Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	CLARIN.SI
Website:	https://www.clarin.si/repository/xmlui/
Certification Date:	30 September 2020

This repository is owned by: **Jožef Stefan Institute**

CLARIN.SI

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, National repository system; including governmental

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

The CLARIN.SI research infrastructure (www.clarin.si) is the Slovenian national node of the European Research Infrastructure for Language Resources and Technology CLARIN ERIC and provides, as one of its services, the CLARIN.SI repository of language resources and tools [1].
The CLARIN.SI digital repository platform is hosted at the Jožef Stefan Institute (JSI [2]), the largest research institute in Slovenia. JSI cooperates in the development of the platform and that of connected services with other institutions that are members of CLARIN ERIC and of the CLARIN.SI consortium.

In the repository we follow the standard principles of a high quality digital repository, such as the usage of persistent identifiers, authorisation and authentication, and sharing of metadata and data.

[1] https://www.clarin.si/repository/xmlui/
[2] https://www.ijs.si

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

The designated community of the CLARIN.SI repository is the national and international research community, in particular researchers involved in computational linguistics, corpus linguistics, digital humanities and other fields that produce language data or utilise such data or natural language processing tools. For resources with appropriate licences (such as CC-BY) the repository is also valuable for companies developing applications in the area of language technologies.

While the data for many more languages is available in the repository, the principal languages covered are Slovenian, Croatian and Serbian, so researchers and companies dealing with these languages are our primary consumers.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

B. Basic curation – e.g. brief checking; addition of basic metadata or documentation

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

In the first step, users deposit resources into the repository by themselves using a web-based submission workflow (i.e., a form with several stages for providing metadata about the submission). When applicable, answers are validated against vocabularies or pre-defined rules after each stage [1].

In the next step, editors review and curate the submission. The editors inspect the metadata and data and either accept the submission or return the submission to the depositor requesting changes or more details. Pre-programmed tasks (e.g., URL checks, metadata completeness) help editors decide if the submission meets the technical requirements.

Editors do not execute file format conversion or enhancement of documentation but return the submission to the depositors with detailed instructions on how to update the submission, if any of these parts are insufficient [2,3].

CLARIN.SI performs regular checks on the metadata and data (e.g., completeness, checksums) and may request additional information from the depositors. Occasionally, minor metadata modifications (e.g., correcting grammar mistakes, unifying keyword lists) can be done also after the item has been published. All the changes, including the ones done by editors, are recorded in the provenance metadata [4].

[1] https://github.com/ufal/clarin-dspace/blob/clarin/dspace/config/input-forms.xml
[2] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[3] https://github.com/ufal/clarin-dspace/wiki/Metadata-info

[4] https://www.clarin.si/repository/xmlui/page/item-lifecycle

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

N/A

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

An overview of the repository can be seen at re3data [1]. We are using DSpace [2] as the basis of our repository system, although in a modified version called CLARIN DSpace, developed by LINDAT/CLARIN, the Czech national node of CLARIN ERIC hosted at the Institute of Formal and Applied Linguistics, Charles University, Prague. CLARIN DSpace is better suited for the needs of storing linguistic data and software and integrating it with the CLARIN network [3]. CLARIN DSpace is also used by a number of other CLARIN centres, with joint development of the common codebase [4].

The repository currently hosts over 130 digital resources, which amounts to around 200 GB of data; the majority of these are language corpora. The datasets cover over 100 languages, with the top 4 being Slovenian, English, Croatian and Serbian.

[1] http://service.re3data.org/repository/r3d100011922
[2] https://duraspace.org/dspace/
[3] http://hdl.handle.net/11372/DOC-78
[4] https://github.com/ufal/clarin-dspace

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The ultimate objective of CLARIN ERIC (which CLARIN.SI is part of) is to advance research in the humanities and social sciences by giving researchers unified single sign-on access to a platform which integrates language-based resources and advanced tools at a European level. This shall be implemented by the construction and operation of a shared distributed infrastructure that aims at making language resources, technology and expertise available to the humanities and social sciences research communities at large.

CLARIN.SI is thus committed to the long-term care and preservation of items deposited in its repository and strives to adopt the current best practices in digital preservation [1], and in particular, to continue as a certified [2] CLARIN B Centre [3].

The national research infrastructure CLARIN.SI focuses on Slovene language resources, but not excluding other languages, in particular other South-Slavic languages, with this mission supported by its CLASSLA Knowledge Centre [4].

The resources can be deposited by associated researchers as well as researchers who are not affiliated with us.

CLARIN.SI is organised as a consortium, which comprises all the main institutions that are involved in the production and use of language resources in Slovenia [5], with all the partners having signed a consortium agreement [6] which formalises the mission of CLARIN.SI.

The CLARIN.SI mission is supported by integration of the repository into the EU CLARIN infrastructure [7]. As part of the CLARIN infrastructure, the repository is included in promotional activities carried out at the national level (CLARIN.SI) as well as at the European level (CLARIN ERIC). We also support and promote our repository at suitable events in Slovenia [8].

The repository implements standard protocols for sharing metadata and data. Public submissions can be easily mirrored. Protected submissions can be mirrored after legal requirements are met.

[1] https://www.clarin.si/repository/xmlui/page/about
[2] http://hdl.handle.net/11372/DOC-115
[3] http://hdl.handle.net/11372/DOC-78
[4] http://www.clarin.si/info/k-centre/
[5] http://www.clarin.si/info/partners/
[6] https://www.clarin.si/info/wp-content/uploads/2014/10/Clarin_konzorcij-podpisano.pdf
[7] http://www.clarin.eu/
[8] http://www.clarin.si/info/events/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All visitors to the repository agree to the repository's Terms of Service [1], which binds them to comply with the licenses attached to repository items.

The license attached to a repository item is displayed prominently on the item page (see [2] for an example) together with colour-coded “openness” of the license (“public”, “academic”, “restricted”).

The license attached to a repository item is chosen during submission by the person submitting it. The licences cover all the CC licences (version 4), the standard open source software licenses and several more licences for submissions of a more restricted nature. Given the large number of licenses, we also provide guidance to select the appropriate license using a graphical license selector tool. While open/public licenses are strongly preferred when possible, we also offer options to put more requirements on the consumer. For instance, we require that consumers have an academic account (which in our setup means they are real people and can be identified with the help of their institute). For restricted submissions, consumers have to authenticate and electronically sign the license before downloading the data. We store the information about licenses signed by each consumer. In case a suitable license cannot be found among the pre-existing licenses [3], submitters can contact the repository staff with a request to create a custom license.

During the submission, the submitter enters a standard contract with the repository (more precisely with the CLARIN.SI consortium), the so-called “Deposition License Agreement” [4], where we describe our rights and obligations, and the submitter(s) acknowledge that they have the right to submit the data and give us the right to distribute the data on their behalf. The repository also offers the option of putting an embargo on submissions, meaning that the submissions will be archived immediately after completion of the curation workflow, but the data will become publicly available only after a specific date.

In case we identify non-compliance with license conditions or terms of use by a registered user, we can identify the real person with the help of his/her Identity provider. We deny such users further access to the repository. We make the research community, at least the part connected to our channels - mailing lists, social media feeds and various other bodies - aware of the misuse. As a last resort, we would take legal action.

[1] https://www.clarin.si/repository/xmlui/page/about#terms-of-service
[2] http://hdl.handle.net/11356/1136
[3] https://www.clarin.si/repository/xmlui/page/licenses

[4] https://www.clarin.si/repository/xmlui/page/contract?locale-attribute=en

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

CLARIN.SI is financed by the Slovenian Research Agency under its ESFRI Infrastructures Programme. This programme has no end-date, but its results and plans are annually evaluated, although, to date, no Slovenian ESFRI infrastructure has had its funding cancelled or interrupted. The current level of funding is sufficient to maintain the repository system and keep up its development and improvements, as well as keeping data security at least at the current level.

CLARIN.SI has measures in place to preserve data access in case of unexpected emergency budget cuts. The CLARIN DSpace repository platform is a very low maintenance system, easy to keep running, while the group members that administer are repository is employed on regular contracts at the hosting institute or at one of the CLARIN.SI consortium partners. Thus, if CLARIN.SI funding would be interrupted, the hosting institute would be able to keep the repository running without dedicated funding for a substantial time, certainly for at least five years, while most likely continuing to

accept new submissions as well.

The CLARIN DSpace repository is open source software and already eight CLARIN centres are deploying it [1], in part also so as to ensure the sustainability of access, as it allows for simple migration of all the data from one CLARIN DSpace repository to another while keeping the records accessible under the same PIDs and with the exact same feature set. Thus if, as the worst case scenario, the funding for the CLARIN.SI infrastructure would be terminated completely, one of the other CLARIN centers would be able to host our data and to reconfigure its permanent identifiers for the CLARIN.SI collection. We have a signed agreement with the Czech LINDAT-CLARIAH-CZ centre [2] for such a migration.

The continuity plan for the CLARIN.SI repository is published at [3].

[1] https://github.com/ufal/clarin-dspace#clarin-dspace-deployments
[2] http://www.clarin.si/info/wp-content/uploads/2020/04/CLARIN-agreement-CZ-SI_signed.pdf
[3] https://www.clarin.si/repository/xmlui/page/about#preservation-policy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The submitters acknowledge during submission that they have the right to distribute the data and that they also have the right to grant the repository permission to distribute the data on their behalf [1]. Acknowledging that the submitter has the right to distribute the data in the first place includes resolving possible legal issues, because if these were not resolved the submitter would not have the right to distribute the data at all.

The submissions are reviewed by the repository staff (editors). For language data, in particular language corpora, several legal issues must be considered [2]. If the editors are in doubt about the compliance of the dataset with applicable laws or regulations, they request more information from the submitter or refuse to publish the submission. If there are special conditions, they can be addressed in a distribution license tailored specifically for the particular item. There is also the possibility, in cooperation with the submitters, to use various tools to modify the data so that they do not infringe laws, such as releasing only samples of integral texts or shuffling their tsentences to protect against copyright infringement, or anonymising texts with the use of named entity recognition tools to protect against personal data disclosure or the right to be forgotten (especially relevant for corpora containing older newspaper texts).

To date the repository has no submissions containing confidential data or data with disclosure risk which cannot be anonymised, and we do not expect this to change in the future. Most of our data is Open Access or distributed under similar public licenses, in particular variants of the Creative Commons licences. Substantially less data is available under custom licenses, which are however still public or relatively permissive (e.g., academic restriction and no redistribution). Given that the mission of the repository is to make data widely available, we do not accept items that would contain confidential data or data with disclosure risks.

It should be noted that the CLARIN Legal and Ethical Issues Committee [3] (which CLARIN.SI is a member of) organises training sessions in the legal and ethical management and distribution of text data.

[1] https://www.clarin.si/repository/xmlui/page/contract
[2] https://www.clarin.si/repository/xmlui/page/about#about-ipr
[3] https://www.clarin.eu/governance/legal-issues-committee

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CLARIN.SI repository is hosted at the Jožef Stefan Institute (JSI) in Ljubljana. JSI was founded in 1949 and is the leading Slovenian scientific research institute, with a staff of over 900. The running of the repository and other services of CLARIN.SI is a joint undertaking of three organisational units of the institute:

- The Department of Knowledge Technologies, which performs research in advanced information technologies. The department is involved in many national and EU projects, some of which it also leads. Established areas of research include intelligent data analysis, text and web mining, language technologies and computational linguistics, decision support and knowledge management. In the area of language technologies, the department is one of the leading (and oldest) Slovenian centres for the development of language resources and annotation tools, especially in the domain of standardisation of resource encoding and linguistic formalisms and in open accessibility of resources. The department was also among the first to develop and promote the field of digital humanities in Slovenia.

- The Artificial Intelligence Laboratory, which is concerned with research and development in information technologies with an emphasis on artificial intelligence. The main research areas are data analysis with an emphasis on text, web and cross-modal data; scalable real-time data analysis; visualization of complex data; semantic technologies; and language technologies. The laboratory has been involved in many EU projects in the area of text analytics and processing, where their task is mainly in providing technologies for knowledge extraction from text and for machine translation.

- The Networking Infrastructure Centre, which manages the networking and hardware infrastructure at JSI, including large high-performance computing clusters used especially by the physics departments of the JSI. It is active in the areas of trust and authentication, maintaining the JSI IdP service and being actively involved in EduGain and other EU efforts in these areas.

While JSI is the hosting institution, CLARIN.SI is a national CLARIN centre and consortium with its own management structure [1]. Prof. Tomaž Erjavec (working at the Dept. of Knowledge Technologies) is the national coordinator of CLARIN.SI, while the Governing Board consists of one representative from each of the 11 consortium member institutions [2]. These include all the main institutions that deal with language data in Slovenia, including the four Slovenian universities and the Institute for Slovenian Language at the Scientific Research Centre of the Slovenian Academy of Sciences and Arts.

Since the establishment of the Slovenian CLARIN national centre CLARIN.SI (2013), it has been financed by the Slovenian Research Agency under its ESFRI Infrastructures Programme. This programme has no end-date, but its results and plans are annually evaluated, although, to date, no Slovenian ESFRI infrastructure has had its funding cancelled or interrupted. The current level of funding is sufficient to maintain the repository system along with its development and improvements, as well as data security at least at the current level.

As explained in R3, even in the case of a discontinuity of funding, the hosting departments of JSI can cover the interim period from their own resources, especially as the staff working with the repository is employed by the JSI or by one of the CLARIN.SI consortium partners. The staff is qualified to manage the repository in all its aspects, from data and metadata curation to the technical maintenance of the software and hardware. The repository is run by the core technical group, which is mostly employed at the JSI. Currently this involves: one person (PhD in linguistics) who works at the Networking Infrastructure Centre and takes care of AAI related issues, connectivity, domain name maintenance, and consulting on new hardware; one person (PhD in computer science) who works as a senior system engineer at the Department of Knowledge Technologies and is in charge of the hardware, including backups, system engineering tasks (such as configuration of the machines and their smooth operation), and maintenance of the off-the-shelf services (such as GitLab) that CLARIN.SI provides; one person (PhD in psycholinguistics) who is employed by the University of Ljubljana at the Centre of Language Resources and Technologies and is in charge of the repository maintenance and upgrades as well as the on-line corpus analysis CLARIN.SI services; and one person (PhD in linguistics) who takes care of the on-line CLARIN.SI annotation tools. They report to and are coordinated by the national coordinator of CLARIN.SI and are supported by the Network Infrastructure Centre IT group. There are currently two persons working as editors for the repository; one is the national coordinator of CLARIN.SI, who currently oversees the majority of submissions, and who makes the final decisions on data acceptance policy, while the other (PhD in linguistics) works at the Artificial Intelligence Laboratory and oversees the submissions where the coordinator is a co-author and in cases when the coordinator is not available. In 2020 it is planned to employ a full time person to take care of editorial work. CLARIN.SI is a member of SLING, the national supercomputing network consortium, enjoying its resources, technical expertise and support, and is included in national repository and open science data priorities.

The CLARIN.SI budget is sufficient to enable the regular attendance of the CLARIN.SI staff in F2F meetings of the CLARIN ERIC, such as annual meetings of committees and task forces. Most staff members also work on other international projects in H2020 and COST, where important expertise is shared too.

For the training of new staff during natural staff exchange, documentation of the key infrastructure and processes is available and significant experience has been already reached in this area.

[1] http://www.clarin.si/info/governance/
[2] http://www.clarin.si/info/partners/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

CLARIN.SI is a member of the EU CLARIN infrastructure and, as such, is in regular touch with its experts. CLARIN.SI is a member of the CLARIN ERIC Standing Committee of Technical Centres, Legal and Ethical Issues Committee, Standards Committee, and User Involvement Committee [1]. In regular tele-conferences and F2F meetings in the scope of these committees, latest developments and subsequent necessary modifications to the workings of the CLARIN repositories are also discussed.

The CLARIN.SI repository team is in regular touch, via email, a dedicated Slack channel and GitHub issues with LINDAT/CLARIN, the developers of the DSpace CLARIN platform. CLARIN.SI not only reports on problems and suggests improvements to the platform, but also contributes to its development, as well as to the development of standards used for encoding language resources.

CLARIN.SI is a member of the “RDA in Slovenia” [2], a project started in 2019 and led by the University of Ljubljana, where open access to research data in Slovenia is a focus of the project.

The technical team and editorial team involved in the CLARIN.SI repository, in the scope of their project involvements, also regularly attend conferences and workshops with a focus on language resources, such as the Language Resources and Evaluation Conference.

The metadata in our repository is regularly harvested by several harvesters including the CLARIN ERIC VLO and OLAC. These perform additional curation tasks with the results regularly inspected by LINDAT/CLARIN. In CLARIN, the progress on these efforts is regularly reported as part of CLARIN’s Metadata Curation Taskforce.

As mentioned, CLARIN.SI is organised as a broad consortium, which contains institutions which are also the main contributors and users of the repository in Slovenia, and can thus advise JSI how to tailor the repository to their needs. CLARIN.SI has an email address where users can ask questions, report problems or suggest improvements [3].

[1] https://www.clarin.eu/content/governance
[2] https://www.rd-alliance.org/groups/rda-slovenia
[3] info@clarin.si

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The general overview is described in [1].

Integrity: To verify that a digital object has not been altered or corrupted we periodically (on a weekly basis) verify the md5 checksums of the objects. The md5 checksum is computed as soon as the user uploads a file, so they can confirm it was not corrupted during transport. Also, the editors check the files before approving an item for publication. The item submission is a (web) form-based process. The item will not pass through submission unless all the metadata fields marked as required are filled in with appropriate values. The editor has tools available that help to further validate the metadata, such as fetching of URLs in the metadata, or showing the level of support (supported / known / unknown) for the submitted file formats. Some of these editors' tools are part of the weekly checks (e.g., whether all required metadata are present, URLs are working). The results of weekly checks are automatically sent to the repository staff.

We do not support changing the data. A change or a new version of a dataset must be created as a new repository item [2]. We do this for the sake of reproducibility (of results using the dataset) and to have a clear meaning of what a PID (persistent identifier) refers to. The new and the old version(s) have a relation added to their metadata, which is visually represented on the web page (e.g., http://hdl.handle.net/11356/1210). Changes to the metadata do happen occasionally (mostly typo fixes), and they are recorded in the provenance metadata.

Authenticity: Only registered users can deposit items to our repository and the registration can be performed only when users have an academic account at one of the member institutions of our identity federation. Thus the academic institutions are responsible for verifying the user identity; see R8 for more details. Provenance information is kept for each repository item from the moment the item is created. Once the item is approved, only the administrators are able to change its (meta)data. The data producers can refer to the Deposited Item Lifecycle [1] mentioned to get acquainted with the details or ask our helpdesk directly.

[1] https://www.clarin.si/repository/xmlui/page/item-lifecycle?locale-attribute=en
[2] https://github.com/ufal/clarin-dspace/wiki/New-Version-Guide

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

CLARIN.SI accepts any language data, including but not limited to language corpora, text and speech collections, machine readable dictionaries, computational lexicons and language processing tool models, as well as language processing software. It provides public guidelines for data submission that include preferred formats and metadata preparation and instructions for preparing and submitting data for publication [1,2,3,4].

The submission interface is separated into several steps. These steps can be slightly different for different types of submissions. Each step has a set of mandatory fields including value checks (e.g., for valid email). Submitters are not allowed to move to the next step unless all required fields are filled in correctly. After submission, the item is reviewed by an editor who checks the quality of the metadata. A thorough check of the data quality is not performed since it is beyond our mission and scope, but if editors understand the data (the NLP field has large variability of specialised data formats), they also check the data. As an example, if the dataset is a morphologically annotated corpus, the editors do not (cannot) check each and every morphological category of each and every word in the corpus. What they do check is if there are morphological annotations in the data submitted. As stated in the Distribution License Agreement, submitters are

responsible for the quality of their data. In case the submission does not comply with our expectations the submission is returned via the editorial workflow for further improvements and re-submission.

The repository relies on the group of emerging metadata standards around CMDI (ISO 24622-1:2015, ISO 24622-2:2019); in particular, the submission interface is based on one particular CMDI profile [1]. This ensures that the metadata required to interpret and use the data are provided and are sufficient for long-term preservation.

The repository recommends using standard data formats during submission. Especially for language resources, depositors are referred to the list of relevant standards [2] during the upload step. However, as stated above, natural language processing is an active research area with many data formats in constant development and CLARIN.SI cannot dictate to the researchers what formats they can or need to use. Thus the policy of the repository is to encourage users to use formats recommended by CLARIN [2], but to accept all data formats, whenever the researchers convincingly explain why they are needed. If the format is unknown or not in the list of the recommended standard formats [3], it must be well documented and the documentation must be either part of the submission or the metadata must contain a link to it. In the case of XML files, which are the most common file type in the repository, the minimal requirement is that files must validate according to an attached XML schema, which must contain explanations of the meanings of the defined elements and attributes. The validity of the submitted data sets is checked both manually and automatically (if the format is supported by our automated checks).

[1]: http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1349361150622/xsd
[2]: https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[3]: https://www.clarin.si/repository/xmlui/page/about
[4]: https://www.clarin.si/repository/xmlui/page/metadata

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

CLARIN.SI is hosted on dedicated infrastructure at Jožef Stefan Institute providing highly available storage, backup and disaster recovery for archival data and software. Backups are regularly performed both of virtual system images and of exported data (files and text-format database exports) on-site and replicated at a secondary location. The virtual machine image backup system is implemented using a client-server solution backing up complete VM images or image snapshots and keeping multiple versions. The file and export backups are performed with a completely different implementation, backing up database export files or filesystem snapshot-derived files, where possible, to ensure atomicity and integrity of files and using different servers and backup locations to exclude the possibility of a single point of failure in the back-up and restoration procedures. Our technical storage procedures are included in our internal sysadmin wiki documentation.

With the use of the Clarin-DSpace version of DSpace [1] as developed mostly by LINDAT/CLARIN, we rely on a stable upstream repository system and a well-maintained fork, both of which meet the requirements of OAIS. Repository systems ensure that in the ingestion process, the Submission Information Packages (SIPs) are received for curating and are assigned to a task pool where editors can process them. The standard way is that the ingestion process is done through our web-based interface, which hides the implementation details [2] but an offline entry point for large file uploads and direct access to files is also available.

For the second step, the archival storage, the submission is curated by an editor. The repository provides a web interface for metadata maintenance (addition, deletion, modification) and validation of submitted files, now represented as bitstreams. In this way the editor provides the guarantee of consistency and quality for each submission. Only upon the approval of the editor do the Archival Information Packages (AIPs) become available in the repository.

We are open to all submissions which meet our standards, but we do require that all data producers be authenticated, which usually means they have an academic background or verified local accounts. A contract is digitally signed during the ingestion process to ensure the availability of submission and the rights attributed are communicated and agreed upon. We use a robust administration interface to provide specific detailed reports on the contents of our repository.

All backups follow standardised backup recommendations, including hashes/checksums for ensuring file integrity and automatic monitoring tools to ensure functionality on various levels. The infrastructure and backups are further described in the Technical Infrastructure and Security sections.

A secondary (testing/beta) installation is maintained to avoid service interruptions in case of upgrades or configuration modification. This installation is independent, fully functional and could be used as a hot spare in the case of unintended behaviour of the main infrastructure.

[1] https://github.com/ufal/clarin-dspace

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

CLARIN.SI has the right to copy, transform, store and provide access to the data [1]. The preservation function encompasses: taking delivery of the dataset ingested, storing it, and ensuring it is archived, accessible and usable to the researcher community, as is the mission of CLARIN [2].

DSpace, and thus CLARIN-DSpace repository software, provides two levels of digital preservation. The first approach is

"bit preservation" which ensures the integrity of both data and metadata over time regardless of possible changes in the physical storage media; the second one is "functional preservation": even if the file may change over time it remains usable in the future by evolving its original digital format and media. Format migration is a straightforward strategy for functional preservation.

The preservation strategy is implemented in all the functional concepts of the Open Archival Information System (OAIS) reference model for digital preservation environments. During the ingest phase, data depositors are presented with a user interface divided into logical blocks. The blocks also include: data upload where data depositors are urged to use formats and standards mentioned in [3]; information about the legal issues including signing the distribution agreement [1]; assisted selection of an appropriate licensing model.

All the information is verified by editors during the review step including file format selection. Refer to [4] for more information. The archival storage phase is referenced in R2 and R7. Data management related to preservation is described in R7 and R12. The general policy of the repository is to disable deleting of dataset metadata [5], which is crucial for long term preservation. In the administration phase, in addition to the common administration tasks (see also R9), we have automated reports that help us identify possible issues with long term preservation. This includes extensive automated weekly reports for the whole repository that are checked by the repository staff. The access phase is described in more detail in R2, R4 and R8. A very important policy for our repository is that the metadata of a resource is always public. In order to follow the best practices in Preservation Planning, the repository staff regularly visits relevant conferences (see R6 for more details).

Language data is complex, as it can be in various modalities, and heavily annotated with complex structures, such as an entry in a comprehensive dictionary, or a syntactically and semantically annotated corpus.

The repository encourages the usage of specific file formats as recommended by CLARIN [3]. The guiding principles for format selection are: open standards are preferred over proprietary standards, formats should be well-documented, verifiable and proven, text-based formats are preferred over binary formats, and in the case of digitization of analogue signal lossless or no compression is recommended.

The number of accepted file formats is small and well documented to make future conversions to other formats more feasible. For textual resources (that account for a significant number of the repository items), text is always encoded in Unicode, and XML encoded whenever possible. For structurally simpler data, lists and TSV/CSV tables are also accepted. XML data must always have a documented XML schema included. Standard schemas are preferred, in particular parametrisations of the Text Encoding Initiative Guidelines (TEI ODDs) that validate against the CLARIN.SI TEI schema [6].

In cases where proprietary or custom formats need to be used, we require detailed and exhaustive documentation, in order to make the implementation of future data converters possible.

While the formats that the repository accepts are all well established, the preferred file formats could still change over

time, and, based on user requests, the repository will make every effort to migrate to other formats, while keeping originals intact for reproducibility purposes (i.e., migrated item will be a new repository record linked to the old one). The conversion can be merely technical (in particular, for audio or video files), for which open source converters will be used, or structural, which needs to be solved on a case by case basis. For XML conversion, which is the most common case, XSLT scripts are used. The more generally applicable ones are maintained on the CLARIN.SI GitHub site [7].

All metadata and data have a persistent identifier (PID) and metadata can be converted to self-explanatory and human readable XML files. The metadata and preservation policies are outlined on our site [8].

[1] https://www.clarin.si/repository/xmlui/page/contract
[2] https://www.clarin.eu/content/clarin-in-a-nutshell
[3] https://www.clarin.eu/content/standards-and-formats
[4] https://www.clarin.si/repository/xmlui/page/deposit
[5] https://www.clarin.si/repository/xmlui/page/item-lifecycle
[6] https://github.com/clarinsi/TEI-schema
[7] https://github.com/clarinsi/
[8] https://www.clarin.si/repository/xmlui/page/about

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:

4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CLARIN DSpace platform used by CLARIN.SI has carefully crafted the submission process in such a way that enough information about the resource is gathered but that it does not overload the submitter with forms.

During the submission, hints, examples and suggestions are provided to get the highest quality metadata. We provide a page [1] summarising the information (metadata) we gather about resources and the metadata formats we can disseminate (i.e., the specific CMDI profile and OLAC Dublin Core).

Sufficient completeness and quality of metadata is assured by requiring certain fields in the submission process (without them being filled in, the submission cannot be completed), by filling in certain fields automatically (e.g., the PID and date of entry into the repository), by automated curation and by the final approval by editors. If the editors are not satisfied with the metadata, they have the option to correct them and ask the submitter for approval, or to return the submission to the submitter requiring them to elaborate some of the fields.

Each submission is given a PID and we strongly encourage people to use it for citation of the resource in publications [2].

As we are harvested by other organisations (CLARIN VLO, OLAC harvester), we incorporate their feedback on potential metadata issues. Occasionally we also get feedback from the end users regarding the metadata via the feedback email [3].

Each entry has the option of including URLs of publications that reference the data, and the description of the data can also include references to publications referencing it.

The data itself is checked, as much as possible, for formatting requirements. For all submissions with data in XML (i.e. the majority of submissions), it is required that an XML schema and documentation is also included with the submission and the editors check that the data validates according to the supplied schema.

[1] https://www.clarin.si/repository/xmlui/page/metadata?locale-attribute=en
[2] https://www.clarin.si/repository/xmlui/page/cite?locale-attribute=en
[3] repo-help@clarin.si

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

After submitting the data, a curation platform, offered by and integrated into the CLARIN-DSpace software, is employed to ensure the quality and consistency of the submission with the possibility to return the data to the submitter for changes. These include automated and manual checking.

After final approval from the editor, the submission becomes visible and retrievable via the repository web interface, as well as interfaces more suitable for machines (OAI-PMH, REST API). Information on the submission and curation workflows can be found here: [1,2].

The complete workflow consists of:

1. Create metadata and upload data. Metadata is filled out for each resource by the submitter in several steps. These steps can be slightly different for different types of submissions. Each step has a set of mandatory fields including value checks (e.g., for valid email). Submitters are not allowed to move to the next step unless all required fields are filled in correctly.

2. Assign persistent identifiers. Persistent identifiers (PIDs) provide a unique identification of the research data and metadata in a location-independent manner. This means that even data migration or metadata will continue to use the

same identifier.

3. Specify licenses. Submitter chooses the appropriate license for the data. The web interface provides guidance to select the appropriate license using a graphical license selector tool.

4. Review data/metadata. In this process step, editors assess the metadata in accordance with the guidelines set by best practices criteria.

5. Publish submission. Through the repository web application, the metadata are publicly accessible and the data are accessible based on the specified license and/or specific conditions described in R4 (this means access to some items might be restricted). After this step, the data are backed up together with the other published submissions. The metadata/data is also immediately available in the other interfaces, namely OAI-PMH and REST API.

Usually, the user interacts with the repository via the web UI which allows them to view/search the metadata and download the bitstreams.

The OAI-PMH is used to disseminate metadata about records; however, some of the metadata formats (OAI-ORE, CMDI) have provisions for linking to the bitstreams, which makes it possible to download those too. The repository administrators have the option to export the AIPs via tools provided with the software.

[1] https://www.clarin.si/repository/xmlui/page/deposit
[2] https://www.clarin.si/repository/xmlui/page/item-lifecycle

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The repository metadata offers an OAI-PMH endpoint and REST API [1] for machine access and exports the following DC elements: creator, date, description, identifier, language, publisher, relation, rights, source, title, and type. The repository metadata is regularly harvested by several other repositories and services, such as the CLARIN VLO, the Virtual Language Observatory [2], OLAC, the Open Language Archives Community [3], OpenAIRE [4], and OpenDOAR [5].

The repository also has browse and search capabilities, and provides faceted search and filter queries on the metadata. All the metadata as well as text files are indexed for full text search [5].

Each repository item is assigned a PID (a handle), a textual hint of how to correctly cite the item is shown prominently on the item page (also providing a bibtex snippet), and we have also written a guide for our users on how to cite the repository items properly [7].

[1] http://www.clarin.si/repository/oai/request?verb=Identify
[2] https://vlo.clarin.eu/
[3] http://www.language-archives.org/archive/clarin.si
[4] https://explore.openaire.eu/search/dataprovider?datasourceId=re3data::fe0d76581a60e1287a93e2ed2cb29339
[5] http://v2.sherpa.ac.uk/id/repository/4302
[6] https://www.clarin.si/repository/xmlui/discover?advance
[7] https://www.clarin.si/repository/xmlui/page/cite?locale-attribute=en

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

CLARIN.SI requires that a set of metadata (both mandatory and recommended) providing information about the submitted data are filled in [1].

The required set is chosen in order to support different metadata profiles/formats (e.g., LINDAT/CLARIN CMDI profile [2], Dublin Core and OLAC). Therefore, we support various formats, including OAI-ORE, METS and others, in our OAI-PMH endpoint. Because the other profiles/formats are dynamically constructed, the sustainability and future evolution of metadata formats can be easily supported.

The user can see these descriptive metadata, together with licensing information covering intellectual property, conditions use and others on the item view page.

The depositors upload files in either standard formats for language resources [3] suitable for long term preservation that are constantly updated by language resource community experts, or other formats. In case the latter happens, editors require a detailed description on how to process the data to be available in the data itself. Changing the format of the data is possible because of the distribution license [4] and the known formats are also supported by the underlying CLARIN-DSpace software [5].

[1] https://www.clarin.si/repository/xmlui/page/metadata
[2] https://catalog.clarin.eu/ds/ComponentRegistry/#/?registrySpace=published&itemId=clarin.eu:cr1:p_1403526079380&_k=qkn920
[3] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[4] https://www.clarin.si/repository/xmlui/page/contract

[5] https://wiki.duraspace.org/display/DSPACE/User+FAQ#UserFAQ-HowdoesDSpacepreservedigitalmaterial?

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

For the digital repository we use the platform which has been developed by the Institute of Formal and Applied Linguistics, Charles University in Prague, where it is used to host LINDAT-CLARIAH-CZ (Centre for Language Research Infrastructure in the Czech Republic) repository [1]. The LINDAT platform Clarin-DSpace is based on DSpace, adapted for archiving and distributing language resources, and is available on GitHub [2]. DSpace itself is based on the OAIS reference model and the implementation follows a list of standards that are relevant for the CLARIN community [3].

In our implementation, the adapted and localised version of the repository is running in a virtualized OS instance on an application cluster consisting of five application servers. Two servers are configured to be able to run the instance at any time and all the servers have access to a data storage with hardware RAID controllers and RAID-6 volumes with support for versioned snapshots and replication. In addition, the system includes a distributed filesystem running on application server local disks for local volumes, which also provides high-availability within the cluster.

Submitted datasets are stored in a DSpace repository bitstream store on a network-attached volume managed by one of the application servers and replicated on another server, while the data itself is stored on a distributed high-availability file store.

Repository and dataset metadata is stored in a virtualized PostgreSQL instance inside the virtual machine instance.

Repository software and configuration is tracked with the Git version control system and exported from the virtual machine instances.

Each active virtual machine instance for the CLARIN.SI web system repository and related services is cloned and backed up regularly and before each software configuration change or update. In addition, configuration changes are tested on a hot beta instance which is also available in case of system failure.

Additional application servers are available through Network Infrastructure Centre support at JSI and at the Department of Knowledge Technologies, where the infrastructure is managed to support contingency in case of failure of the application server infrastructure or technical issues in the server room. Virtual machine image backups, dataset backups and database export backups are cloned to the backup system of the Department of Knowledge Technologies at a different location to allow for full recovery in case of data centre failure. The infrastructure and procedures are also documented on the repository's About page [3].

[1] https://lindat.mff.cuni.cz/repository/xmlui/
[2] https://github.com/ufal/lindat-dspace
[3] https://www.clarin.si/repository/xmlui/page/about

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept.

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CLARIN.SI infrastructure is hosted at the Jožef Stefan Institute. The computer hardware that runs the CLARIN.SI repository is the property of the Jožef Stefan Institute, which also takes care of its configuration and maintenance, but the oversight of the programs that constitute the repository platform is performed under the direction of the CLARIN.SI consortium. The CLARIN.SI consortium agreement (referred to from the governance page [1]) states, in Article 10, that the Jožef Stefan Institute is its technical centre which performs the services as required by the statues of CLARIN ERIC [2], but that, in case the services are inadequate, the CLARIN.SI Management Committee can relocate the technical center to another member of the CLARIN.SI consortium.

Since the CLARIN.SI repository infrastructure is maintained as part of the JSI IT infrastructure, it shares its IT and physical security. IJS IT security officers are responsible for general network security and provide guidelines for secure server and service maintenance. The repository team has a dedicated system administrator responsible for infrastructure security, working closely with the NIC team. Servers and network devices are kept in a dedicated computer centre with physical access limited to authorised personnel, and the backup server and facilities are located in a different computer room in a different building, with similar physical controls. Both buildings are located on the Institute's campus with 24/7 security service as well as remote monitoring and alerting and IT security staff on-call. Physical facilities are equipped with fire alarms and inert gas fire retardant systems, uninterrupted power supply and an automatic stand-by electrical power generator to ensure full operations under adverse conditions.

The Jožef Stefan Institute provides network security, border monitoring and protection (firewalls, logging, security advisory and assessments). The datasets we consider for security and preservation consist of multiple components: (1) submitted datasets (files or bitstreams), (2) metadata for repository and the datasets, (3) digital repository software and its configuration, (4) the underlying operating system instances with configuration and logs for the repository and related

services in their separate operating system instances and (5) exported backups of configuration and databases for related service instances.

Each component has its own data security and backup policy and implementation. Specifically, system images are checkpointed before any configuration changes or updates and regular replication of system images to a secondary location is performed. Files representing datastreams are backed up as independent files. All databases undergo regular daily database exports which are backed up and replicated by a different mechanism from the operating system image backup. The same approach to database consistency is implemented for databases used in service instances, with the exception of specialized databases created from available datasets where the original data and transformation scripts are the main back-up strategy.

Please see R15 (Technical infrastructure) for further description of security measures at the physical, hardware, administrative and operation levels.

Our servers, services and procedures are supervised regularly for security issues by Jožef Stefan Institute Network Infrastructure Centre in collaboration with SI-CERT (Slovenian Computer Emergency Response Team).

We follow the upgrade and development path of CLARIN-DSpace, but we ensure service consistency and availability by monitoring development discussions and security assessment and performing validations on a beta instance before applying each configuration or upgrade change.

[1] https://www.clarin.si/info/governance/
[2] http://hdl.handle.net/11372/DOC-143

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its

relevance to your organization, as well as any other related thoughts.

Response:

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: