Title	CLARIN B Centre Checklist for the TLA assessment (2015)
Date	2015-04-13
Status	Final
Author	DvU

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 http://www.clarin.eu/node/3767 (CE-2013-0137) for a template.

Centre statement:

See separate document from the CLARIN-D technical management

Check procedure: Check that the written statement exists and is signed by the technical manager of the national CLARIN consortium.

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 at least 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.

Centre statement:

See https://corpus1.mpi.nl for the CLARIN logo + link to the assessment report on clarin.eu

Check procedure: Check that a clear reference exists.

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.

Centre statement:

See https://tla.mpi.nl/tla-news/tla-policy-statement/

As for the funding beyond 2016:

The Max Planck Institute for Psycholinguistics will support the archive and development of tools with 3 FTE of permanent staff.

A 3-year project together with the University of Cologne has been granted (funding body: German ministry of education and research, BMBF) that will provide funding for 3 years (as of 2015) for 1 full-time developer to work on the TLA repository system. Within the context of TLA there will also be support until June 2017 (1.5 FTE) from the Dutch royal academy of sciences (KNAW).

Additional proposals via the Max Planck Society to support the TLA repository on the longer term are currently being planned.

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

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

Centre statement:

https://tla.mpi.nl/resources/access-permissions/
https://tla.mpi.nl/resources/archiving-service/

(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.)

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.

2.b Privacy statement

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

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

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]

Centre statement: *https://corpus1.mpi.nl/IMDI/info/privacy_statement.html*

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 dpcoc@clarin.eu)

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 Data Seal of Approval or MOIMS-RAC approaches.

Details: For Data Seal of Approval see http://datasealofapproval.org. The centre cannot be certified as a B Centre until the DSA or the MOIMS-RAC assessment is achieved, but the CLARIN assessment procedure can be completed as long as the DSA or the MOIMS-RAC assessment is applied for.

Centre statement:

previous DSA seal:
https://assessment.datasealofapproval.org/assessment_48/seal/html/

an updated DSA assessment (guidelines 2014/2015) will be submitted before May 2015. Note: The new DSA submission system does not allow anymore to get a publicly viewable URL of the DSA submission, so we cannot include it here.

(Add URL to application here: Centre can login on http://assessment.datasealofapproval.org/assessments/ and provide the link to "Show")

Check procedure:
Is DSA or MOIMS-RAC achieved or applied for? – see
http://www.datasealofapproval.org/en/assessment/

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[4].

Centre statement:

[2] See http://www.csc.fi/english/research/sciences/linguistics/lat-privacypolicy for an example
[3] See https://www.clarin.eu/node/3910 for more information
[4] A complete list:
https://docs.google.com/spreadsheet/pub?key=0Ah-tHXMAwqU3dGx0cGFObG9QM192NFM4UWNBMlBaekE&single=true&gid=1&output=html

https://corpus1.mpi.nl *https://tla.mpi.nl/*

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

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): http://www.clarin.eu/page/3537
3. Joining the national Identity Federation (when available – see https://refeds.terena.org/index.php/Federations)
4. Allow users from the CLARIN IdP to login – see http://www.clarin.eu/page/3398
5. Join the CLARIN Service Provider Federation – see http://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 http://www.clarin.eu/spf

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

TLA has multiple service providers, all fulfilling the 7 requirements above. Its main repository is connected to the SPF and the CLARIN IdP and can be tested as follows: https://corpus1.mpi.nl > Login > External IdP > select home IdP or CLARIN IdP After a successful login the eduPersonPrincipalName is displayed in the upper part, next to "log out" The other SP can be checked via the shib_test.pl script: https://catalog.clarin.eu/secure/shib_test.pl

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.

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.

6. Metadata

Requirement: Centres need to offer component based metadata (CMDI) that make use of elements from accepted registries such as ISOcat 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[5]). It should feature an OAI-PMH endpoint through which the metadata can be harvested. The metadata should be CMDI-compliant (see http://www.clarin.eu/cmdi).

List of sub-requirements:

Computer access to the repository:

1. Setup an OAI-PMH URL of the repository
2. Show that the OAI-PMH URL of the repository validates using http://re.cs.uct.ac.za/

Harvesting of metadata:

3. Show that harvesting by the VLO can be done - see http://catalog.clarin.eu/oai-harvester/ for the results of the harvesting
4. Check at http://catalog.clarin.eu/vlo 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 + ISOcat:

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 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 that are used (http://catalog.clarin.eu/ds/ComponentRegistry) :
    a. Are they public?
    b. Do the elements contain valid ConceptLinks to ISOcat?

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 http://www.isocat.org/clarin/ws/cmd-core/#validation

Centre statements:

(1) OAI URL: http://corpus1.mpi.nl/ds/oaiprovider/oai2?verb=Identify

[5] See http://www.clarin.eu/cmdi

(2) Validates for 41 of the 42 checks, the schema validity (check 1) fails because there is a #-sign in the OAI identifiers, this issue will be resolved soon. It does not block the actual harvesting.

(3) Harvest results:
http://catalog.clarin.eu/oai-harvester/The_Language_Archive_s_IMDI_portal.html

(4) VLO results:
http://catalog.clarin.eu/vlo/search?q=tla:

(5) example
in VLO:
http://catalog.clarin.eu/vlo/record?q=kilivila&fq=collection:TLA:+Language+and+Cognition&docId=hdl_58_1839_47_00-0000-0000-0004-A0C6-2_64_format_61_cmdi

link to resources (video file, EAF transcription):

http://hdl.handle.net/1839/00-0000-0000-0004-A0C7-B (video)
http://hdl.handle.net/1839/00-0000-0000-0004-A0C9-2 (EAF)

(6) all CMDI files validate

(7) they contain a handle in the MdSelfLink, e.g. hdl:1839/00-0000-0000-0004-A0C6-2@format=cmdi

(8) web-accessible (if the permissions allow it), landing page is always available, eg:
http://hdl.handle.net/1839/00-0000-0000-0004-A0C6-2@view

(9) used CMDI profiles:

imdi-session:
http://catalog.clarin.eu/ds/ComponentRegistry?itemId=clarin.eu:cr1:p_1271859438204&registrySpace=published

imdi-corpus:
http://catalog.clarin.eu/ds/ComponentRegistry?itemId=clarin.eu:cr1:p_1274880881885&registrySpace=published

ToolService:
http://catalog.clarin.eu/ds/ComponentRegistry/?item=clarin.eu:cr1:p_1311927752306

(9a) both are public
(9b) both contain valid CCR links

(10) https://corpus1.mpi.nl

(11)
Instances: http://catalog.clarin.eu/metadata/cmdi/services/

These files do validate:
Validation result
*The XML document is a **valid** instance of the CMD core model for CLARIN Web Service descriptions.*

(For sub-requirements 1 to 9 state that the centre fulfils the requirements. If sub-requirement 8 and 9 apply for the centre state that the centre fulfils these requirements as well)

Check procedure:
Check computer access to the repository: Enter the OAI-PMH URL at http://re.cs.uct.ac.za/and see if it validates.

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

CMDI files + profiles + ISOcat:
Validate the harvested CMDI files against their XML schema
Check the profile(s) used at the component registry (http://catalog.clarin.eu/ds/ComponentRegistry):

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

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 http://www.isocat.org/clarin/ws/cmd-core/#validation

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[6] 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

[6] See http://www.clarin.eu/faq/3453

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 mimetypes:

- **text/html** (web-browser, human readable[7])
- **application/x-cmdi+xml** (CMDI[8] 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[9] 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

Centre statements:

(For each sub-requirement state that the center fulfils the requirements)

All the TLA metadata files have a PID, eg:

http://hdl.handle.net/1839/00-0000-0000-0004-A0C6-2@format=cmdi

The (non-metadata) files also have a PID, eg:

http://hdl.handle.net/1839/00-0000-0000-0004-A0C7-B (video)
http://hdl.handle.net/1839/00-0000-0000-0004-A0C9-2 (EAF)

Support for HTML and XML (CMDI) rendering of the metadata PIDs is underway: it is currently deployed on an internal test server and should be made available before the end of March in production.

At that time the PID above (http://hdl.handle.net/1839/00-0000-0000-0004-A0C6-2@format=cmdi) should be rendered as HTML in the browser, and as XML otherwise.

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

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.

[7] A generic CMDI-to-HTML XSLT is available at http://infra.clarin.eu/cmd/xslt/cmdi2xhtml.xsl
[8] See http://www.clarin.eu/cmdi
[9] The need for authentication to access an online file does *not* influence this.

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 (http://www.clarin.eu/fcs).

Centre statements:

The TLA has an FCS-compliant endpoint: http://cqlservlet.mpi.nl/

Because it already uses the newest version of the protocol it cannot be tested yet with the validator mentioned below.

However, the new FCS aggregator shows that this endpoint is compliant at http://weblicht.sfs.uni-tuebingen.de/Aggregator/stats:

MPI for Psycholinguistics - http://cqlservlet.mpi.nl/
Max concurrent scan requests: 4
1 request(s), average:0.655s, max: 0.655s
5 root collection(s):
Corpus Spoken Dutch
The European Science Foundation Second Language Database
IFA Corpus
Child Language Data Exchange System (CHILDES)
Corpus TalkBank

(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)

Check procedure: enter the endpoint URL at http://clarin.ids-mannheim.de/srutest and validate.