CLARIN-LV repository

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

CLARIN-LV is a repository of language resources and tools (language domain repository) created in Latvia. The CLARIN-LV repository is run by the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) [1]. The Ministry of Education and Science appointed IMCS UL as National contact point and service provider of CLARIN ERIC [2].

CLARIN-LV has been registered as CLARIN C-center [3] in March, 2020. As a center, it is committed to the long-term preservation of deposited items, following the best practice in digital preservation [4]. Our aim is to become certified CLARIN B Centre, successful conclusion of CoreTrustSeal assessment is one of requirements for B Centre certification [5]: “the centre cannot be certified as a B Centre until the CoreTrustSeal assessment has been successfully concluded”. The requirement to establish B Centre is one of obligations listed in CLARIN statutes [6]: “6.2. Each member shall: … (f) provide at least one data and service centre; …”.

As a national repository system of language resources and tools CLARIN-LV hosts language resources and tools created by different academic organizations of Latvia, including corpora data that contain governmental data (e.g. Parliamentary debates, legislation of Republic of Latvia). Repository also contains information about language resources of the Latvian Language Agency (a direct administration institution supervised by the Minister of Education and Science).

References

[1] Institute of Mathematics and Computer Science, University of Latvia: https://lumii.lv/
[2] CLARIN ERIC: www.clarin.eu
[3] CLARIN Centres: https://www.clarin.eu/content/clarin-centres
[4] About including Policies: https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en
[5] Checklist for CLARIN B Centres: https://www.clarin.eu/content/checklist-clarin-b-centres
[6] STATUTES OF THE CLARIN ERIC, Approved by the European Commission on 4 April 2018: https://www.clarin.eu/content/checklist-clarin-b-centres

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository’s Designated Community.

The designated community of the CLARIN-LV language resource and tools repository is the national and international research community, in particular scholars and students in digital humanities and social sciences, computational linguists and corpus linguists, language technology developers working with written and spoken language data, scholars and students in communication and media studies, scholars and students in linguistics and translation studies. The repository is also valuable for companies developing applications in the area of language technology. The principal language covered is Latvian, but it also includes language resources for the Latgalian language.

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

CLARIN-LV repository serves as library (or catalogue) of language resources and tools for SSH researchers. Therefore we do not strictly require to upload the data and accept also a metadata-only record, if required [1]. Submission of metadata only records is usually chosen when (1) data or tools are deeply incorporated into Web Service (e.g., NLP pipe toolkit (http://hdl.handle.net/20.500.12574/4.)), or, (2) there are copyright limitations on data distribution (http://hdl.handle.net/20.500.12574/11).

The Repository performs basic curation of the submissions by checking and editing the metadata. CLARIN-LV repository is based on the software developed by LINDAT/CLARIN. Thus in the data curation we follow principles and guidelines established by LINDAT/CLARIN [2]. The submission and curation procedure includes the following steps:
1) At first, users deposit resources into the repository themselves using a web-based submission workflow: 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 [3].

2) Before accepting submission, editors review and curate the submission. The editors have the option to inspect the data, edit the metadata or to return the submission to the depositor requesting changes or more details. Several 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 is insufficient [4, 5].

CLARIN-LV performs automatic weekly checks on the metadata and data and may request additional information from the depositors. Occasionally, minor metadata modifications can be done after the item was published. All the changes (including the ones done by editors) are recorded in the provenance metadata [6].

References

[1] What submissions do we accept? (clarin.lv) :
https://repository.clarin.lv/repository/xmlui/page/faq?locale-attribute=en#what-submissions-do-you-accept
[2] Lindat/CLARIN Core Trust Seal: https://www.coretrustseal.org/wp-content/uploads/2019/08/LINDAT-CLARIN.pdf
[3] clarin-dspace input forms: https://github.com/ufal/clarin-dspace/blob/clarin/dspace/config/input-forms.xml
[4] Standards for LRT: https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf;
[5] clarin-dspace Metadata info: https://github.com/ufal/clarin-dspace/wiki/Metadata-info
[6] CLARIN-LV Deposited Item Lifecycle:
https://repository.clarin.lv/repository/xmlui/page/item-lifecycle?locale-attribute=en

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

Insource/Outsource Partners. If applicable, please list them.

For the sign-in, we use the identification service of the Latvian academic identity federation LAIFE [1] developed and maintained by the University of Latvia, which supports single sign-on. Members of LAIFE [2] are all academic institutions of Latvia.

The CLARIN-LV repository is installed on a universal cloud computing platform (e-Spiets) [3], created and maintained by IMCS UL as part of the research infrastructure in Latvia. Data and curation software is hosted on e-Spiets, supporting the whole process from depositing to curation and access.

We use Handle.Net Registry [4] for reserving and handling persistent links (identifiers) to the datasets and tools. We have a 5-year subscription to Handle.Net, which will be extended in advance.

[1] Latvian academic identity federation LAIFE: https://laife.lanet.lv/
[2] LAIFE Organizations: https://laife.lanet.lv/?page_id=46
[3] e-Spiets universal could http://e-spiets.lv/
[4] Handle.Net Registry https://www.handle.net

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

Summary of Significant Changes Since Last Application (if applicable).

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

Other Relevant Information.

The CLARIN-LV node is part of CLARIN virtual infrastructure. Its metadata is harvested by the CLARIN Virtual Language Observatory (VLO) [1].

An overview of the repository can be seen at re3data[2]. We are using DSpace [3] as the basis of our repository system, although in a modified version called CLARIN DSpace [4], developed by LINDAT/CLARIN [5] .

Since Mach, 2020, when repository was established, 36 digital language resources are registered (about 15-20 resources per year) - 23 corpora, 11 lexical conceptual resources and 2 tools and services. Our repository currently hosts datasets for five languages, most of resources are for the Latvian language. The data storage, provided by a Redundant Array of Independent Disks (RAID), is prepared to scale up rapidly and transparently; it is well monitored and renewed if signals of failure are registered. Off site and on site backups are also kept (see R9). We have more than 40 registered users, including users from Lithuania, Sweden and Greece.

The number of visitors to our repository has increased every month from 229 item views in July, 2020 to more than 1500 item views in Winter, 2022. The most viewed items include Tezaurs.lv, Balanced Corpus of Modern Latvian, and Latvian Treebank.

The state research programme “Digital resources for humanities” has recently started to use the CLARIN-LV repository to register and store outcomes of the programme. The use of the CLARIN-LV repository is planned also in upcoming state research and education programmes in digital humanities and language technology. We have signed agreements with several partner institutions: University of Latvia, Institute of Folklore and Arts, Riga Stradins University, etc. as depositors.

References;

[1] CLARIN VLO: https://vlo.clarin.eu/
[2] re3data: https://www.re3data.org/
[3] DSPACE: https://duraspace.org/dspace/
[4] CLARIN Dspace: https://github.com/ufal/clarin-dspace
[5] LINDAT/CLARIAH-CZ: https://lindat.cz/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
With the new changes I would now have been willing to accept a level 4 here.

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

The mission of CLARIN research infrastructure is to “create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences” (SSH) [1]. This objective is being implemented by the construction and operation of a shared distributed infrastructure.

The CLARIN-LV Centre at the IMCS UL implements the CLARIN mission in Latvia by collecting, documenting, curating and providing easy and sustainable long-term access to the digital Latvian language data and tools for a wide group of users: SSH research community, language technology developers, language teachers and students, and others (e.g. via citizen science), allowing for discovering, exploring, exploiting, annotating, and analysing Latvian language data [2]. CLARIN-LV is thus committed to the long-term care and preservation of items deposited in its repository [3] and strives to adopt the current best practices in digital preservation, and in particular, to become a certified CLARIN B Centre [4].

The CLARIN-LV mission is supported by integration of the repository into the EU CLARIN infrastructure [5]. 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.

The national research infrastructure CLARIN-LV focuses on Latvian (and Latgalian) language resources, but not excluding other languages, in particular morphologically rich languages with this mission supported by CLARIN Knowledge Center SAFMORIL [6] (CLARIN-LV is a member of SAFMORIL).

IMCS UL is conducting research on natural language processing and provides access to different language resources and tools for almost 30 years. The strategy of IMCS UL ([7] (page 10: R2.1, and R2.3) for 2021-2027 lists CLARIN as international research e-infrastructure that will be further developed by IMCS UL. Furthermore, at the national level, the strategy highlights development of modern digital language resources (page 11: U3.1.1) as direction to support development and growth of Latvia. Following strategy, the CLARIN-LV provides access to the digital language data collections and tools, and expertise for researchers to work with them [8].

Overview of the large infrastructures in Latvia, where CLARIN-LV belongs, can be found in the “Report on Participation of Latvia in European Strategy Forum on Research Infrastructures (ESFRI) European Roadmap for Research Infrastructures Consortia” [9]. The Ministry of Education and Science has appointed IMCS UL as the CLARIN National Contact Point and service provider. Research infrastructures included on the National Roadmap are currently supported and financed by the European Regional Development Fund project (2018-2022). The financial support is being continued through the State Research Programme “Digital resources for humanities” and the project "Research on Modern Latvian Language and Development of Language Technology” of the National Research Programme "Letonika – Fostering a Latvian and European Society" (2022-2024).

References

[1] CLARIN Value proposition, 2017 [http://hdl.handle.net/11372/DOC-138]
[2] About including Policies (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en
[3] Preservation Policy (clarin.lv):
https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en#preservation-policy
[4] Checklist for CLARIN B Centres: hdl:11372/DOC-78
[5] CLARIN – European Research Infrastructure for Language Resources and technology: https://www.clarin.eu/
[6] CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages:
https://www.clarin.eu/sites/default/files/k-centre_certificate_-_safmoril.pdf
[7] Institute of Mathematics and Computer Science, University of Latvia. Strategy 2021-2027 (in Latvian).
https://lumii.lv/media/uploads/2021/10/29/110_sl_pielikums_lumii_strategija_W6GxYmn.pdf
[8] CLARIN-LV: https://www.clarin.lv/en-us/
[9] “Report on Participation of Latvia in European Strategy Forum on Research Infrastructures (ESFRI) European Roadmap for Research Infrastructures Consortia”:
https://www.izm.gov.lv/sites/izm/files/media_file/izm_050416_inf_zinojums_par_esfri_cela_kartes_konsorcijos_2016.pdf

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

2. 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
Accept

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 together with the color coded “openness” of the license (“public”, “academic”, “restrictive”). The license for a repository item has been selected during submission by submitter. Open/public licenses are strongly preferred when possible.
At the moment, CLARIN-LV distinguishes three types of contracts:
- For every deposit, we enter into a standard contract with the submitter, the so-called "Distribution License Agreement" [2], in which we describe our rights and duties and the submitter acknowledges that they have the right to submit the data and gives us (the repository centre) right to distribute the data on their behalf.
- Everyone who downloads data is bound by the licence assigned to the item - in order to download protected data, one has to be authenticated and needs to electronically sign the licence. 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 the user further access to the repository. A list of available licenses in our repository can be found at [3].
- For submitters, there is a possibility for setting custom licences to items during the submission workflow.

References

[1] About including Policies (clarin.lv):
https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en#terms-of-service
[2] Distribution License Agreement: https://repository.clarin.lv/repository/xmlui/page/contract
[3] Available Licenses (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/licenses

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

3. Continuity of access

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

CLARIN-LV is financed by the Ministry of Education and Science through the European structural funds project “University of Latvia and its institutes in the European Research Area - Excellence, activity, mobility, capacity” (January, 2018 - December, 2022).

Additional funding is provided through:
(1) the State Research programme “DIGITAL RESOURCES FOR HUMANITIES: INTEGRATION AND DEVELOPMENT” (October, 2020 - September, 2022)
and
(2) recently started the National Research Programme "Letonika – Fostering a Latvian and European Society" project "Research on Modern Latvian Language and Development of Language Technology” (2022-2024)

CLARIN-LV is expected to receive funding also from the planned Language Technology Excellence Center project which has been listed in Latvian Recovery and Sustainability Mechanism Plan [1]. The Language Technology Excellence Center project is expected to start in autumn, 2022 and will continue till 2026.

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. In case of user support, it is also provided through related projects run by IMCS UL, where CLARIN-LV staff involved.

CLARIN-LV has measures in place to preserve data access in case of unexpected emergency budget cuts. The

CLARIN-LV repository platform runs under the software developed for the LINDAT/CLARIN repository for linguistics [2], which is a low maintenance system, relatively easy to install and maintain. In case of unexpected funding problems, this will allow the IMCS UL keep running the repository through the base funding [3] provided to IMCS annually by Ministry of Education and Science, since the development of CLARIN-LV is among priorities of IMCS for 2021-2027 [4].

The LINDAT/CLARIN repository for linguistics repository is open source software based on Dspace allowing to ensure the sustainability of access. It allows 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-LV infrastructure would be terminated completely, one of the other CLARIN centres would be able to host data and to reconfigure its permanent identifiers for the CLARIN-LV collection.

We are considering to sign an agreement with the Czech LINDAT-CLARIAH-CZ centre for such a migration. Moreover, there are at least nine CLARIN centres [6] running this same system.

The continuity plan for the CLARIN-LV repository is published at [5]. CLARIN-LV is hosted in a cloud infrastructure at IMCS UL, providing a highly available storage, backup and disaster recovery facilities for archival data and software (for details see R9).

References

[1] Latvian Recovery and Sustainability Mechanism Plan:
https://likumi.lv/ta/id/322858-par-latvijas-atveselosanas-un-noturibas-mehanisma-planu
[2] LINDAT/CLARIN repository for linguistics: http://lindat.mff.cuni.cz/lindat/
[3] Ministry of Education and Science. Three pillar funding model: https://www.izm.gov.lv/en/three-pillar-funding-model
[4] Institute of Mathematics and Computer Science, University of Latvia. Strategy 2021-2027 (in Latvian).
https://lumii.lv/media/uploads/2021/10/29/110_sl_pielikums_lumii_strategija_W6GxYmn.pdf
[5] About including Policies (clarin.lv):
https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en#preservation-policy
[6] GitHub - ufal/clarin-dspace: clarin-dspace digital repository based on DSpace and LINDAT/CLARIN DSpace:
https://github.com/ufal/clarin-dspace#clarin-dspace-deployments

Translations

From Latvian Recovery and Sustainability Mechanism Plan (approved by Cabinet of Ministers) [1]
(718) Investments in training and related R&D activities, including related research infrastructure, are planned in three areas: quantum technology, high-performance computing and language technology, for a total funding of EUR 17.5 million euro. Funding will be provided for the development of learning modules combining face-to-face and online learning, teacher training, the creation or purchase of teaching materials, resources, tools and digital platforms, innovation activities and training, including transnational workshops, summer schools, interdisciplinary mobility and investment. related R&D activities, including the involvement of foreign professionals and researchers. The research results will form the content

base of the study modules, including materials for the preparation of the study modules, for example, text and speech corpora in language technologies, language resources and tools.

Reviewer Entry

Reviewer 1

Comments:
A formal succession agreement with an organisation like LINDAT will be needed to obtain a level 4 for this requirement.

Reviewer 2

Comments:

4. 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
Accept

Response:

CLARIN-LV repository serves as a library (or catalog) of language resources and tools accepting both - submission of data and metadata, or submission of a metadata-only record [1].

During submission of language data, the submitters acknowledge 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 [2]. 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.

All submissions are reviewed by the repository staff (editors). For language data, in particular language corpora, several legal issues must be considered [3]. 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 custom distribution license tailored specifically for the particular item. We can control access to items and submissions and grant it on a per user basis. If a more restricted access is required, we need to work with the submitter, in person or via email, on defining the target group of users or individuals with access.

To date the repository has no submissions containing confidential data or data with disclosure risk and we do not expect this to change in the future. Our data is Open Access or distributed under similar public licenses, in particular variants of the Creative Commons licences. 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 [4] (which CLARIN-LV is a member of) organises training sessions in the legal and ethical management and distribution of text data.

References
[1] What submissions do we accept? (clarin.lv) :
https://repository.clarin.lv/repository/xmlui/page/faq?locale-attribute=en#what-submissions-do-you-accept
[2] Distribution License Agreement (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/contract
[3] Intellectual Property Rights (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/about#about-ipr
[4] CLARIN Legal and Ethical Issues Committee: https://www.clarin.eu/governance/legal-issues-committee

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

5. 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:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

The CLARIN-LV repository is hosted at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) in Riga. IMCS UL is a stable institution established in 1959 as a computing centre of research character. The staff of IMCS UL consists of more than 200 employees from whom about half are in academic positions (senior researchers, researchers and assistants). IMCS UL is also one of the largest Internet service providers in Latvia.

CLARIN-LV receives moderate but stable dedicated funding through two research infrastructure projects (European structural funds 2018-2022 and a State Research Programme 2020-2022) to operate. In January, 2022 the new project "Research on Modern Latvian Language and Development of Language Technology" (2022-2024) of the National Research Programme "Letonika – Fostering a Latvian and European Society" started providing some funding till 2024. Additional funding from the Recovery Funds is expected in Autumn, 2022, allowing continuation at least till Summer, 2026 (for details see R3).

CLARIN-LV has been previously supported through different projects and activities related to the State Language Policy Guidelines (2015-2020). The task of the digital infrastructure for the Latvian language is listed in the State Language Policy Guidelines (2021-2027) [1], while Science, technological development and innovation guidelines for 2021-2027 include a goal to ensure participation in international ESFRI research infrastructure platforms and ERIC consortia [2].

The 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. The staff consists of about 5 individuals, about 1.5 FTE in total: coordination (0.2), core technical repository management (0.5), network and cloud infrastructure management (0.3), curation and management of data and metadata (0.3), knowledge sharing and user involvement (0.2). Most of the staff involved have permanent positions. The staff is experienced to manage main aspects of the repository (data and metadata curation, the technical maintenance of the software and hardware, knowledge sharing).

CLARIN-LV staff regularly participates in CLARIN committees and task forces. They also regularly participate in training and professional development activities organised and supported by CLARIN ERIC. The CLARIN-LV has sufficient budget to attend all necessary meetings.

Since Latvian language is among less resourced languages that have a rather small amount of language resources available, when compared to widely used languages [3], we do not expect rapid growth of repository size and scalability problems. However, there could be growing interest in knowledge sharing and user involvement. Therefore knowledge sharing and user involvement activities are also implemented through related projects, e.g., State research programmes in Digital Humanities and Letonika, and teaching activities.

References

[1] State Language policy guidelines (2021-2027):
https://likumi.lv/ta/id/325679-par-valsts-valodas-politikas-pamatnostadnem-2021-2027-gadam
[2] Science, technological development and innovation guidelines (2021-2027):
https://likumi.lv/ta/id/322468-par-zinatnes-tehnologijas-attistibas-un-inovacijas-pamatnostadnem-20212027-gadam
[3] ELE___Deliverable_D1_22__Language_Report_Latvian_.pdf (european-language-equality.eu): https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_22__Language_Report_Latvian_.pdf

Translations

State Language Policy Guidelines (2021-2027)

V. Directions and tasks of the state language policy:
…
Task 2.4. To create a strategic infrastructure for the digitization of the Latvian language.
….
Science, Technological Development and Innovation Guidelines (2021-2027):

1.2. Direction of action. R&D infrastructure for research excellence and innovation:

1.2.1.2. Ensure participation in the dissemination and expansion of excellence sub-activity, international ESFRI research infrastructure platforms and ERIC consortia.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

6. 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
Accept

Response:

CLARIN-LV is a member of the CLARIN ERIC infrastructure and, as such, is in regular touch with its experts. CLARIN-LV is a member of:

- the CLARIN ERIC Standing Committee of Technical Centres,
- the Legal and Ethical Issues Committee,
- the Standards Committee,
- the User Involvement Committee [1].

The Committees consist of other CLARIN members and the main focus is on knowledge sharing and creating guidelines for the whole CLARIN. In regular virtual and physical 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-LV repository team is in regular touch, via email, a dedicated Slack channel and a GitHub issue tracker with LINDAT/CLARIN, the developers of the DSpace CLARIN platform.

The technical team involved in the CLARIN-LV 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, and the annual CLARIN conferences.

The metadata in our repository is regularly harvested by several harvesters including the CLARIN ERIC VLO and OLAC. These services 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.

CLARIN-LV has an email address (info@clarin.lv) where users can ask questions, report problems or suggest improvements. CLARIN-LV organizes regular user involvement workshops [2] to inform and teach the SSH community in

Latvia about different aspects of CLARIN-LV, in particular, use of language resources and application of tools for humanities research.

References

[1] CLARIN Gevernance: https://www.clarin.eu/content/governance

[2] Konferences un seminari (clarin.lv); https://www.clarin.lv/lv/clarin-latvija-seminari

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

DIGITAL OBJECT MANAGEMENT

7. 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
Accept

Response:

The general overview is described in Deposited Item Lifecycle (clarin.lv) [1]. CLARIN-LV provides guidelines for data submission that include preferred formats and metadata preparation and instructions for preparing and submitting data for publication [3,4,5].

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, thus they can confirm it was not corrupted during the transport. Also, the editors check the files before approving an item to be published. For certain file formats, these weekly checks also contain a test by additional tools e.g., PNG image files are checked for corruption using `pngcheck` or zip archives using `unzip -t`.

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 e.g., if there are URLs in the metadata they are fetched, or they can see the level of support (supported/know/unknown) for the submitted file formats. Some of these editors' tools are part of the weekly checks, e.g., all required metadata are present, URLs are working. The results of weekly checks are automatically sent to the repository staff. If an error is reported, the error message is analysed and necessary corrections are made by the repository staff, as soon as possible.

One of our policies, coming from our view on persistent identifiers, is that a handle always resolves to one concrete dataset. We do this for the sake of reproducibility of results using the dataset. We do not support changing the data. A changed or a new version of a dataset must be submitted as a new repository item. The New Version Guide [2] documents versioning process from user's and technical perspective. The new and the old version have the relation added to their metadata and are visually represented on the web page (see, for instance, http://hdl.handle.net/20.500.12574/10.). Changes to the metadata happen occasionally (mostly due to 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 [1]. All the changes (including the ones done by editors) are recorded and stored in DSpace (field dc.description.provenance), accessible only by administrator. After the item was approved, only the administrators are able to change its metadata. The data producers can refer to the Deposited Item Lifecycle [1] mentioned above to get acquainted with the details or ask directly our helpdesk.

References:

[1] Deposited Item Lifecycle (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/item-lifecycle?locale-attribute=en
[2] New Version Guide: https://github.com/ufal/clarin-dspace/wiki/New-Version-Guide
[3] Standards for LRT (clarin.eu): https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[4] About including Policies (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en
[5] Metadata policy (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en#metadata-policy

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

8. 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
Accept

Response:

CLARIN-LV accepts any language data, including but not limited to corpora, text and speech collections, monolingual and multilingual machine readable dictionaries and lexicons and language processing tools.

Repository 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 CLARIN Standards Committee [5] maintains and extends the list of recommendations for data deposition formats [6], specified by CLARIN centres.
During submission the web interface allows initial value checks (e.g., for valid email), requesting submitters to fill all required fields correctly (for details of submission and review process see R11).

According to 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, natural language processing is an active research area with many data formats in constant development, and CLARIN-LV 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 explain their choice.

If the format is unknown or not in the list of the recommended standard formats [3], it must be documented and the documentation must be either part of the submission or the metadata must contain a link to it. For example, submission of Latvian AMR Sembank (http://hdl.handle.net/20.500.12574/40.) includes readme file, describing data and metadata formats, while Latvian Treebank records (e.g., http://hdl.handle.net/20.500.12574/55) provides links to corresponding documentation files.

If accepted, these submissions are preserved as submitted. There are currently no policies to deaccession these materials.

In the case of XML files, 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).

References

[1]: ComponentRegistry:
http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1349361150622/xsd
[2]: CLARIN Standards Information System: https://standards.clarin.eu/sis/
[3] CLARIN-LV repository About and Policies: https://repository.clarin.lv/repository/xmlui/page/about
[4] Metadata (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/metadata
[5]: CLARIN Standards Committee:
https://www.clarin.eu/content/standards#clarin-standards-committee%C2%A0%E2%80%93-information
[6]: Format Recommendations (ids-mannheim.de):
https://clarin.ids-mannheim.de/standards/views/recommended-formats-with-search.xq

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

9. Documented storage procedures

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
With the new addition I would have been willing to accept a level 4 here.

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

CLARIN-LV is hosted in an OpenStack cloud infrastructure at IMCS UL [1] 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 supports backing up complete VM images or image snapshots and keeping multiple versions. The file backups and export backups are performed by backing up database export files and filesystem snapshot-derived files, where possible. Backups are stored in a separate FreeNAS cold storage infrastructure at IMCS UL. Additionally, we have implemented automatic remote backups to the Wasabi Cloud Storage infrastructure [3] (the eu-central-1 data centre in Amsterdam, Netherlands) to exclude the possibility of a single point of failure in the back-up and restoration procedures.

With the use of the Clarin-DSpace version of DSpace [2], 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.

For the archival storage, our editor takes the submission. Using the web interface, the metadata are updated (added, deleted, modified), the submitted bitstreams are validated. In general, the editors ensure the consistency and quality of each submission. If an editor approves an item, the Archival Information Packages (AIPs) is available.

Part of the archiving workflow is the integrity check of the data and the metadata by the archive manager. The metadata is parsed for syntactic correctness, completeness and soundness. The object data is tested for syntactic correctness if possible. All datastreams and versions are equipped with a MD5 checksum, which is checked in coordination with the backups as described above. For further details of the ingest part of the archiving workflow see also R12.

CLARIN-LV encourages users to use formats recommended by CLARIN, but to accept all data formats. If the format is unknown or not in the list of the recommended standard formats, it must be well documented and the documentation must be either part of the submission or the metadata must contain a link to it.

We are open to all submissions which meet our standards (Data Producers must be authenticated which means they must have an academic background or have 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 instance is maintained on another virtual machine to avoid service interruption in case of upgrades or configuration modification.

References:
[1] e-spiets universal could: http://e-spiets.lv/
[2] Clarin-DSpace: https://github.com/ufal/clarin-dspace
[3] Wasabi Cloud Storage infrastructure: https://wasabi.com/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

10. Preservation plan

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

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept

Response:

CLARIN-LV has the right to copy, transform, store and provide access to the data [1]. Redundant backups of all data and multiple drives and virtual machines assure long-term preservation. The preservation policy [2] 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 Centre [3]. Similar preservation policy is implemented by certified CLARIN repositories in Czech Republic [9] which stores more than 1300 language resources and tools, Slovenia [10] and Poland [11]. All these repositories are based on Clarin-DSpace [4].

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 [5];
- 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 (for more information see [6]). 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 [7], 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 events (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 text corpus.

The repository encourages the usage of specific file formats as recommended by CLARIN [5]. 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.

For textual resources (that account for a significant number of the repository items), text is always encoded in Unicode, and well known formats, such as XML, JSON, CoNLL , are used 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 cases where proprietary or custom formats need to be used, we require detailed documentation or link to the documentation, in order to make the implementation of future data converters possible. For example, Latvian AMR Sembank dataset (http://hdl.handle.net/20.500.12574/40.) includes readme file, documenting data and metadata formats of this submission. In case of tools documentation needs to be uploaded together with the tool. For example, LVBERT submission (http://hdl.handle.net/20.500.12574/43.) documentation is provided in a readme.txt file. For tools no changes or updates are planned.

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].

The R10 is currently in implementation stage. According to the preservation policy [2] the repository ensures that datasets are ingested and distributed in accordance with their license. Since repository currently contains rather small number of language resources and tools, file formats that could be migrated in future are not defined. Any future file format conversions will be done taking into account recommendations of CLARIN ERIC Standards Comittee [12]. We plan to meet R10 in next 2-3 years, before submission of next application for CTS.

References

[1] repository.clarin.lv/repository/xmlui/page/contract
[2] https://repository.clarin.lv/repository/xmlui/page/about?locale-attribute=en#preservation-policy
[3] https://www.clarin.eu/content/clarin-in-a-nutshell
[4] https://github.com/ufal/clarin-dspace
[5] https://www.clarin.eu/content/standards-and-formats
[6] https://repository.clarin.lv/repository/xmlui/page/deposit
[7] https://repository.clarin.lv/repository/xmlui/page/item-lifecycle

[8] https://repository.clarin.lv/repository/xmlui/page/about

[9] https://lindat.mff.cuni.cz/repository/xmlui/

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

[11] https://clarin-pl.eu/dspace/

[12] Standards Committee: https://www.clarin.eu/governance/standards-commitee

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

11. 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
Accept

Response:

The CLARIN DSpace platform used by CLARIN-LV 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, it is required that an XML schema and documentation is also included with the submission.

References
[1] About Metadata (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/metadata?locale-attribute=en
[2] About Citations (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/cite?locale-attribute=en
[3] info@clarin.lv

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

12. 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
Accept

Response:

After submitting the data, a curation platform, integrated into the CLARIN-DSpace software[1], 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 the final approval by the editor, the submission becomes visible and retrievable via the repository web interface, as well as machine-readable interfaces (OAI-PMH, REST API). Information on the submission and curation workflows is described on the following pages of the CLARIN-LV repository website:
- How to deposit [2]
- Deposited Item Lifecycle [3].

The complete workflow consists of:

1. Creating metadata and uploading 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. Assigning persistent identifiers. Persistent identifiers (PIDs) provide unique identification of the research data and metadata in a location-independent manner. This means that the same identifier will persist even in the case of data or metadata migration.

3. Uploading data and specifying licenses. Submitter uploads data and chooses the appropriate license for the data. The web interface provides guidance to select the appropriate license using a graphical license selector tool. This step is performed only if data is submitted to the repository.

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

5. Publishing the 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.

References:
[1] GitHub - ufal/clarin-dspace: clarin-dspace digital repository based on DSpace and LINDAT/CLARIN DSpace: https://github.com/ufal/clarin-dspace
[2] How to deposit: https://repository.clarin.lv/repository/xmlui/page/deposit
[3] Deposited Item Lifecycle: https://repository.clarin.lv/repository/xmlui/page/item-lifecycle

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

13. 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
Accept

Response:

The CLARIN-LV repository runs on the software developed for the LINDAT/CLARIN repository and thus supports data discovery and identification similarly to the LINDAT repository in Prague. In particular:
- The repository has browse and search capabilities, it provides faceted search and filter queries on the metadata. All the metadata as well as text files are also indexed for full text search [1].
- The repository provides an OAI-PMH endpoint and we are harvested by several organisations (CLARIN VLO [2], OLAC, WOS) and REST API. We support various formats, including OAI-ORE, METS and others, in our OAI-PMH endpoint [3]. The full list of supported formats is listed in [4] but some of the formats might not be applicable to all items.
- Each repository item is assigned a PID (a handle), a textual hint how to correctly cite the item is shown prominently on the item page (also providing a bibtex snippet). The repository issues PIDs via a local Handle.net server running on the repository's server with a global PID prefix, 20.500.12574, registered by the Handle.Net Registry (HNR).
- Data citations are implemented according to the recommendations of the Research Data Alliance's Data Citation Working Group. End-users are asked to acknowledge and cite data sources properly in all publications and outputs. We have also adopted a LINDAT guide for users on how to cite the repository items properly [5].

References:

[1] Advanced search (clarin.lv): https://repository.clarin.lv/repository/xmlui/discover?advance

[2] CLARIN VLO: https://vlo.clarin.eu/

[3] http://repository.clarin.lv/repository/oai/request?verb=Identify

[4] Supported formats (clarin.lv): https://repository.clarin.lv/repository/oai/request?verb=ListMetadataFormats

[5] About citation (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/cite

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

14. 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
Accept

Response:

CLARIN-LV 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 of 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 in other formats. In case of non-standard formats, a detailed description how to process the data is required. This documentation is stored on the repository along with the data and metadata and is provided to everyone who wishes to download/access the resource.

In case of tools, we require documentation (or link to the documentation), providing instructions for installation and running, as well as information about requirements and dependencies (if any). When the tool is submitted to the repository, the documentation needs to be uploaded together with the tool. For example, for LVBERT submission (http://hdl.handle.net/20.500.12574/43) documentation is provided in a readme.txt file.

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].

References

[1] About metadata (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/metadata

[2] CMDI Component Registry: https://catalog.clarin.eu/ds/ComponentRegistry/#/?registrySpace=published&itemId=clarin.eu:cr1:p_1403526079380&_k=qkn920

[3] Standarts for LRT: https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf

[4] Distribution License Agreement (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/contract

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

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

TECHNOLOGY

15. 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

Accept

Response:

CLARIN-LV uses a repository platform developed by the Institute of Formal and Applied Linguistics, Charles University in Prague, which is used to host the LINDAT/CLARIAH-CZ (Centre for Language Research Infrastructure in the Czech Republic) repository [1] as well as many other repositories hosted by CLARIN ERIC members. The LINDAT/CLARIAH-CZ platform is based on DSpace and is adapted for archiving and distributing language resources. Its open source code is maintained and available via 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].

The repository is installed on a universal OpenStack cloud platform e-Spiets [4], maintained by IMCS UL, which is part of Latvia's academic computing and data storage infrastructure. e-Spiets supports the most popular types of data processing: the classic cloud infrastructure as a service (IaaS), platform as a service (PaaS), Hadoop platform for Big Data, Large RAM - 768GB per process, GPU - 4000 threads, CUDA, SMP - 480 threads, x86, HTC - data flow processing up to 20Gb/s, HPC - parallel computing up to 50 Tflops. The broadband Internet connection of e-Spiets and CLARIN-LV is ensured via the IMCS UL connection to the largest Internet Service Providers of Latvia and via the European Academic Network GÉANT. The broadband connections and the network of IMCS UL is monitored and maintained 24/7 by the Academic Network Laboratory of IMCS UL [5].

In our implementation of the CLARIN-LV repository, the adapted and localised version of the repository is running in a virtual machine in the e-Spiets infrastructure [4]. The submitted datasets are stored in a DSpace repository bitstream store on a network-attached volume. The repository and the metadata of the submitted language resources are stored in a PostgreSQL database instance installed in the virtual machine.

A snapshot of the CLARIN-LV virtual machine is made and backed up after each software configuration change or update to ensure that the most up to date version is available for recovery if necessary. In addition, configuration changes are tested on a beta instance. The data and metadata of the CLARIN-LV repository is backed up on a daily basis by an automatic cron job.

The manual virtual machine image backups, the automatic dataset backups and the automatic database export backups are stored onsite, but in a separate FreeNAS cold storage infrastructure. Additionally, we have implemented automatic offsite backups to the Wasabi Cloud Storage infrastructure [6] (the eu-central-1 data centre in Amsterdam, Netherlands) for extra security in case of a major IMCS UL data center failure.

Additional support regarding network administration and security is available from the Academic Network Laboratory (SigmaNet) at IMCS UL. SigmaNet also maintains a separate industry-level cloud data center which can potentially be used to rapidly deploy the CLARIN-LV virtual machine in case of a major failure of the e-Spiets infrastructure.

The CLARIN-LV infrastructure and procedures are also documented on the repository's About page [3].

References

[1] LINDAT/CLARIAH-CZ repository: https://lindat.mff.cuni.cz/repository/xmlui/

[2] CLARIN Dspace on gitHub: https://github.com/ufal/lindat-dspace

[3] About and Policies (clarin.lv): https://repository.clarin.lv/repository/xmlui/page/about

[4] e-spiets universal could: http://e-spiets.lv/

[5] Sigmanet: https://www.sigmanet.lv/about-us

[6] Wasabi Cloud Storage infrastructure: https://wasabi.com/

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

16. 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
Accept

Response:

The CLARIN-LV infrastructure is hosted at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL).

Since the CLARIN-LV repository infrastructure is maintained as part of the IMCS UL infrastructure, it shares its IT and physical security, and follows its risk management procedures. IMCS UL 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, collaborating with the Latvian National and governmental CSIRT CERT.LV which is one of IMCS UL laboratories. Servers and network devices are kept in a dedicated computer centre with physical access limited to authorised personnel, and the offsite backup server and facilities are located in a different hardware infrastructure. Physical facilities are equipped with fire alarms, uninterrupted power supply and an automatic stand-by electrical power generator to ensure full operations under adverse conditions.

The IMCS UL 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 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.

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.

Reviewer Entry

Reviewer 1

Comments:
accept

Reviewer 2

Comments:

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:
Thank you very much for the additions and changes. My suggestion is to approve the application.

Reviewer 2

Comments:
Thank you for the revision and added information. I recommend accepting this application for certification.