
# Implementation of the CoreTrustSeal

The CoreTrustSeal board hereby confirms that the Trusted Digital repository CELR META-SHARE complies with the guidelines version 2017-2019 set by the CoreTrustSeal Board.
The afore-mentioned repository has therefore acquired the CoreTrustSeal of 2016 on November 2, 2018.

The Trusted Digital repository is allowed to place an image of the CoreTrustSeal logo corresponding to the guidelines version date on their website. This image must link to this file which is hosted on the CoreTrustSeal website.

Yours sincerely,

The CoreTrustSeal Board

# Assessment Information

| | |
|---|---|
| Guidelines Version: | 2017-2019 \| November 10, 2016 |
| Guidelines Information Booklet: | DSA-booklet_2017-2019.pdf |
| All Guidelines Documentation: | Documentation |
| Repository: | CELR META-SHARE |
| Seal Acquiry Date: | Nov. 02, 2018 |
| For the latest version of the awarded DSA for this repository please visit our website: | http://assessment.coretrustseal.org/seals/ |
| Previously Acquired Seals: | None |
| This repository is owned by: | • **Center of Estonian Language Resources**<br><br>T +372737 6433<br>E krista.liin@ut.ee<br>W http://www.keeleressursid.ee/ |

# Assessment

## 0. Context

### Applicant Entry

*Self-assessment statement:*

Repository Type:

Subject-based repository for digital resources of Estonian language;

National repository system for linguistic data of digital humanities;

Research project repository for results of projects of National Programme for Estonian Language Technology (NPELT)

Brief Description of the Repository's Designated Community:

The main users of CELR are researchers from Estonian R&D institutions and digital humanities researchers all over the world via the CLARIN ERIC network of similar centers in Europe.

Level of Curation Performed:

A. Content distributed as deposited

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

Outsource Partners:

For technical matters, storage and backup of CELR data is maintained by the High Performance Computing Centre (HPC) at University of Tartu. HPC has declared Security Class and Backup Concepts according to Estonia's IT Baseline Security System (ISKE).

CELR belongs to list of data registries in DataCite Estonia. It enables to (automatically) add DataCite DOI to all language resources registered in CELR registry.

Both partners above belong to same juridical body as CELR - University of Tartu.

There is a Service Level Agreement for Storage and Backup (SLA) between HPC and CELR. Summary in English is added to S9 (Documented Storage Procedures)

Other Relevant Information.

• The usage and impact of the repository data holdings (citations, use by other projects, etc.).

CELR is responsible for archiving and enabling access to the outcomes of the projects of the National Programme for Estonian Language Technology (NPELT)

• A national, regional, or global role that the repository serves.

CELR is SSH infrastructure with national importance belonging to Estonian Research Infrastructures Roadmap

• Any global cluster or network organization that the repository belongs to.

CELR is a consortium representing Estonia as member in the CLARIN ERIC.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 1. Mission/Scope

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

One of the main aims of the Center of Estonian Language Resources (CELR) is to provide access to Estonian language digital resources and to preserve those resources. This is mentioned in the Consortium agreement (in Estonian) and also stated on the consortium webpage: https://keeleressursid.ee/en/ .

Summary of Consortium Agreement is translated into English.

CELR is also responsible for archiving and enabling access to the outcomes of the projects of the National Programme for Estonian Language Technology (NPELT) as stated in text of the programme.

As a representing entity of Estonia in CLARIN ERIC (see art-s 2.1-2.2 and 6.2-6.5 in CLARIN ERIC Statutes), CELR enters CLARIN Agreement to lay down the conditions and specifications under which data and service centers, access to data and user authentication and authorization system are provided.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept

Reviewer 2: Accept

## 2. Licenses

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

CELR's policy of offering data and services is based on assumptions declared in CLARIN ERIC Statutes to ensure a licensing, access and authentication framework that ensures easy access and at the same time protects the reasonable rights of owners of data and tools, and privacy of individuals.

Generally, Open Access, Open Source and Open Data principles are favoured and promoted, but existing licences are respected.

In CELR language resources metadata registry META-SHARE (https://metashare.ut.ee/ ) language resources and tools are made publicly visible by means of core metadata descriptions, including information about depositing and user licences of particular dataset.

For language resources as linguistic data the Creative Commons (CC) types of licences are encouraged, and for tools and software General Public Licences (GPL) and BSD types of licences.

In general CELR uses CLARIN framework of agreements for depositing and using linguistic data and resources. CLARIN licences are available for curating a minimal set of usage conditions to include a resource in the CLARIN PUB, ACA or RES categories.

For data depositors and users workshops and personal assistance are provided, for depositors is available CLARIN license calculator to find most relevant type of licence or usage conditions.

In terms of using language resources CELR relies on copyright research exception in its activities. According to our interpretation, copyright regulation in Estonia allows us to use copyrighted material for our research (§ 19. Free use of works for scientific, educational, informational and judicial purposes). Although our research is based on the approach that we do not need any permissions of IP owners to do research we have cooperation with them.

In terms of privacy policy of user information GÉANT Data Protection Code of Conduct is followed.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept

Reviewer 2: Accept, but non-compliance with the licenses could be more explicit

## 3. Continuity of access

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

The level of responsibility undertaken for data holdings is “custodian”. We promise to provide access, preservation, and/or data storage to quality level what data holder/owner makes available.

Overall long-term preservation policy for research data in Estonia is in progress. Obligations imposed by the Estonian government on the long-term preservation of results of the NPELT program (since 2006) and obligations to represent Estonia in CLARIN ERIC (see S0. Context and S1. Mission/Scope) however, provide certainty in financing for 2018-2027.

CELR belongs to DataCite Estonia consortium and the consortium assures that the open research output of the universities is findable, accessible and reusable via DataCite. University of Tartu Library is the allocator for data centres and repositories for minting DOI.

Metadata about datasets and collections harvested from our META-SHARE registry by CLARIN VLO and META-SHARE top node will remain in mirror servers too.

Negotiations with National Archive (NA) are held for migration and preservation of long-term records. NA is the competence centre for digital archiving in Estonia and ensures archiving (i.e long-term preservation) at first but not accessibility of the data in case of rapid changes of circumstance or cessation of funding.

One of CELR consortium partners, Estonian Literary Museum is a state institution of research and development and functioning by archival regulations at the same time. Institution has long expertise in archiving and long term preservation and in case of an unlikely major business continuity failure of CELR it can take over the responsibility of CELR repository.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept for compliance level 3 (implementation)

Reviewer 2: Accept for compliance level 3 (implementation)

## 4. Confidentiality/Ethics

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

As CELR is a repository providing only basic curation for data, the full responsibility of managing disclosure risks in data collection or creation is data provider's. We have competence to provide guidance in the responsible collection or use of disclosive, or potentially disclosive data, especially suggestions to choose relevant usage conditions with the help of CLARIN license calculator.

CELR staff is trained in recognising and handling not only data with disclosure risk, but also seeking advice from authorities in Estonian Data Protection Inspectorate who consult in such cases.

CELR staff understands very well Estonian legal environment and the relevant ethical practices in digital humanities in Estonia. If the data consists personal data for example, staff asks for relevant consents form data holder. If data needs anonymisation, our staff is able to consult data holder too.

Access to data with disclosure risks is restricted by technical measures and right to share access to data is given to data holder. For all kind of data with restrictions and risks CELR metadata registry META-SHARE provides limited access and asks to contact with data holder directly. If access is granted, data holder himself, or metadata curator, can enable access to restricted data.

In terms of privacy policy of user information GÉANT Data Protection Code of Conduct is followed.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept

Reviewer 2: Accept

## 5. Organizational infrastructure

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

CELR functions as consortium of 4 R&D institutions - University of Tartu, Tallinn University of Technology, Institute of Estonian Language and Estonian Literary Museum. Each consortium member has specific competence and profile of expertise.

CELR as consortium belongs to Estonian Research Infrastructures Roadmap.

The roadmap contains a list of nationally important research infrastructure units and is an input to the investment decisions regarding research infrastructures. In addition, it is the task of the Estonian Research Council to coordinate the implementation of the roadmap objects.

Support for Infrastructures of National Importance is involved in the investment plan for „Research Infrastructures of National Importance“ based on the national research infrastructures roadmap. This activity is funded by the EU structural funds 2014-2020 in total 716 000 euros for CELR.
State financial support measures for NPELT (National Programme for Estonian Language Technology) and Core Infrastructures are relevant for funding CELR. Support for basic activities includes 100 000 euros for 2017 and is planned minimum 100,000 euros per year for 2018 – 2020.

CELR have 2 full-time personnel only in UT, including executive manager and technical staff for servers and web-services, meta-data management and user help-desk. Some part-time personnel work for user involvement and PR, legal issues, specific data and user interfaces, at other consortium members also. All employees are educated in the fields of (computational) linguistics, natural language processing, software development.

The staff have constant opportunities to improve their expertise by participating in relevant courses and workshops organised by consortium members. Participation in seminars and conferences is encouraged. Cooperation with CLARIN series of workshops also offers opportunities for the CELR staff to improve their professional skills.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept

Reviewer 2: Accept

## 6. Expert guidance

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

CELR staff include people who have expertise in the area of language resources and technology, our consortium includes the the four major research institutions dealing with language technology and resources in Estonia: University of Tartu, Tallinn Technical University, Institute of Estonian Language and Estonian Literary Museum. As such we collaborate closely with researchers working in those institutes and elsewhere in the community. For technical matters, storage and backup we also have an agreement with the High Performance Computing Centre in University of Tartu.
As CELR is a member of the European research infrastructure CLARIN, we also rely on the technologies and principles that have been established in CLARIN based on the expertise and user feedback from several European countries.

We regularly attend the conferences and events of the language resources and technology community in Estonia, there are at least two larger conferences per year. As we are related to the National Programme for Estonian Language Technology we meet the researchers participating in programme in order to include the results of their work and if necessary, can refer to its steering committee for advice. CELR organizes workshops to community members, provides support regarding language resource management to participants of the programme (and others). Users can also leave their feedback on our webpage or by e-mail.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept

Reviewer 2: Accept

## 7. Data integrity and authenticity

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

Data integrity

- In order to check that a digital object has not been altered or corrupted the CELR has saved md5hash of resource among metadata of that resource. It is possible to compare saved hash with the calculated hash in order to verify integrity of the resource.

- In the process of depositing a resource, the repository automatically checks the provided metadata is checked automatically for completeness (only certain formats are allowed - according to the CMDIprofiles agreed on in the CLARIN infrastructure). The associated data is checked manually before the resource is published.

- Provenance data and audit trails - automatically logged by Meta-Share (who changed what and when). Every change is saved as a new version of metadata, with information about who and when made the change. Changes to the resource data are not allowed, every change causes a creation of the new version of that resource.

- All resources and their metadata have version numbers, assigned according to principles of Semantic Versioning.

Data authenticity

- CELR has stated in depositors agreement that all deposited resources will be archived as they are accepted from depositors. Following derivated AIP's may be made as separate packages and corresponding metadata includes reference to original package.

- For each version of resource exists a metadata record where are links to the metadata record of the previous version of that resource. If resource is derivative work from some other resource that is existing in the local or remote repository then it is possible to refer to that resource also.

- CELR has seen no need for comparing the essential properties of different versions of the same file.

- Checking identities of depositors - only authorized users who have been given editor privileges can deposit resources. Authentication relies on our national identity provider federation; as an exception a separate account can be created in the repository for checked and trusted users.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept for level 3 (implementation)

Reviewer 2: Accept for level 3 (implementation)

## 8. Appraisal

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

CELR accepts Estonian language resources: texts, corpora, audio and video recordings, lexical data, terminologies, tools for NLP as declared on https://www.clarin.eu/content/depositing-services or https://keeleressursid.ee/en/services. We are also the designated repository for all resources that a produced as part of the Programme for Estonian Language Technology. The results of the programme have been checked by its steering committee. For both these and other data our staff will consult the data owner and consult them on what consists of a (separate) language resource and how to describe it in our repository. Existing users who have previous experience can initiate the depositing process themselves, but the data and its metadata are checked by our staff before publishing inside the metashare repository system. If there are any problems, the depositor is asked to correct them, until then the resource will stay unpublished. So far the quality check is manual, we are planning to automate parts of the process. If the data has been formally validated, it is possible to add reference to the validation procedure and results in the relevant parts of the metadata.

We refer to a list of CLARIN preferred formats for data (https://metashare.ut.ee/site_media/ms-provider-en.pdf, a newer version is being compiled), but since different types of research questions or linguistic annotation can require the choice of a different one, other formats are accepted in our repository if their use is justified. In that case we require the user to refer to the description about the data format in the description (metadata) of the resource. The documentation should be detailed enough to provide for (even lossy) format conversion, if necessary, and include conversion tools, if these are available. However the repository does not do the format conversion for the users. So far we have accepted only text resources in formats that are not in the list of preferred formats, but are still used by some tools - meaning that there is reason to believe that users would be interested in this format.

Metadata is described in the repository system, in a built-in metadata editor (it is also possible to upload XML files there, but those will also go through the editor). The MD editor checks automatically that all required fields are filled in and that metadata is in the correct format. The use CMDI [https://www.clarin.eu/cmdi] format from CLARIN, more specifically the CMDI profiles for corpora, lexical/conceptual resources, tools/services and language descriptions. These profiles are based on the META-SHARE metadata model that has been decided on in collaboration with the META-NET network of excellence. Since all published metadata is harvested by CLARIN, it also goes through an XML validator before being entered into their metadata browser Virtual Language Observatory.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept for level 3 (implementation)

Reviewer 2: Accept for level 3 (implementation)

## 9. Documented storage procedures

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

CELR is using the Meta-Share system [https://metashare.ut.ee/] for metadata management and the Entu Repository Software [http://www.entu.ee] for archiving the files.

Backup and storage of the repository as well as all other important applications and IT-systems in CELR is carried out according to **Storage and Backup Procedures** (in Estonian) [https://entu.keeleressursid.ee/api2/file-14954]

This document declares there is a Data Backup Plan for each system.

For the high priority systems and applications of CELR (including repository) a backup of their data is made at least once a month. The list of those systems is given as Appendix of the Storage and Backup Procedures document.

For the high priority applications a full backup is made at least once a month and at least three latest working copies are kept as backup at any time. The data in file servers is to be backed up regularly.

**Recovery plan**

Working order of the backup system is checked periodically. At least twice a year the procedure of restoring backup is checked. If there is no need to restore real data, a restore exercise will be carried on.

**Long-term Storage**

For long-term storage there are regular backups of the data processed in CELR servers saved to two or more backup servers that are located in server rooms in different physical locations.

Service Level Agreement [https://entu.keeleressursid.ee/public-document/entity-7133] with HPC declares more in details (according to Storage and Backup Procedures):

Regular backups are made daily, outside working hours, without stopping the application.

The type of backup used by HPC is incremental.

Three latest working copies are kept as backup.
The backup of deleted files is kept for one year.

As the documentation of relevant processes we are using CELR workflow description what is basically adapted description from OAIS 4.1.1.3 Archival Storage in Estonian [https://entu.keeleressursid.ee/api2/file-14955].
Backups are made by system administrator.

•

Security level for data storages of CELR is K1T1S2 by ISKE, where

K1 – reliability – 90% (acceptable total interruption per week ~ 24 hours); acceptable increase in the required response time during peak – hours (1÷10)

- 

T1 – information source, the fact of its amendments or termination shall be detectable; controlling the correctness, integrity and being up to date in special cases and according to need;

- 

S2 – classified information: information can only be used by certain user groups; access to the information shall be granted if the person requesting access has legitimate interest.

- 

- 

[https://www.ria.ee/public/ISKE/Regulation-the-system-of-security-measures-for-information-systems-2007-12-20.pdf].

CELR has three backup storages and they are placed in different physical locations. In future we are planning to store one or two backups in other countries.

- 

In the case of main storage failures we have data recovery plan as a part of Storage and Backup Procedures [https://entu.keeleressursid.ee/api2/file-14954].

- 

In order to ensure consistency across archival copies the file hashes are calculated regularly and compared to initial hashes in metadata registry. This procedure is part of the storage media monitoring process in Storage and Backup Procedures.
The storage system is monitored with Spectrum Archive tools and an independent surveillance system. The storage system failures are responded instantly, there is a standby server ready for this.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept for level 3 (implementation)

Reviewer 2: A summary in English of the Storage and Backup Procedures is provided and satisfactory.

## 10. Preservation plan

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

Self-assessment statement: One of the main tasks of the Center of Estonian Language Resources is to preserve materials for long term. This concerns both data itself and metadata.

A META-SHARE node serves as our resource registry: it bears mostly metadata about resources, and some data itself. This system is set up on one of our virtual servers (https://metashare.ut.ee) that is backed up on regular basis. The repository is harvested by CLARIN VLO and META-SHARE top node, that would help to restore metadata on an unlikely event if the repository server and last backup would turn out to be corrupted.

The data itself can be found on Entu system or META-SHARE node itself. Entu system is also backed up on regular basis, see R9, (Documented Storage Procedures).

We take into account physical, technical, organisational and human behaviour aspects of long-term preservation.

As for physical and technical aspects, the hardware running CELR services and all backup facilities are situated in the rooms of HPCC (https://hpc.ut.ee/). They have all necessary infrastructure and trained stuff to manage highly virtualized services and backup-based archiving (see R9 and HPCC Backup Concepts ).

The Depository Agreement is based on Clarin Depository Agreement and META-SHARE Depository Agreement, and the Repository (i.e. CELR) has not only right, but also the obligation to make all necessary copies needed for long-term preservation.

As a rule, CELR will not do any modifications in the deposed resource, this is the responsibility of Deposer. If a modified version of the Resource is uploaded, then older version is preserved too. Different versions get different version labels, and all of them are considered as items to be long-term preserved. If a format conversion is needed,

CELR can do it, but this would be a special arrangement outside Depository Agreement.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Still not a very clear statement on the rights of the repository eith respect to the data it holds. Accept for level 3 (implementation) though.

Reviewer 2: For level 3 (implementarion) OK

## 11. Data quality

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

The data deposited in our repository undergoes a manual check by the CELR staff before being published. The depositor can also refer to other quality check results in the relevant section of the metadata. See R8 for more details.

•

Metadata is described in a built-in metadata editor of our META-SHARE repository that checks automatically that all required fields are properly filled in and

•

We use META-SHARE metadata model, the metadata is automatically converted to CMDI format, validated against its schema. These formats are the recommended ones in two European infrastructures: META-NET and CLARIN. See R8 for more details.

- 

There is no feature to comment on the data or metadata directly in the repository, but in case of problems or questions users can give feedback on our webpage or contact CELR otherwise.

- 

- The metadata profiles in use include sections for relevant documentation, quality assessment references or related projects or publications. An example of reference to project can be seen here, to publication here. All resources have a DOI attached. There are also plans to include a popup citation box next to the DOI to make it easier to cite the dataset, similar to the ones in use in Datacite.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept for level 3 (implementation)

Reviewer 2: OK for level 3 (implementation)

## 12. Workflows

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

CELR is currently in the process of following the guidelines of the OAIS reference model and implementing a most essential workflows to collect, store and distribute the various informational objects (see https://entu.keeleressursid.ee/api2/file-14955). The main workflows in the focus are:

Deposition of information packages

An Submission Agreement is negotiated between CELR and a Data Producer.

- 

The submitted data - Deposited Information Packages (DIP) - are validated by CELR. Resources must be prepared in one of the formats that the CELR accepts. The relevant metadata and the documentation must be included. Requirements for DIP are described here.

- 

Archival Information Packages are generated, the persistent identifiers are assigned to the selected objects.

- 

- 

Archiving

Archival Information Packages are stored in the Entu repository. Entu is a software platform for data management in a object-oriented manner.

- 

Regular backups of the repository are made.

- 

- 

  Access

  A metadata registry is made publicly available for searching and browsing on web. It’s content is shared with the rest of the CLARIN community, by means of metadata harvesting (http://www.clarin.eu/faq-page/275).

- 

  Access to some corpora data is available through KORP software.

- 

  A need for authentication of a Data Consumer is deduced from a licensing policy of the Information Package.

- 

- 

The workflows for Data Management, Administration and Preservation Planning are waiting for the next implementation stage.

CELR staff members are available for user assistance and repository maintenance. The team has expertise in the areas of validation of data, data processing, repository management, using of involved standards, and metadata formats. The employees are educated in the fields of natural language processing, linguistics, software development.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1:accept for level 3 (implementation), although evidence only in Estonian

Reviewer 2: accept for level 3 (implementation)

## 13. Data discovery and identification

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

META-SHARE software developed for META-NET Network of Excellence is used for resource repository at CELR. The address of the repository is https://metashare.ut.ee/

Both free-text search and faceted search are used to find language resources. Users are able to type a search term in the search field or they can search by various criteria.

Metadata produced at CELR is available at CLARIN via Virtual Language Observatory (https://vlo.clarin.eu), META-SHARE top node (http://metashare.elda.org/) and E-varamu (https://www.e-varamu.ee/), the central Estonian portal for culture and science.

Metadata for META-SHARE is provided by META-SHARE software export layer, for CLARIN and E-varamu via OAI-PMH module by CMDI profiles created at CLARIN. Both methods are machine-harvestable.

Url for harvesting: https://metashare.ut.ee/oai_pmh/

The repository offers PIDs for metadata and DOIs for resources, data is citeable via DataCite (https://www.datacite.org/). All PIDs and DOIs are aquired automatically in metadata publishing process.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept

Reviewer 2: Accept

## 14. Data reuse

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

We use the META-SHARE metadata model and its corresponding CMDI profiles: corpora, lexical/conceptual resources, tools/services and language descriptions. The minimally required fields depend on the profile, for all of these it should include the necessary information (and relevant mapping) for Dublin Core. The optional fields are enough to also describe the content and format of the data, with references to relevant files, if necessary. The metadata is checked before the resources are published, and amended if something is missing.

The metadata will be converted to newer versions of CMDI (or its possible successors) automatically as new versions of the profiles come out and we make the relevant changes in the repository metadata editor and conversion scripts.

We recommend that the data should be in one of the preferred formats listed by CLARIN, but also accept other formats if their use is justified and sufficient documentation describing the format is given (see S8). Understandability of the data is ensured by the metadata and additional documentation of the data. CLARIN preferred formats are reviewed by its standards committee. From our side the person who is our representative in the committee is also responsible for keeping an eye on the formats used in our repository. We encourage data holders to provide new or enriched versions of resources as they become available. Should there be a need to convert data to newer formats, the first step would be to notify the resource owner to ask them to submit a new version of their data. If the data provider cannot be contacted or is not willing to undertake the task, we will convert the data to a new format according to recommendations from the relevant field (for now, the standards committee would consolidate different opinions from community experts in that field in regards to which format the data should be converted to). More specific plans will be developed as necessary.

A newer version is linked to an older one (via the field ‘Relation’-’Related Resource’ as seen here), but they are handled as separate resources. Resubmission of reused or enriched data or metadata is subject to the same process as newly deposited data, being checked for quality and sufficient relevant documentation and licences before it is published.

For possible future migrations plans will be developed if the need arises (even for a certain subset of resources, should the data provider ask for it) or if the repository’s policy should change. The data providers who are not part of the CELR consortium have given us the right to make copies for continued access by the depositor’s agreement

(Estonian version derived from the CLARIN depositor's agreements); for resources from inside the consortium it is regulated in the consortium agreement.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept at compliance level 3 (implementation)

Reviewer 2: Accept at compliance level 3. Next review round the promised 'more specific plans' on format migration procedures must be requested and reviewd.

## 15. Technical infrastructure

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

4. Implemented: This guideline has been fully implemented for the needs of our repository.

*Self-assessment statement:*

The Technical infrastructure of CELR consists of computing cluster, Storage Area Network system, backup servers and network swiches. The available hardware is used to run virtual machines, that makes the system flexible and reliable.

**General purpose computing cluster (three nodes):**
**Hardware:**
All machines are based on IBM x3750M4 and each has:

- 4 sockets with 8-core CPU-s (Intel Xeon E5-4640), 32 cores per node.
- 128GB of RAM (16 x 8GB), upradable to 1.5TB per node.
- 4x Gigabit 10/100/1000 interfaces.
- 2x 10Gbit SFP+ interfaces.

Redundant connections to two 10GbE switches.

- VLAN-s for iSCSI, internal network and Internet connectivity

- Integrated Lights out Managemnt (fully featured).

- 6 free PCIe 3.0 x8 expansion slots (3 high-profile and 3 low-profile).

- Next Business Day warranty coverage.

- Integrated RAID controller ServeRAID M5110e SAS/SATA.

- Only used for virtualization host OS installation with two SAS drives in RAID1 configuration.

- 2 redundant high-efficiency power supply units and redundant cooling fans.

**Software:**
All machines are running CentOS 6 and provide KVM virtualization

**SAN (Storage Area Network) system - IBM Storwize V3700:**
Fully redundant system with 48 3TB 7.2k SAS disk drives providing 125TB of RAID6 protected storage to the

computing cluster. Multipath iSCSI connectivity to all computing nodes (no single point of failure).

**Backup servers (3 in total)**
Two identical servers in Tartu, Estonia. In two separate University of Tartu datacenters.
Hardware

- 2 sockets with 6-core CPU-s (Intel Xeon E5-2609v3).
- 64GB of RAM,  4x 16GB DDR4 RDIMM.
- 2x GbE interfaces
- 2x 10GbE SFP+ interfaces
- End-to-end redundant 10GbE connectivity to the compute cluster.
- RAID controller Supermicro AOC-S3108L-H8IR-O-P
- Redundant power supply
- 2x 120GB SATA3 SSD for system and cache.
- 48 4TB SAS3 7,2k RPM hot-swap disk drives.

Third similar server (24 8TB disks) in Tallinn, Estonia for geographic diversity (not in production yet).
All servers use different RAID technologies and software versions to avoid a theoretical simultaneous software failure on all three machines.

**Network switches**
Two IBM RackSwitch G8124-R switches with 24 10GbE SFP+ slots each.
Lights out Management network is provided separately over University of Tartu network.

Links: https://wiki.ut.ee/pages/viewpage.action?pageId=39552576

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 16. Security

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

3. In progress: We are in the implementation phase.

*Self-assessment statement:*

CELR is planning to implement Estonian standard for security - ISKE [https://www.ria.ee/en/iske-en.html]. This task requires quite an amount of work and we are suspending it to the next implementation stage. Our goal is to achieve overall security level K1T1S2 by ISKE, where

- K1 – reliability – 90% (acceptable total interruption per week ~ 24 hours); acceptable increase in the required response time during peak – hours (1÷10);

- T1 – information source, the fact of its amendments or termination shall be detectable; controlling the correctness, integrity and being up to date in special cases and according to need;

- S2 – classified information: information can only be used by certain user groups; access to the information shall be granted if the person requesting access has legitimate interest.

[https://www.ria.ee/public/ISKE/Regulation-the-system-of-security-measures-for-information-systems-2007-12-20.pdf].

Availability

Our applications are deployed on virtual machines in order to minimize possible downtime in case the server fails. Virtual machines also provide the opportunity to use additional resources for quicker response time if needed.

Data integrity

For managing data integrity, logging data operations and carrying out periodic controls, we are depending on the functionality of the chosen software

Confidentiality

Access to systems and data is managed by chosen software, which uses IdP and authentication services (https://taat.edu.ee/main/about/?lang=en). In case of physical access we rely on the physical security of datacentres where our servers are located.

## Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

Reviewer 1: Accept for level 3 (implementation)

Reviewer 2: Accept for level 3 (implementation) and request a more robust evidence and better compiance for recertification.

## 17. Comments/feedback

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

0. N/A: Not Applicable.

*Self-assessment statement:*

As a CLARIN center the DSA/CTS seals are useful. The fact that there are no required implementation levels make it difficult to see the importance of different statements and the relations between them. For example in S0 one should declare the level of curation performed, but that choice seems to have no relation to the answers to other statements.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*