Implementation of the CoreTrustSeal

The CoreTrustSeal board hereby confirms that the Trusted Digital repository The ILC4CLARIN Centre at the Institute for Computational Linguistics complies with the guidelines version 2017-2019 set by the CoreTrustSeal Board.
The afore-mentioned repository has therefore acquired the CoreTrustSeal of 2016 on April 18, 2018.

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

Yours sincerely,

The CoreTrustSeal Board

Assessment Information

Guidelines Version:	2017-2019 | November 10, 2016
Guidelines Information Booklet:	DSA-booklet_2017-2019.pdf
All Guidelines Documentation:	Documentation
Repository:	The ILC4CLARIN Centre at the Institute for Computational Linguistics
Seal Acquiry Date:	Apr. 18, 2018
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.coretrustseal.org/seals/
Previously Acquired Seals:	None
This repository is owned by:	• **Institute for Computational Linguistics "Antonio Zampolli"** Pisa Italy T +39-50-3158379 F +39-50-3152839 E ilc4clarin@ilc.cnr.it W http://www.ilc.cnr.it/

Assessment

0. Context

Applicant Entry

Self-assessment statement:

Context:
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, whose main service is the digital repository of language resources, tools and (web) services available at https://dspace-clarin-it.ilc.cnr.it
CNR-ILC is a centre of reference in the field of Computational Linguistics at both national and international levels. It’s part of the Department of Social Science and Humanities, Cultural Heritage (DSU)[4] of CNR and carries out research activities in strategic scientific areas of the discipline, as well as publishing activities, training and education activities and technology transfer.

(1) Repository Type:
The ILC4CLARIN repository is an institutional repository regularly harvested by the Virtual Language Observatory, 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 resource and technology within the CLARIN research infrastructure and has been implemented/used by several other CLARIN centers all over Europe.
The repository is mainly a collection of linguistic data and NLP tools developed at ILC and by other members of the CLARIN-IT Consortium[1]; its data focuses on the Italian language, but does not exclude other languages and classical/historical languages. Linguistic data and tools cover many subject areas and domains, e.g. legal domain, biology, physics … From the data producer’s point of view, the repository focuses on an user-friendly interface which 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, tools and services.

(2) Repository Designated Community:
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 long term preservation of both data and tools according to the definition of Preservation Description Information (PDI) given in the OAIS reference model[8].

(3) Level of Curation:
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 quality and consistence 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.

(4) Outsourcing:
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] and eduGAIN service [13] (through Géant) and user involvement for the Italian scientific and academic community.

(5) Other Relevant Information:
ILC4CLARIN is listed in the Registry of Data Repositories re3Data.org under ID:r3d100012262 [14] and indexed by SHARE and the Web of Science Data Citation Index.

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://www.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] eduGAIN membership status: http://www.edugain.org/technical/status.php
[14] Re3data.org record URL for ILC4CLARIN: http://doi.org/10.17616/R3W365

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

1. Mission/Scope

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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 researchers who are not affiliated. The focus is on the Italian language, but other languages, especially historical and classic ones, are not excluded. Such mission is 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 unified single sign-on access to language resources and technology. ILC4CLARIN is in fact part of the CLARIN networked federation of centres and is currently applying for B-type status [3]. Explicit statements of such mission can be found here [4].

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 center 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 center is implemented, developed and maintained by the Institute for Computational Linguistics “Antonio Zampolli” (CNR-ILC), a center 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. In fact our institute 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 1990’s[6].

See [5] for statements on the main activities and mission of CNR-ILC and [6] 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?page=4
[7] CNR-ILC: http://ilc.cnr.it/en/content/institute
[8] CNR-DSU: http://www.dsu.cnr.it/?page_id=18

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

2. Licenses

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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 submitter affirms that they have the right to submit the data and gives us (the repository center) the right 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].
Also, in order 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 a number of 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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

3. Continuity of access

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Italy became a member of CLARIN ERIC in October 2015, 3 years after CLARIN was made an ERIC. The Italian Ministry of Education and Research awarded CNR a mandate to represent Italy as a representing 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 mandate awarded by the Ministry has a duration of five years. The Ministry has committed to the payment of the membership fees for five years and 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 under the form of in kind contributions. The continued availability and accessibility of the data in the repository is guaranteed (along with documentation) until 2020. 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), the data contained in the repository will be transferred to one of its sister repositories in another CLARIN data center.
Again, in a worst case scenario, thanks to the recommendations agreed upon by the Italian CLARIN-IT Consortium to use the same software for each CLARIN-IT repository, one 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.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. Confidentiality/Ethics

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

During the submission process, the submitter agrees to and accepts the repository policy leaving him/her all the responsibility regarding his/her submission.
In particular, the distribution license says[1]:
"You represent that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not,[...], infringe upon anyone's copyright. If the submission contains material for which you do not hold copyright, you represent that you have obtained the unrestricted permission [...] "

Additionally, each submission is verified and validated both using automatic tools and manually by a repository editor, whose task includes a check for respect of privacy and ethical issues [2]. If the editor detects problems with the submission he/she will interact with the depositor, via the web interface, and request/suggest modifications or integrations. The submission will not be approved until all the fundamental requirements are met.

Any submitter must be an authenticated user through Shibboleth (where we manage the list of IdPs - Identity Providers), which is important because it makes him/her traceable. This ensures a high level of trust as the submitter comes from a list of "well known" submitters. No anonymous items are allowed.

Disclosure Risk Data: we expect a low proportion of data deposited to the ILC4CLARIN repository to raise confidentiality and ethical issues, and in particular disclosure risks. In any case, data providers 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 may choose (or be advised during the curation steps) to distribute such data under restricted access (i.e. limited to academic use/research). This way the data will be protected via (shibboleth) authentication, so that it will only be available to scholars that are able to log-in through Identity Providers operated at institutions taking part in the CLARIN AAI federations.

The CLARIN Legal Issues Committee (CLIC) [3], set up and run by CLARIN ERIC, organises periodic training sessions in management of data with a disclosure risk especially during the "Clarin Annual Conference".

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 are made available with a licence that requires signing, the user is asked to electronically sign the licence before downloading or accessing the resource. Furthermore, in case of misuse, we can retrieve the exact dates and specific id's of people who have accessed the resources; such users will be blacklisted, they can be denied further access to the repository and the research community might be made aware of the misuse.

URL:
[1] Distribution license: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/contract
[2] Deposited Item Lifecycle: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/item-lifecycle at the section

‘Edited Item’
[3] CLARIN Legal Issues Committee: https://www.clarin.eu/content/legal-information-platform

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

At your next renewal of CoreTrustSeal we would recommend a documented approach to managing disclosure risk and to breaches of licence conditions.

5. Organizational infrastructure

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The 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 supported financially by external funding, 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 organised and supported by CLARIN and CLARIN-PLUS[1]. The expertise and experience of ILC staff is 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 ILC4CLARIN staff currently includes 5 units of personnel, with different roles:

ILC4CLARIN Coordinator (0.3 FTE)
Technical manager (0.3 FTE)
Repository Manager (0.3 FTE)
Repository Developers (0.2 FTE)
Metadata curators (two units, 0.1 FTE each, thus 0.2 FTE)
Helpdesk curator (0.05 FTE)
Preservation managers (two units, 0.05 FTE each, thus 0.1 FTE)

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

As executing center we certainly need further financial support to move from level 3, "in progress", to level 4 of statement of compliance. We have asked the Ministry of Education and Research to support our activities with a National Project in order to have more staff and more dedicated people.

URL:
[1] CLARIN-PLUS: https://www.clarin.eu/content/factsheet-clarin-plus

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

At your next renewal of CoreTrustSeal we would recommend a documented approach to managing disclosure risk and to breaches of licence conditions.

6. Expert guidance

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

ILC4CLARIN 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 TC37 SC4 [1] and expert members of some of its committees, nominated by the Italian standardization body, UNI [2].
CNR hosts and promotes several (self)archiving systems and data management activities/initiatives, for instance: 'People' is a self-archiving platform that offers CNR researchers a virtual environment to describe, organise and self-archive their research outcomes; the repository Open Access PUMA [3] was developed by our colleagues at ISTI in Pisa; Solar (Scientific Open-access Literature Archive and Repository) [4] is developed and maintained by the CNR central library.
Advice can thus be received on various topics by expert staff involved these and other relevant initiatives.

External expert guidance:
ILC4CLARIN has contacts with ICCU Istituto Centrale per il Catalogo Unico within the PARTHENOS project and the CLARIN-DARIAH partnership.

Most importantly, 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 [5].

ILC4CLARIN is also the national data center of the Italian CLARIN Consortium (CLARIN-IT) [6], whose coordinator regularly participates to 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 Centers [7] which coordinates the activities of all CLARIN technical centers 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 virtue of these roles, ILC4CLARIN can seek advice from any expert 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).

URL:
[1] ISO/TC 37/SC 4: https://www.iso.org/committee/297592.html
[2] UNI: http://www.uni.com
[3] PUMA repository: http://www.opendoar.org/find.php?rID=2367 , http://pumalab.isti.cnr.it/index.php/it
[4] Solar repository: http://www.opendoar.org/find.php?rID=1283 ,
https://bice.cnr.it/en/?option=com_content&view=article&id=177&Itemid=185
[5] CLARIN B and K centres: https://www.clarin.eu/content/clarin-centres

[6] CLARIN-IT Consortium: http://www.clarin-it.it
[7] Clarin Standing Committee: https://www.clarin.eu/governance/standing-committee-clarin-technical-centres

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. Data integrity and authenticity

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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 are 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 for whether it is supported, known or unknown. 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.

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 rights 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 the 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) which points to the new version. In case there is a new version it's 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, to ensure authenticity, submitters can only be authenticated users of the repository, thus people authorised by well defined authorities e.g., eduGain using shibboleth.

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

8. Appraisal

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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-IT community. Within both collection 4 types of data can be submitted: Corpora, Lexical-Conceptual Resources, Language descriptions, 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 three human experts that will check for the quality of the metadata. Editors are also responsible of verifying that the data submitted actually corresponds to what is 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, submitters are responsible for the quality of their data.
In case the submission does not comply with our expectations the submission is returned to the data provider for rectification, via the interface.

Metadata: The repository relies on the group of emerging metadata standards around CMDI (ISO-CD 24622-1); in particular the submission interface is based on this CMDI profile[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 to use standard data formats uploaded during submission. Especially for language resources recommended formats, depositors are referred to [3] from the FAQ page of our ILC4CLARIN 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 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) in order 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 acceptance/rejection of such submissions is taken by the ILC4CLARIN metadata committee in collaboration with our repository administrator.

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

9. Documented storage procedures

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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 is a number of pre-configured supported SIP 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 are updated (added, deleted, modified), the submitted bitstreams are validated. In general, the curators ensure consistency and quality of each submission. If a curator approves an item, the Archival Information Packages (AIPs) is available.
We are open to all submissions which meet our standards (Data Producers must be authenticated which means they must have an academic background or have verified local accounts). A contract is signed during the ingestion process. We are using a specific robust administration interface including specific detailed reports on the contents of our repository.
All backups follow standardised ways of using MD5 checksums for determining the consistency and we use automatic monitoring tools at various levels.
Regarding the long-term storage of digital data, the server storage is a raid-5 configuration and we use the backup2l utility for taking daily backups as suggested by Clarin-DSpace software[3].
Currently, we are working on setting up preservation policy for each submission using Eudat replication throught B2SAFE[4]: we will automatically replicate each submission after approval. The submission will be converted to AIP format and uploaded to an iRods server; we will use our PID in the name of each AIP.
In addition to the previous strategies, ILC4CLARIN schedules a 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.

URL:
[1] ILC4CLARIN on DURASPACE: http://registry.duraspace.org/registry/repository/7890
[2] SIP formats:
https://wiki.duraspace.org/display/DSDOC18/Importing+and+Exporting+Content+via+Packages#ImportingandExportingContentviaPack
[3] Clarin-DSpace Backup: https://github.com/ufal/clarin-dspace/wiki/Backup
[4] Clarin-DSpace Eudat replication: https://github.com/ufal/clarin-dspace/wiki/EudatReplication

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. Preservation plan

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

We indicate to all the submitters of data in [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 clearly 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 Clarin-DSpace repository software, provides two levels of digital preservation. The first approch is "bit preservation" which ensures the integrity of both data and metadata over time regardless possible changes in the physical storage media; the second one is "functional preservation": even if the file may change over time it remains usable in the future by evolving its original digital format and media. Format migration is a straightforward strategy for functional preservation.
The Clarin-DSpace repository software provides some default values for Supported, Known and Unknown formats[4], many of them are, ultimately, mapped on formats mentioned in [5] and data depositor is urged to use such formats when submitting his/her data.
Those formats, used by ILC4CLARIN, are widely accepted by LRT community and have been chosen to take into account the challenges of long-term preservation. One of these challenges is format migration. ILC4CLARIN main strategy includes migration (when possible) implementing best practices coming both from CLARIN[6], [7] and other initiatives, such as the PREFORMA research project[8] [9].

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] PREFORMA research project: http://www.preforma-project.eu
[9] Digital curation and quality standards for memory institutions: https://link.springer.com/article/10.1007/s10502-015-9242-8

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

11. Data quality

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

In our complex curation framework with several editors, each submission is verified and validated using automatic tools and manually by a repository editor.
We require a set of metadata attributes providing information about submitted data and the authorship to be filled-in. The submission cannot be completed unless all the required metadata is filled out. The required metadata are different for different types of submitted data (i.e. corpus, lexical/conceptual resource, tool, language description), but the validation of the metadata is automatic for the different types.
During the process appropriate explanations, examples and suggestions are provided to the submitters in order 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].
The basic set of validation is done by our automatic tools 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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

12. Workflows

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

After submitting the data, a curation platform, offered by and integrated into the Clarin-DSpace software, is employed to ensure the quality and consistency of the submission with the possibility to return the data to the submitter for changes.
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, if relevant, receives a Persistent identifier, becomes immediately visible and retrievable via the repository web interface and via 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].

A synthetic workflow for submitting items to ILC4CLARIN is the following (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 helpdesk curator receives 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 ILC4CLARIN repository manager.
6. The repository manager finalizes the submission in ILC4CLARIN and publishes the dataset with a persistent identifier (PID).

The workflow described above is valid for all the 4 types of resources that the repository accepts.

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Accepted with the requirement that the renewal of the CoreTrustSeal in future includes the provision of public links to documented workflows

13. Data discovery and identification

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Our repository 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 several institutions which reuse the metadata provided by our repository, first of all our main research infrastructure CLARIN ERIC with the Virtual Language Observatory (VLO) [2] where language resources may be discovered using a facet browser. Our repository is also registered to different archive initiatives. [3], [4], [5], [6]
Each submission is given a PID and we strongly encourage 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: http://registry.duraspace.org/registry/repository/7890
[6] ROAR: http://roar.eprints.org/12771
[7] How to Cite: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/cite

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

14. Data reuse

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Data reuse is a main component of our repository. ILC4CLARIN requires that a set of metadata (both mandatory and recommended) providing information about the submitted data and the authorship be filled in [1]. Data depositors are asked to fill different sets of metadata according to the specific type of data (e.g., corpus, tool, language description) submitted. Howerer, all metadata comply with CMDI profiles/schemas [2] and most of them map to Dublin Core. A submission cannot be completed unless all the mandatory metadata are filled in.
We support OAI-PMH, OAI-ORE and several other specific protocols of metadata and data sharing. We offer the items to the community in different formats, from the standard Dublin Core to CMDI.[3]

Regarding 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]. In this page, 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 such 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 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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. Technical infrastructure

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The core repository infrastructural software is DSpace 5.x. 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 Centers that grows every year and periodically meets under the aegis of CLARIN. [4]
Regarding the standards that the repository uses for reference, we intend to follow the long 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], connected to form an HA cluster infrastructure using the Corosync Cluster Engine [7] and the data ad replicated using the DRBD storage system. [8]
Regarding network performance, ILC4CLARIN is connected to the GARR Network [9], the broadband network infrastructure dedicated to the italian community of Education and Research. It's connected to GÉANT, the pan-European research and education network. [10]

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] Corosync: https://en.wikipedia.org/wiki/Corosync_Cluster_Engine
[8] DRBD: https://en.wikipedia.org/wiki/Distributed_Replicated_Block_Device
[9] GARR Network: https://www.garr.it/en
[10] GÉANT topology map:
https://www.geant.org/Networks/Pan-European_network/Pages/GEANT_topology_map.aspx

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. Security

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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.
For recovery, we use two application servers, with automatic failover, that provide a safe environment to run virtualized services. A service is defined here as the application and underlying operating system.
For backup, the data and metadata are backed up every week with daily incremental updates. All backups follow standardised ways of using MD5 checksums for determining the consistency and we also use automatic monitoring tools at various levels.
The submissions are replicated using Eudat Replication [1], as recommended by clarin-dspace software.
The repository administrator actively monitors the log stats to prevent malicious behavior such as artificially inflating download counts or systematic attacks.
As part of CLARIN Authentication and Authorization Infrastructure, if there is a security incident we will report it using SIRTFI - REFEDS [2].

URL:
[1] https://github.com/ufal/clarin-dspace/wiki/EudatReplication
[2] https://refeds.org/sirtfi

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

17. Comments/feedback

Minimum Required Statement of Compliance:

0. N/A: Not Applicable.

Applicant Entry

Statement of Compliance:

0. N/A: Not Applicable.

Self-assessment statement:

Compiling of this Data Seal of Approval has helped us to reflect on many aspects of our repository configuration and to work on improving the quality of the data provided to the community and its visibility on various other Open Access Registries.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: