Assessment Information

CoreTrustSeal Requirements 2020–2023

Repository:	The ILC4CLARIN Centre at the Institute for Computational Linguistics
Website:	https://dspace-clarin-it.ilc.cnr.it/
Certification period:	08 August 2023 - 07 August 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022
This repository is owned by:	**Institute for Computational Linguistics "Antonio Zampolli"**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

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

Compliance level:

Not Applicable - 0

Response:

- Domain or subject-based repository
- Institutional repository

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Accepted
This does have however the flavour of a domain repository?

Description of Repository

Provide a short overview of the repository.

Compliance level:

Not Applicable - 0

Response:

The Italian National Research Council (CNR) [1] is the lead institution of the Italian CLARIN research infrastructure [2] and the Institute for Computational Linguistics “Antonio Zampolli” [3] (CNR-ILC) is the host of its main infrastructural node, the ILC4CLARIN B center [15]. CNR-ILC is a centre of reference in the field of Computational Linguistics at both national and international levels. It is part of the Department of Social Science and Humanities, Cultural Heritage (DSU) [4] of the CNR, and carries out research activities in strategic scientific areas of the discipline as well as publishing activities, training, education activities, and technology transfer.
The ILC4CLARIN center hosts an institutional and domain digital repository of language resources and offers linguistic tools and (web) services.
While data resources may be deposited and thus stored and made persistent in the repository, tools, and services are simply "cataloged" (i.e. described with relevant and common metadata) and linked to make them directly accessible
The repository is regularly harvested by the Virtual Language Observatory [16], the central discovery service of the CLARIN - European Research Infrastructure for Language Resources and Technology [5]. The repository is based on the CLARIN-DSpace software [6] developed by the Institute of Formal and Applied Linguistics, Charles University, in Prague [7]. This DSPACE adaptation is specifically tailored to the purpose of archiving and distributing language resources and technology within the CLARIN research infrastructure and has been implemented/used by several other CLARIN centres all over Europe.
The ILC4CLARIN repository is mainly a collection of linguistic data and Natural Language Processing (NLP) tools; currently it hosts both language resources developed at the Institute for Computational Linguistics of the Italian National Research Council (CNR-ILC) and resources developed by other

member institutions of the CLARIN-IT National Consortium [17]. It also offers to deposit services to any authenticated scholars belonging to the CLARIN research network with relevant linguistic data. Although resources for the Italian language are the most represented, the ILC4CLARIN center hosts data and catalogs tools for other languages as well, including classical/historical languages. Linguistic data cover many subject areas and domains, e.g. legal domain, biology, physics ...
From the data producer's point of view, the repository focuses on a user-friendly interface that allows for publishing data easily. From the data consumer's point of view, the repository offers advanced searching and browsing functionalities for retrieving available resources and citing them.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Designated Community

Provide a clear definition of the Designated Community

Compliance level:

Not Applicable - 0

Response:

The aim of a CLARIN repository is to preserve research data sets and make them available for a Designated Community, which is constituted by the scholars of disciplines where language plays a central role. In particular, a CLARIN repository helps researchers, working in the Humanities and the Cultural and Social Sciences, to access, prepare and analyze research data. The community is subdivided into producers and consumers. Typical producers are Computational Linguists, Information and Communication Technologies (ICT) experts, and Language Engineers who produce language data and digital tools to work with such data. Typical consumers of the infrastructure include students and researchers of all stages who are working in the fields of the Humanities (linguists, philologists, historians...) and in the Social and Cultural Sciences (sociologists, political scientists, theologians, anthropologists) who are interested in analyzing language data and using text processing tools available in the CLARIN infrastructure. Due to the nature of CLARIN, there is no neat distinction within the community: members can be both data producers and data consumers.
We ensure the long-term preservation of the deposited data according to the definition of Preservation Description Information (PDI) given in the OAIS reference model [8]. As for tools, we can only ensure preservation of the respective descriptive metadata.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

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

Compliance level:

Not Applicable - 0

Response:

- A. Content distributed as deposited
- B. Basic curation – e.g. brief checking; addition of basic metadata or documentation

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Level of Curation - explanation

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

Compliance level:

Not Applicable - 0

Response:

We perform a basic curation of the submissions mostly by checking and editing the metadata. Our repository uses the DSPACE submission workflow which foresees several steps before completing a submission. The submitter goes through each of them guided by the software indications. After the data is submitted to the repository, some basic curation is performed which implies assessment and revision of metadata by local experts and a minimum quality check of the data to be stored if any. The curation workflow should ensure the quality and consistency of the data and it offers the possibility to return the metadata and /or the data to the submitter for additional changes before it enters the repository and is visible and harvestable by users and other services. We also have automatic tools helping the editors to verify and validate metadata [9][10] and the integrity of the submitted data which are performed by every editor during the curation step and automatically at regular time intervals.
This repository distributes data as deposited. Third parties retain full ownership of the data they deposit in the ILC4CLARIN repository and are responsible for its quality and migrations to new data formats. However, for all deposited resources at least basic curation is provided (see [20]).
The depositors submit data by themselves and are guided all through the process by a web-based submission workflow: a form with several stages for providing metadata about the submission.
When applicable, answers are validated against vocabularies or pre-defined rules after each stage (cf. [21]).

Once the depositors have submitted an item, before it definitively enters the repository and becomes findable and usable, the repository curators (experts affiliated with the ILC4CLARIN Center) assess and revise its metadata for correctness, appropriateness, and expressivity: minor metadata modifications (e.g., correcting grammar mistakes, unifying keyword lists, unifying organization names). Curators then perform some basic checks (e.g. URL checks, completeness, validity, and checksums) assisted by automatic procedures made available by the software, which help them decide if the submission meets the technical requirements (details at [22], [23]).
Additionally, curators also make sure that the data deposited corresponds to the associated metadata and fits the repository scope, but do not perform

any file format conversion or enhancement of documentation. If deemed necessary curators return the submission to the depositor with detailed instructions on how to update the submission (see [24], [23]) and will support the depositor upon request.

Tools and services (software) are not deposited/distributed, but only "cataloged" (i.e. described according to the CLARIN common metadata framework) and linked via URIs, and thus made more easily findable and accessible. Although basic curation of their metadata is performed as for other submission types, they are not the object of this CTS request.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

We do not outsource any service. The repository is located, configured, and managed internally. The “Italian Research & Education Network” (Consortium Garr) [11] is our main technical partner: an organizational relationship on several aspects such as network connection [12], other network services [13] and user involvement for the Italian scientific and academic community.
The same GARR Consortium is our consultant on storage and High-Performance Computing (HPC) topics.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Compliance level:

Not Applicable - 0

Response:

-

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Other Relevant Information

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

Compliance level:

Not Applicable - 0

Response:

ILC4CLARIN is listed in the Registry of Data Repositories re3Data.org under ID:r3d100012262 [14] the language archive, http://www.language-archives.org/archive/dspace-clarin-it.ilc.cnr.it, and it is indexed by SHARE and the Web of Science Data Citation Index.
The repository is built on DSPACE5.11. It hosts (September 2022) 881 (457 items with physical files and 424 items with metadata only) and more than 80 registered users from Italy and Europe [19].

Although the repository allows for the deposition of resources under restricted access, the currently deposited resources with physical data are publicly available under some Creative Commons license. Registration to the repository is needed only for depositing resources.

URL:
[1] CNR: https://www.cnr.it/en
[2] CLARIN-IT: http://www.clarin-it.it
[3] CNR-ILC: http://www.ilc.cnr.it/en
[4] CNR-DSU: http://www2.dsu.cnr.it/
[5] CLARIN ERIC: https://www.clarin.eu
[6] CLARIN-DSpace: https://github.com/ufal/clarin-dspace
[7] UFAL Institute: http://ufal.mff.cuni.cz
[8] The OAIS reference model: https://public.ccsds.org/pubs/650x0m2.pdf
[9] CLARIN-DSpace metadata info: https://github.com/ufal/clarin-dspace/wiki/Metadata-info
[10] DSpace curation system: https://wiki.duraspace.org/display/DSDOC5x/Curation+System
[11] Consortium GARR: https://www.garr.it/en
[12] GARR and GÉANT: https://www.geant.org/News_and_Events/CONNECT/Pages/Aliens-Our-allies-on-the-optical-network.aspx
[13] https://www.servizi.garr.it/en/cloud
[14] Re3data.org record URL for ILC4CLARIN: http://doi.org/10.17616/R3W365
[15] ILC4CLARIN: https://dspace-clarin-it.ilc.cnr.it/
[16] VLO: https://vlo.clarin.eu/
[17] http://www.clarin-it.it/en/content/consortium
[18] ILC DATA CENTRE (page 16, nr 48): https://www.cnr.it/it/trasparenza/delibere-cda/documento/104615
[19] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/statistics Login is needed
[20] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/item-lifecycle
[21] https://github.com/ufal/clarin-dspace/blob/clarin/dspace/config/input-forms.xml
[22] https://github.com/ufal/clarin-dspace/wiki/Metadata-info
[23] https://wiki.lyrasis.org/display/DSDOC5x/Curation+System
[24] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

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 repository is in the implementation phase - 3

Response:

The mission of our repository is to provide reliable archiving and/or documentation as well as easy access to language-based data, tools, services, and associated metadata for research purposes (e.g. corpora, lexicons, audio and video recordings, grammars, language models, parsers, format converters, lexicon extraction tools, etc.) to scholars in the fields of Social Sciences and Humanities and beyond. Resources and tools can be deposited by CNR-ILC associated researchers as well as by researchers who belong to other academic institutions.

The focus is on the Italian language, but other languages, especially historical and classic ones, are not excluded.

At the moment of resubmitting (September 2022), 457 of 881 entries contain physical data, mainly for historical and classical languages. Moreover, the consortium is increasing, meaning that the repository itself will grow in terms of both data and metadata. Indeed, as a CLARIN center, we believe that recording the data (that is to say creating the metadata item in a
certified national repository) by allowing them to be widely visible (cf. the Virtual Language Observatory, VLO [11]) is important as well as accepting their deposit.

The mission of our repository is therefore supported by the integration of the repository into the national and international CLARIN infrastructure [1], [2], [3] whose ultimate objective is to advance research in the humanities and social sciences by giving researchers access to language resources [11] and technology [13], [14] even through single sign-on, when required.

ILC-CNR for CLARIN-IT is committed to the long-term care of items deposited in our repository and strives to adopt the current best practice in digital preservation, [4].

ILC4CLARIN is part of the CLARIN networked federation since 2015 and a B-certified Centre since May 3, 2018 [9]. ILC4CLARIN got the first CTS certification on April 18, 2018 [10].

CNR has received the mandate from the Ministry of Education and Research to represent Italy within the CLARIN-ERIC research infrastructure and to implement the national infrastructure by 1) coordinating the national effort and 2) implementing the first technical centre of this research infrastructure which provides the necessary services, including a system of user authentication and authorization and a repository of relevant data and tools.

At CNR, the CLARIN-IT technical centre is implemented, developed, and maintained by the Institute for Computational Linguistics “Antonio Zampolli” (CNR-ILC), a centre of reference in the field of Computational Linguistics at both national and international levels. The Institute is part of the Department of Social Science and Humanities, Cultural Heritage (DSU) and carries out research activities in strategic scientific areas of the discipline, as well as publishing activities, training and education activities, and technology transfer. Its main areas of competence are Text Processing and Computational Philology; Natural Language Processing and Knowledge Extraction; Resources, Standards, and Infrastructures; Computational Models of Language Usage. The studies carried out within each area are highly interdisciplinary and involve different professional skills and expertise that extend across the disciplines of Linguistics, Computational Linguistics, Computer Science, and Bio-Engineering. CNR-ILC activities range from innovative research in the field of Digital Humanities to the definition of representation standards and distributed research infrastructures. Research is carried out within a consolidated network of national and international collaborations with research institutes, universities, and public bodies, as well as companies involved in European, national, and regional research projects. It is part of our institute, CNR-ILC's, mission to ensure that resources managed by the institute remain usable in the long term. CNR-ILC has a long history [5] of ensuring that resources remain accessible and usable many years after their creation, as in the first networks in the early, such as the ELSNET projects in the 1990s [6]. See [5] for statements on the main activities and mission of CNR-ILC and [8] for the mission of CNR-DSU.

