Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	CLARIN-PL Language Technology Centre
Website:	https://clarin-pl.eu/dspace
Certification Date:	17 December 2019
This repository is owned by:	**Wrocław University of Science and Technology**

CLARIN-PL Language Technology Centre

Notes Before Completing the Application

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

True

Reviewer Entry

Reviewer 1

Comments:
Accept

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:

Other (Please describe below)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

CLARIN-PL Language Technology Centre is the leader of the CLARIN-PL national consortium, the Polish node of CLARIN ERIC. It is a B-type centre and as such applies for the CoreTrustSeal. The establishment of one B-centre within the CLARIN-PL consortium was a decision of the CLARIN-PL board and is stated in CLARIN-PL consortium agreement. The centre is hosted by Wrocław University of Science and Technology.

The repository maintains the full set of language tools and resources created and managed by the members of CLARIN-PL consortium. Yet, it is first and foremost, an open repository for data storing and sharing addressed mainly to researchers from the Humanities and Social Sciences. The centre also has a rich collection of data sets uploaded by CLARIN-PL users.

Main website: http://clarin-pl.eu/en/home-page/
https://clarin-pl.eu/dspace

Other relevant websites:
https://www.clarin.eu/
https://www.clarin.eu/content/certified-centres
(https://centres.clarin.eu/centre/25)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository’s Designated Community.

CLARIN-PL Language Technology Centre is open to new users, mainly from the Humanities and Social Sciences research communities from Poland, but also from other countries.

CLARIN-PL Language Technology Centre has been established in response to user demands in areas such as Humanities and Social Sciences to support their research work. We offer natural language processing tools that allow to take either faster or completely new approach to research problem in areas such as corpus analysis, lexicography, stylometrics, text mining, information extraction, statistical semantic analysis and more. CLARIN-PL centre has two main user types. Type One (A) are users who use services under formal scientific cooperation agreements with institutions participating in CLARIN-PL consortium, or who benefit from CLARIN-PL long-term support without signing a formal agreement of cooperation. The close contact with such users enables a thorough diagnosis of the needs and expectations of researchers in a specific field of science.

The services provided by the CLARIN-PL centre are also designed with a spontaneous user - B-type- in mind, who uses them via a web browser or after installing the necessary software made available by the centre on its website. Such users do not need to contact the centre staff (in some cases a user account - e.g. at DSpace repository - or a survey is required).

The repository serves a central function in the activities of users of CLARIN-PL infrastructure: it ensures the space for the storage of the research material (corpora, annotations, databases etc.), allows for publishing and and disseminating research results (papers, databases, corpora, reports etc.) and functions as an access point for many services provided by the consortium (e.g. a corpus search engine Kontext; a tool for the extraction of multi-word expressions MeWeX or a system for corpus management annotation Inforex).

The repository also plays an important role in the works of the CLARIN-PL team, whose members deposit digital tools and resources there.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

Every published dataset is reviewed by a CLARIN-PL data reviewer. From the data repository point of view, after submitting the data a complex curation platform is employed to assure quality and consistence of the data with the possibility to return the data to the submitter for additional changes. Data and metadata are regularly replicated at various levels to several different deposits ensuring robustness and sustainability.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

1. Handle.Net Registry - reserves handle link for the dataset as a support for the persistent identifiers https://www.handle.net
2. PIONIER.Id - Identity Federation https://aai.pionier.net.pl/en/
3. CLARIN-PL consortium members:

a). Institute of Computer Science, Polish Academy of Science (ICS PAS) www.ipipan.waw.pl.
b). Institute of Slavic Studies, Polish Academy of Science (ISS PAS) www.ispan.waw.pl.
c). Polish-Japanese Academy of Information Technology (PJAIT) www.pjwstk.edu.pl (infrastructure responsibilities: additional bitstream preservation and data back-up systems).
d). University of Łódź (ULodz) http://ia.uni.lodz.pl/englang/.
e). University of Wrocław (UWr) www.uni.wroc.pl (infrastructure responsibilities: bitstream preservation, system backup, curation, and access services).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

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 repository mission is to collect and share Polish language data and provide support for researchers from Humanities and Social Sciences in terms of data storage, management, and processing by NLP tools (http://clarin-pl.eu/en/about-project/clarin-pl/). We mainly focus on the Polish language data, but we also include and are open to bi- and multilingual resources as well as resources for other languages. We aim to provide long-lasting access to the stored data to a wide group of users with a particular focus on the scientific community. The mission has been approved by CLARIN ERIC as demonstrated in the CLARIN Consortium Agreement.

https://clarin-pl.eu/dspace/page/about

https://clarin-pl.eu/dspace/page/item-lifecycle

CLARIN-PL Language Technology Centre is financed by the Polish Ministry of Science and Higher Education. The Ministry pays the CLARIN ERIC membership fee and provides funding for the employment of the centre staff.

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:

For each submission, the submitter signs “Deposition License Agreement”, in which the repository’s rights and duties are described, while the submitter acknowledges that they have the right to submit the data and gives the repository centre the right to distribute the data on their behalf. Each submission is monitored by a dedicated CLARIN-PL staff member. Before its publishing the identity of the submitter and the content of each submitted item are verified with respect to its compliance with the submission contract: https://clarin-pl.eu/dspace/page/contract.

When depositing a dataset to the repository, Data Producers must select a license governing its access and usage from a set of license options, which include Creative Commons licences, software licences, and hardware licences. Our licensing policy is based on the license selected by the submitter. A license can be either free or restricted. If the data are licensed with a license that requires signing, the user is asked to electronically sign the license before downloading the data. The process is carefully monitored to make sure that authenticated users are real people and the signed license agreements are stored. The data consumer is made aware of usage restrictions using clear visual indicators (see e.g. https://clarin-pl.eu/dspace/handle/11321/47). In case of misuse, the user is denied further access to the repository and the research community is made aware of the breach of license agreement.

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:

The financing of CLARIN-PL Language Technology Centre was secured for the years 2018-2021 by the Polish Ministry of Science and Higher Education. The funding is based on a 3-year project model. A positive evaluation is an important element in the project extension process. In the unlikely event of the suspension of funding by the Ministry, Wrocław University of Science and Technology, as an official owner of the whole infrastructure, is obliged to guarantee the management and maintenance of the CLARIN-PL centre. Already in the first phase of the CLARIN-PL project (2013-2015), it was decided that the infrastructure became the fixed asset of Wrocław University of Science and Technology.

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:

CLARIN-PL centre aims to ensure datasets posted to and made available by the repository comply with confidentiality and ethics guidelines:
https://clarin-pl.eu/dspace/page/about#privacy-policy,
http://clarin-pl.eu/en/privacy-policy/.

We also accept confidential data with disclosure risks. In such case, the submitter has to contact our HelpDesk. For such a user, a private collection will be created. They will be granted exclusive access to this collection. They will be responsible for the management of this collection assisted by the Repository administrator.

The CLARIN-PL Centre has implemented the rules from the GDPR regulations. Two Data Protection Officers have been installed who monitor all data handling that might involve personal data.

We follow the rules guidance described in the documents:

https://www.clarin.eu/governance/legal-issues-committee

https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations_en

https://www.fnp.org.pl/assets/FNP-Code-of-Ethics-2018.07.13-en.pdf

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 funding is project-based, currently guaranteed till 2022. The whole CLARIN-PL centre staff amounts to 8,7 FTE, out of which 2,5 FTE of the repository staff:

- 0,75 FTE senior system administrator (PhD in Computer Science with experience in data center management and commercial projects)
- 0,75 FTE junior system administrator (MSc in Computer Science with experience in databases management and

application development)

- 0,5 FTE content manager (PhD in Linguistics with the knowledge of legal issues)

- 0,5 FTE user involvement officer (PhD in Linguistics and experience in user involvement activities)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept

VI. Expert guidance

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

CLARIN-PL Language Technology Centre created PolLinguaTec - a CLARIN certified knowledge sharing centre http://kcentre.clarin-pl.eu/cert.php?en.

PolLinguaTec - CLARIN Knowledge Centre for Polish Language Technology is the part of the CLARIN-PL Language Technology Centre (LTC) at Wrocław University of Science and Technology. The K-Centre offers technical support, training, and advice in the field of natural language technology, especially in relation to the Polish language. CLARIN-PL Language Technology Centre (LTC) is an institution ensuring access to knowledge useful in the application of tools and systems for natural language analysis, especially Polish, within Digital Humanities and Social Sciences. LTC has at its disposal the documentation (instructions, guidelines, tutorials) and experienced employees able to solve problems connected with the use of tools, resources, and systems. LTC also offers a number of research applications built for the purposes of specific types of research tasks and in close cooperation with researchers from the areas of H&SS. LTC has been one of the leading Polish centres building technologies of natural language processing for many years now.

The centre offers expertise in the following areas described at:
http://clarin-pl.eu/en/what-are-we-working-on/

New users can contact the centre via the contact form or by phone:
http://kcentre.clarin-pl.eu/kontak.php?en

Communication procedure:
- A new user contacts the centre.
- The user’s request is handled by the user involvement officer.
- The user involvement officer assesses the request and assigns it to a domain expert.
- The domain expert analyses the request and replies to the user.

The employees of the Centre regularly take part in CLARIN ERIC conferences, workshops and other meetings to ensure the exchange of expert knowledge and stay up-to-date with the recent developments.

CLARIN ERIC committees regularly provide us with updates in the area of digital data curation and user involvement. The CLARIN B Centre coordinator is also the National Coordinator for Poland in the CLARIN ERIC, meeting his European colleagues every month, either through video conferencing or face-to-face.

Since the Czech DSPACE LINDAT CLARIN centre offers support to CLARIN centres who decided to use DSpace type of repository, we also take advantage of this support. In case problems, we provide feedback to LINDAT via as issues on the repository Github tracker https://github.com/ufal/clarin-dspace/issues.

The employees of the Centre also monitor the developments and needs of users in different SSH domains by taking part in domain conferences, workshops as well as maintaining cooperation with individual researchers. The CLARIN B Centre communicates directly with researchers in Poland, primarily using email. All of this is supplemented by regular on-site activity by crucial staff members.

The centre is also supported by domain experts from all the remaining CLARIN-PL partners listed in R0.

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:

CLARIN-PL ensures the integrity and authenticity of the data posted to the repository.

Authenticity:
The depositor’s identity is verified either by a local account or a shibboleth account (providing their name and email address and attributes). The dataset is then linked to the Data Producer’s account, and only the Data Producer may edit, manage and publish changes to their dataset.

Once a dataset is published, the dataset version is fixed and immutable. The Data Producer can edit their dataset, but these edits create a new version. We do require CMDI metadata. We support OAI-PMH, OAI-ORE and several other specific protocols of metadata and data sharing. We offer different formats from the standard Dublin Core to CMDI. We are currently regularly harvested by several institutions which reuse the metadata provided by our repository (e.g., http://www.clarin.eu/vlo/, Google Scholar).

CMDI profiles used are published here: http://catalog.clarin.eu/ds/ComponentRegistry/#

Integrity:
When the dataset is created, the data files are stored on our cluster, which ensures the integrity of files by calculating

checksums and immediately repairing corrupted files using redundant data. All backups follow standardized ways of using MD5 checksums for determining the consistency and we use automatic monitoring tools at various levels.

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:

We show a recommendation to use standard formats when uploading files during a submission workflow e.g. for language resources we show http://www.clarin.eu/node/2320. The usage of standardized formats is encouraged, but not enforced. In the case of non-standard formats, the validity is checked manually by a content manager. If the format is unknown, it must be well-documented and the documentation must be either the part of the submission or the metadata must contain a link to it. The repository automatically performs regular checks on the integrity and the file formats of data. The report is sent to the content manager and system administrator who keep track of all the formats used. If there appears a new emerging and more commonly used format, we can add it to the recommendation.

See https://clarin-pl.eu/dspace/page/deposit or https://clarin-pl.eu/dspace/page/about for a description of the submission workflow from the data producer point of view.

The collection development policy is quite broad with respect to purpose: all Polish language materials that might be of interest to researchers of any discipline. As to the quality of the data sets, there must be clear evidence of usefulness reflected by properties like completeness, size, structure, annotation, etc. We do not have a clear set of rules to measure quality. The data sets need to be supplied with all the information that is essential for sustainable data management and future use. Data producers are encouraged to supply additional data description documents or links to publications (using persistent identifiers) about the data. CMDI descriptions are created and maintained by repository staff using a number of standard components sufficient to assure discoverability. We have the policy to provide data as much as possible in open standards. The data from other producers is accepted if it complies to these standards. It will be rejected if it does not comply.

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:

Data Producers may create datasets via a web interface. They may edit and privately share a draft dataset version before publishing, at which point the version becomes fixed, and the dataset is accepted from the Ingest function by the repository and is assigned to Storage. After the dataset has been reviewed by the CLARIN-PL reviewer, it is forwarded to long-term archiving and preservation.

The dataset Access function is fulfilled by CLARIN-PL and VLO, which provides dataset record view and file downloads at the dataset URL (resolvable to via the dataset Handle.net - DOI), dataset listings, and search capability. The data is stored on our cluster server.

Any corruption of datasets creates error logs and backups are kept to restore data. Automated database backups happen every day, with a retention period currently set to 7 days. CLARIN-PL D-SPACE uses MD5 checksums to verify the datasets and data files received are exactly those which were sent. See https://clarin-pl.eu/dspace/page/deposit or https://clarin-pl.eu/dspace/page/about for a description of the submission workflow from the data producer point of view.

In a nutshell, the storing procedure includes:
- Local backups on cluster level as described above.
- Monthly full back-ups external server (PJA Long term storage server). The full back-ups are followed by incremental back-ups on a weekly basis.
- Quarterly full back-ups to the external device, to be stored at another location. A restore can be carried out upon request.
- Back-ups are carried out on two levels:
- Whole virtual machine full backup
- Repository data files and database backup.
- The essential information for disaster recovery is stored on paper and digitally in a vault at a great distance from the data centre.
- Installation of security patches and updates on a monthly basis.
- Daily and automated monitoring

At the infrastructure level, we have three components:
- Software: High Availability Application Cluster using XenServer
- Hardware: HA Cluster provided by IBM Storwize V7000
- Hardware: Backup by ProtecTIER 6710 IBM System with reduplication mode.

The detailed hardware description is provided in section R15.

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:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

To ensure long-term data preservation one of CLARIN-PL consortium members, Polish-Japanese Academy of Information Technology provides off-site backup long-term storage described in the following document: https://mul.pja.edu.pl/archive.pdf. The repository system administrator monitors the backup procedure which allows for incremental export of the newly deposited data (on monthly basis). The administrator also has the access to tools which allow making manual full data backup or the backup of their selected part. The transfer of custody and responsibility handover between the depositor and the repository are specified in Deposition License Agreement: https://clarin-pl.eu/dspace/page/contract. All datasets are treated equally in terms of preservation level.

Our plan is long-term preservation. We are currently working on detailed procedures for data sustainability management in terms of long-term secure data storage. This procedure will include how to deal with different, more detailed scenarios of data security and sustainability risks.

All issues connected with data migration between CLARIN-PL partners are regulated by CLARIN-PL Consortium Agreement.

As already mentioned in Section 3, Wrocław University of Science and Technology is obliged to guarantee the management and maintenance of the CLARIN-PL centre (in the unlikely event of suspension of its funding). It became the fixed asset of Wrocław University of Science and Technology already in the first phase of the CLARIN-PL project (2013-2015), therefore no custody transfer is needed.

The repository encourages to use file formats as listed by CLARIN [https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf]. The number of accepted file formats is limited and well documented to make future conversions to other formats easier. The use of open (non-proprietary) file formats is recommended. The significant number of our resources is textual 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/customized formats need to be used, we demand detailed and exhaustive documentation so as to make the implementation of future data converters possible.

Since the preferred file formats may change over time, 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.

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:

From the perspective of the repository, after submitting the data a complex curation platform is employed to assure quality and consistency of the data with the possibility to return the data to the submitter for additional changes. The data and metadata are regularly replicated at various levels to several different deposits ensuring robustness and sustainability. The system is based on D-SPACE which tries to follow the OAIS (Open Archival Information System) reference model. It follows the standard principles of a high quality digital repository such as the usage of Persistent Identifiers via handle.net, authorization and authentication using shibboleth authentication, sharing of metadata and data. We do encourage submitters to use open licences, such as Creative Commons, but for legacy and other exceptional reasons, we allow data to be associated with older types of public or private licences. This policy of maximal openness allows for any party to assess the scientific and scholarly quality of data as much as possible, which is common practice in the area of language resources. Clarin-PL requires a set of metadata attributes providing information about submitted data and the authorship to be filled in.

The submission cannot be completed unless all the required metadata is filled in. The required metadata are different for different types of submitted data (e.g., corpus, tool, language description). During the submission process, the submitter agrees and accepts our policy leaving them the responsibility for the correctness of their submission, their legal status and accessibility and all related ethical issues, if any. Nevertheless, the basic set of validations is done by our automatic tools and the editor is responsible for particular submissions. The reviewer checks the quality of the content and if there are unclarities they either return the data to the submitter for additional information or ask the research community connected with the repository for help. Our main channel for feedback on quality for users is the repository email: dspace@clarin-pl.eu.

The users can also contact us via the repository HelpDesk.

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:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase

Reviewer 2

Comments:
3 – The repository is in the implementation phase

Response:

The CLARIN-PL D-SPACE submission workflow is internally configured in our repository and the submitter goes through all its stages. We have automatic tools helping the editors to validate metadata and the integrity of the submitted data which are performed by every editor during the curation step and automatically at regular time intervals:
https://clarin-pl.eu/dspace/page/deposit

We also work on the tools that make creating series of metadata files for a large number of files.

The repository requires submitters to electronically sign the confirmation of their right to archive the data. The responsibility of the content lies with them. After submitting an item, the content manager validates the submission before making it public. We are currently implementing a subsystem of tracking reviewer's decisions concerning the publication of the dataset and document corrections. This is done because D-SPACE does not have a functionality of reviewer's checklist nor does it document or log changes.

The repository enables the submitters to restrict the access to their resources at various levels.

These include assigning licenses to submissions which must be electronically signed by authenticated users. The signature information is archived. For every deposit, we enter into a standard contract with the submitter, the so-called "Deposition License Agreement", in which we describe our rights and duties and the submitter acknowledges that they have the right to submit the data and gives us (the repository centre) right to distribute the data on their behalf. Everyone who downloads data is bound by the license assigned to the item – in order to download protected data, one has to be authenticated and needs to electronically sign the license.

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:

Each data submitted by the user receives a unique persistent identifier (PID via handle.net). We strongly encourage our users to use these identifiers when they cite a resource (see https://clarin-pl.eu/dspace/page/cite). We support OAI-PMH, OAI-ORE and several other specific metadata and data sharing protocols. Our metadata resources are available in various formats from the standard Dublin Core to CMDI. The metadata available through the repository is now regularly collected and reused by several institutions such as http://www.clarin.eu/vlo, Google Scholar. The data collection is available via a repository webpage and CLARIN Virtual Language Observatory(https://vlo.clarin.eu) which offers metadata keyword search as well as file content search that enable effective data discovery.

This keyword search capability will be powered by DataSearch, a data search engine.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

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

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

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

Reviewer 2

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

Response:

CLARIN-PL aims to ensure that sufficient metadata (CMDI) in high enough quality are used to support understanding and (re)use of data. The description of the resource is required to meet the (minimal) metashare schema (http://www.meta-net.eu/meta-share/metadata-schema) and/or our CMDI profile https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1349361150622&registrySpace=public. To help ensure understandability of the data to consumers, Data Producers must provide all workflow metadata like the title, description, research which led to the data, software used; links to any software or other datasets used in generating the data, associated articles. In addition, Data Producers may provide a description for individual files within the dataset, for instance describing the contents, related findings, the format, or the processes which led to the creation of an individual file. Data Producers must select an appropriate licence from a predefined range of options, so that they can set their desired conditions for reuse of the data. To address the issue of future evolution of formats, and any future migrations therefore needed, our long-term archive undertakes to ensure all files uploaded in preferred formats are usable in perpetuity, while all formats will be preserved.

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:

At the infrastructure level, we have three components:

1) Software: HA (High Availability) Application Cluster using XenServer

We use complete automation tool for managing XenServer pools which utilize the XAPI management interface and tool stack. Our software suite provides complete HA features within a given pool. The overall design is intended to be lightweight with no compromise of system stability. HA is provided with built-in logic for detecting and recovering failed services. We have two virtual machine servers, with automatic failover, which provide a safe environment to run services. The service is defined as the application and the underlying operating system.

The features of our scripts for XenServer are as follows:

- Auto-start of any failed VMs
- Auto-start of any VMs on after reboot

- Detection of failed hosts and automated recovery of any affected VMs
- Detect and clean up of orphaned resources after a failed host is removed
- Removal of any failed hosts from the pool with the takeover of services

2) Hardware: HA Cluster
The data storage subsystem is built on IBM Storwize V7000 with redundant dual-active intelligent controllers. The storage is using RAID10 volumes (Redundant Array of Independent Disks; in this mode each chunk of data is repeated).

3) Hardware: Backup
The data backup is implemented on DS3500 Storwize V7000, ProtecTIER 6710 IBM System with reduplication mode. The system is configured to create a complete data snapshot every Sunday. Our D-Space repository is stored on XenServer virtual machine. Single point failure at the data storage subsystem does not affect running D-Space repository service instance at all. Single point failure of the primary application server will initiate reconnecting to a redundant second controller to another application server and restarting of the D-Space repository service. The policy described above applies to the digital repository and the data and metadata as well. The digital repository software source code is publicly available and is stored in multiple places on multiple machines. The content of the digital repository is backed up to the ProtecTIER every week (of the last month) including daily incremental updates using standard backup tools and can be restored using automatic tools. All backups follow standardized ways of using MD5 checksums for determining the consistency and we use automatic monitoring tools at various levels. All warnings and logs are send to administrators via an email every Sunday.

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:

CLARIN-PL takes a proactive approach to security of research data and user data. Regular penetration testing is carried out to ensure service is secure against attacks. All previous penetration tests have failed to breach the service; recommendations issuing from tests have been implemented. Our data is stored on Wroclaw University of Science and Technology computer cluster. Our services have 99% annual availability. Wroclaw Centre for Networking and Supercomputing (https://www.wcss.pl/en/?c=static_contact) is our ISP. Additionally, our long-term archive partner provides multiple backups and redundancy.

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:

This is our second application, please to consider to look on
https://assessment.datasealofapproval.org/assessment_149/seal/html/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: