Title	Minutes SCCTC 18 November 2020
Version	1
Author(s)	LS
Date	2020-11-18
Status	Approved
Distribution	Centre Committee
ID	CE-2020-1786

Participants

Martin Matthiesen (Chair, FI), Jan Hajič (CZ), Simon Gray (DK), Bart Jongejan (DK), Krista Liin (EE), Christophe Parisse (FR), Riccardo Del Gratta (IT), Daan Broeder (NL), Luís Gomes (PT), Martin Hennelly (SA), Leif-Jöran Olsson (SE), Cyprian Laskowski (SI), Brian MacWhinney (USA), Dieter Van Uytvanck (CLARIN ERIC), Linda Stokman (CLARIN ERIC, minutes)

Excused: Vanja Štefanec (HR)

0 Action points

#	Action	By whom	By when
1	Recommendations on PIDs	Daan & Dieter	March 2021
2	Follow up on CNRI/EPIC Handle issue	Dieter	End of Feb 2021
3	Set-up a Google doc for exchanging information relevant for the CTS certification process and gathering and registering useful information resources.	Daan	Nov 2020

1 Agenda

1. Agenda: the agenda is approved.
2. Approval minutes last meeting (14.10.2020) & action point status
    a) Update on EPIC unresponsiveness (DVU)
3. Centre assessments
4. FIN-CLARIN approached by Wikitongues to take over archive
    a) Short videos, 500 languages, 1-2TB, Metadata situation unclear
    b) Lack of funding in Finland, are other Centres interested?
5. Outcomes of a preliminary "VLO in Google Data Set Search" evaluation: https://office.clarin.eu/v/CE-2020-1776-Data-set-search-for-records-in-the-VLO.pdf (DVU)
6. Status update per country/member (please provide a short bullet-wise summary in the Google Doc)
7. AOB

2 Approval minutes last meeting & action points

The meeting minutes of 14 October 2020 (CE-2020-1758) are approved.

a) AP2: Update on EPIC unresponsiveness (DVU)

In short: Bulgaria requested an EPIC account but never received an answer. This was escalated and a response was finally received, and Bulgaria confirmed they received the information needed. In the end, it turned out the person assigned to this ticket retired and the request was not forwarded to another colleague. This specific case is solved but the situation in which the terms of responses and support is sub-standard remains.

Daan comments there are more EPIC providers besides this one, GRnet but also SURFsara. Daan was in contact with SURFsara concerning a prefix or getting a quote for that and received a response within a week. Daan recommends doing a sort of test , the next one that needs a prefix could perhaps request one from SURFsara.

Dieter comments we're currently paying a membership fee of 5000 euros or so per year. If we see a possibility to do this test without too much admin overhead, it would be good to do so. This is something to investigate, since there are a lot of questions that need to be answered first. We should give it about three months' time to find out about all the details here and in the meantime, try to already start to document this so that we have the other action point (#AP1 recommendation on PIDS), providing documentation on persistent identifiers in reasonable shape.

3 Centre assessments

Currently no updates.

4 FIN-CLARIN approached by Wikitongues to take over archive

Martin Matthiesen: Wikitonguess approached us directly in Finland. They are a non-profit organization with a focus on providing tools and resources for language sustainability. They have quite a few videos and about two terabytes of data. At the moment their archive is with the Library of Congress in the US and they're looking for a second base in Europe. Due to a lack of funding in Finland, we are unable to accommodate, but wonder if other centres are interested.

Jan Haijc has already indicated an interest and Martin will bring him in contact with Wikitongues.

5 Outcomes of a preliminary "VLO in Google Data Set Search" evaluation (DVU)

Link: https://office.clarin.eu/v/CE-2020-1776-Data-set-search-for-records-in-the-VLO.pdf

State of affairs:
Efforts that have been made so far to adjust the VLO to expose the metadata in such a way that it is embedded in the webpages and can be harvested by the Google Data Set Search haven't delivered a lot of effect. The main problem seems to be that only a small portion of the whole VLO is indexed/crawled by Google and secondly that within this small portion, the number of records that actually contain embedded information is absolutely minimal.

Dieter attended a session of a working group that is dedicated to this topic at the last RDA meeting. There it turned out others have similar issues and experiences. (More info: meeting and meeting minutes).

Two general recommendations were made by some people in the group.

1. Provide the embedded metadata at a lower level. So, at the individual repositories, which probably look a bit more like the average website because they have fewer pages, and they have more incoming links. This will probably be good for the index ability algorithms from Google.
2. Provide the metadata at the higher lever through big players. The ones where we are 100% sure that they are indexed and that's for instance the DataCite metadata catalogue.

It seems that right now, after 3 to 4 months after implementing, it is not a successful strategy to get metadata into Google Data Set Search. There is an idea to invite someone from Google to the next RDA plenary in the spring. But it is very difficult to get additional information on these inhouse secrets from Google where they prefer to keep it a black box.

Martin Matthiesen: Instead of having to resort to different centres everything is handed to Google via the VLO. Is the VLO not big enough?

Dieter comments that this has to do with the nature of the platform. The VLO is not a typical website but a search portal. And it does not have or hardly has any incoming links. People are linking to the original repository, not to the VLO. At the same time, it contains lots of pages, over a million right now. A lot of these pages are looking very similar. And there have been some speculations that the similarity between many of the pages in the VLO are actively discouraging indexation by Google, because it then does not see enough entropy. From a technical perspective, if you want to optimize two things that both have incoming links and are perceived as being more of a traditional repository or website, it does make sense to try to get the repositories directly harvested.

The question now is how far we want to invest in something that is in the end a black box. It does not make sense to continue on this track, not knowing how things happen and having hardly anyway to find out why thing is happening or not. It would be best to think about other strategies. This also relates to the long-term development plans for the VLO. For instance, there is some potential for instead of having one monolithic portal, having one simple backend that still has all the metadata and all the information that the VLO has but at the same time establishing smaller portals that are based on the information from VLO, but do look more like a traditional website. In that sense they are easier to grasp, both by individual visitors, but also by search engines. But this is something that we need to explore in more detail. This will change the architecture of the VLO. These are all potential options that have been put in our update of the strategy at the ERIC and are being discussed with the VLO developers. It will be part of the next major update of the VLO, version 5, which is planned for 2021.

6 Status update per country/member

Austria

- No report

Bulgaria

- No report

Croatia

- No report

Czech Republic

- Continued work on “Evaluation 2021”, to renew LINDAT from 2023 on
- Presentations at the SSHOC/FREYA conference
- Got a new INFRA project - “CLS INFRA” (with DARIAH EU and 11 other partners), and awarded another H2002 project in call on Language Equality - project “ELE”, coord. By Andy Way (Dublin, Ireland)

Cyprus

- No report

Denmark

- Workflow manager system Text Tonsorium made available in CLARIN-DK.
- MUMIN multimodal annotation specifications (ANVIL and ELAN) uploaded in CLARIN-DK.
- Maintenance/bugfixes for CLARIN-DK repository and KORP.

Estonia

- KORP performs faster after adding RAM, except when computing statistics with more rows to show.
- Slow transfer to Kubernetes.
- User involvement - virtual seminars on one topic seem to work well.

Finland

- Korp will be moved away from CentOS6 before updating (deadline 30.11.)
- LAT will be shut down, first replaced by Download service only.
- Waiting for funding decisions, first to come end of November

France

- Participation in SSHOC activities related to CLARIN (Vocabularies meeting, EOSC/FREYA/EOSC-HUB, SSHOC poster during CLARIN AM etc.)
- Finishing of the Tour de CLARIN
- Discussions at the ministry level as the national roadmap for infrastructures is to be totally rewritten in January 2021 which is closely correlated to the French participation into European infrastructures

Germany

- Fedora Commons: first successful test migrations FC3 -> FC6, more contact with Lyrasis/Duraspace about identified bugs

Greece

- No report

Hungary

- No report

Iceland

- No report

Italy

- ILC4CLARIN
    - Release 2020.01
    - Attempt to change login method for release 2020.02
    - CTS is approaching its deadline. Re-certification in Spring

Latvia

- Participation at DHN 2020 conference with presentation "CLARIN in Latvia: From the Preparatory Phase to the Construction Phase and Operation."

Lithuania

- No report

The Netherlands

- CLARIAH has been included in landscape inventory of large-scale research infrastructures. This is a precondition for being eligible for inclusion on the *National Roadmap Large-scale Research Infrastructures*, which in itself is a precondition for being eligible for funding in a new round
- Antal van den Bosch (director of the Meertens Institute) has taken over the role of Principal Investigator of the CLARIAH-PLUS project from Lex Heerma van Voss (director Huygens Institute)
- We are closely collaborating with EHRI in the Netherlands (European Holocaust Research Infrastructure)
- We are working hard on a new website and a new portal, both to be launched early next year.

Norway

- MENOTA (Medieval Nordic Text Archive) launched its integration in CLARINO on Nov. 17, 2020

Poland

- CLARIN-PL in research practice - 2 days of workshops (280 registered users)
- November weekly series of seminars for students and PhD students of sociology at the Warsaw University

Portugal

- New resources added to the repository
- Implementation and testing of file processing (continued)
- Implementation and testing of web services API (continued)
- Dissemination at the largest national linguistics conference

Slovenia

- First version of ParlaMint 4-language corpus finished and submitted to repository (http://hdl.handle.net/11356/1345)
- First resources for processing Macedonian added to repository (http://hdl.handle.net/11356/1373, http://hdl.handle.net/11356/1374, http://hdl.handle.net/11356/1359)
- Re-doing the documentation (internal and external) of the technical aspects of CLARIN.SI

South Africa

- In flight to integrate CLARIN SP to our identity management system; Metadata delivered; environment setting for test ongoing

Sweden

- "Digital humanities" declared a priority RI area by the Swedish Research Council for its next call (spring 2021)
- The annual autumn workshop of the National Language Bank was held (online) on 16th October
- Swedish diachronic corpus v. 1.0 released. Swe-NERC v1.0 will be released next week for SLTC 2020.
- Webinar series on crowdsourcing and automatic transcription, arranged by the Institute for Language and Folklore (October–November, 2020)
- New project grant: SweTerror (A multimodal study of terrorism), led by KTH and including CDH/SBX (U. Gothenburg)

United Kingdom

- Four new consortium members (Cardiff, Edinburgh, King's, Huddersfield), and renewed contacts at the British Library.
- Launch of CorCenCC, the Welsh national corpus, available from Cardiff University, visible in the OTA repository and VLO
- Publication online of UK Tour de CLARIN posts
- Ongoing discussions regarding a possible application for full membership, and national funding.

USA (third party TalkBank - CMU)

- Waiting for CTS approval, working on collaborative commentary browser, new CLAN, and TalkBankDB (FCS)

7 AOB

Screencast

A blog and a screencast on how the switchboard can be used has been created and published. It is mostly about CLARIN materials and services. It can function as promotion

and educational material for outsiders to get a quick idea about how things work. Feel free to use this for promotional purposes (presentations etc.) and to add it to your website.

If you have feedback, please inform Dieter. The second iteration of the screencast will be next year.

Link to the screencast: https://www.clarin.eu/blog/clarin-services-european-open-science-cloud

Next meeting: A doodle for the next virtual meeting will be sent out shortly.