URL
[1] CLARIN-IT: http://www.clarin-it.it
[2] CLARIN ERIC: https://www.clarin.eu
[3] CLARIN Short Guide: http://www.clarin.eu/files/centres-CLARIN-ShortGuide.pdf
[4] About ILC4CLARIN: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/about
[5] CNR-ILC History: http://www.ilc.cnr.it/en/content/history
[6] CNR-ILC Ended Projects: http://www.ilc.cnr.it/en/content/ended-projects
[7] CNR-ILC: http://www.ilc.cnr.it/en/content/institute
[8] CNR-DSU: http://www2.dsu.cnr.it/missione/
[9] ILC4CLARIN B-Centre: http://www.ilc.cnr.it/en/content/ilc4clarin-officially-recognized-clarin-b-centre
[10] CTS: http://hdl.handle.net/11372/DOC-144
[11] CLARIN VLO: https://vlo.clarin.eu
[12] Selected Papers from the CLARIN Annual Conference 2020: https://doi.org/10.3384/ecp180
[13] Language Resource Switchboard: https://switchboard.clarin.eu/
[14] WebLicht: http://weblicht.sfs.uni-tuebingen.de/

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:

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:

ILC4CLARIN distinguishes three levels of License agreements:

1) For every deposit, we enter into a standard contract with the submitter, the so-called “Distribution License Agreement”[1], in which we describe the rights and duties of the repository and the submitters affirm that they have the right to submit the data and give the repository center the rights to distribute the data on their behalf.
The repository requires submitters to electronically sign the right to archive the data and the agreement that the responsibility of the content lies with them. The author of the work will always remain the owner of the data. The repository stores a copy of the data which it must take good care of, according to the terms of the contract and the terms and conditions for use.

2) Everyone who downloads data is bound by the license assigned to the item: by using the search functions offered by the repository web interface and accessing or downloading the archived data the user agrees to the Terms of Service of the ILC4CLARIN CLARIN-DSpace repository available here [2]. Additionally, to download protected data, one has to be authenticated and needs to electronically sign a license.

3) The repository licensing policy is based on the license selected by the depositor when submitting his/her data.
There are several available open licenses a depositor can choose from directly in the interface within the submission workflow (e.g. Creative Commons, GNU licenses, ... for a list of all available licenses see [3]). In case none of these suits the needs of the depositor, there is also the possibility of contacting the repository staff for setting-up custom licenses. The repository also enables the submitters to restrict access to their resources at various levels. This includes the possibility of assigning licenses that must be electronically signed by authenticated users
before they can get access to them. This means that only authenticated users can access it after submitting a form where they agree to adhere to the

specific terms of the license. The repository keeps track of those signatures and because the authenticated users must be real people, this process is well defined. The repository also offers the option to put an embargo on submissions, which means that the submissions will be archived immediately after completion of the curation workflow, but they will become publicly available after a specific date.

URL:
[1] Distribution License Agreement: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/contract
[2] CLARIN Terms of Service: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/terms-of-service
[3] Available Licenses: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/licenses

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Believe that this compliance has been achieved and would accept.

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:

Italy became a member of CLARIN ERIC in October 2015, 3 years after the establishment of CLARIN ERIC. The Italian Ministry of Education and Research awarded CNR a mandate to represent Italy as a representative Entity and a mandate to the Director of the Department of Social Sciences and Humanities to act as National Delegate in the General Assembly of the ERIC. The Institute for Computational Linguistics “A. Zampolli” was chosen as the Executing institution due to its key role in the development of Language Resources and Technologies over the last 50 years as well as the primary part that it has played in research and development in Computational Linguistics both nationally and internationally. It is therefore very unlikely that its mission will change in the future.

The mandates awarded by the Ministry have durations of five years. The initial mandate committed by the Ministry (payment of the membership and fees for five years, 2015-2019) has been renewed for additional five years (ending October 2025, prot m_pi.AOODGRIC.REGISTRO UFFICIALE.U.0021661.05-12-2019). The Ministry will provide funding to sustain the implementation of the repository as described above.

The staff at ILC will ensure the ongoing development of the repository and the management of
all the activities connected with it in the form of in-kind contributions. The continued availability and accessibility of the data in the repository is guaranteed (along with documentation) until 2025. However, the mission of ILC is to provide long-term preservation of its diverse and extensive range of digital resources, thereby ensuring continued access to these resources even after this date.

In the worst case, i.e. in case a decision is taken not to maintain the ILC4CLARIN repository any longer (although this is very unlikely to happen, being the ILC4CLARIN hosted at the ILC-CNR, hence part of the National Council of Research (CNR)), the data contained in the repository will be transferred to one of its sister repositories in another CLARIN data centre. This is made easier by the recommendation agreed upon within the Italian CLARIN-IT Consortium to use the same software for each CLARIN-IT repository. In this way, any of the CLARIN-IT partners would be able to upload the metadata hosted by ILC4CLARIN, although a case-by-case analysis of the licensing rights would have to be followed for the resources themselves.

However, this scenario is unlikely to happen since the MUR (Ministry of University and Research) recently launched the National Plan for Research Infrastructures PNIR, [1], and the PNRR [2], dedicated to 1) the strengthening of the Italian nodes of the ERIC infrastructures, recognized as "projects of international significance" (such as CLARIN), and to 2) the creation of infrastructural clusters in the various ESFRI disciplinary sectors aims at promoting the federation of infrastructures for Human Sciences and Cultural Heritage, with strong propulsion to the continuity plan.

[1]
https://www.mur.gov.it/sites/default/files/2021-10/Decreto%20Ministeriale%20n.1082%20del%2010-09-2021%20-%20PNIR%202021%20-%202027.pdf

[2] https://www.mur.gov.it/sites/default/files/2021-12/Avviso%20n.%203264%20del%2028-12-2021.pdf

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Compliance Level Comment: Accepted

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 depositors (data producers) are responsible for the data’s adherence to relevant ethical norms and standards in the discipline in question.
During the submission process depositors are required to accept the "Distribution License Agreement" [2] by which they acknowledge that they have the right to distribute the data and that they also have the right to grant the repository permission to distribute the data on their behalf, thus leaving to depositors all the responsibility regarding the content of
his/her submission. Acknowledgment that depositors have the right to distribute the data in the first place includes resolving all the privacy issues including GDPR because if these were not resolved the depositor would not have the right to distribute the data at all.
The deposited item lifecycle is explained on the website [5]. Information about the requirements and handling of disciplinary and ethical issues are available in the text of the Terms of Service [1].

Any depositor must be authenticated through the federation of ID providers, which ensures a high level of trust. No anonymous upload of items is allowed. Additionally, all submissions are validated automatically and reviewed manually by an expert Repository editor, whose tasks include a check if privacy and ethical issues are respected. Should metadata editors detect ethical privacy problems with the submission (e.g. he/she detects or suspect the lack of suitable
anonymization) he/she would interact, via the web interface, with the depositor asking for modifications to the data to be deposited or more information to clarify the compliance. The curator may also refuse to publish the submission and indicate to the depositors how they can prepare their data for a suitable submission. A submission will not be approved until all the fundamental requirements are met and compliance with applicable laws or regulations is ascertained. The metadata editors can also take advantage of internal standard curation tools [7].

Repository editors are data experts who are careful not to make changes that in themselves may create privacy or ethical issues.
Special conditions may be addressed in a distribution license tailored specifically for the particular item, which can be negotiated directly with the repository.
Furthermore, if required by the special case, the repository can control access to items and submissions and grant them on a per-user basis. In such cases, the repository staff will work together with the depositor, in person or via email, on defining the target group of users or individuals with access.

For instance, a depositor may choose (or be advised during the curation steps) to distribute her/his data under restricted access (i.e. limited to academic use/research). The CLARIN-DSPACE we are using [6] is configured to protect restricted resources with a federated (shibboleth) authentication. Those resources will only be available to trusted authenticated scholars, i.e. users who can login through Identity Providers operated at institutions taking part in the CLARIN AAI federations.

So far no data resource deposited (mostly text corpora and lexica) contains personal nor confidential data and the repository does not expect this to change sensibly in the future. In any case, the repository system and editorial staff can handle stricter requirements, as explained above.

As for the protection of the privacy of the repository users, personal data is protected according to the Code of Conduct for Service Providers, a common standard for the research and higher education sector to protect users' privacy. Thus, users’ personal data are safeguarded by an explicit commitment to non-disclosure of personal data to anyone outside of the ILC-CNR or CLARIN-IT team.

Also, the repository publishes a statement in which it explicitly states:
- what personal information is fetched from the Identity Provider server of the home organizations;
- what log files are created at various levels (i.e. email addresses, name, timestamp, and the performed action; IP address; signed licenses if any);
- what purpose personal data are processed for (i.e. user authentication and identification when signing licenses; user authorization in special cases; automated sending of email messages necessary for use of the services --password reset, submission information, etc.--; statistics and development of the service; etc.)

Personal data can be deleted upon request of the user, alternatively, it will be deleted automatically after 5 years of inactivity of the user. All details are explicitly reported here [4].

Measures in case of misuse: The repository system/interface does not allow the depositing of data without providing an appropriate license for its access and use. These license conditions are available via CMDI metadata and the data consumer is made aware of usage restrictions also in the interface via clear visual indicators of the applicable license. If the data is made available with a licence that requires signing, the user is prompted to sign the license electronically before accessing the resource. Furthermore, in case of misuse, because detailed logs are generated and maintained by the system, the repository can retrieve the exact dates and specific ID's of the user who might be blacklisted and denied further access to the repository; the research community might also be made aware of the misuse.

The CLARIN Legal Issues Committee (CLIC) [3], set up and run by CLARIN ERIC, organizes periodic training sessions in management of data with a disclosure risk". ILC4CLARIN organized a seminar on "Legal Issues and Language and Resources Technologies", held in 2019 [4], to make researchers aware of the use and misuse of language data in
research projects, and may organize similar events if needed.

URL:
[1] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/about#terms-of-service
[2] Distribution license: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/contract
[3] CLARIN Legal Issues Committee: https://www.clarin.eu/governance/legal-issues-committee
https://ilc4clarin.ilc.cnr.it/en/news/seminar-on-legal-issues-and-language-and-resources-technologies-at-ilc-cnr/
[4] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/privacypolicy
[5] Deposited Item Lifecycle: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/item-lifecycle at the section 'Edited Item'
[6] https://github.com/ufal/clarin-dspace
[7] https://wiki.lyrasis.org/display/DSDOC5x/Curation+System

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:

Compliance Level Comment: Accepted

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 repository is in the implementation phase - 3

Response:

The ILC4CLARIN repository is hosted by the Institute for Computational Linguistics "A. Zampolli" (ILC), which is part of the Italian National Research Council (CNR). This repository is financially supported by the Ministry of Education through a part of the "Ordinary Fund for Research Entities -Fondo

Ordinario Enti di Ricerca (FOE)", with the staff at ILC ensuring its ongoing development and the management of all the activities connected with it under the form of in-kind contribution. The staff members of ILC also regularly participate in training and professional development activities organized and supported by CLARIN-ERIC.

The expertise and experience of ILC staff are extensive also thanks to their involvement in numerous other international research projects as well as within national and international bodies such as UNI and ISO.

The importance of belonging to the CLARIN-ERIC network and its relevance for the Italian research community of Human and Social Sciences has resulted, over the years, in the participation of Italy and the CNR in partnerships of major projects European infrastructure: ELEXIS [1], which combines the sectors of LanguageTechnologies, Digital Humanities, and Computational Lexicography; TRIPLE [2], for the creation of an integrated platform of innovative tools and advanced services to combine bibliographical data, projects and people to advance research in the Human and Social Sciences sectors; SSHOC [3], which aims to integrate the various infrastructures of the Human and Social Sciences sector into a single virtual environment to foster high-quality research and realize the vision of the Open Science Cloud in the sector.

These projects, which extend temporarily until 2023, see a massive presence of CLARIN consortia and other infrastructures in the sector. The involvement in innovative and strategic European projects of this caliber testifies the central importance, for Italy, of the active participation in CLARIN, ensures the alignment of national policies for research in the sector with the strategies defined at the European level in terms of open science and FAIR data and, finally, it guarantees the sustainability of CLARIN-IT in future years. See R3 for new national initiatives such as PNIR and PNRR that will contribute to the enhancement of the Italian node of CLARIN in terms of both funds and human resources.

The ILC4CLARIN staff consists of 11 units of personnel with different roles and with different temporal involvement in the management of the ILC4CLARIN center.
The total amount of time sums up to 2.4 FTE/Y, meaning that most of the staff is working part-time for ILC4CLARIN. Despite this, there has been an improvement wrt the previous certification period in both the number of personnel (11 vs 9), in role definition, and in FTE/Y (on average 0.22FTE per person vs 0.16) .

ILC4CLARIN Coordinator (0.3 FTE), 1 senior researcher
Technical manager, 1 senior technologist (0.3 FTE)
Repository Manager, 1 researcher (0.3 FTE)
Web manager, 1 technologist (0.1)
Repository/Services Administrators and Developers, 1 technician + 1 researcher +1 technologist (0.9 FTE total)
Metadata curators, 1 technician + 1 researcher (0.3 FTE total)
Community engager and curator, 1 researcher + 1 technician (0.2 FTE total)

The FTE is fairly evenly distributed between technological and technical staff (1.1) and research staff (1.3) as well as among the various professionals involved: technological and technical staff (0.4 vs 0.7), senior staff (0.3), research staff (1.0).

In addition to the above list, ILC4CLARIN has recruited (in 2021) a research fellow dedicated to the project activities (study and recognition of existing resources for the population of the repository, enhancement of outreach activities, and contacts with CLARIN-ERIC).

The activities under each role are described in R12 (Workflows).

URL:
[1] ELEXIS: https://elex.is/
[2] TRIPLE: https://www.gotriple.eu/
[3] SSHOC: https://sshopencloud.eu/about-sshoc

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Accept
Minor comment: please explain UNI when it appears in the second paragraph. For the moment it is explained in R6.

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

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:

IILC4CLARIN has in-house potential advisors and can reach out to other experts worldwide. In-house expert guidance: In addition to the repository staff, expert guidance and advice can be sought in-house within CNR-ILC and CNR more generally. CNR-ILC is active in the standardization of linguistic data formats: it hosts the chairperson of the ISO TC37 SC4 [1] and expert members of some of its committees, nominated by the Italian standardization body, UNI [2].

External expert guidance
By being part of the CLARIN federation of technical centers, ILC4CLARIN is in constant contact with experts in all the CLARIN ERIC member countries, in particular with those working at B and K centers [3]. ILC4CLARIN is also the national data centre of the Italian CLARIN Consortium (CLARIN-IT) [4], whose coordinator regularly participates in the CLARIN ERIC coordination activities and major events. Furthermore, the ILC4CLARIN repository Technical manager is the Italian representative to the Standing Committee for CLARIN Technical Centres [5] which coordinates the activities of all CLARIN technical centres Europe-wide and takes decisions on implementation. Communication with experts can also take place during meetings or seminars within the above-mentioned activities or by email in personal communication exchanges. By these roles, ILC4CLARIN can seek advice from any experts of the European CLARIN network for every relevant aspect of the repository and data management. We do not believe that it is necessary to involve international experts because we feel that the support provided by CLARIN ERIC members is sufficient for our needs. Finally, a help desk is also active which also serves the purpose of collecting feedback from users (dspace-clarin-it-ilc-help@ilc.cnr.it)

Also, we ask the GARR Consortium [6] consultancies on storage and HPC topics.

URL:
[1] ISO/TC 37/SC 4: https://www.iso.org/committee/297592.html
[2] UNI: http://www.uni.com
[3] CLARIN B and K centres: https://www.clarin.eu/content/clarin-centre
[4] CLARIN-IT Consortium: http://www.clarin-it.it
[5] CLARIN Standing Committee: https://www.clarin.eu/governance/standing-committee-clarin-technical-centres
[6] Consortium GARR: https://www.garr.it/en

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: Accepted

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 guideline has been fully implemented in the repository - 4

Response:

Integrity - The repository is based on the DSPACE software. To verify that a digital object has not been altered or corrupted the repository uses MD5 checksums for all objects and checks it periodically. The repository automatically performs regular checks on the integrity and the file formats of data. There is a list of supported and known formats whose consistency is regularly checked using existing tools (e.g., integrity testing of bzip format is done using bzip -t). Files are checked three times (not necessarily by editors). The file extension (file format) is checked and marked as supported, known, or unknown [8, 9]. The file integrity is checked for several supported and known types regularly. Finally, md5 checksums are checked regularly to ensure the consistency of submissions. A report is sent to the editors and administrators who keep track of all used formats. If there is a new emerging and more commonly used format, we can add it to the recommendation.
The repository employs a number of standards:
OAI-PMH (v2) protocol standard for metadata harvesting; the repository makes metadata formats available via the repository's OAI-PMH endpoint. Metadata are made available according to the CMDI 1.2 metadata specification [3] and are harvested via the repository's OAI-PMH endpoint. CLARIN ERIC harvests our CMDI metadata to a central registry, which can be seen at CLARIN VLO [4].
CMDI Profile schemas are based on the W3C XML Schema standard and refer to the CLARIN Concept Registry [5] (previously ISOcat), a concept scheme model based on SKOS (Simple Knowledge Organization System). CMDI profiles used are published here:
http://catalog.clarin.eu/ds/ComponentRegistry/#
The repository database and HTML web pages use the UTF-8 encoding standard, for correct character encoding.
The repository uses ISO language codes in metadata and allows for a variety of data formats for the submitted digital objects, as described on the repository website, see [6,7].

Authenticity - Once deposited and archived, the submitted data sets can not be changed by the submitter nor by editors. As stated in the Distribution License Agreement, within the repository no alteration of the submitted data will be made. This ensures that data is authentic and it is also important for the assigned persistent identifiers, which must always refer to the same content. Only the administrators of the repository have the right to make changes, thus submitters should contact the help-desk for requesting changes (as is indicated at [1]). Our overall policy in this respect is to allow for changes in case minimal corrections to the metadata are needed, e.g. for typos. For non-trivial changes, a new version of the submission will be required. Anyway, each request of modification will be evaluated case-by-case.

At present, if a submission is superseded by a new version our preferred policy is to withdraw the old one but to keep the PID URL working and add a special metadata value ("isreplacedby") that points to the new version (that has the metadata "replaces"). In case there is a new version it is still possible to download the previous version using a link that appears with relations dc.relation.replaces [2]. For each change, anyway, the provenance metadata are stored including appropriate log messages.

Additionally, while for search and consultation purposes the repository is open, for depositing and/or describing resources access is restricted to authenticated users (see R4).

URL:
[1] Deposited Item Lifecycle:https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/item-lifecycle#modifying-item
[2] CLARIN-DSpace, New Version Guide: https://github.com/ufal/clarin-dspace/wiki/New-Version-Guide
[3] CMDI1.2 https://www.clarin.eu/content/cmdi-12
[4] VLO https://vlo.clarin.eu/
[5] CCR https://www.clarin.eu/ccr
[6]Frequently Asked Questions
[7] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/metadata
[8] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/faq#what-submissions-do-you-accept
[9] https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf

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:

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:

Collection development policy - The repository is structured in two main collections: one collection represents the institutional/project repository where data from inside the Institute of Computational Linguistics (CNR-ILC) are described and/or stored; the other collection is for describing and/or depositing data by any user of the CLARIN-community. Within both collections, 4 types of data can be submitted: (i) Corpora, (ii) Lexical-Conceptual Resources, (iii) Language descriptions, (iv) Software, tools, web-services (see [1]).

Quality control checks - The submission interface and workflow guide the user in providing relevant and complete (meta)data. If the minimum information required is not provided, the interface will not allow the user to complete the submission. After submission, the item is then reviewed by the metadata curators that will check for the quality of the metadata. Editors are also responsible for verifying that the data submitted corresponds to what it describes. A thorough check of the quality of the data however is not performed since it is beyond our mission and scope. As stated in the Distribution License Agreement, depositors are responsible for the quality of their data. The depositor is expected to deliver data living up to
academic standards. In case a submission does not comply with the expectations of the repository, it is returned to the data provider for rectification, via the interface.
The same quality checks (i.e. evaluation of the relevance of the data for the repository, metadata completeness and correctness, verification that the data described corresponds to the actual accessible/downloadable data), apply also to data that is not physically held by ILC4CLARIN but only cataloged.

As specified above, depositors are responsible for the quality of their data. This is especially true for services and tools that are not on the ILC4CLARIN servers. In such cases, depositors have to guarantee the operability and functionality of services and tools running on proprietary servers.

Metadata - The repository relies on the group of emerging metadata standards around CMDI (ISO-CD 24622-1) [5]; in particular the submission interface is based on this CMDI profile (use curl) [2]. This ensures that the metadata required to interpret and use the data are provided and are sufficient for long-term preservation.

Preferred formats - The repository recommends using standard data formats uploaded during submission. Especially for language resources recommended formats, depositors are referred to [3] from the FAQ page of ourILC4CLARIN repository[4] which provides the list of formats recommended by CLARIN. The validity of the submitted data sets is checked manually by an expert editor.

About the risk assessment approach to the recommended formats of submitted items, when the submitted items contain attached bitstreams, metadata curators manually verify whether they meet the requirements of integrity, authenticity, availability, and/or their restrictedness. Metadata curators are, obviously, in contact with the depositor(s) to obtain missing information if there is any.

If the format is unknown or not in the list of the recommended standard formats [3], it must be well documented and the documentation must be either part of the submission or the metadata must contain a link to it. However, the final decision on the acceptance/rejection of such submissions is taken by the ILC4CLARIN metadata committee in collaboration with our repository administrator.

We want to emphasize that the described procedure applies for submissions that include either metadata only or both data and metadata. For both of them, the quality checks apply to metadata, and when files are submitted --see R7 for integrity tools-- additional checks are performed. When submissions link to externally-maintained resources, according to R2, R3, and R4, the quality of such resources is due to the depositors. However, ILC4CLARIN has many academic partners, which, by their manifestos, guarantee the quality of the submitted data.

URL:
[1] Type of data: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/deposit
2] CMDI profile:http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1349361150622/xsd
[3] Standards for LRT: https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[4] FAQ:https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/faq?locale-attribute=en#what-submissions-do-you-accept
[5] CMDI ISO: https://www.iso.org/standard/37336.html

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 repository is in the implementation phase - 3

Response:

With the use of the DSPACE (one of the leading digital repository systems [1]) and the defined workflow supported by the repository's interface, the ILC4CLARIN repository meets the requirements of OAIS. For the first step, the ingestion process, the Submission Information Packages (SIPs) are received for curating and are assigned to a task pool where our curators can process them. There are several pre-configured supported IP formats (see [2]). However, the default way is that the ingestion process is done through our web-based interface which hides the implementation details. For the second step, the archival storage, one of our curators takes charge of the submission. Using a web interface, the metadata is updated (added, deleted, modified), the submitted bitstreams are validated. In general, the curators ensure the consistency and quality of each submission. If a curator approves an item, the Archival Information Package (AIP) is generated. We are open to all submissions which meet our standards (Data Producers must be authenticated which means they must have an academic background or have verified local accounts). A contract is signed during the ingestion process. We are using a specific robust administration interface including specific detailed reports on the contents of our repository. All backups follow standardized ways of using MD5 checksums for determining the consistency and we use automatic monitoring tools at various levels.

Regarding the long-term storage of digital data, the server storage configuration is a Redundant Array of Independent Disks, RAID5, which by itself ensures data redundancy.

In addition to that, ILC4CLARIN schedules night-based replicas of the repository, with automatic data consistency check, on a second twin server computer located in a private network environment. In case of failure of the repository, the replica can be on-line in less than 3 hours.

Moreover, the ILC4CLARIN centre implemented at CNR-ILC will receive funding in 2021 and 2022 from the Ministry of Economy and Finance (through the central CNR) devoted to strengthening the national nodes of the Research Infrastructures. An IT data center will be implemented -- based on the Converged Infrastructure approach -- to support long-term digital preservation and AI activities with high-performance storage solutions and high-performance computing [3].

URL:
[1] ILC4CLARIN on DURASPACE: https://duraspace.org/registry/entry/6015/
[2] SIP formats: https://wiki.lyrasis.org/display/DSDOC5x/Importing+and+Exporting+Content+via+Packages#ImportingandExportingContentviaPack
[3] ILC DATA CENTRE (page 16, nr 48): https://www.cnr.it/it/trasparenza/delibere-cda/documento/104615

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:

Compliance Level Comment: Accepted

R10 Preservation plan

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

Compliance level:

The repository is in the implementation phase - 3

Response:

According to [1] and [3], ILC4CLARIN has different rights on the submitted data. These rights include how to store, provide, and access the submitted data.

Indeed, we indicate to all the submitters of data [1] that “ILC4CLARIN is committed to the long-term care of items deposited in our repository and strives to adopt the current best practice in digital preservation”. This preservation function encompasses: taking delivery of the dataset ingested, storing it, and ensuring it is archived and accessible, and usable to the researcher community as is the mission of a CLARIN Centre[2]. Any other responsibility regarding custody or rights to copy, transform and store are presented to the depositor in the Distribution License Agreement [3]. During the submission process, the submitter agrees to and accepts our policy which leaves him or her the responsibility for the correctness and quality of his/her submission, its legal status and accessibility, and all related ethical issues, if any. DSPACE, and thus the CLARIN-DSpace repository software, provides two levels of digital preservation. The first approach is "bit preservation" which ensures the integrity of both data and metadata over time regardless of possible changes in the physical storage media; the second one is "functional preservation": even if the file may change over time it remains usable in the future by evolving its original digital format and media. Format migration is a straightforward strategy for functional preservation.

Our repository suggests using standard formats, which, by themselves, are well documented and widely used. Formats such as texts, images are easily preserved and migrated. The same can't be said for proprietary formats. In conclusion, the first type of format is classified as either supported or known [4], while the second one is classified as Unknown (generally, application/octet-stream). Many of either Supported or Known are, ultimately, mapped on formats mentioned in [5]. Proprietary formats which are harder to be functionally preserved undergo bit preservation.
In addition, the ILC4CLARIN main strategy includes migration (when possible) implementing best practices coming both from CLARIN [6], [7] and other initiatives, such as the SSHOC thematic cluster [8].

URL:
[1] Preservation policy: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/about#preservation-policy
[2] CLARIN Centre mission: https://www.clarin.eu/sites/default/files/centres-CLARIN-ShortGuide.pdf
[3] Distribution License Agreement: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/contract
[4] How does DSPACE preserve digital materials:
https://wiki.duraspace.org/display/DSPACE/User+FAQ#UserFAQ-HowdoesDSpacepreservedigitalmaterial?
[5] Standards for LRT: https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
[6] CLARIN preservation short guide: https://www.clarin.eu/sites/default/files/preservation-CLARIN-ShortGuide.pdf
[7] Data Management Plan: https://www.clarin-d.net/en/preparation/data-management-plan
[8] SSHOC: https://sshopencloud.eu/d16-sshoc-data-management-plan

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: Accepted

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 DSPACE repository implements an automatic curation system [3]. To the automatic curation system, is added manual curation. On one hand, the repository editors verify and validate each submission using automatic tools, on the other, they check the consistency between the values submitted and the submission itself.

The repository system requires a set of metadata attributes to provide information and the authorship about submitted data. The submission cannot be completed unless all the required metadata is filled out. The required metadata is different for different types of submitted data (i.e. corpus, lexical/conceptual resource, tool, language description).

All datasets/resources added to any collection are assessed for quality in the same way. Collections are simply virtual containers for resources belonging to a conceptually homogeneous pool (e.g. an institution, an archive, etc.). All resources within a collection are individually described with the full stack of metadata as described in [1].

During the process, appropriate explanations, examples, and suggestions are provided to the submitters to get high-quality metadata, and we have provided a web page to provide information about what metadata we require and how we disseminate it [1].

In conclusion, the basic set of validation is done by our automatic tools [3] and by the editor(s) responsible for the curation of the submission. The editor checks the quality of the content and if there are things that are not clear he/she either returns the data to the submitter for additional information or asks the research community connected with the repository for help.

Each submission is given a PID and we strongly encourage people to use it for citation of the resource in other works [2].

URL:
[1] About metadata: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/metadata
[2] How to Cite: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/cite
[3] https://wiki.lyrasis.org/display/DSDOC5x/Curation+System

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:

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:

After submitting the data, a curation platform, offered by and integrated into the CLARIN-DSpace software, is used to ensure the quality and consistency of the submission with the possibility to return the data to the submitter for changes.

Our curation framework requires three stages of manual metadata checks: a first basic quality check which ensures that the (meta)data is appropriate to the repository; a second stage where expert metadata curators thoroughly check the quality and appropriateness of the descriptions added by the depositor and edit them if necessary. The metadata expert(s) also check that the data attached, if any, corresponds to what is described, has an appropriate license, and follows the recommended standards. At this stage the experts may interact with the depositor via the curation platform, asking for integrations or changes. The software also integrates automatic tools that verify and validate the metadata according to the adopted scheme. Finally, at the third and final stage, an administrator of the repository performs a final formal check before definitively approving/promoting an entry into the repository.

After this stage, the data described and uploaded receives a Persistent identifier, becomes immediately visible, and retrievable via the repository web interface and the Virtual Language Observatory [1] at the next scheduled OAI-PMH harvesting.

Information on the submission and curation workflows can be found here: [2,3].

The workflow described below is valid for all 4 types of resources that the repository admits. Please note the involvements of the organization as in R5 (Organizational Infrastructure):

1. Depositors submit their data to ILC4CLARIN using the (customized) submission workflow defined in CLARIN-DSpace[2].

2. ILC4CLARIN community (helpdesk) curator receives a notification and performs a pre-acceptance appraisal; if necessary he or she contacts the depositor.

3. After performing all the checks provided in step 1 of CLARIN-DSpace, the helpdesk curator passes the submission to the ILC4CLARIN metadata curators.

4. Metadata curators open and review the submitted items, checking their quality at different levels such as the content of the metadata and any information related to Language Resources and Technology (LRT).

5. Curators work closely with the depositor in case of lack of quality, and/or licensing issues; at the end of the curation phase, the item is passed to the ILC4CLARIN repository manager.

6. The repository manager finalizes the submission in ILC4CLARIN and publishes the dataset with a persistent identifier (PID).

URL:
[1] CLARIN VLO: https://vlo.clarin.eu
[2] Deposit: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/deposit
[3] Item Lifecycle: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/item-lifecycle

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: Accepted

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

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:

Our repository, ILC4CLARIN, has an advanced tool for browsing and searching items powered by Solr based on full-text indexing of all text-based files in the repository as well as for faceted browsing of the repository metadata [1].

We are regularly harvested by various institutions that reuse the metadata we provide:
CLARIN ERIC, the reference RI, with the Virtual Language Observatory (VLO) [2] where
language resources may be discovered using a faceted search engine. ILC4CLARIN is also registered to different archive initiatives, see OLAC [3], Open Archives [4], DURASPACE [5], ROAR [6].

Each submission is given a PID and we strongly recommend people to use it for citation [7].

URL:
[1] Browsing interface: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/discover?advance
[2] CLARIN VLO: http://vlo.clarin.eu
[3] OLAC: http://www.language-archives.org/archive/dspace-clarin-it.ilc.cnr.it
[4] Open Archives:
http://www.openarchives.org/Register/BrowseSites?viewRecord=http://dspace-clarin-it.ilc.cnr.it/repository/oai/request
[5] DURASPACE: https://duraspace.org/registry/entry/6015/
[6] ROAR: http://roar.eprints.org/12771
[7] How to Cite: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/cite

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Compliance Level Comment: Accepted

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:

Data reuse is one of the main components of ILC4CLARIN. The repository requires that a set of metadata (both mandatory and recommended), providing information about the submitted data and the authorship, be filled during the submission phase. Mandatory and suggested metadata are described in [1] (many of those fields see are regex managed and/or pre-filled with autocompletion). In [8] (login required) is possible to see and manage the set of metadata used to describe the DSPACE items. DSPACE items are essentially qualified dublincore and dcterms (to which other schemas are added).

In addition, the submission form of the repository graphically distinguishes required and optional metadata. The descriptive metadata, the license information, how to cite, and so on are visible on the item view page (cf. [6])

Data depositors are asked to fill different sets of metadata according to the specific type of data (e.g., corpus, tool, language description) submitted. All these metadata comply with CMDI profiles/schemas [2] and are mapped to various metadata schemas [3]. The formats listed in [3] are automatically generated by the software so that new schema and/or future evolution of metadata formats can be easily sustained. Users are informed by repository curators that their items have been ingested into the repository. [9] is an example of the full list of metadata about a DSPACE item.

If a submission contains a physical file, this can be either uploaded according to [5] (which are updated by the community) or to other formats (see R10). In such a case, the repository manager and the metadata curators must be in touch with the submitters to understand how to manage the data. The ILC4CLARIN staff has the possibility of changing the data format (see [7]), taking into account supported/known/unknown formats managed by the underlying
CLARIN-DSPACE software [6].

The supported harvesting protocols such as OAI-PMH, OAI-ORE, METS... rely on the metadata formats provided in [3]

Regarding the understandability of the data, we believe that the binding set of information required during the deposit ensures that we have items with clearly legible data [4]. We encourage our depositors to upload files in LRT standard formats [5] suitable for long-term preservation and constantly updated by LRT experts.

Regarding our plans for the future migration of formats, ILC4CLARIN makes heavy use of standard formats. This makes ILC4CLARIN aware of emerging international standards and community-approved data formats and enables us to keep up to date with the current best practices for migrating data to new formats when it happens to be necessary and feasible, see R10.

ILC4CLARIN supports resubmission of data sets (new versions and/or enriched versions of data); the repository software is in charge of keeping track of relations between different versions and/or different data sets through a subset of dedicated metadata.

URL:
[1] About metadata: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/metadata
[2] CLARIN Component Registry:
https://catalog.clarin.eu/ds/ComponentRegistry?registrySpace=published&itemId=clarin.eu:cr1:p_1403526079380
[3] ILC4CLARIN list of metadata formats:
http://dspace-clarin-it.ilc.cnr.it/repository/oai/request?verb=ListMetadataFormats
[4] ILC4CLARIN How to Deposit: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/deposit
[5] Standards for LRT: https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf

[6]
https://wiki.lyrasis.org/display/DSPACE/User+FAQ#UserFAQ-HowdoesDSpacepreservedigitalmaterial?
[7] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/contract
[8] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/admin/metadata-registry
[9] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-548?show=full

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:

At the time of writing, the core of the repository infrastructural software is DSpace 5.11.
The LINDAT/CLARIN Centre [1] has developed and maintained a modified version named CLARIN-DSpace for the CLARIN community [2].

A software inventory has been maintained and the system documentation is available [3]. The software is supported and used by a community of CLARIN Centres that grows every year and periodically meets under the aegis of CLARIN [4].

Customization for the local repository, including licenses, logos is available at [12].

Regarding the standards that the repository uses for reference, we follow the list of standards that are relevant for the CLARIN community [5].

ILC4CLARIN is hosted, and research data stored, on two Dell PowerEdge R630 Rack Servers [6] in the active/passive configuration. The servers are configured in a High Availability (HA) arrangement [9], with the addition of data replication using the Distributed Replicated Storage System DRBD [10]. The images of the virtual machines where the whole stack runs are backed up.

Regarding network performance, ILC4CLARIN is connected to the GARR Network [7], the broadband network infrastructure accountable for ensuring availability, bandwidth, and connectivity to the Italian community of Education and Research, which is our Designed Community. It’s also connected to GÉANT, the pan-European research and education network [8].

The local infrastructure is completed with a Synology NAS [11] hosted at a different CNR institute in a different building.
Both the location of the servers and of the backups are hosted within the CNR Research Area of Pisa [13] which provides network security, monitoring, and protection (such as firewalls).

Moreover, the CLARIN-IT National Consortium manages storage on the GARR Cloud [14], where our data are weekly replicated.

The disaster and recovery plans take into account the following aspects:
1) Software repository GitHub
2) Customization GitHub
3) Periodic Backups
4) Distinct locations for server and backups
The points above allow us to restore the software and the data in case of failure.
Additional details on the actions above are reported in [15]

URL:
[1] LINDAT/CLARIN: https://lindat.mff.cuni.cz/en
[2] CLARIN-DSpace: https://github.com/ufal/clarin-dspace
[3] CLARIN-DSpace wiki: https://github.com/ufal/clarin-dspace/wiki
[4] CLARIN workshop on DSpace: https://www.clarin.eu/event/2016/clarin-workshop-dspace-digital-repository
[5] Standards and Formats: https://www.clarin.eu/content/standards-and-formats
[6] Dell PowerEdge R630: http://www.dell.com/en-us/work/shop/productdetails/poweredge-r630
[7] GARR Network: https://www.garr.it/en/garr-en
[8] GÉANT topology map:
https://www.geant.org/Networks/Pan-European_network/Pages/GEANT_topology_map.aspx
[9] Corosync Cluster Engine http://corosync.github.io/corosync/
[10] DRDB https://linbit.com/drbd/
[11] https://www.synology.com/en-global/products/RS819
[12] https://github.com/cnr-ilc/ilc4clarin-overlays
[13] http://www.area.pi.cnr.it/
[14] https://cloud.garr.it/
[15] https://ilc4clarin.ilc.cnr.it/en/wp-content/uploads/sites/1/2023/07/ILC4CLARIN_Draft-Disaster-Plan.pdf

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:

In ILC4CLARIN we have taken the necessary precautions to ensure that the data housed in our repository is protected and secure.

Following the GARR network guidelines, regular penetration testing is carried out to ensure service is secure against attack. Also, we periodically perform the https://www.ssllabs.com/ssltest tests to maintain the correct security level (for example Apache patches).

For recovery, we use two application servers (cf. R15) in active/passive configuration that provides a safe environment to run virtualized services.

Data, metadata, and software are fully backed up every week using a Synology NAS [4] hosted at a different CNR institute.
Both the location of the servers and of the backups are hosted within the CNR Research Area of Pisa [13] which provides network security, monitoring, and protection (such as firewalls).

The repository administrator actively monitors the log stats to prevent malicious behavior such as artificially inflating download counts or systematic attacks. In addition, the system is monitored by icinga2. [1]

We use single sign-on (SSO) to log the users on the system. This implies that the details on the logging users are stored in their identity providers, as described in more detail in our privacy policy [2].

As part of CLARIN Authentication and Authorization Infrastructure, if there is a security incident we will report it using SIRTFI - REFEDS [3].

URL:
[1] icinga2: https://monitoring.clarin.eu/monitoring/service/show?host=ILC4CLARIN
[2] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/privacypolicy
[3] https://refeds.org/sirtfi

[4] https://www.synology.com/en-global/products/RS819

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:

We have
(i) revised the answers according to the reviewers’ comments;
(ii) added clarifications to some of the reviewers’ comments, see the Comment area below
(iii) attached is the XSD file downloaded from http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1349361150622/xsd, for Reviewer 2 (R8)

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:

-