Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository CLARIN-D Resource Center Leipzig 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 30, 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-D Resource Center Leipzig
Seal Acquiry Date:	Apr. 30, 2015
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.datasealofapproval.org/seals/
Previously Acquired Seals:	Seal date: May 3, 2013 Guidelines version: 2010 | June 1, 2010

This repository is owned by:

NLP Group, Department of Computer Science, University of Leipzig

04109 Leipzig
Saxony
Germany

T +49-(0)341-9732230
E pgamrath@informatik.uni-leipzig.de
W http://asv.informatik.uni-leipzig.de/en/

Assessment

0. Repository Context

Applicant Entry

Self-assessment statement:

CLARIN-D Resource Center Leipzig (http://clarin.informatik.uni-leipzig.de/repo/)is part of CLARIN-D (Common Language Resources and Technology Infrastructure Deutschland) - a web and centres-based research infrastructure for the social sciences and humanities. The aim of CLARIN-D and its service centres is to provide linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities. The research infrastructure is rolled out in close collaboration with expert scholars in the humanities and social sciences, to ensure that it meets the needs of users in a systematic and easily accessible way. CLARIN-D is funded by the German Federal Ministry for Education and Research.

CLARIN-D is building on the achievements of the preparatory phase of the European CLARIN initiative as well as CLARIN-D's Germany-specific predecessor project D-SPIN. These previous projects have developed research standards to be met by the CLARIN services centres, technical standards and solutions for key functions, a set of requirements which participants have to provide, as well as plans for the sustainable provision of tools and data and their long-term archiving.

This repository offers resources such as a set of corpora of the Leipzig Corpora Collection (http://wortschatz.uni-leipzig.de/), based on newspaper, Wikipedia and Web text. Furthermore several REST-based webservices are provided for a variety of different NLP-relevant tasks.

Within CLARIN-D this resource centre is a certified centre of type B. CLARIN distinguishes a number of different centre types that have different impact for the language resources and tools infrastructure. Type B centres offer services that include the access to the resources stored by them and tools deployed at the centre via specified and CLARIN compliant interfaces in a stable and persistent way.

Within CLARIN-D the following requirements hold for centres of type B (https://www.clarin.eu/node/3542) and are fulfilled by this resource centre:

- Centres need to offer useful services to the CLARIN community and to agree with the basic CLARIN principles (own architecture choice, explicit statement about quality of service, usage of persistent identifiers,

adherence to agreed formats, protocols and APIs).

- Centres need to adhere to the security guidelines, i.e. the servers need to have accepted certificates.

- Centres need to join the national identity federation where available and join the CLARIN service provider federation to support single identity and single sign-on operation based on SAML2.0 and trust declarations. In case all resources at a centre are open, setting up a Service Provider is optional.

- Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the Data Seal of Approval or MOIMS-RAC approaches.

- Centres need to offer component based metadata (CMDI) that make use of elements from accepted registries such as CLARIN Concept Registry in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI PMH.

- Centres need to associate PIDs records according to the CLARIN agreements with their objects and add them to the metadata record.

- Each centre needs to make clear statements about their policy of offering data and services and their treatment of IPR (intellectual property rights) issues.

- Each centre needs to make explicit statements to the CLARIN boards about its technological and funding support state and its perspectives in these respects.

- Centres need to employ activities to relate their role in CLARIN to the research community in order to guarantee a research based status of the infrastructure and allow researchers to embed their services in their daily research work.

- Centres that are offering infrastructure type of services need to specify their services for CLARIN and the terms of giving service.

- Centres are advised to participate in the Federated Content Search with their collections by providing an SRU/CQL endpoint. This content search is especially suitable for textual transcriptions and resources.

A short overview of all requirements for centres of type B is also given in the form of a checklist (https://www.clarin.eu/content/checklist-clarin-b-centres).

List of outsource partners:

1) Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

The repository makes use of a common CLARIN PID service (https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf) based on the Handle System (http://www.handle.net/) and in cooperation with the European Persistent Identifier Consortium (EPIC). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs.

CLARIN-D has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2. The following document lists the services which were stipulated:
http://de.clarin.eu/mwiki/images/0/0b/GWDG_PID.pdf

This outsource partner offers relevant functionality for guideline 10: "The data repository enables the users to utilize the research data and refer to them.".

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

Good, comprehensive context-setting information.

