new_text = text.replace('<td>', '<td>
return(new_text)
def new_file(filename):
newfile = open(path + '/' + filename +
newfile.write(new_text)

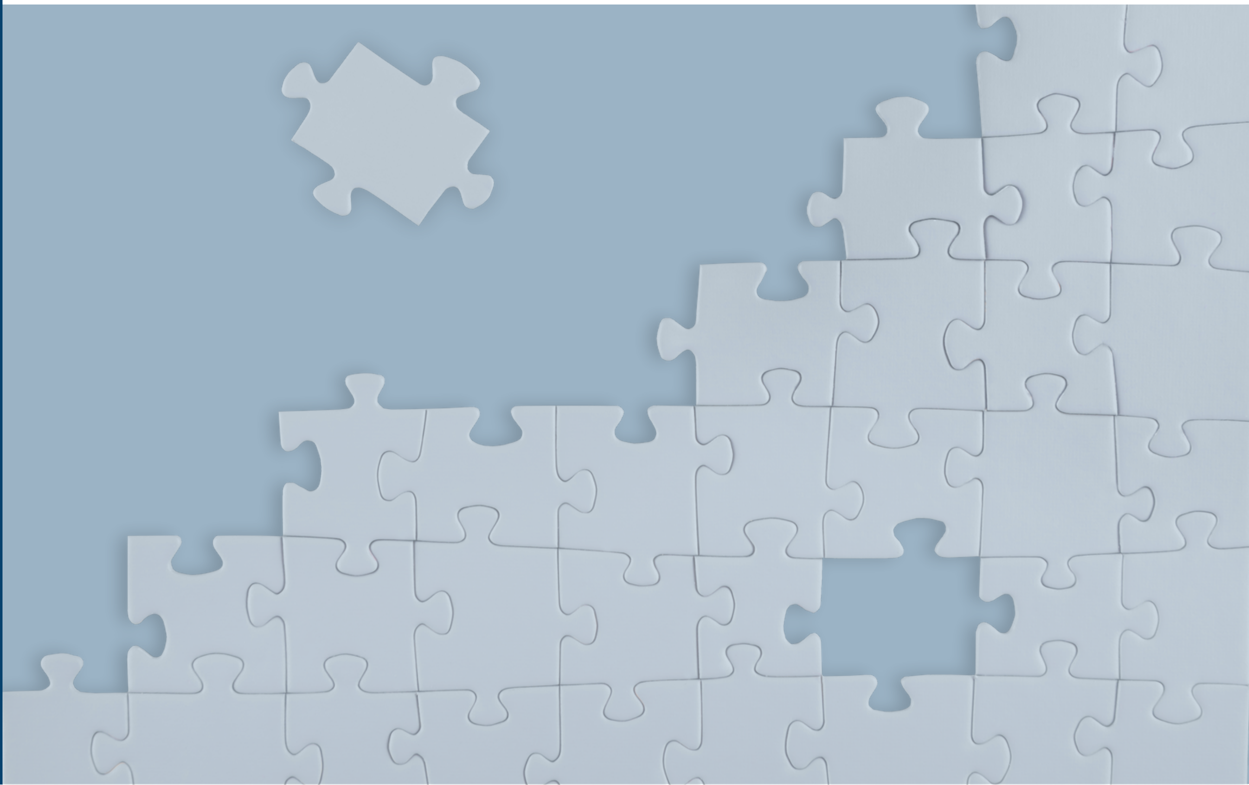
Language as Social and Cultural Data

Enhanced Infrastructural Support for Research Using Language Materials in the Era of AI

CLARIN.SI

Strategy 2024-2030

TABLE OF CONTENTS



January 2024

Background

This document presents the CLARIN.SI Strategy for the period 2024–2030 and will form the basis for the work for this phase. The Strategy closely follows the CLARIN ERIC Strategy 2024–2026 (CE-2023-2333) but is focused on the Slovenian node of CLARIN and on cooperation of CLARIN.SI with and within CLARIN ERIC. The CLARIN.SI Strategy is aligned with Slovenia's Research Infrastructure Roadmap 2030, i.e. it covers a longer time period than the CLARIN Strategy. Most of the goals will likely remain the same in the additional time-frame, although, given the fast pace of development of language data and technologies, certain aspects might have to be modified in the later years covered by this Strategy.

Introduction

CLARIN's Vision: All digital language resources and tools from Slovenia and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities, social sciences and other language-related disciplines

CLARIN's Mission: Create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities, social sciences, and other language-related disciplines.

CLARIN's Strategic Orientation: Language is the reflection of scientific and societal knowledge, as an instrument for human communication and persuasion, as one of the central aspects of the identity of individuals, groups, cultures and nations, and as an instrument for human cognition and creative expression. Moreover, language materials form a considerable part of the historical records which are considered cultural heritage. By giving access to language material, in the first instance to Slovenian, as well as other South Slavic languages, CLARIN.SI facilitates the comparative research perspective for complex phenomena, and enables the development of data-driven analytics and computational modelling of these phenomena. The resources CLARIN.SI makes available also fuel the training and development of large language models, a key technology of Artificial Intelligence (AI), as well as the development of methodological frameworks for the analysis of heterogeneous data.

CLARIN's Key Scholarly Domains: Linguistics, Language Studies, Language Technology, Artificial Intelligence, Literary Studies, History, Ethnic Studies, Journalism and Media Studies, Communication Studies, Ethnography and Anthropology, Migration Studies, Political Studies, Culture Studies, Sociology and Psychology.

This document outlines the CLARIN.SI strategy to fulfil its commitment to enabling high-quality research and innovation, as well as the measures aimed at strengthening its accessibility, sustainability and resilience for the period 2024–2030. The strategic planning of CLARIN.SI is aligned with the strategy of CLARIN ERIC 2024–2026 (CE-2023-2333), to the ERA Policy Agenda 2022–2024 and to the Slovenian National Strategic Plan for the Digital decade, and will contribute to the way science in Slovenia is performed, with an emphasis on collaboration, inclusiveness and open access.

CLARIN.SI was established in 2015 with the vision to provide first-class resources for researchers that critically rely on the availability of language resources, as well as services that give access to relevant digital language data and tools for processing language materials from Slovenia, other South-Slavic speaking countries, and beyond.

CLARIN.SI's mission is rooted in the wide acknowledgement of the role of language as societal and cultural data, and the increased potential for comparative research of cultural and societal phenomena, also across different languages. With its richly faceted nature, range of potential uses and inherent connection to questions of identity and origin, language is an integral part of the humanities and social sciences research, as well as an interesting phenomenon from the perspective of information science, data science, language technology and Artificial Intelligence. This makes CLARIN.SI a crucial pillar for the support of researchers from a very broad spectrum of disciplines, with different needs and skill sets.

In its 8 years of existence, CLARIN.SI has reached a mature operational stage. In the European context, according to the ESFRI Landmark Monitoring Review, the performance of CLARIN is fully satisfactory and meets its objectives, provides valuable services to the academic community on a truly pan-European, multilingual and multidisciplinary level, has contributed significantly to the development of natural language processing tools and techniques that have been widely adopted in the research community, and is located in the right intellectual space to ensure that European scholars can make a positive and important contribution to the development of AI. In Slovenia, the Research Infrastructure Roadmap 2030 (NRRI 2030) published in 2022 includes CLARIN.SI and stresses that it plays an important role in the relevant national infrastructure.

However, further attention is needed to ensure CLARIN.SI's long-term technical, financial, and organisational sustainability, including risk assessment and management. To ensure its continuous, effective role in the landscape, it is essential to have an ambitious plan for change and development that addresses the very dynamic landscape in which CLARIN.SI operates and takes into account other infrastructure initiatives, the GLAM sector (Galleries, Libraries, Archives, and Museums), as well as industry, with its radically increasing need for language data and linguistic benchmarking.

How CLARIN.SI Works

CLARIN.SI has been established as a long term infrastructure project as part of **CLARIN ERIC**. CLARIN.SI is organised as a **consortium**, which has been established as a special form of a society, which is not a legal entity. Membership is open to institutions, corporations, associations and other legal entities from Slovenia which are engaged in the development or use of language resources, tools and technologies, primarily for Slovenian and other South Slavic languages.

The CLARIN.SI consortium has currently **12 member institutions**, which develop or use language resources and technologies in Slovenia:

- **Universities**: University of Ljubljana, University of Maribor, University of Nova Gorica, University of Primorska
- **National research institutes**: ZRC SAZU, Jožef Stefan Institute, Institute of Contemporary History, Science and Research Centre Koper
- **Library**: National and University Library of Slovenia
- **Companies developing natural language processing technologies**: Alpineon, Amebis
- **Society**: Slovenian Language Technologies Society

The consortium's **Management Committee** makes decisions regarding CLARIN.SI and represents its members. Each member of the consortium is on the Management Committee with one vote.

CLARIN.SI is situated at the **Jožef Stefan Institute** (JSI), which is the technical centre of this research infrastructure. It maintains the CLARIN.SI repository of resources and tools and other services, as defined by the Statues of CLARIN ERIC. At JSI, the CLARIN.SI technical centre is being maintained and developed by three units: the **Department of Knowledge Technologies** (JSI E8), the **Artificial Intelligence Laboratory** (JSI E3), and the **Center for Network Infrastructure** (JSI CMI).

CLARIN.SI participates in the work of **CLARIN ERIC** via the membership of the representative of Slovenia's government in the **General Assembly** (GA), as the highest governing body of CLARIN ERIC, and representatives of CLARIN.SI in the **National Coordinators' Forum** (NCF), as the main liaisons between CLARIN ERIC and national nodes, the **Standing Committee for CLARIN Technical Centres** (SCCTC), which coordinates the activities of the technical centres, and Thematic Committees that help to implement the strategy. Currently, CLARIN.SI representatives participate in the **CLARIN Legal and Ethical Issues Committee**, **CLARIN Committee for Standards and Interoperability**, and the **CLARIN User Involvement Committee**.

Strategic Orientation of CLARIN.SI

Supporting Scientific Excellence and Innovation

Language is the carrier of cultural content, and is the lens through which social, cultural and scientific dynamics can be studied, both in the present time and historically. By giving access to language material, CLARIN.SI facilitates the comparative research perspective for complex phenomena, such as human communication and cognition, the identity of individuals, groups, cultures and nations, and the expression of ideas and theories. It also provides digital resources for the development of data-driven analytics and computational modelling of these phenomena. The resources, which CLARIN.SI makes available, fuel the training and development of large language models, a key technology of **Artificial Intelligence**, as well as the development of methodological frameworks for the analysis of heterogeneous data.

Promoting data registries and data management services that comply with the **FAIR principles** (Findable, Accessible, Interoperable, Reusable) underpins all aspects of the CLARIN.SI strategy, and the interoperability paradigm of what is now known as the **Open Science agenda** has been one of CLARIN.SI's distinguishing features from the outset. Their importance will only increase in the coming period as the 2023 Lund Declaration on Maximising the Benefits of Research Data calls for reinforcing, accelerating, and maximising the benefits of FAIR and open research data in Europe, within scientific communities and through research infrastructures (RIs), in order to increase the overall research and innovation performance of the ERA and strengthen the outreach to and impact on industry and society. The declaration further points out that access to reusable high-quality research data is crucial for strengthening and advancing knowledge within and across disciplines, as well as enabling comparative research agendas. What is more, it determines how efficiently new challenges and emerging crises, such as the COVID-19 pandemic and the Russian aggression in Ukraine, can be tackled by the research community, further underlining the importance of Open Science, and its ability to support effective and rapid responses to future crises.

In addition, FAIR and open research data can also bring increased societal benefit during non-crisis times. CLARIN.SI is committed to actively promoting open data, open source code, and open standards in order to support **reproducible and replicable research**. With this, CLARIN.SI contributes to **sustainable**, **cross-disciplinary** and **responsible (re)use of research data**, thereby facilitating new knowledge discovery and maximising the benefit of investment by academia, industry, and society.

In combination with the inherent **multilinguality** of Europe and the growing attention to **language equality** and **digital inclusion**, it is CLARIN.SI's ambition to consolidate its role in supporting the emerging research agendas for the Social Sciences and Humanities (SSH) domain and to contribute to the innovation potential of advanced models for interaction between people, data and technology for data processing. This is facilitated by the strong embedding of the developers of tools and data collections in the local, culturally specific context of Slovenia, and the interoperability paradigm for the model of collaboration between the centres involved. With this, CLARIN.SI is well positioned both in terms of computing resources and availability of qualified experts to meet the current and future needs of the research and innovation community, including Artificial Intelligence, as called for in the 2023 Tenerife Declaration on the Global Dimension and Sustainability of RIs.

Fostering Strategic Partnerships

CLARIN.SI's core community consists of **academic researchers**, **developers** and **lecturers** from a range of disciplines who work with language data and who have made crucial contributions to the construction and operation of the CLARIN.SI infrastructure in the form of resources, technology, and knowledge. Over the next years, the collaboration with academic parties will be reinforced and broadened.

Even though CLARIN.SI's main focus is to enable curiosity-driven fundamental research and scientific excellence, it also has significant innovation potential. It cooperates with a variety of stakeholders from outside of academia, including **industry**, **governmental organisations**, and the **GLAM sector** (Galleries, Libraries, Archives, and Museums) in the role of contributors as well as users of data, tools and know-how. Two commercial companies and a GLAM organisation (the National and University Library) have even formalised their collaboration as members of the CLARIN.SI consortium. The rising importance of language and speech technology in member states' digital transformation offers CLARIN.SI exciting opportunities for cooperation with industry, which will be informed by the newly established CLARIN Working Group tasked to prepare an Industry-Liaison Strategy and Innovation Action Plan for 2023–2024.

In the broader RI ecosystem, CLARIN.SI is, via CLARIN ERIC, well positioned in the **ESFRI** cluster of Social and Cultural Innovation. It closely cooperates with the Slovenian nodes of the **DARIAH** and **CESSDA RIs**, i.e. DARIAH-SI and ADP, with all the institutions that are members of these RIs also being members of the CLARIN.SI consortium, and, furthermore, also participating in joint projects with the two RIs, such as the development of parliamentary corpora or the RDA Node Slovenia project.

Forging Impact Pathways

The primary goal of CLARIN.SI is to contribute to scholarly excellence and scientific progress by offering access to state-of-the-art resources and tools, as well as to best practices, training and expertise. However, CLARIN.SI's position in the very centre of the **data science** and **Artificial Intelligence** communities, whose most influential innovations are entering most academic fields as well as our everyday lives at an unprecedented speed, also helps to position CLARIN.SI as an enabler of research to address societal challenges, for instance by using improved processing of language and speech signals in order to develop assistive technologies for people with a language or speech disorder. But while big data and algorithms based on machine learning are of significant value for research, the economy, and society in general, they also pose a threat for social justice. Overcoming these difficulties requires researchers and industry alike to adopt guidelines for **responsible data science practices** and **ethical AI**. CLARIN.SI aims to take a role in supporting the development of use cases that can help furthering the understanding of the pitfalls of data-driven methods.

The language resources offered by CLARIN.SI's repository and its inclusion into the CLARIN resource discovery service, the Virtual Language Observatory (VLO), forms a critical component for (the realisation of) the European Language Data Space (LDS). This concept is envisaged as a vehicle for language data provision, based on a model that allows monetising the available service offer and addressing the needs of the industry sector. The LDS is part of the European strategy for data, aiming at a single market for data in order to ensure Europe's **global competitiveness** and **data sovereignty**.

Due to CLARIN.SI's open-access policy, the collection of quantitative data to track the socio-economic impact (SEI) of resources and CLARIN.SI at large is far from straightforward. While a quantitative causal chain from (access to) resources to an end result (e.g. a publication or policy) is not obvious, in qualitative terms CLARIN.SI's history already provides evidence of its socio-economic impact. For example, the Institute for Slovenian Language "Fran Ramovš" regularly uses the CLARIN.SI concordancers in the scope of its consulting services and in the process of dictionary compilation, several translation agencies use the CLARIN.SI parallel corpora to help in their translation process, while the industry sector uses the CLASSLA annotation tool-chain developed by CLARIN.SI to annotate Slovenian and other South Slavic texts.

In this strategy period, CLARIN.SI will develop quantitative and qualitative indicators for scientific and socio-economic impact that align with its core value of open access, including the formulation of impact-specific Key Performance Indicators (KPIs)/Key Impact Indicators (KIIs). Following the ESFRI Policy Brief on Impact Assessment of RIs, CLARIN.SI will take a customised, mixed-method approach, including access and bibliometric analysis, case studies, and stakeholder survey.

CLARIN.SI’s Strategic Agenda

The implementation agenda of the CLARIN.SI strategy 2024-2030 has been organised into three pillars that are key to operations of CLARIN.SI:

Pillar 1: Our Users

Existing and potential new users with academic and non-academic interests.

Pillar 2: Our Offer

Technical and knowledge infrastructure with interoperable data centres and collaborative knowledge centres.

Pillar 3: Our Landscape

Programme for organisational development, coordination of action lines and collaboration with stakeholders and embedding in the broader context.

Each pillar is organised into **focus areas** (What is important to us?). Focus areas are broken down into **goals** (What do we want to achieve?), which are then further divided into individual **strategies** (How will we achieve it?). They will form the basis for work plans for each of the coming years in order to ensure that the descriptions of the envisaged actions are sufficiently precise and match the available human and financial resources. All of the strategies presented below share three common objectives: **increasing the potential for uptake**, **increasing the potential for impact** and **increasing the sustainability** of CLARIN.SI’s role and service offer in the RI landscape.

Pillar 1: Our Users

Focus Area 1: Academic User Base

Goal 1.1: Improve outreach to CLARIN.SI's existing academic user base.

- **Strategy 1.1.1:** Continue to identify CLARIN.SI's key user groups and their needs in terms of research data, (large) language models, tools, documentation, expertise and training.
- **Strategy 1.1.2:** Improve the website and other communication channels to better reflect CLARIN.SI's mission, network, offer, as well as impact, and to better cater for the diverse communities served by CLARIN.SI.
- **Strategy 1.1.3:** Improve central and national support services, as well as offer of expertise, via a newly created training, workshops, and online tutorials, with special attention to multidisciplinary research agendas.

Goal 1.2: Broaden CLARIN.SI's academic user base portfolio.

- **Strategy 1.2.1:** Extend and improve the use of web analytics to better understand how users find and interact with CLARIN.SI's offer and develop plans to optimise their experience.
- **Strategy 1.2.2:** Survey the academic landscape and identify priority new and emerging user communities for this strategy period and understand their needs via surveys, focus groups, and interviews.
- **Strategy 1.2.3:** Review and adjust plans to reach out to new users via CLARIN.SI communication instruments.

Focus Area 2: Non-Academic User Base

Goal 2.1: Develop systematic engagement with non-academic communities that have an interest in CLARIN.SI's outputs and can fuel technological or societal innovations.

- **Strategy 2.1.1:** Identify and extend existing relationships and collaborations with non-academic R&D communities from industry, the GLAM sector, the public sector, NGOs, educators, science journalists, policy makers and the wider audience interested in topics such as language diversity and multilinguality, digital revolution and digital language equality, language technology and Artificial Intelligence.

Focus Area 3: Quality and Interoperability of Data, Tools, Models and Metadata

Goal 3.1: Improve the quality of metadata served by CLARIN.SI and thereby increase the quality of discoverability and reusability of resources and tools.

Strategy 3.1.1: Develop and implement policies and workflows for quality assurance and monitoring for aggregated and curated metadata through a combination of automatic procedures (e.g. link checking) and manual curation, and by maximising synergies with CLARIN curation efforts and discovery portals (e.g. VLO, Switchboard, CLARIN Resource Families).

Goal 3.2: Improve the coverage and quality of data, models and tools offered through the CLARIN.SI infrastructure.

- **Strategy 3.2.1:** Perform a coverage and gap analysis as well as a technical check on availability, single-sign-on support etc. on resources and tools that will serve as a basis for defining priorities for future development as well as training, outreach and communication initiatives, especially considering possible new application areas and fields/topics that are strategically important to CLARIN.SI.
- **Strategy 3.2.2:** Develop procedures for collecting, creating and documenting test suites for tools, models and data sets, as well as procedures for their quality assurance and quality control that are fully aligned with the Open Science agenda and FAIR data principles.
- **Strategy 3.2.3:** Develop a more coordinated approach to attract more deposits of language resources, possibly including sensitive data (e.g. from the health domain, oral history). Improve the existing depositing guidelines and actively approach researchers with interesting and relevant datasets that are not yet available in the CLARIN.SI infrastructure.
- **Strategy 3.2.4:** Position CLARIN.SI as an infrastructure focusing on high-quality language data sets. Share definitions of quality (in terms of format and content), best practices and experiences with other CLARIN centres with regards to achieving such elevated quality standards. Publish data quality success stories and practices.
- **Strategy 3.2.5:** Highlight quality and coverage of the NLP tools hosted by CLARIN.SI by joining the planned CLARIN ‘BLARK meets Test Automation’ dashboard that will show the tools that are available in combination with a test run on standard test files.
- **Strategy 3.2.6:** Act to improve existing citation guidelines and practices in Slovenia by approaching authors, journal editors and conference chairs to improve citations of CLARIN.SI datasets and tools in scientific publications.

Goal 3.3: Improve interoperability of metadata, data, models and tools.

- **Strategy 3.3.1:** Cooperate with CLARIN on interoperability of metadata at several levels: between CLARIN.SI and other CLARIN centres (e.g. via core metadata recommendations, facilitating the use of common vocabularies, and a general CLARIN gateway service for FAIR digital objects), and explore interoperability with popular repositories like Zenodo, Open Science Framework (OSF), HuggingFace and Github.
- **Strategy 3.3.2:** Explore possibilities to ensure a more unified access to NLP tools and models provided by CLARIN.SI, e.g. by providing standard programming interfaces. Highlight the added value in comparison with existing, widely used NLP frameworks such as SpaCy.
- **Strategy 3.3.3:** Cooperate with CLARIN to explore possibilities to provide a more unified access to the key corpora provided by CLARIN.SI, e.g. by offering them through a common concordancer that would importantly flatten the learning curve for users and facilitate comparable analyses.

Focus Area 4: Accessibility and Usability of Data, Models, Tools and Services

Goal 4.1: Foster a high level of accessibility, usability and interoperability between technical services offered by CLARIN.SI vis a vis other CLARIN national nodes and centres.

- **Strategy 4.1.1:** Cooperate with CLARIN to adopt best practices on CLARIN-wide accessibility, usability and interoperability of data, models, tools and services, including facilitating the use of processing services for sensitive data.

Focus Area 5: Collaboration and Knowledge Exchange

Goal 5.1: Consolidate the CLARIN.SI community for sharing knowledge and experience.

- **Strategy 5.1.1:** Identify systematic opportunities and develop practices to improve interaction between CLARIN.SI and other CLARIN centres.
- **Strategy 5.1.2:** Help review and adjust the Language Technologies and Digital Humanities Conference of which CLARIN.SI is a co-organiser, so that it remains a rich and inspiring source of knowledge exchange for the local and international community.
- **Strategy 5.1.3:** Improve, translate to English and disseminate the collection of CLARIN.SI documents, such as overview and best practice papers to better reflect the knowledge and experience of the CLARIN.SI network, also outside Slovenia.
- **Strategy 5.1.4:** Develop best practices and templates for Data Management Plans (DMP) for researchers collecting language data as a new means of knowledge sharing.

Goal 5.2: Integrate CLARIN.SI K-centres more into the ‘bigger picture’.

- **Strategy 5.2.1:** Cooperate with CLARIN to promote increased use of CLARIN.SI K-centres’ expertise by researchers via a variety of means, such as Tour de CLARIN and Impact Stories.

Pillar 3: Our Landscape

Focus Area 6: Sustainability and Impact

Goal 6.1: Extend membership in the CLARIN.SI consortium

- **Strategy 6.1.1:** Develop a growth strategy to attract new members to the CLARIN.SI consortium, with the ultimate goal of including all relevant academic institutions from Slovenia, in particular research institutes, as well as more industry partners.

Goal 6.2: Monitor sustainability and impact.

- **Strategy 6.2.1:** In cooperation with CLARIN, develop a methodology for socio-economic impact assessment.
- **Strategy 6.2.2:** Identify and address current and foreseen scalability issues and risks as well as a continuous evaluation loop of infrastructure components of the technical infrastructure, such as AAI (e.g. growing number of SPs which need to be supported, impact on architecture) and metadata curation (e.g. support, issue handling, link checking capacity).

Goal 6.3: Strengthen personnel and governance.

- **Strategy 6.3.1:** Pro-actively recruit and employ support staff to better enable continuous operation and upgrades to the CLARIN.SI technical and knowledge infrastructure.
- **Strategy 6.3.2:** Better involve members of the CLARIN.SI consortium to actively participate in the running and decision-making process of the infrastructure.

Goal 6.4: Reinforce CLARIN.SI as a key enabler of language-based state-of-the-art research and innovation.

- **Strategy 6.4.1:** Identify opportunities and improve CLARIN.SI offer with regard to emerging technology, such as speech technologies and resources, Large Language Models (LLMs), and AI, all focused on the Slovenian and other South Slavic languages.
- **Strategy 6.4.2:** Continue to offer the annual CLARIN.SI mini-grants and to improve the annual project call to respond to changing circumstances.

Focus Area 7: Collaboration beyond CLARIN.SI

Goal 7.1: Reinforce alignment and collaboration with other Slovenian infrastructures.

- **Strategy 7.1.1:** Continue close collaboration, including at the operational level, with sister RIs in Slovenia, in particular DARIAH-SI and ADP, as well as initiate collaboration with new relevant RIs, to ensure optimal visibility, representation and participation with reference to the EC, EOSC and ESFRI, as well as in relevant projects and working bodies.
- **Strategy 7.1.2:** In cooperation with CLARIN, develop CLARIN.SI's position as an intermediary node in the EOSC ecosystem. Among other benefits, this could importantly expand CLARIN.SI's user base and widen its disciplinary engagement.
- **Strategy 7.1.3:** Monitor the development of the Alliance for Language Technologies (ALT-EDIC) and prepare relevant action plans.

Goal 7.2: Strengthen collaboration across borders.

- **Strategy 7.2.1:** In the scope of the CLARIN.SI CLASSLA K-centre identify systematic opportunities and develop practices to better support collaboration with partners or networks in other South-Slavic speaking countries, i.e. Croatia, Serbia, Bosnia and Herzegovina, Montenegro, North Macedonia and Bulgaria.
- **Strategy 7.2.2:** In the scope of the CLARIN.SI ELEXIS K-centre identify opportunities to support collaboration at the European level with centres working on digital lexicography.

CLARIN.SI
Jožef Stefan Institute
Jamova cesta 39
SI – 1000 Ljubljana
Slovenia

January 2024