understand the requirements of researchers, groups, and projects that need the support of digital content, tools, and resources.
- Promote and facilitate the participation of research communities that are working in the humanities, arts, and social sciences in CLARIAH-ES.
- Produce the required annual reports and documentation describing the progress of the infrastructure.

4. Methology and Work Plan

The activities of the research network are grouped into six work packages:

WP1 Project management (Leader: HiTZ-UPV/EHU). General project management to cover all administrative activities, supervision and coordination of the other work packages, as well as the promotion and dissemination of the project results. This WP includes the management and coordination of the CLARIAH-ES strategic network and the coordination and collaboration with CLARIN-EU and DARIAH-EU.

WP3 CLARIAH-ES Infrastructure (Leader: HiTZ-UPV/EHU). Design and coordination of the CLARIAH-ES infrastructure: committees, managers, work packages, centers, activities, CLARIAH-ES office, etc. This WP includes the design of CLARIAH-ES infrastructure design. The deployment and coordination of the CLARIAH-ES infrastructure (see WP3 and WP4). Continuous training in the CLARIAH-ES infrastructure. Surveys to collect data and unite needs and objectives and the assessment of the work carried out in the CLARIAH-ES network.

WP3 CLARIN (Leader: HiTZ-UPV/EHU). Coordination of CLARIN-EU activities in CLARIAH-ES. The coordination, management and deployment of CLARIN-EU activities is one of the main objectives of the CLARIAH-ES network. In addition, CLARIAH-ES must participate in the different committees and working groups of CLARIN-EU, as well as in its annual meetings, workshops and conferences. In particular, this WP covers the design and specification of CLARIN-EU in-kind contributions. The deployment and coordination of CLARIN-EU activities in CLARIAH-ES. The participation in CLARIN-EU committees and work-

[5] http://ixa2.si.ehu.eus/clarin-es/en/node/41

[6] https://www.upf.edu/es/web/e-noticies/assetlang/-/asset_publisher/s63mYpRtW3TT/content/presentacio-de-l-iula-upf-centre-de-competencies-clarin/10193

[7] https://www.digitisation.eu/

[8] https://linhd.uned.es/?lang=en

ing groups and the participation in annual meetings, workshops and conferences organized by CLARIN-EU.

WP4 DARIAH (Leader: LINHD-UNED). Coordination of DARIAH-EU activities in CLARIAH-ES. The coordination, management and deployment of DARIAH-EU activities is another main objective of the CLARIAH-ES network. Likewise, CLARIAH-ES must also participate in the various committees and working groups of DARIAH-EU, as well as in its annual meetings, workshops and conferences. In particular, this WP covers the design and specification of DARIAH-EU in-kind contributions. The deployment and coordination of DARIAH-EU activities in CLARIAH-ES. The participation in DARIAH committees and working groups and the participation in annual meetings, workshops and conferences organized by DARIAH-EU.

WP5 Workshops (Leader: BVMC-UA). Organize on an annual basis a series of workshops on research that can be developed thanks to the services, resources, and tools of research infrastructures in the humanities, arts and social sciences. The workshops are an essential space for the community to learn about the possibilities offered by the CLARIN-EU and DARIAH-EU infrastructures, carry out a needs analysis of CLARIAH-ES, establish connections and collaborations between research groups, publish the most relevant works and projects of the community, and seek collaboration on open problems jointly. In particular, this WP plans to organize several meetings and two workshops.

WP6 Communication and dissemination (Leader: HiTZ-UPV/EHU). A strategic communication plan will be established, focusing on impact and visibility in different areas and for different communities and types of users. Communication channels will be established through which the activities and results of the project will be disseminated: web pages, mailing lists, social networks, blogs, and other appropriate media.

5. Network Members

The ambitious goals of CLARIAH-ES can only be achieved by bringing together the necessary resources in terms of data, computing facilities and knowledge that are not available to any one research group in Spain. CLARIAH-ES is formed by a multidisciplinary group of **ten** leading research centers in Language Technologies (TL), Artificial Intelligence (AI), High Performance Computing (HPC), linguistic experts in the official languages in Spain (Spanish, Catalan, Basque and Galician), and experts in digital transition in the areas of humanities, arts and social sciences.

The current CLARIAH-ES consortium is made up of the UPV/EHU (HiTZ), the University of Santiago de Compostela (Instituto da Lingua Galega and CiTIUS), the University of Alicante (Miguel de Cervantes Virtual Library Center), the UNED (LINDH and LENAR), the Barcelona Supercomputing Center (BSC), the Complutense University of Madrid (UCM), the University of Jaén (CEATIC), ULPGC (IATEXT), the CSIC (Spanish National Research Council) and the National Library of Spain (BNE) (See Figure 1).

HiTZ - Basque Center for Language Technology[9] (UPV/EHU) is a multidisciplinary research center with members from different departments of the University of the Basque Country. The aim of the center is to research language and speech technologies. It is formed by two research groups Aholab and Ixa, both with extensive experience in the field of language and speech technologies, performing basic research, creating resources and tools, and launching several commercial products to the market. HiTZ is also a center of reference for endangered and under-resourced languages. The center currently has 80 members. It is a leader in the application of deep learning techniques to language and speech processing. HiTZ is also a member of the European Erasmus Mundus+ Masters Program in Language and Communication Technologies (LCT) and offers a Doctoral Program in Language Analysis and Processing. From 2017 to 2021 HiTZ participated in CLARIN's Knowledge Sharing Infrastructure (KSI) commission and is a coordinating member of the Spanish CLARIN-K Centre and the central office of CLARIAH-ES.

BVMC - Biblioteca Virtual Miguel de Cervantes[10] (UA) was created in 1999 and was established as the Center for Digital Humanities at the University of Alicante in 2021. Its purposes are to promote research, knowledge transfer, design and development of technologies related to the humanities and digital libraries. The Center makes relevant cultural works in the different Hispanic languages available to users around the world free of charge, as well as the most relevant research and studies surrounding them. It also has the missions of promoting research in the different areas of the digital humanities and carrying out the development and analysis of technological tools and services that facilitate the use and exploitation of the growing set of digitized materials. The BVMC is a member of the IMPACT Competence Centre, a CLARIN K-Centre.

BSC-CNS - Barcelona Supercomputing Center - Centro Nacional de Supercomputación[11] is a leading

[9] https://www.hitz.eus/
[10] https://www.cervantesvirtual.com/
[11] https://www.bsc.es/

Figure 1: Map of the groups participating in the network.

multidisciplinary research center, supported by a public consortium formed by the Spanish and Catalan public administration and the Polytechnic University of Catalonia. The BSC-CNS hosts high-performance computing infrastructures serving the international scientific community and is a Tier 1 member of the European Association for Advanced Computing in Europe (PRACE) research infrastructure. BSC also manages the Spanish Supercomputing Network (RES), a Singular Scientific-Technical Infrastructure and supports the international biomedical community, coordinating the Elixir and INB-ISCII infrastructures. The different research areas, grouped in four Departments (Computer Sciences, Life Sciences, Earth Sciences and Computer Applications), have European, national and regional funding, mostly competitive, as well as through collaboration with leading companies. Several of the research projects currently underway belong to the fields of Natural Language Processing, Artificial Intelligence, Social Sciences and Humanities and Digital Art.

ILG[12] and **CiTIUS**[13] (USC) has carried out since 1971 intense research activity in the fields of Galician linguistics and philology and, at the same time, in the development of technological tools and applications that make available to the academic community and society in its the knowledge generated within the framework of research activities. CITIUS develops its research in ten areas, including Linguistic Technologies. The center stands out for its work in Natural Language Processing, both in basic research related to the creation and evaluation of linguistic resources and models as well as in the development of applications.

CLARIAH-CM, led by Universidad Complutense (UCM),[14] in which all public universities of the Comunidad de Madrid participate: Universidad de Alcalá, Universidad Autónoma de Madrid, Universidad Carlos III, Universidad Rey Juan Carlos and Universidad Politécnica.

LINDH[15] and **LENAR**[16] (UNED) are both leading research centers in the area of digital humanities and natural language processing. UNED is characterized by teaching and research excellence and by promoting the transfer

[12] https://ilg.usc.gal/en
[13] https://citius.gal/
[14] https://www.ucm.es/loep/equipo
[15] https://linhd.uned.es/
[16] https://sites.google.com/view/nlp-uned/home

of knowledge in all areas of knowledge, but especially in the field of language technologies applied to science and digital humanities. UNED is a cooperating partner of DARIAH and is part of the Spanish CLARIN-K Center. UNED has participated since 2014 in H2020 projects together with DARIAH-EU and has technological services aimed at the dissemination of the results of digital humanities in its central library.

CCHS - Centro de Ciencias Humanas y Sociales[17] (CSIC). The Consejo Superior de Investigaciones Científicas (Spanish National Research Council, CSIC) is the largest research institution in Spain. It organizes its research projects across 8 Global Areas. Within the Humanities and Social Sciences Global Area, there are 14 institutes and schools. The Center for Human and Social Sciences (CCHS) in Madrid serves as an integrated service center supporting researchers from six of these institutes: the Institute of Languages and Cultures of the Mediterranean and the Middle East (ILC), the Institute of Philosophy (IF), the Institute of Economy, Geography, and Demography (IEGD), the Institute of Public Goods and Policies (IPP), the Institute of History (IH), and the Institute of Language, Literature, and Anthropology (ILLA). The CCHS comprises 18 departments and 62 research groups, maintaining constant interaction and collaboration with other CSIC institutes and schools. Several Humanities and Social Sciences projects are among the 72 research groups participating in the CSIC's AIHub.

CEATIC - The Center for Advanced Studies in Information and Communication Technologies [18] (UJA) aims to bring together research groups, resources and instrumental means that allow the advancement of knowledge, development and innovation in the field of information and communication technologies through education, scientific research and technological development of excellence. The research group on Intelligent Information Access Systems (SINAI) at the University of Jaén has more than twenty years of experience in NLP research, with uninterrupted funding through competitive calls (European, national, regional, local) and around twenty industrial technology transfer projects.

IATEXT - The Research Institute of Text Analysis and Applications[19] (ULPGC) is currently made up of nine divisions distributed between different areas within the fields of humanities and computer science. Its research focuses on the revision and analysis of different types of texts from interdisciplinary perspectives (linguistic, literary, historical, computational, heritage, etc.), as well as the computational treatment and digitization of any type of data in these areas. IATEXT's general objective is to produce results in basic research and develop multimedia computer applications for research in addition to educational, cultural and professional fields. IATEXT has participated in more than twenty funded projects and has developed more than twenty tools and applications in the fields of digital humanities and social sciences.

BNE - La Biblioteca Nacional de España [20] is the central depository for the Spanish bibliographic and documentary heritage that is produced within Spain and about Spain abroad in any type of support or medium. Its mission is to gather, catalog, preserve, increase, manage, disseminate and transmit, in compliance with its purposes, this heritage as a source of knowledge for the whole of society. The BNE participates in numerous initiatives and projects in collaboration with national and international institutions and research groups. The digital transformation processes at the institution have led to the availability of new digital resources, textual corpora, and data sets, creating great potential for research. The BNE actively promotes dissemination and use, as stated in its Strategic Plan, within the framework of the BNElab program to encourage innovation and digital reuse, including through support for the development of textual technologies, Artificial Intelligence or digital humanities.

6. Expected Results and Impact

The consequences of the lack of research infrastructures for the humanities and social sciences in Spain are evident. As we confirmed in the INTELE strategic network, researchers require data (difficult to access or dispersed) as well as tools and work methods (often specific for Spanish and other co-official languages such as Galician, Catalan and Basque) that can only be produced by their own infrastructures for their own needs. Without these, Spanish researchers risk losing their presence in eHumanities publications and are unable to participate in competitive projects and European research infrastructures.

Fortunately, Spanish researchers have demonstrated clear signs of interest in exploiting these distributed infrastructures, fully operational in Europe since 2016 (ESFRI LANDMARK 2016).[21] Thus, the objective of the CLARIAH-ES network is to design, coordinate and deploy the CLARIAH-ES infrastructure, promoting the participation of Spanish researchers in the European CLARIN-ERIC and DARIAH-ERIC infrastructures. Through these efforts, we anticipate an increase in research projects and scientific production, as well as an improvement of national and international positioning in the areas of the humanities, arts, and social sciences.

The main results we expect from the CLARIAH-ES strategic network are:

[17] https://www.cchs.csic.es/es
[18] https://www.ujaen.es/centros/ceatic/en
[19] https://iatext.ulpgc.es
[20] https://www.bne.es/en
[21] https://www.esfri.eu/

- Increased visibility of the new CLARIAH-ES infrastructure (data, services and tools, success stories, research communities, etc.).
- Increased sustainability and visibility of the results of national research projects.
- Increased interdisciplinarity and multidisciplinarity.
- Increase in opportunities for national, European, and Ibero-American research and collaboration.
- Increase in funding and the possibility of obtaining projects at the European level through the articulation of Spain, as a result of this network collaboration, within the infrastructures.
- Increased visibility of national research in the eHumanities at the European level, along with CLARIN-EU and DARIAH-EU and the rest of the European SSHOC infrastructures.
- Increased interaction with the cultural and creative industry (GLAM).
- Increased interaction with the agents of the new language economy.

In summary, the CLARIAH-ES strategic network seeks a tangible impact on society. Through the greater exchange of knowledge, data, technologies, infrastructures, skills, and best practices, we aim to amplify the potential of research projects. This collaborative synergy will not only ensure the sustainability of tools and services, but also foster collaborative environments. Our aspirations transcend borders, encompassing Europe, Ibero-America, and the global stage as we endeavour to increase funding opportunities for vital infrastructures. The interdisciplinarity of the participating research groups, which includes areas of research as diverse as computer science, philology, social sciences, history, etc., ensures a broad contribution from different perspectives. By properly weaving together these perspectives, we can develop research results that are useful for promoting high-impact digital tools and artificial intelligence applications in different social science and digital humanities scenarios, including research and cultural infrastructures such as libraries and museums [5].

Acknowledgments

CLARIAH-ES is a strategic research network funded by the Spanish Ministry of Science, Innovation and Universities within the framework of the State Program to Promote Scientific-Technical Research and its Transfer (*Plan Estatal de Investigación Científico-Técnica y su Transferencia*) (RED2022-134527-E).

References

[1] M. Iruskieta, A. Estarrona, A. Farwell, G. Rigau, INTELE: promoviendo la participación en las infraestructuras: CLARIN y DARIAH, in: The International Congress on Libraries & Digital Humanities: Projects and Challenges, 2022.

[2] E. Hinrichs, S. Krauwer, The CLARIN research infrastructure: Resources and tools for ehumanities scholars, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC2014), European Language Resources Association (ELRA), 2014, pp. 1525–1531.

[3] J. Edmond, F. Fischer, M. Mertens, L. Romary, The DARIAH ERIC: Redefining research infrastructure for the arts and humanities in the digital age, ERCIM News (2017).

[4] N. Bel, E. González-Blanco, M. Iruskieta, CLARIN Centro-K-Español, Procesamiento del Lenguaje Natural 57 (2016) 151–154.

[5] OECD, Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, 2023. URL: https://www.oecd-ilibrary.org/content/publication/a8d820bd-en. doi:https://doi.org/https://doi.org/10.1787/a8d820bd-en.