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:

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

Self-assessment statement:

The minimal requirements for data/tools to be deposited in the repository are:

1. the data/tool is provided in a standardized format or with an exhaustive documentation of the proprietary format
2. metadata is available in CMDI
3. contact information on the data depositor / data producer is present in the metadata
4. a statement on the legal status of the resource is available

The data that is put into the repository is checked for compliance with internal and CLARIN guidelines concerning scientific and scholarly quality. Only data that:

1. is the result of research projects,

2. comes with exhaustive metadata,

3. which's data structure is described by a sophisticated documentation (PDF/A),

4. comes with information on how the data was originally created,

5. was reviewed by a third party

will be added to the repository. The data itself, the metadata and additional documentation is an obligatory part of each repository entry. Currently these guidelines are not in a fixed state and subject to minor changes. A preliminary version is available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/). Until a final version is released, no data created by external data producers will be added to the repository. The data stored in the repository is limited to well known and documented content created by our own institution as the result of a long running research project. According to these guidelines metadata on data/tools always contains information on the resource producer (name and URL of the institution, information on contact persons that allows interested users to obtain further information. Adding references to publications to the metadata (or adding the papers to the repository), on how the data was created and in which scenarios it is intended to be used is encouraged but not enforced. Data sharing and reuse is promoted by providing free access to the data (download, webservices) and metadata (via the OAI-PMH protocol). The CLARIN infrastructure contains software components such as the VLO (http://www.clarin.eu/vlo/) which enables users to browse and search through combined (metadata of all CLARIN repositories) catalogs.

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:

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

Self-assessment statement:

It is recommended to use formats listed in the CLARIN standard recommendations (http://www.clarin.eu/recommendations). In addition relevant standards and formats in the context of CLARIN are listed (http://www.clarin.eu/content/standards-and-formats). Manual checks are performed by CLARIN members before data is added to the repository. Usage of standardized formats is encouraged but not enforced.
In case no recommended/well known and documented format is used, an exhaustive documentation on the syntax and semantic of the data (e.g. database dumps: names of tables and columns; specifications and examples on the contents of each column; examples on how to retrieve different types of data) has to be provided by the data producer. This documentation (English, PDF) is stored on the repository along with the data and metadata and is provided to everyone who wishes to download/access the resource. The repository maintainers keep track of all formats already used by the depositors and commit themselves to work on updates of the CLARIN standard recommendation if new formats gain in popularity.

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:

Metadata for all CLARIN repositories has to be provided in the CMDI format. There is exhaustive documentation (http://www.clarin.eu/cmdi) available on how to create CMDI compliant metadata profiles and instances. Additionally a set of tools is provided that allow data producers to easily create new or adapt existing metadata to the CMDI standard.

Resources must be accompanied with valid CMDI metadata in order to be considered for deposit. Metadata is checked for compliance according to CMDI standards in the following way:

1. Check if XML metadata is well-formed and valid.

2. Check if the used CMDI components and profiles are stored in the Component Registry (http://catalog.clarin.eu/ds/ComponentRegistry/) (public, PID present).

3. Check if the data categories used in those components/profiles are present in the CLARIN Concept Registry (https://openskos.meertens.knaw.nl/ccr/browser/).

4. Check if the provided CMDI files contain enough and consistent information (e.g. consistent specification of the data producer's “name”) according to the needs of the VLO (http://www.clarin.eu/vlo/).

The granularity of CMDI metadata is up to the (meta)data producer. The repository itself is able to handle a high granularity of metadata. The creation of metadata files (instances) is supported via the XML Editor ARBIL (https://tla.mpi.nl/tools/tla-tools/arbil/) that comes with CMDI support. Metadata elements need to be compliant to the standards set in CMDI. Since CMDI is a component based approach which allows (meta)data producers to create custom tailored metadata profiles there is no limit to the usage of established standards etc. In order to be visible and useable in the CLARIN infrastructure CMDI metadata added to the repository needs to contain a minimum set of attributes (linked to data categories stored in the CLARIN Concept Registry) which is enforced by the quality checks described above. The usage of metadata elements that are accepted by a research community is encouraged and technically supported via re-use of existing metadata components, but is not enforced.
This information is part of the resource depositor guide which is available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

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:

The mission of the repository is to serve as the repository of a CLARIN-D resource center of type B (http://www.clarin.eu/files/centres-CLARIN-ShortGuide.pdf). The mission of CLARIN-D is to provide "linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities" (http://de.clarin.eu/en/home-en.html). Therefore a repository in which data, tools and associated metadata is archived on a long term basis must be operated by such a resource center.
The repository is part of the CLARIN infrastructure and thus does not carry out promotional activities on it's own, but is embedded into such activities on CLARIN-D and the European CLARIN level. These activities do include but are not limited to:

1. Providing exhaustive information on the CLARIN mission through websites (http://www.clarin.eu, http://de.clarin.eu).

2. Operation and maintenance of the Virtual Language Observatory (VLO) which provides means to search for data/tools to the end user (based on the metadata provided by the resource centers/repositories that are part of CLARIN).

3. Presenting data, tools and services provided by CLARIN on conferences.

4. Organization of dissemination conferences that aim at getting in touch with the user communities of CLARIN.

5. Organization of training courses.

Many parts of the CLARIN infrastructure do address the migration of data from one resource center / repository to another. Since the usage of these infrastructure services (e.g. a PID system, CMDI) is obligatory, every CLARIN center is, to a certain extent, ready to move it's digital assets to another center.

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. It is run by the University of Leipzig which is an institution governed by public law. Depositors need to sign an agreement stating that they own all necessary rights required to deposit the data and that during the creation of the resource the data producer respected IPR (Intellectual Property Rights) and privacy issues. Data depositors are themselves responsible for compliance with any national or international legal regulations. Since no data with disclosure risk will be added to the repository, depositors also have to state that the deposited resource does not contain any data with disclosure risk. The repository staff maintains a checklist of cases in which resources containing data with disclosure risk have previously been rejected or modified (and if so, how they were modified) in order to be compliant to the repository regulations. This list is intended to help in cases in which the depositors are unsure about the status of their resource regarding disclosure risk.
In case a violation of conditions is observed, the original data provider is contacted. In case the violator can be identified, further access by this person/institution will be prevented if technically possible (Shibboleth => the home institution will be informed).

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:

Data is stored on a RAID system and backups are created on a regular basis (every time the content of the repository changes, since ingests are done by the repository maintainers). These backups are held on separate hardware. Deterioration of storage media is monitored via Nagios probes which do a regular check of the used hardware (e.g. S.M.A.R.T. - Self-Monitoring, Analysis and Reporting Technology - data) and report drastic changes or imminent failures. In case of failures/problems/... the administrators of the repository are notified and will take appropriate actions. For further information please refer to the preservation policy provided on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

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:

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

Self-assessment statement:

By encouraging data depositors to use standardized formats (UTF-8, documented XML formats, ...) we try to minimize the cases in which obsolescence of file formats will occur in the near future. By enforcing a detailed and exhaustive documentation in case proprietary / "custom" formats are used we ensure that exhaustive documentation is available under all circumstances. Thus it will, at least, be possible to specify and implement data converters.

Long term data usability is ensured by the following measures:

1. We make sure that all data formats, also proprietary ones, are well documented.

2. We enforce provision of information on authorship of the data and encourage adding references to scientific papers describing the data and usage scenarios.

3. Access to data and metadata is provided via widely used open source software stacks (MySQL, Tomcat, Fedora Repository) that are installed on virtual machines. This maximizes the probability of long term support (updates, security fixes) for the tools being used and improves the ability to run installations of these software stacks independent from the underlying hardware/operating system/... .

For further information please refer to the preservation policy provided on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

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:

Currently there are no established workflows that define how to integrate/archive data provided by external data providers. We are currently working on a documentation on how to archive some types of resources that we will add to the repository on our own. Based on this work this documentation will be extended in order to address similar kinds and other types of resources we expect to be added by external depositors. Once all currently open questions mentioned below are part of the documentation, these documents will be available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

Currently there is no documentation or process for transformations on archival data.
Up to now there is no established and documented selection process. While some questions still need to be answered (e.g. up to which scale are we able to handle big data of external depositors) some outlines are already clear from a CLARIN perspective:

1. Only data that is freely available or data that is free to be used for research and teaching (non commercial purposes) will be added to the repository in the first stage.

2. There needs to be a plausible usage scenario and a user community for the deposited data/tools.

3. If possible, access to the data has to be provided via webservices on a level of granularity that matches these use-cases.

4. Metadata in CMDI needs to be present.

The handling of requests to deposit data that does not fall within the (CLARIN) mission will be decided on a case by case basis. Data that supports the CLARIN mission will be prioritized.

On a technical level access to the data can be limited to users working in research institutions (CLARIN-AAI, DFN-AAI). According to the CLARIN rules access to metadata is not limited. Only data that comes with licenses that fit these rules (e.g. texts/audio/video must not be shared freely on the web but are free to be used in research and teaching) will be added to the repository. In case privacy of subjects is a concern, this needs to be addressed by contracts signed by those subjects (e.g. interviewed people explicitly state that the data may be provided freely to researchers/teaching purposes). In case new usage scenarios are supported by the CLARIN-AAI we will adapt these rules.

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:

In a first stage only data that is available for free (and states this in a license) or comes with “compatible” licenses (free for research/teaching) will be added to the repository. Thus, currently there are no licences / contractual agreements with data producers since only data created by the institution which runs the repository is present. In the future external depositors will have to sign a depositor agreement. These contracts contain statements on:

1. the involved parties
2. licenses and copyright
3. rights and responsibilities of the depositor and the repository
4. the content to be deposited
5. removal of content and access conditions
6. availability to third parties

7. provisions relating to use by third parties

8. death of the Depositor

9. liability

10. term and termination of the Agreement

A preliminary version is already available on the repository website (http://clarin.informatik.uni-leipzig.de/repo/).

Enforcing licenses by data users in the case of misuse is conducted by the property rights owner. Crisis management concerning the availability of the digital objects is addressed on a technical level. Since a PID 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). Our setup consists of virtual machines which may be moved to other CLARIN partners . In case virtual machines are moved internally (inside the CLARIN-D center in Leipzig) this will be possible without severe impact to user experience (live migration is supported). In case the machines need to be moved to other CLARIN partners a limited downtime will occur.

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:

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

Self-assessment statement:

As described previously in chapter 2 the decision which formats are used is up to the data provider. Since usage of standardized formats is encouraged, data usually will be available in formats used by the research community.

Harvesting of metadata is possible via OAI-PMH. Search facilities are currently not provided by the repository itself. Instead CLARIN operates OAI-PMH harvesters which collect CMDI-metadata from all repositories run by CLARIN centers. The collected metadata is used in the back-end of web applications such as the VLO (http://www.clarin.eu/vlo/), which provide a central starting point when searching for resources in the CLARIN infrastructure. In cases of some resources “deep search” is supported by the means of the CLARIN Federated Content Search (http://www.clarin.eu/fcs) interface.

The repository uses the common CLARIN PID service (https://www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf) based on the Handle System (http://www.handle.net/) and in cooperation with the European Persistent Identifier Consortium (EPIC). The usage of PIDs is mandatory for resources in CLARIN thus all resources added to the repository may be referenced using PIDs.

CLARIN-D has a contractual relationship with GWDG concerning the provision of PID-services via EPIC API v2 as mentioned in section 0 on repository context.

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:

Currently, we are implementing the use of checksums in order to monitor the integrity of digital objects stored in the repository. We are still in the process of checking if the software stack that is already in place allows this kind of monitoring based on existing functionality or if an external solution needs to be implemented. Since we are in the implementation phase there also is no fixed guideline on when to perform these kinds of checks. Currently we plan to do checksum-tests when:

1. new data is added to the repository (data retrieved from the repository needs to be identical to the data that was ingested)

2. periodically (once a week/month/every time a backup is created) in order to ensure that no data was changed unintentionally

For this, Apache Subversion (SVN) has been set up as part of our repository. Within this SVN the data of the repository is already stored as a backup mechanism. These backups are created once the integrity of the data in the repository was ensured after ingestion. In addition a checksum of the original data is created and stored. A mechanism for regular comparison of the state of the resources in the SVN and the repository to checksums created upon insertion into the version control system still has to be implemented.

Additionally the integrity of the data is ensured by the version control capabilities that are part of the Fedora Commons repository which is operated in the backend. Metadata is a data stream within the digital object, and as such is version controlled like object data.

Access to data and metadata is provided via webservice interfaces. The availability of these webservices is monitored via Nagios (http://www.nagios.org/) / Icinga (https://www.icinga.org/) probes. Some of these probes are run in local installations at the center while others are operated by CLARIN-D (http://de.clarin.eu/images/ap3/ap3-005-monitoring.pdf). The frequency of checks depends on the type of service that is monitored.

Multiple versions of data are valid. CLARIN propagates the idea of reproducible research. Thus updates/new versions of existing data is handled like any other resource with the exception of setting and storing a reference to the previous version. Access to metadata and data of all versions is provided at the same time and is handled in the same way:

1. access is provided via OAI-PMH/webservices

2. and a unique PID is assigned.

However, updates of metadata for existing resources are possible without considering the result to be a new version.
Part of the archiving workflow is the integrity check of the data and the metadata by the archive manager. This is done both manually and automatically. The metadata is parsed for syntactic correctness and manually evaluated for completeness and soundness.

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:

In case data that is present in the repository “changes” this data is considered to be a new version of the existing data. Thus data producers need to provide the same type and scale of information (metadata, documentation) that was provided for the previous version (at least in case changes occurred).

The metadata provided for each resource to be added to the repository needs to contain basic information on the data depositor (e.g. name of the institution, contact address) and the provided data (e.g. name, date or version, description of the resource itself and of the data format being used, obligatory links to papers). Adding further information (e.g. change logs) is encouraged but not enforced. In case multiple version of a resource are present in the repository, at least references to previous/newer versions needs to be present in the metadata.

Data and metadata are essential and mandatory parts of the digital objects that represent a resource in the repository. This can be considered to be an implicit link between data and metadata. In CMDI metadata is explicitly linked to data and additional metadata via the ResourceProxy-section (https://www.clarin.eu/faq/3462) in a CMDI file.

Currently we do not intend to compare essential properties of different versions of the same file/resource. Keeping track of changes that occurred in between different versions of the same file/resource will be up the data producers. In order to improve the usability we will encourage but not enforce data producers to provide change-logs in case new versions of already existing data are ingested into the repository.

Currently there is no explicit check of the identity of depositors since especially in the first phase of CLARIN only data that is provided by known partners will be added to the repository. Once this changes an explicit procedure for the check of depositor identities and “ownership” of the ingested data needs to be specified. External deposits will only be accepted after a due diligence process involving a check of the identity of the depositor and a clarification of all legal issues.

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:

For metadata we rely on the group of emerging standards around CMDI (ISO-CD 24622-1). With the use of the Fedora Commons system the repository aims to be as conformant to OAIS as possible. The repository complies with the OAIS reference model's tasks and functions. Moreover, the repository is based one the Fedora Commons software, which is compliant with the Reference Model for an Open Archival Information System (OAIS). Besides of the integration of the repository into the CLARIN infrastructure, there are currently no further plans for infrastructural development. For this repository the OAIS

Submission Information Package (SIP) consists of:

1. the (binary) data to be stored in the repository

2. metadata in CMDI that further describes the resources

3. a documentation or specification on the formats beeing used (links to documentation in case of standardized formats or exhaustive documentation on the format in case a proprietary one is use)

Archival Information Package (AIP) consists of:

1. all information/data that is part of the SIP

2. a persistent identifier for the resource (usually obtained by the repository)

3. an overall checksum

Metadata is available in CMDI via OAI-PMH. The CMDI file of a resource contains links to documents stored in the repository, interfaces - usually webservices in CLARIN – or webapplications that facilitate usage of the resource. The CMDI file tied together with these resources can be seen as a representation of a Dissemination Information Package (DIP).

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:

Currently there is only data stored in the repository that is available for free in case it is used for research or teaching purposes. In a first phase, only this kind of data will be added to the repository. This will change in the near future. Template contracts will be used which enforce resource depositors to specify an appropriate licence (free, free for academic use/research; see http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.3.pdf for details).
Implicitly the regulations of the DFN-AAI (https://www.aai.dfn.de/ and https://www.aai.dfn.de/en/der-dienst/degrees-of-reliance/) need to be mentioned since access to resources stored in the repository that shall only be available for academic use/research will be granted using this AAI infrastructure (based on Shibboleth).

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/events/... in texts) are respected in their deposited data. Access restricted resources (limited to academic use/research; see http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.3.pdf for details) are protected via Shibboleth and are only available to persons that are able to log-in through IDPs operated at institutions taking part in the DFN-AAI or similar AAI federations that are part of CLARIN.

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 system does not allow the integration of data into the repository without the specification of access criteria and without providing an appropriate license. These license conditions are available to the users via CMDI metadata. 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.
Background information: The repository is part of a CLARIN-D center. CLARIN aims at "Providing linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities." (http://de.clarin.eu/en/home-en). The resources provided by the repository are intended to be used for scientific purposes (teaching, research, …). An example of misuse may be the usage of these resources in a commercial context without knowledge/consent of the resource owner.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: