DSI Progress Report 01 April – 31 July 2015

This form is for each partner to report on work actually carried out from 01 April – 31 July 2015 for all work packages.

Please provide the following information, per institution:
- Tasks due this period, according to the work plan
- What you have accomplished (last four months)
- Report on any deviations from the workplan and their impact on other tasks as well as on available resources and planning. If applicable, propose corrective actions.

Reporting in bullet-points is acceptable, but always make sure to report about your activities in a way that allows people outside your institution to completely understand what you did.

Name *

Christian | Thomas

First Last

Email *

thomas@bbaw.de

Institution *

CLARIN-ERIC / CLARIN-D team at the Berlin-Brandenburgische Akac

Work Package *

WP3 Re-user services ▼

Tasks due this period, as listed in the Description of Work

preliminary remark: please note that the work described here by me, Christian Thomas, for the CLARIN-ERIC / CLARIN-D team at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) has by contract/actually been done by Axel Herold, herold@bbaw.de

* aggregate europeana metadata (formats: Dublin Core and Europeana Data Model (EDM)) from OAI-PMH Handler at http://data.theeuropeanlibrary.org/oaipmh/OaiPmhHandler
* convert these metadata into CMDI format
* integrate selected test dataset of above-mentioned europeana metadata into CLARIN's Virtual Language Observatory (ongoing)
* develop (specific) CMDI profile for (historical) newspapers (ongoing)
* give feedback on metadata quality, depth and structuring
* give feedback on newspaper OCR full-text from experimental text dumps at https://github.com/nfreire/HistoricalNewspapersCorpus/ (cf. http://research.europeana.eu/blogpost/experimental-text-dumps-from-europeana-newspapers): full-text quality, depth and structuring
* prepare report on Europeana Research distribution plan

Achievements against the tasks listed above

* aggregated europeana metadata (formats: Dublin Core and Europeana Data Model (EDM)) from OAI-PMH Handler at http://data.theeuropeanlibrary.org/oaipmh/OaiPmhHandler
* developed (specific) CMDI profile for (historical) newspapers (ongoing): prototype is ready and currently discussed within CLARIN team and the VLO task force
* gave feedback on metadata quality, depth and structuring, available at http://research.europeana.eu/blogpost/experimental-text-dumps-from-europeana-newspapers
* harvest newspaper OCR full-text from experimental text dumps at https://github.com/nfreire/HistoricalNewspapersCorpus/ (cf. http://research.europeana.eu/blogpost/experimental-text-dumps-from-europeana-newspapers)
* gave feedback on newspaper OCR full-text from experimental text dumps at https://github.com/nfreire/HistoricalNewspapersCorpus/ (cf. http://research.europeana.eu/blogpost/experimental-text-dumps-from-europeana-newspapers): full-text quality, depth and structuring
* prepare report on Europeana Research distribution plan

Deviations from the Description of Work?

none.

Work planned next three months?

* further improvements on the prototype CMDI profile for (historical) newspapers
* release CMDI profile for (historical) newspapers in September 2015 at CLARIN Concept Registry Browser (https://openskos.meertens.knaw.nl/ccr/browser/)
* convert europeana's EDM/DC metadata into that CMDI flavour
* integrate a (curated) subset from europeana's newspaper collection into CLARIN's VLO
* evaluate, discuss and propose possible (and realistically reachable) improvements on the way OCR full-text data is shared via europeana: more in-depth structuring, standardized formats like ALTO, tighten the connection of each document/article to its (standaradized) metadata, etc.
* evaluate and, if promising, implement an instance of CLARIN's Federated Content Search (FCS) for full-text data from europeana's newspaper collection

Please also fill out the Meetings, Presentations and Publications spreadsheet

(careful: this opens in the same windows, please submit this survey first!)

If you have any problems or questions about filling out this reporting form, please get in touch with me at nicole.emmenegger@europeana.eu.

Thanks for taking the time!