Title	CLARIN B Centre Checklist
Version	6
Author(s)	Peter Wittenburg, Dieter Van Uytvanck, Thomas Zastrow, Pavel Straňák, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel, Lene Offersgaard
Date	2018-02-07
Status	Approved by the Centre Committee
Distribution	Public
ID	CE-2013-0095

The following guidelines are meant as practical checks for the requirements mentioned in the centre requirements document (CE-2012-0037).

1. General requirements

1.a Centre compliancy

Requirement: Centres need to offer useful services to the CLARIN community.

Details: The technical management of the national CLARIN consortium of the centre has to give a written declaration of centre compliancy. The centre should attach or give a URL to this document. See https://www.clarin.eu/node/3767 (CE-2013-0137) for a template. This is only required when a centre is assessed for the first time.

Check procedure: Check that the written statement exists and is signed by the technical manager of the national CLARIN consortium. This check is not relevant in case of re-assessment.

Centre statement:
(Attach the declaration as a separate document together with this document)

1.b Visibility of connection to CLARIN

Requirement: Each centre needs to refer to CLARIN in a visible way on its website.

Details: Each centre has to have a clear reference to the CLARIN website or in other ways clearly refer to CLARIN. Another acceptable reference can be the logo and link to the national CLARIN consortium. If this requirement is not met, a good explanation should be given.

Check procedure: Check that a clear reference exists.

Centre statement:
(Add the URL with reference to CLARIN ERIC)

1.c Funding support

Requirement: Each centre needs to make explicit statements about its funding support state and its perspectives in this respect.

Details: Each centre has to give a short description of the funding situation and the future funding expectations.

Check procedure: Check that description guarantees reasonable funding support for at least two years.

Centre statement:
(Add description here)

1.d Resources and services provided

Requirement: Each centre needs to make explicit statements about CLARIN compliant resources and services available at the centre.

Please note that documents and web pages referred to as background information for the assessment must be in English or must be accompanied by a summary in English.

Details: The centre should offer online data access/sharing and services for users from other CLARIN ERIC countries. Therefore, each centre has to give a short description of the resources offered to the CLARIN community.

Check procedure: Check that the centre states it is offering data and services for users from CLARIN ERIC countries - either public resources, or via login using the CLARIN IdP and national AAI services.

Centre statement:
(Add description here)

2. Intellectual Property Rights and Privacy

2.a Data offering & IPR

Requirement: Each centre needs to make clear statements about their policy of offering data and services and their treatment of IPR issues.

Details: The centre has to give a short description (preferably on its website) of its policy of offering data and services and the treatment of IPR issues[1] including a description of how licenses are presented to users. The centre should offer data access/sharing for users from other CLARIN ERIC countries.

Check procedure: Check that the centre gives a clear statement about its data offering policy and about the IPR issues regarding data sharing.

Check that the centre states it is offering data for users from CLARIN ERIC countries - either via login using the CLARIN IdP or national AAI services.

Centre statement:
(If the policy of offering data and treatment of IPR issues can be found on a webpage, then stating which page contains the information is sufficient, otherwise add description here.)

[1] See https://tla.mpi.nl/resources/access-permissions/as an example.

2.b Privacy statement

Requirement: The centre has to implement the GÉANT Data Protection Code of Conduct (DP-CoC) for each of its federated Service Providers.

Details: The centre has to provide a URL to a webpage where its privacy policy is described[2]. It must also add this in a machine-readable way to its SAML metadata[3]

Check procedure: Inspect the provided Privacy Policy URL(s). If the SPs have also joined eduGAIN, compliance can be easily tested via http://monitor.edugain.org/ , otherwise the AAI taskforce will check the SAML metadata manually (contact the taskforce via tf-aai@lists.clarin.eu)

Centre statement:
(Add URL here)

3. External assessment of data centre

Requirement: Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the CoreTrustSeal[4].

Details: For CoreTrustSeal see https://www.coretrustseal.org . The centre cannot be certified as a B Centre until the CoreTrustSeal assessment is achieved, but the CLARIN assessment procedure can be completed as long as the CoreTrustSeal assessment is applied for.

Check procedure: Is the CoreTrustSeal achieved or applied for? – see https://www.coretrustseal.org/certified-repositories or check the file with the application provided by centre.

Centre statement:
(Add URL to application or a PDF version)

4. Server Certificates

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

Details: The SSL-certificates of the web servers at a centre should **not be self-signed** but have to provide a full trust-chain up to one of the root certificates as accepted by Mozilla Firefox[5].

Check procedure: Load an HTTPS URL at the centre. Check in your browser if the certificate is valid.

---

2 See http://hdl.handle.net/11113/00-0000-0000-0000-19BA-5@view for an example
3 See https://www.clarin.eu/node/3910 for more information
4 The CoreTrustSeal is replacing the Data Seal of Approval.
5 See https://wiki.mozilla.org/CA/Included_Certificates

Note for the centres: Although not a strict part of this assessment procedure, it is strongly recommended to use the SSL labs test at https://www.ssllabs.com/ssltest/ to optimize your SSL configuration.

Centre statement:
(Add URL(s) to web servers)

5. Federated Identity Management

Requirement: 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.

Details: Several sub-requirements (in the most logical order):

1. Setup a SAML 2 Service Provider
2. Install the attribute debug script (shib_test.pl) at your Service Provider server: https://www.clarin.eu/page/3537
3. Joining the national Identity Federation (when available – see https://refeds.org/federations)
4. Allow users from the CLARIN IdP to login – see https://www.clarin.eu/page/3398
5. Join the CLARIN Service Provider Federation – see https://www.clarin.eu/spf
6. Allow users from at least one other country to login through their national identity provider
7. Enable login through the other Identity Federations in the CLARIN Service Provider Federation or specify planning for enabling the other Identity Federations – see https://www.clarin.eu/spf

Check procedure: Check if the centre states that sub-requirements 1 to 7 listed above are fulfilled.

Login to the SP from the CLARIN IdP. Check with shib_test.pl if the right attributes are available.

Try to login to the SP from a national IdP from another country than the centre's. See if login from more identity providers are allowed. Check with shib_test.pl if the right attributes are available from a national IdP you have access to.

Check at https://centres.clarin.eu/spf what is provided for the centre. If possible, login to the SP with an IdP from each of the national identity federations that are member of the SPF. Check with shib_test.pl if the right attributes are available.

Centre statements:
(For each sub-requirement state if the centre fulfils the requirement)

6. Metadata

Requirement: Centres need to offer component based metadata (CMDI) that make use of elements from accepted registries such as the CCR[6] in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI-PMH.

Details: Each centre should setup a repository (a web-accessible server that offers human and machine-readable access to language resources/services and their metadata[7]). It should feature an OAI-PMH endpoint through which the metadata can be harvested. The metadata should be CMDI-compliant (see https://www.clarin.eu/cmdi).
The CMDI profiles, that a centre uses for their published metadata, have to be public, with preferably the status *production* (but *draft* or *deprecated* are acceptable), to be accepted in assessment. It is also preferable that the elements contain valid ConceptLinks to the CCR. Proposals for ConceptLinks to be added to a published profile or component can be submitted to the Component Registry administrator (via cmdi@clarin.eu). The evaluation of such a request might require a discussion with various stakeholders to assess the semantic fit of the proposed ConceptLinks.

List of sub-requirements:

Computer access to the repository:

1. Setup an OAI-PMH URL of the repository and give a link to it
2. Show that the OAI-PMH URL of the repository validates using https://clarin.eu/oaivalidator

Harvesting of metadata:

3. Show that harvesting by the VLO can be done - see https://vlo.clarin.eu/data/ for the results of the harvesting
4. Check at https://vlo.clarin.eu whether the metadata shows up correctly
5. Give links to metadata for a few resources as examples on the CMDI-compliant metadata.

CMDI files + profiles + CCR:

6. State if the harvested CMDI files validate against their XML schema
7. State if the harvested CMDI files contain a PID in the MdSelfLink header field
8. State if the harvested CMDI files refer to web-accessible files or a landing page with a ResourceProxy
9. State which profile(s) at the component registry are used (https://clarin.eu/componentregistry):
    a. Are they public? Do they have the status *production, draft* or *deprecated*?
    b. To which extent do the elements contain valid ConceptLinks to the CCR?

In case there is a front-end for end users, which is not a strict requirement but very advisable:

10. State the URL of the web interface of the repository

If the repository offers metadata about web services:

11. Check if the CMDI files validate against the webservice core model via https://clarin.eu/cmd-core

Check procedure:

---

[6] CLARIN Concept Registry: https://www.clarin.eu/ccr/ and https://www.clarin.eu/conceptregistry

[7] See https://www.clarin.eu/cmdi

1) Check computer access to the repository: Enter the OAI-PMH URL at https://clarin.eu/oaivalidator and see if it validates.

Harvesting by the VLO:
Check https://clarin.eu/harvester for the results of the harvesting
Check at https://vlo.clarin.eu/ if the metadata shows up correctly

CMDI files + profiles + CCR:
Validate the harvested CMDI files against their XML schema
Check the profile(s) used at the component registry (https://clarin.eu/componentregistry):

- Are they public?
- Do the elements contain valid ConceptLinks to the CCR?

If offering user access to the repository:
Browse to the web interface of the repository. Inspect some of the metadata records. Try to access some of the resources that are described. (Check for broken links and non-shibbolized password protection. Also check for access to either landingpages or resources)

If offering metadata about web services:
Check if the CMDI files validate against the webservice core model via https://clarin.eu/cmd-core

Centre statements:
(For sub-requirements 1 to 9 state that the centre fulfils the requirements. If sub-requirement 10 and 11 apply for the centre state that the centre fulfils these requirements as well)

7. Persistent Identifiers

Requirement: Centres need to associate (handle) PIDs with their metadata records. These PIDs should be suitable for both human and machine interpretation, taking into account the HTTP-accept header.

Individual files (e.g. a text, zip or sound file) can be referred to with either the PID of the describing metadata record in combination with a part identifier[8] or with another PID.

Details: A metadata record of a digital publication (e.g. a corpus, a treebank, a video file) contains information that is of high importance when citing it (e.g. the author, publication date, information about the corpus design, download links). To reach its maximal potential such important information needs to be available:

- for "classic" citations in e.g. a paper, where the end user is presented a web page with all relevant information
- for automatic processing, by e.g. an application or web service

To cope with both scenarios, CLARIN requires that URLs to which metadata PIDs point support the HTTP-accept header ("content negotiation") with minimally the following mime types:

- **text/html** (web-browser, human readable[9])

[8] See https://www.clarin.eu/faq/3453

[9] A generic CMDI-to-HTML XSLT is available at https://infra.clarin.eu/cmd/xslt/cmdi2xhtml.xsl

- **application/x-cmdi+xml** (CMDI[10] metadata, for machine interpretation)

There is no strict requirement in (the rare) case no HTTP-accept header is given by the client; however it is recommended to return in such a case a human readable version.

Non-metadata files should receive a PID or a PID in combination with a part identifier, if these files:

- are accessible[11] via internet
- are considered to be stable by the data provider
- are considered to be worth to be accessed directly (not via metadata records) by the data provider

For (non-metadata) files there are in general 2 ways of issuing PIDs:

- with a separate PID for each file, pointing directly to the binary object on a web server
- with a part identifier, which in addition to the PID of the related metadata record points to the binary object on a web server

Check procedure: Try to resolve a PID for *a metadata record*.
Check if:

- it redirects to a CMDI file for the HTTP-accept header "application/x-cmdi+xml"
- it redirects to an HTML file when accessing it from a browser

If non-metadata files have PIDs, try to resolve a PID (with or without a part identifier), for *a (non-metadata) file*. Check if it redirects to an existing online resource.

Centre statements:
(For each sub-requirement state that the center fulfils the requirements and give example PIDs)

8. Federated Content Search

Requirement: Centres can choose to participate in the Federated Content Search with their collections by providing an SRU/CQL Endpoint.

Details: A centre can expose its content search engine via SRU/CQL to participate in CLARIN's Federated Content Search (https://www.clarin.eu/fcs).

Check procedure: enter the endpoint URL at https://www.clarin.eu/fcsvalidator and validate.

Centre statements:
(State if the centre provides an SRU/CQL Endpoint. If not then describe the plans for joining the Federated Content Search or explain why there are no plans to implement an SRU/CQL Endpoint)

10 See https://www.clarin.eu/cmdi

11 The need for authentication to access an online file does *not* influence this.