Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository Clarin-PL Repository complies with the guidelines version 2014-2017 set by the Data Seal of Approval Board.
The afore-mentioned repository has therefore acquired the Data Seal of Approval of 2013 on April 15, 2015.

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

Yours sincerely,

The Data Seal of Approval Board

Assessment Information

Guidelines Version:	2014-2017 | July 19, 2013
Guidelines Information Booklet:	DSA-booklet_2014-2017.pdf
All Guidelines Documentation:	Documentation
Repository:	Clarin-PL Repository
Seal Acquiry Date:	Apr. 15, 2015
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.datasealofapproval.org/seals/
Previously Acquired Seals:	None
This repository is owned by:	• **Wroclaw University of Technology** Poland T +48603999456 E marcin.pol@pwr.edu.pl W http://clarin-pl.eu/

Assessment

0. Repository Context

Applicant Entry

Self-assessment statement:

CLARIN-PL, Institute of Informatics, Wroclaw University of Technology digital library is available at https://clarin-pl.eu/dspace. The library has been developed by the CLARIN-PL department.

Using D-SPACE UI, from the data producer point of view, the repository focuses on an easy to use user interface for e-publishing.

From the data consumer point of view, the repository offers advanced searching and browsing of the available resources. The submissions are regularly harvested by several other projects using OAI-PMH (OAI-ORE) protocol in order to offer additional ways to find the resources in the repository.

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

The system is based on D-SPACE which tries to follow the OAIS (Open Archival Information System) reference model. It follow the standard principles of a high quality digital repository like usage of the Persistent Identifiers with handle.net, authorization and authentication using shibboleth authentication, sharing of metadata and data.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data, and compliance with disciplinary and ethical norms.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

In our curation framework with several editors, each submission is verified and validated using automatic tools and manually by a repository editor.

A submitter is an authenticated user either through Shibboleth or using a local account which is create only after validation. This ensures a basic level of trust which can be further increased if the submitter is from a list of “well known” submitters.

If the submitter is outside of the repository’s well known submitters, special care is taken to validate the input.

We encourage submitters to use open licences, such as Creative Commons, but for legacy and other exceptional reasons, we allow data to be associated with older types of public or private licences. This policy of maximal openness allows for any party to assess the scientific and scholarly quality of data as much as possible, which is common practice in the area of language resources.

Clarin-PL 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 (e.g., corpus, tool, language description).

During the submission process, the submitter agrees and accepts our policy leaving him the responsibility for the correctness of his/her submission, their legal status and accessibility and all related ethical issues, if any. Nevertheless, basic set of validation is done by our automatic tools and the editor responsible for particular submissions. The editor checks the quality of the content and if there are unclarities he/she either returns the data to the submitter for additional information or asks the research community connected with the repository (Institute of Formal and Applied Linguistics, Charles University in Prague) for help.

Each submission is given a PID and we strongly encourage people to use it when citing (see https://clarin-pl.eu/dspace/page/citate). We support OAI-PMH, OAI-ORE and several other specific protocols of metadata and data sharing. We offer different formats from the standard dublin core to CMDI. We are currently regularly harvested by several institutions which reuse the metadata provided by our repository (e.g., http://www.clarin.eu/vlo/, Google Scholar).

We allow for browsing and searching in the submission content using our internal search platform.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

2. The data producer provides the data in formats recommended by the data repository.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

An item can be exported in BibTeX format and in CMDI format.

We show a recommendation to use standard formats when uploading files during submission workflow e.g., for language resources we show http://www.clarin.eu/node/2320. Usage of standardised formats is encouraged but not enforced. The validity is checked manually by an editor.

If the format is unknown, it must be well documented and the documentation must be either part of the submission or the metadata must contain a link to it. The repository automatically performs regular checks on the integrity and the file formats of data. The report is sent to the 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.

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

3. The data producer provides the data together with the metadata requested by the data repository.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

https://clarin-pl.eu/dspace/browse

Data are submitted to the repository using D-Space submission graphical user interface (https://clarin-pl.eu/dspace/page/deposit). The submission worflow consists of several steps where the data producer must enter mandatory metadata otherwise user is not able to process to the next step. There are exceptions when submissions can make sense without real data but in this case, the submitter must clearly state the reasons and link to the place where the data can be acquired. The input metadata format is hidden from the user in the graphical user interface.

Another option is to automatically import metadata (data) from repositories which support standard protocols for sharing (e.g., OAI-PMH, OAI-ORE, DSpace Archival Information Package).

During the submission we require that the user provides at least the following information:

type of the resource - currently allowing only 4 types (corpora, tools, language conceptual resources, language descriptions)

title

list of authors

issue date

description

publisher

(if applicable) resource language(s) code(s)

contact person (the responsible person for the submission information) - at least surname and email

distribution information - access rights, license information, license restrictions, distribution media

content information - type of media (eg. text/audio/...), (if applicable) further classification of the resource (eg. onthology/thesaurus for lexical conceptual resources)

size information - size in bytes/words/n-grams/... (if applicable)

Description of the resource as required by the (minimal) metashare schema (http://www.meta-net.eu/meta-share/metadata-schema) and/or our CMDI profile (schema) - http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1349361150622/xsd

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

We have an explicit mission to archive language resources from all around the world.

This mission is supported by integration of the repository into the national and international CLARIN infrastructures (http://www.clarin.eu/files/centres-CLARIN-ShortGuide.pdf). As part of the CLARIN infrastructure, the repository is included by all promotional activities carried out at the national level of CLARIN-Lindat as well as the European level of CLARIN.

The repository implements standard protocols for sharing metadata and data. Public submissions can be easily mirrored. Protected submissions can be mirrored after legal requirements are met. One of the case studies of mirroring submissions from our repository is mirroring to repository provided by META-Share project.

http://www.clarin-pl.eu/en/clarin-pl-4/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The repository is not a legal entity on its own but is a part of CLARIN-PL.

The repository requires submitters to electronically sign the right to archive the data and the that the responsibility of the content lies with them.

After submitting an item, the editors validate the submission before making it public. The repository enables the submitters to restrict the access to their resources at various levels. This include assigning licences to the submissions which must be electronically signed by authenticated users. The signature information is archived.

For every deposit, we enter into a standard contract with the submitter, the so-called "Deposition License Agreement", in which we describe our rights and duties and the submitter acknowledges that they have the right to submit the data and gives us (the repository centre) right to distribute the data on their behalf.

Everyone who downloads data is bound by the licence assigned to the item – in order to download protected data, one has to be authenticated and needs to electronically sign the licence.

For submitters, there is a possibility for setting custom licences to items during the submission workflow.

Contracts are available at http://www.clarin-pl.eu/en/contact/.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

6. The data repository applies documented processes and procedures for managing data storage.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The data and metadata of submissions, the digital repository software and the underlying OS (Operating System) are crucial components of CLARIN-PL digital repository. Each component has specific backup policy.

The preservation policy relates to backup policies above and to the fact that our digital repository uses DSpace software which defines the preservation policy like this https://wiki.duraspace.org/display/DSPACE/User+FAQ#UserFAQ-HowdoesDSpacepreservedigitalmaterial?

At the infrastructure level, we have three components:

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

We use complete automation tool for managing Xen server pools which utilize the XAPI management interface and toolstack.

Our software suite provides complete HA features within a given pool. The overall design is intended to be lightweight with no compromise of system stability.

HA is provided with built in logic for detecting and recovering failed services.

We have Two virtual machine servers, with automatic failover, provide safe environment to run services. Service is defined as the application and underlying operating system.

Features of Our scripts for XenServer

- Auto-start of any failed VMs

- Auto-start of any VMs on after reboot

- Detection of failed hosts and automated recovery of any affected VMs

- Detect and clean up orphaned resources after a failed host is removed

- Removal of any failed hosts from pool with takeover of services

2) Hardware: HA Cluster

Data storage subsystem is build on IBM Storwize V7000 with redundant dual-active intelligent controllers. Storage is using RAID10 volumes (Redundant Array of Independent Disks; in this mode each chunk of data is repeated).

3) Hardware: Backup

Data backup is implemented on DS3500 Storwize V7000, ProtecTIER 6710 IBM System with deduplication mode. System is configure to create complete data snapshot every Sunday.

Our D-Space repository is stored on XenServer virtual machine. Single point failure at the data storage subsystem does not affect running D-Space repository service instance at all. Single point failure of the primary application server will initiate reconnecting to redundant second controller to another application server and restarting of the D-Space repository service.

The policy described above applies for the digital repository and the data and metadata as well.

The digital repository software source code is publicly available and is stored in multiple places on multiple machines. The content of the digital repository is backed up to the ProtecTIER every week (for the last month) including daily incremental updates using standard backup tools and can be restored using automatic tools.

All backups follow standardized ways of using MD5 checksums for determining the consistency and we use automatic monitoring tools at various levels.

All warnings and logs are send to administrators via an e-mail every Sunday.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. The data repository has a plan for long-term preservation of its digital assets.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Our repository is based on D-Space repository system which is one of the leading software in this category.

D-Space supports state-of-the-art preservation tools in various forms. From simple replication to standard backup formats and easily manageable collections.

The metadata can be exported into various formats suited for long time preservation including self describing ones like XML. Multilingual support is secured by using Unicode at every level. The XML format is used at several occasions e.g., when exporting to specific CMDI (Component MetaData Infrastructure) profile or when archiving AIP (Archival Information Packages).

The format validation is done regularly using external harvesting service (http://validator.oaipmh.com/) and it is available at http://catalog.clarin.eu/vlo/search?5&fq=collection:CLARIN-PL

We support standard metadata/data sharing protocols (e.g., OAI-PMH, OAI-ORE, DSpace AIP) which allows for duplicating our repository easily. This has been proved by a use case to mirror the contents of our main repository in META-Share repository node.

After submission, the editors validate each submission and check the uploaded files. The editors also allow binary formats if the submitter provides good reasons. Automatic summary of the file formats present in our repository gives us a good overview of what file formats are really used.

Editors have several tools available which help them to validate the submission. Firstly, the submission metadata are listed and can be edited. Then, the standard DSpace curaton framework was made available (https://wiki.duraspace.org/display/DSDOC18/Curation+System#CurationSystem-StarterTasks) which include checks for known/supported file formats, requred metadata, link checkers and our internal checks.

Files are checked three times (not necessarily by editors). The file extensions (file format) is checked and marked whether it is supported, known or unknown. The file integrity is checked for several supported and known types regularly. Finally, md5 checksums are checked regulary to ensure the consistency if submission.

The item lifecycle is described at https://clarin-pl.eu/dspace/page/item-lifecycle

We try to minimize the cases in which obsolescence of file formats occur in the near future by encouraging data submitters to use standardized formats. By enforcing a detailed and exhaustive documentation in case proprietary/"custom" formats are used we ensure long-term preservation sustainability. Thus it will, at least, be possible to specify and implement data converters.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

8. Archiving takes place according to explicit work flows across the data life cycle.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The D-SPACE submission workflow is internally configured in our repository and the submitter goes through each of them. We have automatic tools helping the editors to verify and validate metadata and the integrity of the submitted data which are performed by every editor during the curation step and automatically at regular time intervals.

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

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The licences with its full text must be signed at the end of each submission.

Our licensing policy is based on the licence selected by the submitter. Each licence can be either free, or a data consumer must sign it which means, that only authenticated users can access it after submitting a form where they agree to adhere to the licence. We keep track of those signatures and because the authenticated users must be real people this process is well defined.

Crisis management concerning the availability of the digital objects is addressed on a technical level. Since a PID (handle.net) system is used in CLARIN, moving resources from one CLARIN resource center to another one is possible without affecting the validity of references (e.g. PID of a resources used in a paper).

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. The data repository enables the users to discover and use the data and refer to them in a persistent way.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The public submissions in repository are being indexed by google scholar. All metadata can be harvested e.g., via the OAI-PMH protocol and free data using the OAI-ORE protocol (unless copyright issues are resolved, than we can export all of the data).

Unique persistent identifiers according to the Handle system are provided for each archived object using EPIC handles.

Link: https://clarin-pl.eu/oai/request?verb=Identify

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

11. The data repository ensures the integrity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

Repository use MD5 checksums for all objects and checked it periodically.

The integrity of data and metadata are monitored by funcionality that files in data sets can not be changed by submitter but only by administrators for e.g., typos in metadata. All data has assigned persistent identifiers; they always refer to the same content.

The repository deal with multiple versions of the data by PID url is working showing the submission with special metadata value shown (isreplacedby) which points to the new version.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

12. The data repository ensures the authenticity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

After submitting data, producers do not have any other option to change the metadata but to directly contact the editors. As described in 11), for non trivial changes a new version of the submission is suggested.

For each change, the provenance metadata are stored including appropriate log messages.

As described in 1), the submitters can be only authorised people by well defined authorities e.g. eduGain using shibboleth.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

We use standards CMDI (ISO-CD 24622-1) for metadata standards. With the use of the DSpace (one of the leading digital repository systems, http://registry.duraspace.org/registry/repository/2326) and the defined workflow supported by the repository's interface, the Clarin-PL repository meets the requirements of OAIS as described below.

1) Ingest: The Submission Information Packages (SIPs) are received for curating and are assigned to a task pool where our curators can process them. The default way is that the ingestion process is done through our web based interface which hides the implementation details.

2) Archival Storage: After the Ingest step, one of our curator takes charge. Using the web interface, the metadata are updated (added, deleted, modified), the submitted bitstreams are validated.

3) Data Management: This function is executed during the creation of the metadata (descriptive, administrative and structural), as seen on the prior step.

4) Preservation Planning: As described in 6), we monitor and backup our system.

5) Administration: We use developed a specific robust administration interface including specific detailed reports on the contents of our repository.

6) Access: The available Dissemination Information Package types (https://wiki.duraspace.org/display/DSDOC18/Importing+and+Exporting+Content+via+Packages#ImportingandExportingContentviaPacl

query responses and reports are delivered to CONSUMERS. All metadata are publicly available.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

14. The data consumer complies with access regulations set by the data repository.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

An account is necessary to access protected data. When an item is protected, the data consumer must sign the appropriate licence in order to be able to download the data. The metadata themselves are always public.

Each submission is clearly marked with its licence and if the licence requires signature, only authenticated users can sign. We rely on the standard academic network which must assure that each authenticated user is a person. We offer local accounts too and in this case we perform the verification manually.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

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. Access to restricted resources are protected via authentication. The licence of each item is clearly visible.

If the licence is not adhered to, we can retrieve the exact dates and specific id's of people which have accessed the resources.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. The data consumer respects the applicable licences of the data repository regarding the use of the data.

Minimum Required Statement of Compliance:

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

Applicant Entry

Statement of Compliance:

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

Self-assessment statement:

The data consumer is made aware of usage restrictions using clear visual indicators (see e.g., https://clarin-pl.eu/dspace/handle/11321/47) . If the data are licensed with a licence that requires signing, the user is asked to electronically sign the licence before downloading.

In case of misuse, the only thing that can be practically done is to deny the user further access to the repository and to make the research community aware of the misuse. Each signing is stored.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: