State of the technical infrastructure

Dieter Van Uytvanck
Technical Director CLARIN ERIC
dieter@clarin.eu

CLARIN Annual Conference 2019
Leipzig
2 October 2019
An overview of the technical infrastructure highlights

By Florian Koppe - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=25767605
Short round-up 2019: Centres

• All newly (re)certified centres now in sync with the CoreTrustSeal (validity of 3 years)
• 1 new B-centre: CELR (EE)
• Re-assessment of 9 centres: ASV (DE), BBAW (DE), EKUT (DE), IDS (DE), IMS (DE), LINDAT (CZ), MPI (NL), UCPH (DK), UdS (DE)
• 7 more assessments pending
• in total 22 certified B-centres, 57 registered centres
Short round-up 2019: Federated login

• New countries in Service Provider Federation: South Africa and Cyprus
• Total number of organisations that can login: **1870**
• On average **2082 logins per month** via the central discovery service
• Major upgrade of the CLARIN Identity Provider, paving the way for future improvements
Short round-up 2019: VLO

- Two releases (4.6 and 4.7) brought
  - collapsing of similar records
  - link checking
  - better visibility of contributing centres
Short round-up 2019: Curation (1)

• The completely revamped Curation Module provides
  – a faster and more complete overview
  – insight in the metadata provided by the CLARIN centres

• It is now actively used during the centre assessments

• [www.clarin.eu/curation](http://www.clarin.eu/curation)
Short round-up 2019: Curation (2)

- Link check results are now also integrated into the VLO

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>lat:1839_00_0000_0000_0009_295C_5</td>
<td>landing page</td>
</tr>
<tr>
<td>HDL content</td>
<td>audio file</td>
</tr>
</tbody>
</table>

Resource details

- **Mime type**: audio/x-wav
- **Link**: https://hdl.handle.net/1839/00-0000-0000-0009-295D-C

Automatic link checking

- **Authentication and/or special permissions may be required in order to access the resource at this location.**
- **Result of automatic checking**: access to the resource is restricted (401 Unauthorized)
- **Last check**: 28 August 2019 17:37
Short round-up 2019: Content Search

• New
  – version of the protocol and search engine
  – endpoints by IvdNT and ILC4CLARIN

• Currently there are 15 centres with an endpoint
  – 9 have at least one version 1 endpoint
  – 9 have at least one version 2 endpoint
Short round-up 2019: Virtual Collection Registry

- New version with
  - Improved user interface
  - Better connectivity to the switchboard
Short round-up 2019: Switchboard

• Major overhaul of the back-end lead to a more robust and stable version
• Updated integration into the EUDAT cloud storage platform (B2DROP),
• Integration into the Parthenos VRE (D4Science) completed
• Please share your experience!
  – beta-switchboard.clarin.eu
  – switchboard@clarin.eu
Short round-up 2019: Back-end

• The average uptime for the period September 2018 to August 2019 for 9 central services is 99.84% (99.91% last year)

• Deployment of a central log server
  – Fluentd, Kibana, Elastic Search

• Improvements to the geographically redundant setup.

• Continuous code-based security scans.

• Upgraded Monitoring setup (Icinga2)
Tactical Overview

Host Summary
- 8 Hosts Down

Service Summary
- 31 Services Critical
  - 229 Ok
  - 1 Warning (Handled)
  - 25 Warning
  - 5 Critical (Handled)
  - 31 Critical
<table>
<thead>
<tr>
<th>System</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha-vlo-clarin_esc_rzg_de</td>
<td>1.6 kbps</td>
</tr>
<tr>
<td>beta-vlo-clarin_esc_rzg_de</td>
<td>4.9 Mbps</td>
</tr>
<tr>
<td>centres-clarin_esc_rzg_de</td>
<td>632.7 bps</td>
</tr>
<tr>
<td>centres2-clarin_esc_rzg_de</td>
<td>945.6 bps</td>
</tr>
<tr>
<td>cesge-vps1</td>
<td>2.7 kbps</td>
</tr>
<tr>
<td>clarin-vcr_ids-mannheim_de</td>
<td>1.2 kbps</td>
</tr>
<tr>
<td>clarineric-backups</td>
<td>1.9 kbps</td>
</tr>
<tr>
<td>clarineric-vps2</td>
<td>1.4 Mbps</td>
</tr>
<tr>
<td>clarineric-vps3</td>
<td>7.0 kbps</td>
</tr>
<tr>
<td>clarineric-vps4</td>
<td>48.5 kbps</td>
</tr>
<tr>
<td>clarineric-vps5</td>
<td>295.3 kbps</td>
</tr>
<tr>
<td>clarineric-vps6</td>
<td>7.1 kbps</td>
</tr>
<tr>
<td>clarineric-vps7</td>
<td>6.0 kbps</td>
</tr>
<tr>
<td>clarineric-vps8</td>
<td>1.9 kbps</td>
</tr>
<tr>
<td>clarinvvm</td>
<td>105.5 kbps</td>
</tr>
<tr>
<td>dev-idp-clarin_esc_rzg_de</td>
<td>901.7 bps</td>
</tr>
<tr>
<td>idp1-clarin_esc_rzg_de</td>
<td>1.6 kbps</td>
</tr>
<tr>
<td>idp2-clarin_esc_rzg_de</td>
<td>163.5 kbps</td>
</tr>
<tr>
<td>rs236235_rh_hosteurope_de</td>
<td>24.2 kbps</td>
</tr>
<tr>
<td>rs238144_rh_hosteurope_de</td>
<td>416.5 bps</td>
</tr>
</tbody>
</table>
Bridging CLARIN and the rest of the universe

https://commons.wikimedia.org/wiki/File:Brandenburger_Br%C3%BCcke_Leipzig_August_2013_018.JPG
European Open Science Cloud (EOSC)

- See the demonstration by Maciej Ogrodniczuk
- www.clarin.eu/eosc
FAIR

• FAIR is in CLARIN’s DNA
• www.clarin.eu/fair
Europeana

- CLARIN and Europeana make discovery and processing quick and easy for 135,000 cultural heritage objects
- clarin.eu/europeana
Under the surface of the technical infrastructure
Resource families links

- A first round of 1040 link checks for the 8 resource families in June
- In June, around 20 broken links (2%)
- About 10 different people were contacted to follow-up
- These conversations gave some good insight in the dynamics behind a simple link
Broken links – technical causes

• 503 Service Unavailable (temporary maintenance)
  – Still needs a check later on

• Real technical issues were found
  – “our handle server is running but cannot open connections to the database”
  – “firewall misconfiguration of our handle server”
  – unreachable DOI: “machine was running out of space; fixed now”
Broken links – other causes

- URL changes & DNS failures
  - “corpus project page got a new URL, that one should be used”
  - “data now available at new subdomain”
  - “… told me that the server has been discontinued, data still available at our repository”

- Observations:
  - Often a URL change reflects an organisational change – it cannot be avoided
  - Interaction with resource providers lead also (unexpectedly) to improvements of the description (fixes, additions, …)
Link checking: some first conclusions

• Automated link checking is of utmost importance
• Following-up requires time, but this is a good investment since improvements often come with a broader impact:
  – Fixes of otherwise “hidden” technical issues
  – Better metadata, reflecting e.g. a new structure of a dataset
• This is a theme that is also relevant on a broader scale (think EOSC)
Thank you for your attention!

... and credits for the following CLARIN superheroes:

- The assessment committee:
  - Lene, Cyprian, Daan, Jozef, Riccardo, Tomas

- The CLARIN ERIC central developer team:
  - André, Hendrik, Menzo, Michał, Twan, Willem

- Developers and contributors from the national consortia:
  - Alexander, Can, Claus, Dirk, Emanuel, Leif-Jöran, Matej, Nathanael, Oliver, Tariq, Thomas, Wolfgang

- The taskforces:
  - AAI, CMDI & Curation, FCS, PIDs

- And all the others who are contributing to the construction & operation of the infrastructure!