| | |
|---|---|
| **Title** | CLARIN B Centre Checklist |
| **Version** | 7.4.1 |
| **Author(s)** | Peter Wittenburg, Dieter Van Uytvanck, Thomas Zastrow, Pavel Straňák, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel, Lene Offersgaard |
| **Date** | 2023-02-08 |
| **Status** | Approved by SCCTC |
| **Distribution** | Public |
| **ID** | CE-2013-0095 |

## Changelog

- Version 7.4.1 (2024-11-05): optional subsection web services in section 6 has been removed
- Version 7.4 (2023-02-08): addition of sections 9 and 10, containing optional recommendations about the attribute checker and attribute aggregator
- Version 7.3.1 (2019-09-30): inclusion of updated and practical checks for the requirements mentioned in the centre requirements document (CE-2012-0037). It relates the requirements, which are mostly unchanged, to the FAIR principles more explicitly than was the case earlier.

## Practical notes

- Each centre should setup a repository (a web-accessible server that offers human and machine-readable access to language resources/services and their metadata[1]), where data and metadata are licensed.
- When the checklist mentions providing some information, this can be done as a separate file or via a URL. In any case please mention where the information can be found.

## 1. 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.

**Details:** For CoreTrustSeal see https://www.coretrustseal.org . The centre cannot be certified as a B Centre until the CoreTrustSeal assessment has been successfully concluded, but the CLARIN assessment procedure can be completed as long as the CoreTrustSeal assessment is applied for. The application for the CoreTrustSeal, or proof that the CoreTrustSeal has been awarded, has to be provided.

**Check procedure:** Is the CoreTrustSeal achieved or applied for? – see the evidence provided by the centre

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

## 2. General information

### 2.a Description of the repository's context, mission and scope

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

**Requirement:** The centre's repository context, mission and scope needs to be clearly defined.

**Details:** The centre statements, given in R0 and R1 of the CoreTrustSeal application, are used. The centre can optionally provide more information.

**Check procedure:** Check statements given in R0 and R1 of the CoreTrustSeal application.

**Centre statement:**
*(Not required. Optionally, additional information can be added here.)*

### 2.b Centre compliancy (updated)

**Requirement:** Centres have to be recognized by a CLARIN ERIC member or observer country or have to be established as third party[2].

**Details:** The centre should provide a recommendation letter[3] from the national coordinator or the national SCCTC member. This is only required when a centre is assessed for the first time.

**Check procedure:** Check that the written recommendation letter exists and is signed by the national coordinator or the national representative in the SCCTC. This check is not relevant in case of re-assessment.

**Centre statement:**
*(Provide the letter as a separate document together with this document)*

### 2.c Visibility of connection to CLARIN (updated)

**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.

**Check procedure:** Check that a clear visual reference exists.

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

### 2.d Continuity of access and funding support (updated)

**Requirement:** Each centre needs to make explicit statements about perspectives of continuity of access and funding support to continue activities as an active CLARIN centre.

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

- The information from section R3 in the CoreTrustSeal application is seen as part of documentation for this requirement.

[2] See https://www.clarin.eu/content/clarin-eric-statutes , article 18
[3] See https://www.clarin.eu/node/3767 (CE-2013-0137) for a template.

- The centre also needs to prove there is funding to stay an active CLARIN centre.

**Check procedure:** Check that description guarantees reasonable funding support for at least two years. Check the financing situation for activities as an active CLARIN centre.

**Centre statement:**
*(Provide input on funding for being an active CLARIN centre, section R3 in the CoreTrustSeal is also seen as part of the documentation)*

### 2.e Details about resources and services provided (updated)

**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 and/or services for users from other CLARIN ERIC countries. Therefore, each centre has to give a short description of the resources and services offered to the CLARIN community, including which resources that are accessible without restrictions and which have access restrictions. Add information about which data that are **not** harvested by vlo.clarin.eu.

> **Why?**
> To ensure the Accessibility of FAIR: either the data or services are openly available, or they are easily accessible via Federated Login. In the latter case users from at least all CLARIN ERIC member countries should be able to login. Seamless cross-border authentication is part of CLARIN's mission.

**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 Identity Provider and national Identity Federations.

**Centre statement:**
*(Add description here)*

| | *Description* | *Access level* |
|---|---|---|
| *Data* | | |
| | | |
| *Services* | | |
| | | |
| | | |

*As illustrated in the example below:*

| | *Description* | *Access level* |
|---|---|---|
| *Data* | *Text corpora* | *Requires federated login* |
| | *Text corpora* | *Open* |
| | *Text corpora* | *Requires federated login and approval by resource owner* |
| *Services* | *Treebank search engine* | *Open* |

| | | |
|---|---|---|
| | *Speech Recognition web application* | *Requires federated login* |

### 2.f Registration in the Centre Registry

**Requirement:** Each centre should be registered in the Centre Registry.

**Details:**

- Register a CLARIN account at https://user.clarin.eu/user/register
- Fill in the form at https://www.clarin.eu/webform/registration-form-centre-registry (requires login with your CLARIN account)

**Why?**
The Centre Registry provides a uniform data store for all information about CLARIN centres. The data in the centre registry can be openly accessed via a web interface (for human consumption) or via a REST-API (for programmatic access). Several other infrastructure components (monitoring, the Virtual Language Observatory, Federated Content Search) use the API for retrieving (meta)data from the centres.

On top of this, the Centre Registry information is automatically transmitted to re3data.org – a registry of research repositories.

All in all, registration in the Centre Registry ensures improved Findability, Accessibility, Interoperability and Reusability

**Check procedure:** Check the provided URL from the Centre Registry

**Centre statement:**
(Provide the https://centres.clarin.eu/centre/[id] Centre Registry URL for your centre)

## 3. Intellectual Property Rights and Privacy

### 3.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. The centre should offer data access/sharing for users from other CLARIN ERIC countries.

**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[4] including a description of how licenses are presented to users. Note that this description could be the same as the one provided for the CoreTrustSeal R2.

**Why?**
FAIR Re-Usability requires a clear and accessible license information.

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

**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 Identity Provider or national Identity Federations.

**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.)*

### 3.b Privacy statement (unchanged)

**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[5]. It must also add this in a machine-readable way to its SAML metadata[6]

> **Why?**
> Adherence to privacy best practices leads to higher trust from Identity Providers and thus improved access for researchers (enabling smoother opt-in and attribute release).

**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.c Licenses on data and metadata (new item)

**Requirement:** Data and metadata are licensed.

**Details:** The centre has to provide a URL to a webpage where its information about licenses is described, enabling users of the repository to get information on how data and metadata are licensed.

**Check procedure:** Inspect the provided license URL(s).

[Note: at the moment, licensing for metadata is discussed by the legal issues committee. Awaiting the outcomes of this process, it is optional to provide a license for metadata]

## 4. Server Certificates

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

[5] See http://hdl.handle.net/11113/00-0000-0000-0000-19BA-5@view for an example

[6] See https://www.clarin.eu/node/3910 for more information

**Detail:** 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[7]. Use the SSL labs test at https://www.ssllabs.com/ssltest/ to optimize your SSL configuration.

> **Why?**
> The HTTPS protocol ensures a higher degree of security and privacy. Safe web access should work out-of-the-box in all standard browsers.

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

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

## 5. Federated Identity Management

[Note: if a centre only provides and will provide fully open resources, this requirement is not applicable]

**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. Optionally, you can connect to the CLARIN attribute aggregator as described at https://github.com/ufal/clarin-sp-aaggregator
3. Optionally, you can install the attribute debug script (shib_test.pl) at your Service Provider server: https://www.clarin.eu/page/3537
4. Joining the national Identity Federation (when available – see https://refeds.org/federations)
5. Allow users from the CLARIN IdP to login – see https://www.clarin.eu/page/3398
6. Join the CLARIN Service Provider Federation – see https://www.clarin.eu/spf
7. Allow users from at least one other country to login through their national identity provider
8. 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 the attribute aggregator[8] or the shib_test.pl if the right attributes are available.

[7] See https://wiki.mozilla.org/CA/Included_Certificates

[8] Available at https://lindat.mff.cuni.cz/services/aaggreg

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 should offer human and machine-readable access to language resources/services and their metadata. Centres need to offer component-based metadata (CMDI) that make use of elements from accepted registries such as the CCR[9] in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI-PMH. 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) and valid.

### 6.a Metadata harvestable via OAI-PMH to the VLO

**Requirement:** Computer access to the repository: Metadata harvesting to VLO should work - see https://vlo.clarin.eu/data/.

**Details:** The Metadata are harvested by the OAI-PMH protocol to the VLO https://vlo.clarin.eu/data/. The centre has to setup an OAI-PMH endpoint of the repository and give a link to it. The OAI-PMH endpoint should validate using the https://clarin.eu/oaivalidator. Harvesting metadata by the VLO should be working and 'search' URL(s) to the harvested metadata should be stated.

**Check procedure:** Validate the OAI-PMH endpoint; check that metadata are at least provided in DC and CMDI formats, and that the metadata shows up in the VLO using the specified URL(s).

**Centre statements:**
*(State if the centre fulfils the requirement)*

### 6.b CMDI metadata validation and use of CMDI profiles with CCR ConceptLinks

**Requirement:** 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. The CMDI metadata should validate.

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

**Details:** Use the curation module report output to state that:

1. The harvested CMDI files validate against their XML schema
2. The 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?

**Check procedure:** Check the curation module report for the specified collection using https://curation.clarin.eu/collection.

**Centre statements:**
*(State if the centre fulfils the requirement:* Use the curation module to check that CMDI files validate and that CMDI profiles are public*)*

### 6.c Metadata PID and references to resources.

**Requirement:**

1. State if the harvested CMDI files contain a PID[10] in the MdSelfLink header field
2. State if the harvested CMDI files refer to web-accessible files or a landing page with a ResourceProxy

**Check procedure:** Check if the harvested CMDI files contain a PID in the MdSelfLink header field. State if the harvested CMDI files refer to web-accessible files or a landing page with a ResourceProxy.

**Centre statements:**
*(State if the centre fulfils the requirement, give a URL/PID to a CMDI file, and a URL/PID to a* web-accessible file or a landing page*)*

### 6.d User access to the repository (updated)

**Requirement:** Each centre should setup a repository (a web-accessible server that offers human access to language resources/services and their metadata). Specify how human access is enabled.

- If access is possible via a web front-end for end users, then state the URL of the web interface of the repository.

- If access is only available via the VLO, state the URL of the VLO that can be used to browse through the (meta)data from the repository.

**Check procedure:** If offering user web-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. (Check for broken links and non-shibbolized password protection. Also check for access to either landing pages or resources.)

**Centre statements:**
*(State if the centre fulfils the requirement)*

[10] See section 7 for the details on PID requirements.

## 7. Persistent Identifiers (updated)

**Requirement:** Centres need to associate PIDs (handles or DOIs) 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 handle of the describing metadata record in combination with a part identifier[11] or with another handle.

**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[12])
- **application/x-cmdi+xml** (CMDI[13] 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 handle or a handle in combination with a part identifier, if these files:

- are accessible[14] 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 handles:

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

  **Why?**
  FAIR Findable requires unique identifiers on metadata and data.

**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

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

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

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

[14] The need for authentication to access an online file does *not* influence this.

On the command line this can be done as follows:

```
curl -L -H "Accept: text/html" http://hdl.handle.net/11372/VC-1000

curl -L -H "Accept: application/x-cmdi+xml”
https://doi.org/10.34733/vc-1000
```

If non-metadata files have handles, try to resolve a handle (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 examples of handles and/or DOIs)

## 8. Federated Content Search (optional)

**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:**
*(Optional requirement: 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)*

## 9. Attribute Checker (optional)

**Requirement:** Centres can opt to configure the Shibboleth SP Attribute Checker which assists in case of failed logins.

**Details**: The Shibboleth SP attribute checker automatically creates meaningful error messages to IdP Admins in case attribute release fails for a given SP. Personal data is not transmitted or logged. The Checker uses the same hook as the Aggregator below, there is documentation on how to use both.

The Checker cannot be meaningfully tested in production, since that requires a non-compliant IdP. The Centre can test the checker by requiring a non-existent attribute to be present during testing and triggering a warning during login. Local IdP admins should be informed about the testing, since they also will get the warning.

**Check procedure:** None. The statement of the Centre is sufficient.

**Centre statements:** The Centre acknowledges the implementation of the Attribute Checker or lack thereof.

## 10. Attribute Aggregator (optional)

**Requirement:** Centres can opt to configure the Attribute Aggregator provided by Lindat which provides insights into failed login attempts.

**Details**: The Attribute Aggregator provided by Lindat can provide valuable insights into failed login attempts from specific Identity Providers to various CLARIN services. Personal data is not transmitted or logged. The Aggregator uses the same hook as the Checker above, there is documentation on how to use both.

**Check procedure:** Log into the Centre's SP and then log into the Aggregator at https://lindat.mff.cuni.cz/services/aaggreg/. You should see the SP Login on the front page.

**Centre statements:** The Centre acknowledges the implementation of the Attribute Aggregator or lack thereof.