Notes meeting - Curation Module - 2016-04-08

Agenda

- Curation module
  - status
  - score calculation
  - web application UI
- VLO
  - 3.4 status
  - 4.0 milestone
- Curation workflow - integration of curation module and VLO
  - Sustainable maintenance of the mappings (including complex
  - Differentiate between types of mappings
  - Using quality score for ranking?
- Any other business

Commented [1]: postpone to next meeting

latest version of curation module: https://clarin.oeaw.ac.at/curate/. (unfortunately - some technical difficulty as of today morning :(sorry, for that)

ui mockups for web app.
https://moqups.com/ostojic.davor@gmail.com/L9GiyDLr

See also: trac page https://trac.clarin.eu/wiki/Curation%20Module

Davor: Status:
- Backend finished
- Normalisation missing yet

Facets currently only
Menzo: could you call the postprocessors
Davor: Where to get the mappings
Twan: Where
We should have both the original and the normalized value in
Dieter: would move mapping from importer to curation
Matej: should we make all the mappings/normalisation explicit/externalised/ made more declarative?

Dieter: Do we really need to take out the normalisation out of vlo-importer?

Menzo: have processors on both places
MAtej: same implementation on both sides - share post-processers

Thomas: format of outputted curated md-records?
Matej:
- For Vlo: solr-simplified records
- For metadata author: original cmdi-records (optionally normalised/original)
-

Twan:

- To what degree are preparation for the VLO and improving the quality of a source record actually the same task?

Dieter: do we really need a machine-readable output, or would a human-readable suggestions for fixing be enough.

Matej: For now enrich the facet-section in the curation report with normalised values.
Only later see if we can/shall also provide curated/normalised records.

Matej: PARTHENOS JRR as consuming registry for CMDI-records
Dieter: we need to see, if it really comes. Not invest too much.

Dieter: integration harvester curator

Menzo: shell scripts would invoke curator and curation-report is stored next to the data
And will be available in the web application next to the datasets

Score calculation
Dieter: make it by means of a configuration
Matej, Dieter: needs calibration based on experiments, on real data.
Twan: not all facets should be taken into account

MAtej: Davor will run curation for the whole harvest (by individual providers) and make it available as some html.

Matej: **integration with VLO**? Vlo-importer would read a curation report. (either invoking curation module for every record; or create a set of reports parallel to the records collection)
Dieter: Do not really see the case. Should we actually add the score into the VLO?
Twan: reward for the authors, better user experience(?)

VLO
3.4 status
Twan: maintenance release (performance hit due to ranking) -> 3.4.1
Statistics-app - grafana, stastd (collection size, over time, status of the harvester)
Currently being tested

4.0
Twan: Priorities
- New visual style at LREC (Prague working on that)
    - https://github.com/clarin-eric/interface_guidelines
    - http://www.clarin.studiovelmans.nl/styleguide.html
- Release of CMDI 1.2

Number of records

stats.gauges.vlo.dev.index.nrRecords

Number of records per collection

- Meertens_collection__Liederenbank
- UBU_Clarin_Set
- TalkBank
- Nederlands_Instituut_voor_Beeld_en_Geluid_Academia_collectie
- WomenWriters_documents
- TLA__DoBeS_archive
- TLA__Language_and_Cognition
- Bavarian_Archive_for_Speech_Signals__BAS_
- Institut_f_r_Deutsche_Sprache__CLARIN_D_Zentrum__Mannheim
- e_BNM_tekstdragers
- TLA__Acquisition
- TLA__MPI_CGN
- Unnamed_provider_at_cocoon_huma_num_fr
- Leipzig_Corpora_Collection
- TLA__CLARIN_NL
- Other collections

Number of values per facet

Harvesting rate

harvested

Import rate

imported

Harvester Log

Info — Warning — Error

Importer log

Info — Warning — Error

Total records imported

Records Imported

Combined harvester and importer