Title	CLARIN-PLUS Improved harvesting workflow work plan
Version	1
Author(s)	Menzo Windhouwer (CLARIN ERIC/Meertens Institute)
Date	2016-04-04
Status	Final
Distribution	Public
ID	CE-2016-0762

1 Introduction

For its core catalogue service, the Virtual Language Observatory, the CLARIN infrastructure relies on an OAI harvester that periodically harvests all CMD records from the CLARIN centers. The OAI harvester has been developed in house and has been improved over the years. But still it lacks some functionality, which will help the harvesting process to keep up with the scale of growth of both the number of CLARIN centers and the number of CMD records.

2 Endpoint specific logging

During a harvest a log is created. The log tells which endpoints will be harvested and, when it is the endpoint's turn, how the interaction with the server went, *i.e.,* which records were retrieved or which errors occurred. When a centre's system administrator has questions about the harvest the log is manually analysed by one of CLARIN's system administrators. If logging messages could be targeted at endpoint specific logs it would be relatively simple to make them available via the OAI harvester viewer and would enable the center's administrator to quickly determine if there have been any problems with the endpoint during the harvest.

3 Incremental harvesting

The current practice of the OAI harvester is to request all the CMD records from an OAI endpoint. However, the OAI-PMH standard, which describes the protocol used by the harvester, also has provisions for incremental harvesting. This could potentially speed up the harvesting process considerably, but several requirements need to be met:

1. Endpoint needs to support this part of the protocol, and as some will not it should always be possible to fallback to a full harvest of such an endpoint.
2. The harvester needs to keep information about previous runs, *e.g.*, timestamps.
3. To be able to offer still all CMD records to the VLO the previously harvested records need to be available still, or the VLO needs to be able to process the delta information (*i.e.,* which records were deleted).

There has been some work in the OAI harvester on incremental harvesting, *e.g.,* an overview file is kept containing information on the last run to be used by a subsequent incremental run (see requirement 2). However, the full workflow to use incremental harvesting in production was never realized.

4 Integration with metadata quality and curation

Within the work page (WP 2) a tool is developed that can assess the quality of a metadata record (T 2.2.1). This tool needs to be integrated into the harvesting workflow, *i.e.*, when a harvest is finished the tool needs to be run and the quality report per endpoint should be available to the centre's administrator preferable by the harvest viewer.

5 Planning

Year	**2016**																			
Month	**April**				**May**				**June**					**July**				**August**			
Week	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
M 1 Endpoint specific logging	0.75 PM																				
M2 Incremental harvesting							2.5 PM														
M3 Integration with metadata quality and curation																			0.75 PM		