Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository CLARIN Centre Vienna 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 4, 2014.

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 Centre Vienna
Seal Acquiry Date:	Apr. 04, 2014
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:	**ICLTT** Austria T +43 51581 2300 E icltt-tech@oeaw.ac.at W http://www.oeaw.ac.at/icltt

Assessment

0. Repository Context

Applicant Entry

Self-assessment statement:

CLARIN Centre Vienna is Austria's main connection point to the European network of CLARIN Centres. Representing one contribution of the national consortium CLARIN-AT to CLARIN-ERIC, it is run by the Institute for Corpus Linguistics and Text Technology (ICLTT) at the Austrian Academy of Sciences and is jointly funded by the Academy and the Federal Ministry of Science, Research and Economy.

CCV is embedded in the Digital Humanities Austria (DHA) initiative which has started in January 2014. DHA represents the umbrella under which the DH infrastructure activities CLARIN and DARIAH are conducted in Austria.

The primary mission of the centre is to provide easy and sustainable access to digital language resources and technology for research communities in the social sciences and humanities, as well as depositing services for language resources created in Austria. To this end, CCV hosts the Language Resource Portal (LRP), a repository dedicated to archiving and publishing various digital language resources.

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:

The repository holds language resources provided by CLARIN-AT member institutions and affiliated Austrian academic institutions. All resources made available in the repository are provided with metadata records in standard formats (TEI, CMDI) which detail aspects such as provenance, format and size of the resources, access modalities (licensing).

Depositors are required to provide metadata in line with the requirements of the repository. The deposition process involved human interaction between depositor and repository manager, to ensure that the deposited resources meet the requirements.

The information about the deposition procedure is available in the FAQ-document on the website of the centre.

LRP is run and hosted by the ICLTT, an institute of the Austrian Academy of Sciences. With respect to compliance with disciplinary and ethical norms, we therefore rely on The European Code of Conduct for Research Integrity promulgated by ALLEA (ALL European Academies) European Science Foundation

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:

Data producers are strongly encouraged to provide the resources in standard formats acknowledged by the respective international research communities. The primary guideline for the recommended formats are the CLARIN standard recommendations (http://www.clarin.eu/recommendations). Use of these formats will ensure that the data is interoperable within the CLARIN infrastructure and consequently other international communities networks. The primary format for textual data in the repository is TEI/XML (with metadata in CMDI and Unicode character encoding).

The data producers are supported during the depositing process by the repository management team to assure that the data is available in recommended formats, if feasible. Otherwise, the data may only be archived “AS IS” with possible implications regarding the long-term availability of the deposited material.

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:

Data depositors are required to provide metadata alongside the resources to be deposited. The primary metadata format is CMDI (http://www.clarin.eu/cmdi), but other formats are also accepted as long as sufficient resource description is assured (teiHeader or Dublin Core are regarded as baseline requirements). The repository management team supports depositors in creating the metadata records, selecting the most suitable CMDI-profile for given resource types.

Particular attention is attached to the the availability of structural metadata in the case of data collections. If not available, it is created in collaboration with the depositor before the ingest into the repository is performed. The default format for structural metadata is METS/XML.

The repository supports multiple concurrent metadata records in different formats for one resource. If possible, these resources are produced from a single source, i.e. the information is stored in the most comprehensive format, other formats are generated automatically, to ensure consistency. Metadata records are validated against corresponding XML Schemas before ingest.

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:

It is an explicit mission of CLARIN-AT (in the context of Digital Humanities Austria; see 1) to foster the digital paradigm in the humanities. Among others, it pursues this goal through providing a sustainable and reliable technical infrastructure for the research community on the institutional and national levels. The repository is a cornerstone and integral part of this emerging infrastructure. It is to ensure the long-term availability of digital resources in order to preserve the knowledge generated in digitally supported research projects.

As part of the CLARIN infrastructure, the repository is included in all promotional activities carried out at the national level of CLARIN-AT as well as the European level of CLARIN.

http://www.clarin-dariah.at, http://clarin.oeaw.ac.at/ccv/about

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 in its own right. It is run by the Institute for Corpus Linguistics and Text Technology(ICLTT) at the Austrian Academy of Sciences and is jointly funded by the Academy and the Federal Ministry of Science, Research and Economy, representing one contribution of the national consortium CLARIN-AT to CLARIN-ERIC.

Every submission of resources is handled by the repository management team and dealt with in direct communication with the depositor. Licensing, IPR and issues regarding ethical norms are cleared before the resources are accepted for deposition. Especially, resources containing personal information have to be deposited in anonymised form except for cases of explicit consent of the involved persons.

The depositor decides the mode of access and applicable licence for the deposited resource. If required, the data is accepted for archiving without being made publicly available. Practically, this is ensured in such a way that the repository website is built on top of the repository system proper which is not directly available. The access mode for the resources has to be set explicitly.

The depositor must sign an agreement acknowledging that they have the right to deposit the data and give CCV the right to distribute the data (except it is meant only for archiving).

Users accessing resources are bound by CCV's general terms of use and the specific licence associated with a given resource. For resources with restricted access, the user has to be authenticated and needs to electronically sign the licence.

Deposition Licence Agreement , Terms Of Use

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 repository is managed jointly by the ICLTT and the Computing Centre of the Austrian Academy of Sciences (ARZ). ARZ is responsible for the secure operation and maintenance of the hardware of the Austrian Academy of Sciences and has well established and documented procedures for server management with regard to hardware and data redundancy (backups).

The repository is based on the well established and widely used system fedora-commons that performs fine-grained versioning of digital objects (on the level of individual data streams) and allows for exporting, archiving and migrating the digital objects in XML-based formats. The repository system proper is only accessible by designated repository administrators.

The data in the repository is backed up in a regular manner: daily on-site and weekly off-site. Backups are checked for integrity (via MD5 checksums). We keep at least three copies at all times, one of the them off-site. The availability and functioning repository is being automatically monitored.

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:

The goal of long-term availability is pursued through technical and organisational strategies. With respect to technical aspects, the CCV enforces the use of standardised data formats as proposed by the CLARIN recommendations (http://www.clarin.eu/recommendations) to ensure the long-term availability of data. Data providers are encouraged to use recommended data formats and encodings. The use of commonly accepted XML vocabularies is meant to lower the risk of obsolescence and to minimize curation efforts in the future.

With respect to the repository system itself, the decision in favour of the OAIS-compliant open-source system fedora-commons has been taken in order to minimize the risk of technical obsolescence.

As for the institutional setting, CCV/LRP has been set up as part of the Austrian infrastructure projects CLARIN-AT and DARIAH-AT and the stable institutional context of the Austrian Academy of Sciences.

The CLARIN-AT team has been advocating sustainable archiving solutions for research data in digital humanities not only within the institution but also on the national level and beyond. They have been cooperating with university libraries and other national stakeholders promulgating the agenda and working towards a national federation of repositories.

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:

During the deposition process, the repository team checks the format of the data (with respect to recognized standard formats, validation is performed by standard tools) and the availability and completeness of the metadata in accepted formats. This is accomplished in close interaction with the depositor. During ingest, every resource and metadata record is assigned a PID (done automatically by the repository system).

If new versions of resources become available, they are stored within the digital object and identified by timestamp, allowing to refer to every version consistently. The PID refers to the last version.

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 data provider retains all intellectual property rights to their data. The depositor must grant distribution rights to the repository provider (Deposition Licence Agreement) and choose an access model (public, academic, individual) and applicable licence.

The fedora-based system is compatible with the repositories of other CLARIN-AT or CLARIN-EU partners which allows migration of data in worst case scenarios. In the case of a migration of the data to another system, the consistent use of PIDs ensures continuity in the validity of references.

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:

A web application built on top of the repository (http://clarin.oeaw.ac.at/ccv) features entry points to all resource collections. Resource collections are for the most part available via web applications allowing searching and browsing in the content of the resource. Additionally, the repository exposes a public OAI-PMH endpoint that is regularly harvested by the main CLARIN harvester and could be harvested by any other interested third party.

The systematic assignment of PIDs makes sure that digital objects will be referenceable in a persistent way irrespective of their future location (example of PID for a Metadata Record (in CMDI format) for a Resource collection : http://hdl.handle.net/11022/0000-0000-001B-2)

Example of an simple web application allowing to search in dictionaries that are part of the repository http://clarin.oeaw.ac.at/lrp/dict-gate/

OAI-PMH endpoint

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:

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

Self-assessment statement:

We utilise MD5 checksums to verify the integrity of digital objects stored in the repository. The procedure is performed when new data is added to the repository, periodically (every week) in order to ensure that no data was changed unintentionally and every time a backup is created. In addition, data integrity is ensured by the version control capabilities of the underlying fedora commons system.

If new versions of resources become available, they are stored within the digital object and identified by a timestamp, allowing to refer consistently and unequivocally to every version. The PID refers to the last version of the resource. In case of major changes, the depositors are advised to create a new digital object.

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:

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

Self-assessment statement:

As a basic principle, all objects in the repository must have metadata. The deposition is executed in personal cooperation between the data producers and the repository management team. Once in the repository, the Fedora Commons version control governing the repository ensures data authenticity by monitoring changes to the data and its metadata.

The authenticity of the digital objects is vouchsafed by the assignment of PIDs. New versions of digital objects are stored separately and identified by a timestamp-based fragment identifier. The basic PID of the digital objects always refers to the object as a whole allowing to retrieve every version. By default, the last version of the object is provided.

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:

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

Self-assessment statement:

CCV/LRP intends to support the tasks and functions as required by OAIS. The Fedora Commons system (http://www.fedora-commons.org) underlying the repository is compliant with the OAIS reference model. It supports ingest of Submission Information Packages (SIP) and processing of Archival Information Packages (AIP).

Currently, the main mode of ingestion is performed through direct communication between the data producer and the repository manager during which (as described in statement 3) depositors are required to provide sufficient descriptive, administrative and structural metadata about the resources in standard formats recognized by the community to ensure that it is understandable by the designated community. With respect to metadata, the repository relies on the emerging standards around CMDI (ISO-CD 24622-1).

As put forward in statements 6 and 7 the repository is embedded in a stable institutional environment and data in the repository is backed up regularly.

The deposited material is accessible together with the descriptive information via a web-interface, the metadata is also exposed via the described OAI-PMH endpoint (cf. 9) and harvested by the central CLARIN aggregator as an additional dissemination channel.

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:

As a matter of principle, all metadata are openly accessible. All data consumers are bound by the general terms of use and the licences specific to individual resources which regulate the access to the resources. Depending on the access mode (public, academic, individual) the repository system enforces appropriate restrictions (i.e. only users who authenticated by means of an identity federation can access academic resources).

Currently, most of the data are deposited under Creative Commons licences. The depositors are encouraged to do likewise. For some data sets, explicit permission from the depositor will be required in which case a login and explicit electronic sigining of the licence is required. In case of non-compliance with the terms and regulations, the user can be excluded from accessing the resources and general legal consequences according to national and international laws are applicable.

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:

During the deposition process issues regarding potential confidentiality of the data or personal data are settled with the depository.

For the use of the repository and access to the resources Terms Of Use as well as the resource specific licences apply.

The user agrees to comply with the disciplinary and ethical norms as specified in The European Code of Conduct for Research Integrity promulgated by ALLEA (ALL European Academies).

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:

For every resource the access mode and licensing is clearly specified. The depositors are encouraged to choose from standard licences (such as Creative Commons), the default licence being Creative Commons licence CC-BY-NC-SA (Austrian version)

For resources the use of which is limited to academic use or by particular licences, these restrictions are enforced by the repository system, users will be required to authenticate via Shibboleth or acquire a separate account and to sign corresponding licence for given resource.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: