Assessment Information

CoreTrustSeal Requirements 2020–2024

Repository:	PORTULAN CLARIN
Website:	https://portulanclarin.net
Certification period:	March 11, 2024 - 11 March 2027
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**Faculdade de Ciências da Universidade de Lisboa**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

Please provide context for your repository. You can select one or multiple options.

Response:

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

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Description of Repository

Provide a short overview of the repository.

Response:

Domain or subject-based:

PORTULAN CLARIN is a repository of scientific resources for the science and technology of language and related disciplines.

National repository system:

It is a national repository of scientific resources for the science and technology of language with a special emphasis on the Portuguese language. It was founded in 2014, it is a distributed research infrastructure based on and supported by a range of organizations and contributors from Portugal and Brazil, indicated here https://portulanclarin.net/who/#network, and pursues the mission described here https://portulanclarin.net/rationale/#mission .

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Designated Community

Provide a clear definition of the Designated Community

Response:

The designated community for PORTULAN CLARIN are scholars, researchers, innovators, students and teachers, language professionals and users in general, from all over the world, that in order to pursue their activities, need to distribute or to resort to data and tools related to human languages, with special emphasis on the Portuguese language.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Level of Curation

Select all relevant types of curation.

- Content distributed as deposited
- Basic curation – e.g., brief checking, addition of basic metadata or documentation
- Enhanced curation – e.g., conversion to new formats, enhancement of documentation
- Data-level curation – as above, but with additional editing of deposited data for accuracy

Response:

- A. Content distributed as deposited
- B. Basic curation – e.g. brief checking; addition of basic metadata or documentation
- C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Level of Curation - explanation

Please add the description for your Level(s) of Curation.

Response:

Resources accepted for archiving and distribution are distributed as eventually accepted for deposit.

Basic curation of the resources submitted is performed by completeness and well formedness checking of the metadata. Our repository uses the METASHARE submission online forms and workflow, which ensures that the depositor is prompted, and several steps are performed for a submission to be completed and accepted. After the data is submitted to the repository, basic curation is continued by the repository staff by means of manual assessment of the metadata and by means of checking its correspondence to the data to be deposited.

Enhanced curation of the resources is offered by means of the processing services also provided by PORTULAN CLARIN, on a par with the repository. This permits to convert the resources to be deposited to formats other than its original formats, including standard formats. All format versions of a given resource are bundled to be distributed together.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Insource/Outsource Partners

If applicable, please list them.

Response:

The PORTULAN CLARIN repository is located at, technically configured and managed internally by the consortium coordinator institution, the Faculty of Sciences of the University of Lisbon (https://ciencias.ulisboa.pt).

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Response:

-

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Other Relevant Information

You may provide other relevant information that is not covered by the requirements.

Response:

PORTULAN CLARIN https://portulanclarin.net is part of the European Research Infrastructure Consortium (ERIC) for the European Research Infrastructure for Language Resources and Technology (CLARIN), https://clarin.eu. CLARIN is an international distributed research infrastructure with over twenty national nodes, and PORTULAN CLARIN repository is its Portuguese national node.

PORTULAN CLARIN belongs to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance (https://www.fct.pt/apoios/equipamento/roteiro/index.phtml.en), set up by Fundação para a Ciência e Tecnologia (FCT), which is the Portuguese national funding agency for science and technology and a body of the Portuguese Ministry of Science, Technology and Higher Education (MCTES http://www.mctes.pt).

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Organizational Infrastructure

R1 Mission/Scope

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The mission of PORTULAN CLARIN infrastructure (https://portulanclarin.net/rationale/#mission), in which the repository is included, is to support researchers, innovators, students, language professionals and users in general whose activities resort to research results from the Science and

PORTULAN CLARIN

Technology of Language by means of the distribution of scientific resources, the supplying of technological support, the provision of consultancy, and the fostering of scientific dissemination.

PORTULAN CLARIN ensures the preservation and fostering of the scientific heritage regarding the Portuguese language, supporting the preservation, promotion, distribution, sharing and reuse of language resources for this language, including text collections, lexicons, processing tools, etc.

It supports activities in all scientific and cultural domains with special relevance to those that are more directly concerned with language — whether as their immediate subject, or as an instrumental mean to address their topics —, including among others, the areas of Humanities, Arts and Social Sciences, Computation and Cognitive Sciences, Artificial Intelligence, Healthcare, Language teaching and promotion, etc.

It serves all those whose activity requires the handling and exploration of language resources, including language data and services:
• in all sorts of modalities – spoken, written, sign, multi-modal, etc.
• in all types of representations – audio, text, video, records of brain activity, etc.
• and in all types of functions – instrument for communication, symbolic object, cognitive ability to be stimulated through formal education in native language, knowledge vehicle, ability to be exercised in the acquisition of a second language, reflection of mental activity, natural form of interaction with artificial agents and devices, etc.
• etc.

The infrastructure is resorted to by its users when it is necessary, for example:
• to use a language processing tool – e.g. conjugators, terminology extractors, concordancers, part-of-speech taggers, deep linguistic processing grammars, etc.
• to access data sets – e.g., linguistically interpreted corpora, terminology data bases, EEG records of neurolinguistic experiments, collections of literary texts, etc.
• to obtain a data sample – e.g. video recording of deaf children sign language, words for concepts in the Organization subontology, etc.
• to use specific research support applications – e.g. lemma frequency extractors, treebank annotators, etc.
• to use an appropriately equipped online workbench – to support field work on the documentation of endangered languages, to do research on translation, etc.
• etc.

It represents an asset of utmost importance for the technological development of the Portuguese language and to its preparation for the digital age, contributing to ensure the citizenship of its speakers in the digital age.

The inclusion of PORTULAN CLARIN in the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance (https://former.fct.pt/apoios/equipamento/roteiro/index.phtml.en) - with the mission described above - was submitted to and approved by Fundação para a Ciência e Tecnologia (FCT https://www.fct.pt), which is the Portuguese national funding agency for science and technology, and from which it receives its mandate to pursue the objectives set out in its mission statement.

Links:

- mission of PORTULAN CLARIN
- Portuguese National Roadmap of Research Infrastructures of Strategic Relevance

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Changes accepted

R2 Licenses

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

While PORTULAN CLARIN favors and promotes Open Science, Open Access, Open Data and Open Source policies, in order to ensure the distribution of and access to the widest possible collection of scientific resources, the licenses for the scientific resources in the repository are established by the depositors of these resources out of the license set of their choice.

For a depositor to deposit his resources during the depositing process he is prompted and, to continue, has to accept the deposit agreement. The deposit agreement can be inspected also before the submission process, from the button “depositing” in the footer of the front webpage of the site, which leads to the deposit agreement in https://portulanclarin.net/agreement/#depositing .

If the depositor needs help to find a suitable license for his resources, PORTULAN CLARIN provides support via its help desk (https://portulanclarin.net/helpdesk/), and resorting to CLARIN License Category Calculator (https://www.clarin.eu/content/clarin-license-category-calculator) as well as to other license selection helpers that are made available to users here: https://portulanclarin.net/agreement/#licensing . These helpers encompass both open and non open licenses, whose details can be consulted therein.

The license of a resource is stored as part of its metadata and is presented to any user attempting to have access to it. To eventually get access to a resource, a user has to explicitly register that he accepts the respective license. In order to obtain a copy of a resource with special restrictions or sensitive data, the user is directed to the respective depositor in order to arrange for the compliance with the specific terms of that licensing.

The terms of use can be inspected from the button “terms” in the footer of the front webpage of the site, which leads to the terms of use in https://portulanclarin.net/legal/#terms .

Links:

- Terms of Use
- List of license selection helpers
- CLARIN License Category Calculator
- PORTULAN CLARIN help desk
- Deposit Agreement

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R3 Continuity of access

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

Compliance level:

The repository is in the implementation phase - 3

Response:

PORTULAN CLARIN belongs to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance, set up by Fundação para a Ciência e Tecnologia (FCT), which is the Portuguese national funding agency for science and technology and a body of the Portuguese Ministry of Science, Technology and Higher Education (MCTES). PORTULAN CLARIN is funded by this agency and its continuity is ensured by its inclusion in this Roadmap (Roadmap: https://former.fct.pt/apoios/equipamento/roteiro/index.phtml.en).

In case of an unlikely major business continuity failure of PORTULAN CLARIN, the data contained in the repository will be made available by its sister repositories in another CLARIN ERIC node. By default this framework encompasses succession safeguard by means of which other infrastructure node

would take over and continue the preservation of the resources in the unlikely event of a discontinuation of the PORTULAN CLARIN operation.

The resources in the PORTULAN CLARIN repository are deposited under a non-exclusive license for their distribution, and they can be, and usually are, distributed by the respective depositors via other distribution platforms. In case of an unlikely major business continuity failure of PORTULAN CLARIN, those resources that are distributed solely by PORTULAN CLARIN can be deposited in and distributed through other distribution channels by their depositors.

Links:

- Portuguese National Roadmap of Research Infrastructures of Strategic Relevance

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R4 Confidentiality/Ethics

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The PORTULAN CLARIN repository provides storage and distribution of data. The responsibility of following disciplinary and ethical norms for data storage and distribution lies with the repository. The responsibility of following disciplinary and ethical norms for creation/gathering of data lies with the depositor of the data. To eventually deposit a resource in PORTULAN CLARIN, the depositor has to sign a depository agreement by means of which he has to explicitly confirm that disciplinary and ethical norms were complied with when the resource was created. The depositor also has to specify whether the resource contains confidential data that could potentially be disclosed and the presence of such data will restrict the possible license and end users.

PORTULAN CLARIN has competence to provide guidance, via its help desk (https://portulanclarin.net/helpdesk/), 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 Category Calculator (https://www.clarin.eu/content/clarin-license-category-calculator).

A low proportion of data deposited to the repository is expected to raise confidentiality and ethical issues, and in particular disclosure risks. In any case, depositors need to make sure that IPR and personal rights (e.g. mentioning of people in context with personal information or events in texts) are respected in their deposited data. If needed, anonymization can be asked to the data provider during the curation workflow steps. Also, the depositor is required during submission to distribute such data under restricted access (e.g. limited to academic use/research).

For a depositor to deposit his resources, during the depositing process he is prompted and, to continue, has to accept the deposit agreement. The deposit agreement can be inspected also before the submission process, from the button “depositing” in the footer of the front webpage of the site, which leads to the deposit agreement (Deposit agreement: https://portulanclarin.net/agreement/#depositing).

The terms of service of PORTULAN CLARIN include appropriate provisions if conditions are not complied with. If a rights holder is concerned that he has found resources on PORTULAN CLARIN repository whose license does not allow its distribution, for which he has not given permission, or is not covered by a limitation or exception in law, such rights holder is asked to contact the infrastructure at the email address indicated in the Contacts page of the repository website (https://portulanclarin.net/contact/) in writing, stating the following:

- His contact details;
- The full bibliographic details of the resource;
- The exact and full URL where he found the resource;
- If applicable, proof that he is the rights holder and a statement that, under penalty of perjury, he is the rights holder or an authorized representative.

A rights holder finds the information on the procedure that the respective issue or query will go through here (Procedure document: ‘Notice and Take Down’ section in https://portulanclarin.net/legal/#terms).

Upon receipt of notification, the ‘Notice and Take Down’ procedure is thus invoked as follows:

- The infrastructure will acknowledge receipt of the complaint by email or letter and will make an initial assessment of its validity and plausibility.
- Upon receipt of a valid complaint, the resource will be temporarily removed from the services provided by the infrastructure pending an agreed solution.
- The Service Provider will contact the individual or organization who deposited the material upon the reception of a valid complaint. The depositor will be notified that the material is subject to a complaint, under what allegations, and will be encouraged to assuage the complaints concerned.
- The complainant and the depositor will be encouraged to resolve the issue swiftly and amicably and to the satisfaction of both parties, with the following possible outcomes:
- The resource is replaced on the PORTULAN CLARIN repository unchanged.
- The resource is replaced on the repository with changes.
- The resource is permanently removed from the repository.

The terms of use can be inspected from the button “terms” in the footer of the front webpage of the site, which leads to the terms of use in https://portulanclarin.net/legal/#terms.

Links:

- Contact PORTULAN CLARIN
- Terms of Use
- CLARIN License Category Calculator
- PORTULAN CLARIN help desk
- Deposit Agreement

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R5 Organizational infrastructure

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:

The guideline has been fully implemented in the repository - 4

Response:

The repository is physically hosted at the cutting edge data center (https://ciencias.ulisboa.pt/pt/datacenter) of the Faculty of Sciences of the University of Lisbon (https://ciencias.ulisboa.pt), the coordinator institution of the consortium running the repository, ensuring the long-term stability and reliability of the hardware and of the physical and functional infrastructure.

PORTULAN CLARIN belongs to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance (https://www.fct.pt/apoios/equipamento/roteiro/index.phtml.en), set up by Fundação para a Ciência e Tecnologia (FCT), the Portuguese national funding agency for science and technology. PORTULAN CLARIN is supported by this agency and its continuity is ensured by sufficient funding contracted with it. This funding had its inception in mid 2017 and is scheduled to be renewed for periods of 3-4 years, with the next renewal procedure to happen in 2023.

The staff members of PORTULAN CLARIN (https://portulanclarin.net/who/#staff) belong to its Board of Directors or to its Management Team. The Board of Directors comprises three Associate Professors, contributing 1.5 full time equivalent (FTE). The Director-General is affiliated with the coordinator institution, the Faculty of Sciences of the University of Lisbon, and is also acting as President of the European Language Resources Association, with

office in Paris (http://www.elra.info).

The Management Team comprises three full-time Managers with PhD degree, thus contributing 3 FTE, with the roles of Scientific Resources Manager, Users Support Manager and Communication and Administrative Manager. They are helped by junior IT professionals, contributing 2 FTE.

All employees are educated in the fields of (computational) linguistics, natural language processing, computer science, whose access to ongoing training and professional development is provided by PORTULAN CLARIN. The staff have constant opportunities to improve their expertise by participating in relevant courses and workshops organized by consortium members. Participation in seminars and conferences is encouraged. Cooperation with CLARIN series of workshops (https://portulanclarin.net/outreach/) also offers opportunities for the staff to improve their professional skills.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R6 Expert guidance

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The PORTULAN CLARIN staff is composed of members with experience in data curation, natural language data processing, technical maintenance and software development. Most of them are also experts in the field of Language Technology who publish on, and attend, top-ranked scientific conferences in their domains of expertise.

The PORTULAN CLARIN network includes twenty research centers in the scientific domains addressed by the repository (https://portulanclarin.net/who/#network). These centers are actively involved in depositing their data and in the enhancing of the repository. Their members are active users of the repository providing constant feedback and their representatives for the network provide scientific guidance and expert advice. The repository help desk is active (https://portulanclarin.net/helpdesk/) and also serves the purpose of collecting feedback from all users wishing to contribute their advice or remarks.

The repository builds upon an improved version of the METASHARE repository software. These improvements were developed in-house, which left our staff with the necessary accumulated technical expertise to maintain the repository.

By being a part of CLARIN ERIC, we also have access to expert advice from other partners on best practices, technological advances, legal matters, etc shared through the events and workshops (https://www.clarin.eu/events) organized on a regular basis.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Digital Object Management

R7 Data integrity and authenticity

The repository guarantees the integrity and authenticity of the data.

Compliance level:

The repository is in the implementation phase - 3

Response:

PORTULAN CLARIN operates a data and metadata management system that is suitable for ensuring integrity and authenticity during the processes of depositing, archival storage, and data access. This system is an in-house solution specifically developed for the repository as an enhancement of the previously available METASHARE software, which is built on top of Django. Further enhancements will unfold taking into account eventual suggestions of users.

Data integrity is ensured by means of a number of procedures.

In order to check that a digital object has not been altered or corrupted, the MD5 hash of the resource is saved among its metadata, being possible to compare saved hash with the calculated hash in order to verify the integrity of the resource. The repository automatically performs regular checks on the integrity and the file formats of the resources, namely by checking the match between the saved and the calculated hashes. If a discrepancy is identified, a warning is sent to the repository manager and manual inspection and resolution will follow accordingly.

In the process of depositing a resource, the repository automatically checks the provided metadata automatically for completeness (only certain formats are allowed - according to the CMDI profiles agreed on in the CLARIN infrastructure https://www.clarin.eu/content/component-metadata). Provenance data and audit trails are automatically logged, with every change being saved as a new version of a metadata record with information about who made the change and when.

Data authenticity is also ensured by means of a number of procedures.

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

Only registered users who have been given editor privileges can deposit resources and/or alter the metadata records relating to the resources deposited by him, with authentication relying on our national identity provider federation. Additionally, a separate user account can be created in the repository for other trusted users, that are checked and acknowledged by the repository staff.

Before being published in the repository and made available to its users, the resources are manually screened by members of the infrastructure team in order to check whether the content of the resource correspond to what is described in the respective metadata, following these guidelines: https://portulanclarin.net/agreement/verification/

Links:

- Verification guidelines
- Component MetaData Infrastructure (CMDI) profiles

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R8 Appraisal

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository accepts data based on the scope of its mission, as part of a research infrastructure that aims at supporting the science and technology of language and related disciplines.

Depositors provide the metadata on their resources through online forms that, in order to allow being submitted, enforce that the metadata are well formed and complete, with all required fields are filled in.

The PORTULAN CLARIN repository adheres to the METASHARE metadata model (http://www.meta-net.eu/meta-share/metadata-schema). The use of CMDI (https://www.clarin.eu/content/component-metadata) profiles from CLARIN ensures that the metadata required to interpret and use the data are provided and are sufficient for long-term preservation. The following profiles are used:

META-SHARE v3.0 - corpora
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1361876010571&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1361876010571/xsd

META-SHARE V3.0 - tools/services
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1360931019836&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1360931019836/xsd

META-SHARE V3.0 - language descriptions
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1361876010554&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1361876010554/xsd

META-SHARE V3.0 - lexical/conceptual resource
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1355150532312&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1355150532312/xsd
All these profiles are public and have production status.

This also allows interoperability with the rest of the CLARIN ecosystem. In concrete, the metadata is automatically harvested using OAI-PMH (https://www.openarchives.org/pmh/), validated and integrated into the Virtual Language Observatory (VLO https://vlo.clarin.eu/), the central CLARIN service for metadata browsing.

Regarding data formats, PORTULAN CLARIN caters to a wide range of areas (Linguistics, Social Studies, Natural Language Processing, Psycholinguistics, etc.) and accepts data in a range of modalities (text, audio, video, etc). Depositors are encouraged to follow the CLARIN recommended formats (https://www.clarin.eu/node/3060) or the standard data formats for their field, since this facilitates dissemination, reuse and the interoperability of processing tools.

As for risk assessment of the formats of submitted resources, when the submitted data contain attached bit streams, repository staff handling the submission process manually verify whether they meet the requirements of integrity, authenticity, availability and/or their restrictedness. If the format is unknown or not in the list of the recommended standard formats, it must be well documented and the documentation must be either part of the submission or the metadata must contain a link to it.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R9 Documented storage procedures

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

To prevent data loss and storage deterioration, the storage media is arranged in a RAID setup on the machine hosting the repository and on the backup machine. The SMART report of the disks is regularly checked by the technical manager monthly.

Backups of the data in the repository are regularly performed onto separate servers, by automatic means onto a backup machine (cf. details in R15 and R16). The SQlite database as well as the directory containing the deposited resources are automatically backed up on a daily basis onto external hard disks (three disks with weekly rotation). One of the disks is stored outside the campus. Check sums of the data dumped into the backups are checked against the original data.

The Management team includes a designed Manager who is responsible for ensuring and monitoring that the above measures and procedures are being carried out.

The PORTULAN CLARIN repository follows the risk management procedures of the data center where the servers are hosted (description of these procedures: https://ciencias.ulisboa.pt/pt/datacenter#toc0). Its facilities include around-the-clock connectivity to public networks at a sufficient bandwidth, physical access control to the hardware, a fire prevention system, connectivity and a backup power supply. This is a data center of the Faculty of Sciences of the University of Lisbon (https://ciencias.ulisboa.pt/pt/datacenter#toc1), which supports the digital operations of research centers from diverse scientific fields, from Atomic Physics to Bio-computation, thus including those with the highest standards and requirements for archival storage.

Links:

- Description of risk management procedures of the data center where the servers are hosted
- Data center of the Faculty of Sciences of the University of Lisbon

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R10 Preservation plan

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

PORTULAN CLARIN

The preservation policy of PORTULAN CLARIN is structured according to the guidelines and standards set forth in the Special Issue on Digital Preservation of the Journal on Information Standards Quarterly, volume 22, issue 2, ISSN 1041-003, and is publicly available to any interested party, including to the depositors and users of PORTULAN CLARIN, from a pointer in the front page of the respective website, available here: https://portulanclarin.net/agreement/#preservation. In the text below, implemented measures enacting this preservation policy are detailed.

To deposit a resource, while filling in online the respective metadata record, a depositor is prompted and, to continue, has to accept a depository agreement granting the non exclusive right to PORTULAN CLARIN to archive the resource and to distribute it under the license indicated by the depositor. It is made clear to the depositor in the depository agreement that PORTULAN CLARIN ensures the long-term preservation of deposited resources.

Each resource deposited for distribution receives a Persistent Identifier (PID) for reference and citation, fostering its long term identification and preservation. As a rule, PORTULAN CLARIN will not do any modifications in the content of the deposited resource, which fall under the responsibility of the depositor. If a modified version of the resource is deposited, the older version is preserved too. Different versions get different version labels, and all of them are considered as items to be preserved.

The PORTULAN CLARIN encourages the usage of specific file formats as recommended by CLARIN ERIC. 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 digitalization of analogue signal, lossless or no compression is recommended.

The preferred file formats will change over time, in which case the PORTULAN CLARIN 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).

An audit trail is automatically maintained by the repository for all operations on a deposited resource. The repository verifies the integrity of the stored data through the use of checksums maintained as part of the system metadata. The built-in version control system ensures that the data is never overwritten but instead new versions are created on every update operation.

To strengthen redundancies and enable fallback recoveries, backups of the relevant virtual machines are created via dedicated mechanisms of our virtualization solution regularly, always preserving the five most recent versions, and backups of the underlying repository are created, storing the three most recent versions, from which, the entirety of the repository can also be recreated. To enable the option of restorations of older snapshots, besides these continuous backup versions, semiannual backups are preserved. Backups are held on hardware that is situated on locations that are separated from the live system and is monitored for deterioration.

PORTULAN CLARIN utilizes widely used open source software stacks to facilitate all repository services. This maximizes the conditions of long term support, including with respect to updates and security fixes, for the tools being used and improves the ability to run installations of these software stacks independent from the underlying hardware and/or operating system. The update status of installed software is monitored regularly and available updates are installed.

All CLARIN ERIC centers commit to ensuring long-term availability, access and to preservation of datasets submitted to their repositories, as set out in their mission statements. CLARIN is setup as a distributed research infrastructure, where each national center brings its own financial resources, which ensures continued availability. In the case of a withdrawal of funding, the repositories content would be transferred to another CLARIN ERIC center. While the legal and technical aspects of the process of relocating data to another center are underway, in articulation with the Faculty of Sciences of the University of Lisbon data center, where the repository is hosted, Foundation for the National Scientific Computation of the Portuguese Ministry for Science and Technology, along their mission of supporting national research infrastructures, ensures a timeframe of at least 10 years of hosting for the PORTULAN CLARIN Repository, in which period preservation of and access to the data will be provided.

To keep being a trustworthy research infrastructure for the preservation of scientific resources, PORTULAN CLARIN undergoes periodical assessments and certifications by independent bodies, including by CLARIN ERIC and CoreTrustSeal.

Links:

- Preservation Plan

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R11 Data quality

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:

The guideline has been fully implemented in the repository - 4

Response:

The PORTULAN CLARIN staff (https://portulanclarin.net/who/#staff) is composed of members with long experience in data curation, natural language data processing, technical maintenance and software development. Most of them are also experts in the field of language resources and technology who publish on, and attend, top-ranked scientific conferences in their domains of expertise.

Metadata record is filled in by the depositor in an online metadata editor that checks automatically that all required fields are properly filled. The data submitted and its correspondence to the respective metadata information undergoes a subsequent manual check by the PORTULAN CLARIN staff to be accepted for deposit and distribution. The depositor can also refer to other quality check results in the relevant section of the metadata. (Verification guidelines: https://portulanclarin.net/agreement/verification/)

In case of problems or questions users are contacted by the staff and can give feedback via the PORTULAN CLARIN help desk to comment on the data or metadata and on eventual mismatches, aiming at responding to clarification questions or resolving eventual mismatches.

The metadata profiles in use include sections for relevant documentation, quality assessment, and references to related projects and to canonical publications or way to refer to the relevant resource. Every resource is assigned a PID in case it does not already have one, and we strongly encourage people to include it in citation of the resource.

Links:

- Verification guidelines
- PORTULAN CLARIN staff

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Changes accepted

R12 Workflows

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository is based on METASHARE software, which supports an ingestion-publication workflow. The major procedures along this workflow are as follows.

1. Access: Help about accessing the online forms and workflow for depositing a resource is requested via the repository help desk (https://portulanclarin.net/helpdesk/). The depositor is directed to the webpage for submission (https://deposit.portulanclarin.net/).

2. Filling in: After being authenticated into the system and requesting to submit a new resource, the depositor fills in the required metadata fields and indicates a file to upload. The depositor is appropriately guided by the self-explanatory headers and/or information associated to each field to be filled in in the online forms.

3. Ingestion: After the depositor accepts the deposit agreement and indicates the license for the resource, the resource and metadata are ingested into the system. The depositor is then made aware that the next steps in the workflow is the validation by the repository's staff of the metadata and data submitted.

4. Archival: The correspondence of the metadata to the material uploaded is manually verified by the repository staff. The repository accepts any data, as long at it is related to human language and falls under the purview of PORTULAN CLARIN. The resource is stored and a persistent identifier is assigned if that verification is successful. (Verification guidelines: https://portulanclarin.net/agreement/verification/)

5. The depositor is informed by email about the successful archival and publication of the resource, or in alternative he is prompted for corrections if these are found to be necessary.

6. Publication: The metadata record for the resource is made available in the online browsing UI. Search can be done at the PORTULAN CLARIN repository browser (https://portulanclarin.net/repository/search/) and also at the Virtual Language Observatory (VLO https://vlo.clarin.eu), the central CLARIN browsing service from where all CLARIN repositories can be searched, since metadata is automatically harvested from each national node/repository via the VLO.

Links:

- PORTULAN CLARIN helpdesk
- Verification guidelines
- CLARIN Virtual Language Observatory (VLO)
- PORTULAN CLARIN repository browser
- Deposit page

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Changes accepted

R13 Data discovery and identification

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The repository offers search facilities on metadata via Apache Solr. The items in the repository can be search by looking into its listing (https://portulanclarin.net/repository/search/), that can be organized according to several sorting criteria. It is possible to undertake a faceted search by filtering the listing under a wide range of values for metadata fields. The resources can also be browsed via keyword search.

Periodically, the metadata is automatically harvested using OAI-PMH (https://www.openarchives.org/pmh/) to the Virtual Language Observatory (VLO https://vlo.clarin.eu/), the central CLARIN browsing service from where all CLARIN repositories can be searched.

Depositors are prompted, as it is in their best interest, to provide in relevant metadata field a canonical citation for the resource being deposited. Every resource in the repository is assigned a persistent identifier provided by the ePIC consortium of type handle (hdl), resolved by handle.net.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R14 Data reuse

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:

The guideline has been fully implemented in the repository - 4

Response:

To foster the re-usability of metadata, PORTULAN CLARIN rely on the CMDI (https://www.clarin.eu/cmdi) format. CMDI metadata sets are made human readable by the use of XSL style sheets in the repository.

The following profiles are used:

META-SHARE v3.0 - corpora
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1361876010571&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1361876010571/xsd

META-SHARE V3.0 - tools/services
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1360931019836&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1360931019836/xsd

META-SHARE V3.0 - language descriptions
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1361876010554&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1361876010554/xsd

META-SHARE V3.0 - lexical/conceptual resource
Info: https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1355150532312&registrySpace=public
XSD: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1355150532312/xsd

All these profiles are public and have production status.

The metadata on each resource follows the METASHARE schema (http://www.meta-net.eu/meta-share/metadata-schema) and is automatically harvestable via the OAI-PMH protocol (https://www.openarchives.org/pmh/).

To foster the re-usability of data, in turn, in order to mitigate the effect of file format obsolescence, depositing data in widely used, raw and open formats is encouraged (see the CLARIN recommended formats - https://www.clarin.eu/node/3060). The number of recommended file formats is limited to make future conversions to other formats more feasible. For textual resources, XML formats are recommended whenever possible, to ensure future interpretability of the files independent of the tool used to create them. Text is recommended to be encoded in Unicode to enhance future interpretability.

The resources in the repository exist in a range of modalities (text, audio, video, etc.) and they serve a wide range of fields (Linguistics, Social Studies, Natural Language Processing, Psycholinguistics, etc). As a consequence of this, the resources exist, and will continue to be deposited, in a variety of formats. Nonetheless, most resources adhere to the standard formats of their respective fields. Keeping several formats in parallel for the same resource may also be accepted if it is justified. Formats are monitored and recommendations regularly updated.

Should the depositor wish to convert their data into a (new/emerging) standard format, but lack the technical expertise to do so, the PORTULAN CLARIN help desk (https://portulanclarin.net/helpdesk/) provides the support needed.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Technology

R15 Technical infrastructure

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:

The guideline has been fully implemented in the repository - 4

Response:

The PORTULAN CLARIN repository is an improved version of the METASHARE repository software, with updated dependencies and added features, maintained on a private GitHub repository. The repository runs in a virtual machine. The metadata is stored in a SQLite database while the resources themselves are stored in the filesystem. All software used is open-source and runs on a Linux operating system.

The PORTULAN CLARIN infrastructure (i.e. the repository and the web services) is currently hosted on a machine with the following specifications:
• Dell PowerEdge R740 Rack Server
• 2x Intel Xeon Gold 6152
• 256 GB RAM
• 240 GB RAID1 for system
• 10 TB RAID5 for storage
• Dual redundant power supply

A second machine with similar specifications replicates the first and provide load-balancing for the repository and web services.

Besides these two machines, there is a third one, which stands as a backup server, with the following specifications:
• Dell PowerEdge R720 Rack Server
• 2x Intel Xeon E5-2640 v2
• 256 GB RAM
• 134 GB RAID1 for system
• 3 TB RAID5 for storage
• Dual redundant power supply

Regular backups of the repository are performed onto a fourth machine, with the following specifications:
• Dell PowerEdge R520 Rack Server
• 1x Intel Xeon E5-2407 v2
• 8 GB RAM
• 136 GB RAID1 for system
• 18 TB RAID5 for storage
• Dual redundant power supply

All machines are physically located in the internal data center the Faculty of Sciences of the University of Lisbon (https://ciencias.ulisboa.pt/pt/datacenter), the coordinator partner of the PORTULAN CLARIN consortium. This is a cutting edge data center that supports the digital operations of research centers from diverse scientific fields, from Atomic Physics to Bio-computation, thus including those with the highest technical standards. Its facilities include around-the-clock connectivity to public networks at a sufficient bandwidth, physical access control to the hardware, a fire prevention system, connectivity and a backup power supply.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R16 Security

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

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

PORTULAN CLARIN adopts the following data security and integrity measures.

The repository web application runs within a virtual machine that only accepts inbound network connections. In case of a security breach, this virtual machine cannot be used as a proxy to connect to other sites.

The SQlite database as well as the directory containing the deposited resources are backed up on a daily basis onto external hard disks (three disks with weekly rotation).

A snapshot of the whole virtual machine is made on a daily basis and saved to a different computer. These snapshots permit a promptly recovery in case of a hardware or software failure.

Security updates available to the operating system (Ubuntu) are applied every night, after the virtual machine snapshot is taken, and the machine is rebooted.

After updating and rebooting the machine, a test script is executed to check that the repository is online and responding properly. If for some reason, the updates to the operating system caused the repository application to stop working, the virtual machine is reverted to the snapshot taken prior to the update, and the system administrator is notified to take action.

Backups of the data in the repository are regularly performed onto separate servers (cf. R9 and R15). The sqlite database as well as the directory containing the deposited resources are automatically backed up on a daily basis onto external hard disks (three disks with weekly rotation). One of the disks is stored outside the campus. Checksums of the data dumped into the backups are checked against the original data.

As mentioned in the previous R15, all machines are physically located in the internal data center of the Faculty of Sciences of the University of Lisbon (https://ciencias.ulisboa.pt/pt/datacenter), the coordinator partner of the PORTULAN CLARIN consortium. This is a cutting edge data center whose facilities include 24/24 surveillance, physical access control to the hardware, a fire prevention system, connectivity and a backup power supply. The PORTULAN CLARIN repository follows the risk management procedures of the data center where the servers are hosted (description of these procedures: https://ciencias.ulisboa.pt/pt/datacenter#toc0). This is a datacenter that supports the digital operations of research centers from diverse scientific fields, from Atomic Physics to Bio-computation, thus including those with the highest standards and requirements for archival storage.

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Applicant Feedback

R17 Applicant Feedback

We welcome feedback on the CoreTrustSeal Requirements and the Certification procedure.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

-

Links:

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

All comments have been addressed and now ready for approval

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments: