Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	LINDAT/CLARIN
Website:	http://lindat.cz
Certification Date:	28 August 2019

This repository is owned by: **Institute of Formal and Applied Linguistics**

LINDAT/CLARIN

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, Institutional repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

Charles University through its Faculty of Mathematics and Physics and through its Institute of Formal and Applied Linguistics is the coordinator of a national project of large research infrastructure called LINDAT/CLARIN. The infrastructure is a part of the European Research Infrastructure for Language Resources and Technology network (CLARIN ERIC - https://www.clarin.eu/). Czech Republic is a founding member of CLARIN ERIC (and CLARIN preparatory EU grant before that).

LINDAT/CLARIN was established in 2010 with the goal of becoming a national infrastructure for collection and creation of language data; removing obstacles in open access to language data and related technologies.

As such, the digital repository is crucial for the goal of the infrastructure. The LINDAT/CLARIN repository is run by the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic (https://ufal.mff.cuni.cz).
There are other entities participating in the infrastructure (see https://lindat.mff.cuni.cz/en/about-lindat-clarin). Despite none of them being directly involved with the repository, we greatly value their technical expertise (see R6).
The data and software being deposited are outputs of language research projects conducted by LINDAT/CLARIN member institutions, outputs of research in other Czech digital humanities projects, and a substantial part of our collection is also the “Language Resources and Tools Inventory” which we look after for CLARIN ERIC (https://www.clarin.eu/content/language-resource-inventory). This collection is completely open to submissions of language resources from anywhere in the world.

We are a certified CLARIN Centre (http://hdl.handle.net/11372/DOC-99), and we have acquired the first Data Seal of Approval in 2014 (https://assessment.datasealofapproval.org/assessment_92/seal/html)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

The primary consumers would be the international research community: computational linguists (for example machine translation, morphology, syntax, or speech recognition and synthesis), or other humanities' researchers producing and preserving language data or looking for such data or Natural Language Processing (NLP) tools.

More widely we have worked with several other repositories on providing our metadata to them. To name a few: OLAC: Open Language Archives Community (http://www.language-archives.org/archive/lindat.mff.cuni.cz), The CLARIN Virtual Language Observatory (https://vlo.clarin.eu/), OpenAIRE (https://explore.openaire.eu/search/dataprovider?datasourceId=re3data_____::a507cdacc5bbcc08761c92185dee5cab), Web of Science Data Citation Index.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

A. Content distributed as deposited, 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: 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 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 [3,4].

LINDAT/CLARIN 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) can be done also after the item was published. All the changes (including the ones done by editors) are recorded in the provenance metadata [2].

Documentation:
[1] https://github.com/ufal/clarin-dspace/blob/clarin/dspace/config/input-forms.xml
[2] https://lindat.mff.cuni.cz/repository/xmlui/page/item-lifecycle
[3] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[4] https://github.com/ufal/clarin-dspace/wiki/Metadata-info

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

As part of our robust backup solution, we use data storing/backup services offered by CESNET (https://www.cesnet.cz/?lang=en) [1] to keep bit level backups of our systems (data included). CESNET is also running the Czech academic identity federation eduID.cz (https://www.eduid.cz/en/index) which plays a role (together with CLARIN Service Provider Federation and eduGAIN) in our single sign on solution.

Several CLARIN ERIC staff members are editors of submissions in one of the
collections of the repository (LRT Inventory).

We also used EUDAT's (eudat.eu) B2Safe and B2Stage to provide an additional level of the backups. This acted on repository item level, where the AIP (Archival Information Package) was sent to EUDAT. As of the beginning of 2019, this is being reevaluated.

[1] CESNET is an association of universities of the Czech Republic and the Czech Academy of Sciences. It operates and develops the national e-infrastructure for science, research and education which encompasses a computer network, computational grids, data storage and collaborative environment.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

An overview of the repository can be seen at re3data (http://doi.org/10.17616/R30G6W). We are using DSpace (https://duraspace.org/dspace/) as the basis of our repository system, though we have modified it heavily to better suit our needs storing linguistic data and software and integrating with CLARIN network (http://hdl.handle.net/11372/DOC-78). Now this DSpace with our changes is used by 10 other CLARIN centres, and we have been working closely with them on running our repository systems and enhancing the (now) common codebase (https://github.com/ufal/clarin-dspace). We also work with DSpace developers and some of the features we have developed (mostly helping administrators get a better overview of their repository) were merged back upstream (to “original” DSpace).

The repository currently hosts over 1000 digital resources which amount to around 1.5TB of data; the majority of these are language corpora. There are datasets for 300 languages; the top 3 being English, Czech, and German; on the other hand, there are also resources in Kurdish, Amharic, or Burmese.

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. On top of that, real time duplicates automatically available when a failure is detected are kept. Furthermore, off site and on site backups are also kept (see R9).

We have 583 registered users. The number of visitors to our repository has increased every year since the beginning. In 2017, we have come close to 300000 Page Views and recorded almost 140000 (data) file downloads. In 2018, we will exceed these numbers again (https://lindat.mff.cuni.cz/en/statistics).

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 LINDAT/CLARIN is part of) is to advance research in 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.

LINDAT/CLARIN is committed to the long-term care of items deposited in its repository and strives to adopt the current best practice in digital preservation [1], in particular, to continue as a certified (http://hdl.handle.net/11372/DOC-99) CLARIN B Centre (http://hdl.handle.net/11372/DOC-78). A more in depth mission statement can be found in the LINDAT/CLARIN project proposal [2]. Overview of the large infrastructures in the Czech Republic where LINDAT/CLARIN belongs can be found in the “Roadmap of Large Infrastructures for Research, Experimental Development and Innovation of the Czech Republic for the years 2016–2022” [3]. The Roadmap is being updated every 3-4 years (2020-2022 update is under preparation, with LINDAT/CLARIN’s continued presence) by the Ministry of Education, Youth and Sports (MEYS) of the Czech Republic and individual inclusions or research infrastructures on the Roadmap have to be approved by the government of the Czech Republic. The Ministry (MEYS) suggests additions to or deletions from the national Roadmap on the grounds of regular evaluation of all infrastructures by international panels for each research area. The last evaluation, based on which LINDAT/CLARIN continues to be included, took place in 2017, with the next one planned for 2021. The criteria include excellence, development of user base, external publications, internal publications, and alignment with the National policy on priorities in oriented research, valid until 2030 [4].

The Research Infrastructures programme is governed (after government's approval and inclusion of all Research Infrastructures on the National Roadmap) by the Ministry of Education, Youth and Sports, which is also apportioned sufficient budget (as part of the national budget's part assigned to the Ministry). The responsible deputy minister, who is assigned the Research Infrastructures agenda, heads a Council for Research Infrastructures (CRI), of about 30 members, who represent other ministries of the government, the national Research Council, innovation agencies and experts (2 each from each supported area of research, such as Life Sciences, Physics, eInfrastructures, or Social Sciences and Humanities). The CRI supervises the execution of the programme, and also acts as an advisory body to the deputy minister and the department which runs the programme at the Ministry.

Research Infrastructures included on the National Roadmap, i.e. those approved by the government, conclude multi-year contracts between the Ministry and the coordinating institution of each Research Infrastructure, describing the rules and obligations of both the Ministry and the recipient(s).

[1] https://lindat.mff.cuni.cz/repository/xmlui/page/about?locale-attribute=en
[2] https://lindat.mff.cuni.cz/bits/documents/LINDAT-CLARIN-2010-2015-Attachment1-EN.pdf
[3] http://www.msmt.cz/file/37456_1_1/
[4] https://www.vyzkum.cz/FrontClanek.aspx?idsekce=653383&ad=1&attid=669651

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 Terms of service (https://lindat.mff.cuni.cz/repository/xmlui/page/about#terms-of-service), this binds them to comply with licenses attached to repository items.

The license attached to a repository item is displayed prominently on the item page (see for example http://hdl.handle.net/11858/00-097C-0000-0001-4914-D) together with colour coded “openness” of the license (“public”, “academic”, “restrictive”).
The license attached to a repository item is chosen during submission by the person submitting it. We provide guidance to select the appropriate license using a graphical license selector tool that we developed [1]. Open/public licenses are strongly preferred when possible. However, we also offer options to put more requirements on the consumer e.g., require that the consumer has 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 signed licenses by each consumer.
In case a suitable license cannot be found in already existing licenses [2], 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 Charles University which is the legal entity behind the repository), the so-called “Deposition License Agreement” (https://lindat.mff.cuni.cz/repository/xmlui/page/contract?locale-attribute=en), where we describe our rights and duties 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 to put an embargo on submissions, which means that the submissions will be archived immediately after
completion of the curation workflow, but they will become publicly available 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 the user 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 the last resort, we would take legal action.

[1] https://github.com/ufal/public-license-selector
[2] https://lindat.mff.cuni.cz/repository/xmlui/page/licenses

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:

Since the establishment of the CLARIN national centre LINDAT/CLARIN (2010) it has been financed from the Czech programme of "Large Research Infrastructures", which is specifically dedicated for infrastructures that provide long-term access to data and services. Currently running period of LINDAT/CLARIN grant ends by the end of 2019 and continuation has already been approved both by evaluators and Ministry of Education until 2022. Funding of this new proposal is sufficient to maintain the repository system and keep up its development and improvements, as well as data security at least at the current level.

At the same time, LINDAT/CLARIN took measures to preserve data access even in case of some unexpected catastrophic situation, like some emergency budget cuts we cannot anticipate now. We developed our repository solution CLARIN DSpace (formerly LINDAT DSpace) as a very low maintenance system, easy to install and keep running. Thus even if LINDAT/CLARIN had to end, the repository hosting department, Institute of Formal and Applied Linguistics, would be able to keep running the repository with very low budget requirements for a substantial time. The agreed upon timeframe is 10 years at a minimum, if no other CLARIN centre is willing to take the data.

CLARIN DSpace[1] was developed as an open source software under permissive MIT license and we supported other CLARIN centres in deploying this solution in part also to ensure the sustainability of access. At this point, there are at least 7 CLARIN centres[2] running this same system. Although there is no formal agreement with other CLARIN centres, this allows for the possibility of simple migration of all the data from one repository to another and keeping the records accessible under the same PIDs and with the exact same feature set.

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

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 the 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. Acknowledgement that submitter has the

right to distribute the data in the first place includes resolving all the privacy issues including GDPR, 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). If they doubt the compliance 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.

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. So far we have not received many submissions containing confidential data or data with disclosure risk and we do not expect this to change in the future. Most of our data is Open Access or distributed under similar public licenses. Substantially less data is available under custom licenses, which are however still public or rather permissive (e.g. academic restriction). Very few records have stricter requirements, but the repository system and editorial staff can handle them.

The CLARIN Legal and Ethical Issues Committee (which LINDAT/CLARIN is a member of) organises training sessions in the management of data with a disclosure risk. Some of LINDAT/CLARIN staff also have substantial experience managing confidential/disclosure risk data from managing holocaust survival data in Malach Centre for Visual History (https://ufal.mff.cuni.cz/malach/en).

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 LINDAT/CLARIN repository is hosted at the Institute of Formal and Applied Linguistics (UFAL), Charles University in Prague. Charles University is a stable institution in the long term, since 1348. UFAL is a large department (about 70 persons) doing research in Language Technology and participating in many research grants nationally, but also in EU (H2020) and USA (NCF, Mellon). UFAL is the hosting institution, but LINDAT/CLARIN is a national CLARIN centre and a project with its own management structure: prof. Jan Hajic is the coordinator of the LINDAT/CLARIN Large Research Infrastructure (a programme of the Ministry of Education and Youth). This means funding is ensured on a national level. Large research infrastructures are long-term projects directly confirmed by a decree of the cabinet.

LINDAT/CLARIN has sufficient funding and staff resources to operate in the long-term. The staff of the infrastructure is about 70 individuals working for the project in the amount of 13 FTE. 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: this is 3 persons coordinated by the technical director. Their work is fully dedicated to the development and management of the repository and related services.

The project LINDAT/CLARIN is not only about running the repository, as an infrastructural scientific project, but it also has an appropriate budget to be able to attend all meetings as necessary. Current funding makes sure of this at least until the end of 2022 and we expect it not to change after that.

LINDAT/CLARIN staff regularly participate in international CLARIN committees and task forces, as well as CLARIN developers' Slack channels for daily communication, where crucial knowledge is continuously shared, There are also special training and professional development activities organised and supported by CLARIN and LINDAT staff attends them, sometimes in learning, and sometimes in teaching capacity. Most staff members are also in other international projects in H2020, COST, Marie-Curie, and more, where important expertise is shared too.

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

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:

LINDAT/CLARIN partners [1] offer a wide range of experts that already have been or can be used as consultants. There are several events like conferences or workshops where the members meet every year and share their knowledge. LINDAT/CLARIN is a dedicated CLARIN knowledge centre (https://www.clarin.eu/node/4210).

LINDAT/CLARIN staff, there's an overlap with the hosting institute IT team, frequently attends CESNET [2] workshops and conferences. Which offers the possibility to consult advances in data storage, networking, etc. with various experts participating in the national research and education network (see R0 and R15 for the details about the technologies used). We also regularly participate at the Open Repositories conference and some of the RDA (Research Data Alliance) workshops. Our repository is based on DSpace; therefore, we are often in touch with DSpace developers. As a CLARIN member, we also attend CLARIN workshops and conferences, so we are up to date on what other centres are up to and what are their user requirements.

We usually receive feedback through our mailing lists [3] or as issues on the repository Github tracker [4]. The Github tracker includes issues from all installations of our (CLARIN DSpace) repository system. Currently, there are more than 10 installations in diverse institutions [5], which gives us considerable feedback.

On an international level, LINDAT/CLARIN is represented in all important CLARIN committees and task force initiatives. They consist of other CLARIN members and the main focus is on knowledge sharing and creating guidelines for the whole CLARIN.

All digital 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.

LINDAT/CLARIN also has an international advisory board that includes experts from both within and without the user community: from heritage institutions to commercial research giants in our field. As such, the advisory board is able to provide both feedback as for the desired functionality, and a realistic look at the sustainability of development of the repository to meet these requirements.

[1]: see R0
[2]: see R0 outsource partners
[3]: lindat-help@ufal.mff.cuni.cz, clarin-list@ufal.mff.cuni.cz
[4]: https://github.com/ufal/clarin-dspace/issues
[5]: This includes not only CLARIN centres, but also humanities research institutions: https://github.com/ufal/clarin-dspace#clarin-dspace-deployments

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, 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.

We do not support changing the data. A change or a new version of a dataset must be created as a new repository item (https://github.com/ufal/clarin-dspace/wiki/New-Version-Guide). 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 have the relation added to their metadata and are visually represented on the web page (see http://hdl.handle.net/11234/1-2837).

Changes to the metadata occasionally happen (mostly typo fixes), 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. After the item was approved only the administrators are able to change its data. The data producers can refer to the Deposited Item Lifecycle

mentioned above to get acquainted with the details or ask directly our helpdesk.

[1]: https://lindat.mff.cuni.cz/repository/xmlui/page/item-lifecycle?locale-attribute=en

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:

LINDAT/CLARIN 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 repository is structured in two main collections: one collection represents the data and tools from the LINDAT/CLARIN consortium, and the other collection represents CLARIN LRT Inventory, where all CLARIN members, as well as users from outside CLARIN, can submit their data [5].

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 that will check for the quality of the metadata. A thorough check of the quality of the data is not performed since it is beyond our mission and scope, but when editors understand the data (NLP field has big 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-CD 24622-1); in particular, the submission interface is based on one 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 LINDAT/CLARIN can't dictate researchers how they do their research and 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, when the researchers insist 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. 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://lindat.mff.cuni.cz/repository/xmlui/page/about
[4]: https://lindat.mff.cuni.cz/repository/xmlui/page/metadata
[5]: either with an account at IdP in EDUGAIN (usually a university) or after successfully applying for a CLARIN IdP account (which requires an individual proof of an academic status).

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:

In general, LINDAT/CLARIN's infrastructure together with external partners provide highly available storage, backup and disaster recovery for archival data and software. Backups are regularly done on-site but also off-site to one of our partners - CESNET.

With the use of the DSpace (one of the leading digital repository systems) as the underlying software for CLARIN DSpace [1] developed mostly by LINDAT/CLARIN, the repository meets the requirements of OAIS. For the first step, 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].

For the second step, the archival storage, one of our editors 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.

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 signed during the ingestion process. We are using a specific robust administration interface including specific detailed reports on the contents of our repository.
All backups follow standardised ways of using MD5 checksums for determining the consistency and we use automatic monitoring tools at various levels.

The infrastructure and backups are further described in the section Technical infrastructure and section Security.

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

[2] https://wiki.duraspace.org/display/DSDOC18/Importing+and+Exporting+Content+via+Packages#ImportingandExportingContentviaPackages-SupportedPackageFormats

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:

LINDAT/CLARIN has the right to copy, transform, store and provide access to the data [1]. Redundant backups, multiple drives, and off-site storage of all data assure long-term preservation [2]. This preservation function encompasses: taking delivery of the dataset ingested, storing it, and ensuring it is archived, and accessible and usable to the researcher community as is the mission of a CLARIN Centre [3].

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 [4];
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 the file format selection. Refer to [5] 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 datasets [6] which is crucial for long term preservation. In the administration phase, except 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 conferences related to the subjects (see R6 for more details).

The repository encourages the usage of specific file formats as recommended by CLARIN [4]. The number of accepted file formats is small and well documented to make future conversions to other formats more feasible. As much as possible, open (non-proprietary) file formats are used. For textual resources (that account for a significant number of our resources), XML formats are used whenever possible or other well documented formats, to ensure future interpretation of the files even if the tool that was used to create them no longer exists. Text is encoded in Unicode to ensure future interpretability. In cases where proprietary/"custom" formats need to be used, we require detailed and exhaustive documentation, in order to make the implementation of future data converters possible.

The preferred file formats will change over time, in which case 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 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 where possible, in the case of digitization of analogue signal lossless or no compression is recommended.

All metadata and data have a persistent identifier (PID) and metadata can be converted to self explanatory and human readable XML files.

[1] https://lindat.mff.cuni.cz/repository/xmlui/page/contract

[2] https://lindat.mff.cuni.cz/repository/xmlui/page/about

[3] https://www.clarin.eu/sites/default/files/centres-CLARIN-ShortGuide.pdf

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

[5] https://lindat.mff.cuni.cz/repository/xmlui/page/deposit

[6] https://lindat.mff.cuni.cz/repository/xmlui/page/item-lifecycle

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:

We have carefully crafted the submission process in such a way that we get enough information about the resource but do not overload the submitter with pages of forms.
During the submission hints, examples and suggestions are provided to get the highest quality metadata. We provide a

page (https://lindat.mff.cuni.cz/repository/xmlui/page/metadata?locale-attribute=en) summarising the information (metadata) we gather about resources and various metadata formats we can disseminate to (e.g., Dublin Core, OLAC, specific CMDI profile, METS).
The sufficient completeness and quality of metadata is assured by requiring certain fields in the submission process (without them filled in the submission cannot be completed), by filling in certain fields automatically (PID is assigned automatically, dates of entry into the repository, etc.), by automated curation and final approval by editors. If the editors are not satisfied with the metadata, they have the option to correct them on their own or to return the submission back to the producer requiring they 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 other works [1].
The underlying software was developed at LINDAT/CLARIN and is publicly available [2].

Furthermore, as we are harvested by other organisations (CLARIN VLO, OLAC harvester, Data Citation Index) we are incorporating their feedback on potential metadata issues. Occasionally we also get feedback from the end users regarding the metadata on the feedback/hotline email.

[1] https://lindat.mff.cuni.cz/repository/xmlui/page/cite
[2] https://github.com/ufal/lindat-common

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 and 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.
As mentioned, the application provides also a REST API which is aimed towards machines interacting with repositories. 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://lindat.mff.cuni.cz/repository/xmlui/page/deposit

[2] https://lindat.mff.cuni.cz/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 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 (https://lindat.mff.cuni.cz/repository/xmlui/discover?advance).
The repository provides an OAI-PMH endpoint and we are harvested by several organisations (CLARIN VLO, OLAC, OpenAIRE, WOS) and REST API.
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) and we have also written a guide for our users on how to cite the repository items properly (https://lindat.mff.cuni.cz/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:

LINDAT/CLARIN 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 all these 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 either upload files in 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 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 supported/known formats are also supported by the underlying CLARIN-DSpace software [5].

[1] https://lindat.mff.cuni.cz/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://lindat.mff.cuni.cz/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:

The repository has 5 servers at its disposal 3 of which are in a so called primary server room the remaining 2 are in a so called backup server room.
The server rooms are connected primarily through a 10Gbps link and another 1Gbps link serves as a backup connection.
The primary server room is backed by a UPS together with a diesel aggregate, the other server room is so far backed just by a UPS. The servers themselves have redundant power supplies and are equipped with ECC memories.
Also a 10Gbps link is provided to the outside world.

The cluster is running on Proxmox 5.2 (https://www.proxmox.com/en/proxmox-ve) platform, which is built on top of Debian and other standard open-source components.
ZFS with lz4 compression is used for all the filesystems in the cluster. The primary servers have an SSD accelerated cache.

The proxmox platform provides us with the following functions (among others):
monitoring and history of load of all the cluster nodes and individual virtual systems
LXC and QEMU replication of the virtual systems using ZFS delta-snapshots
snapshot backups for both VMs and containers
High Availability (HA) regime for a chosen virtual system
fast migration of virtual systems (even live migration in some cases) between nodes

The repository system itself is based on DSpace. We've tailored it to our needs as a data repository and shared (https://github.com/ufal/clarin-dspace) this modified version with the CLARIN community. DSpace itself is based in the OAIS reference model.
We follow a list of standards that are relevant for the CLARIN community (https://www.clarin.eu/content/standards-and-formats).

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 infrastructure described in the previous section (see R15) is designed with sufficient redundancy in mind; thus outages caused by hardware failures should be rather rare.

In addition to the monitoring offered by proxmox, we are using Munin and Icinga2 to monitor the real-time performance/status of our services. We are alerted in case of issues.

We are running weekly integrity checks (see section Data integrity and authenticity) to guarantee fixity. Disaster recovery of data is implemented via a multi level backup scheme:
- first level is replication of the virtual systems between cluster nodes used for the HA regime
- second level are weekly dumps of the virtual systems to a shared NFS volume
- third level is a weekly off-site differential backup on an external hierarchical storage.

We store only a minimal amount of information about users (importantly, no passwords are stored) as we are using single sign on. The user details are stored within their home organizations (identity providers). This is described in more detail in our privacy policy (https://lindat.mff.cuni.cz/privacypolicy.html).
As a part of CLARIN Authentication and Authorization Infrastructure, if there is a security incident we will report it using SIRTFI - REFEDS (https://refeds.org/sirtfi)

The physical access to the university building, where the production servers are, is limited to holders of an RFID chip/card or through a reception desk. The access to server rooms is further limited to the IT staff.
The IT of the hosting institution are the only people with access to the virtualization platform (ie. to the hypervisor/host). There are 5 people with administrator privileges in the repository application, the management and the core technical team (see R5). These privileges will be revoked if a contract expires.

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:

We have removed the confusing technical details and extended the description of our preservation plan.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: