Curation TF meeting - notes

Centre Meeting 10-12 May, 2016

Date	11. 05. 2016
Time	15:10 - 17:00
Location	Utrecht, Netherland

Participated:	Topics:
Claus Zinn	Curation Module
Davor Ostojic	Availability / License Facet
Francesca Frontini	Normalisation Workflow
Henk van den Heuvel	ResourceType Facet
Menzo Windhouwer	
Mitchell Seaton	
Neeme Kahusk	
Pavel Stranak	

Discussion[1]

Curation Module

Menzo: add to the collection view a column for the profiles with scores and link profiles to the view page. It will be helpful for the Meerents institute to identify critical profiles

Pavel: OLAC has similar tool and result representation is very good. It offers hints how to improve the score.

Davor: there is a plan to develop such an application (mockups)

Davor: For profiles assessment both xsd and xml are needed. What happens when schemalocation attribute points the different location then component registry? In that case only xsd is available but for fetching xml url can reconstructed with profileID from CMDI's header. Which schema to use then, from component registry or originally specified? (example: http://hdl.handle.net/11858/00-203Z-0000-0027-536B-3)

Menzo: this situation should be considered as invalid. Schemas could be out of sync. Give penalty to the score for this. Schema from component registry should be used.

Davor: There are situations when <MdProfile> element has an url (component registry by the way) instead of profileId (clarin.eu:cr1:p_...). Hard to handle all of them

Menzo: There are no strict rules what is allowed for <MdProfile> element but id is prefered. It should be reported by the tool. Schematron could be use for such a validation.

Davor: Should Schematron be consider in curation module?

Menzo: why not it is an ISO standard. CMDI Validation tool from Oliver Schonefeld is using it already. Validation rules can be externalised.

Davor: Score calculation still needs to be coded (maybe this can be externalised as well)

[1] Remark: This document gives an overview over discussed topics during the meeting. Order of the topics in this document may not correspond to the original order from the meeting! Some additional resources might be added.

Availability / License Facet

Participants are informed about what has been changed.

Pavel: it is very confusing which values to use for availability / license. Would be nice to have guidelines

Claus: CLARIN provides licence category calculator

Pavel reported problems for LINDAT's data regarding this facet. Needs to be investigated whether VLO importer interprets original values in a wrong way or necessary fields are not specified by LINDAT. It seems that curation instance of VLO (with old mappings) gives correct result.

Agreement is to try to fix LINDATs data together and then from gathered experience to try to create some guidelines or best practices for others.

Email from Florian Schiel sent on 29/04/2016:

The module report a missing facet 'rightsHolder' on instances of profile 'media-corpus-profile', but the profile (and the instances I tested) contains the element <Owner> which is linked to concept http://hdl.handle.net/11459/CCR_C-2956_519a4aab-2f76-0fd3-090e-f0d6b81a7dbb

Email from Hanna Hedeland sent on 02/05/2016:

Since I put this facet in the facetConcepts.xml sometime during the VLO-TF time, I was quite sure we should have this information, but in fact it's now represented by http://hdl.handle.net/11459/CCR_C-2956_519a4aab-2f76-0fd3-090e-f0d6b81a7dbb in our profiles, and this concept is not in facetsConcepts. On the other hand, I didn't know that there was a decision to use this facet (listed among the test facets), but I won't tell anyone if we just incluide it ;) So, if it should be regarded, could you please add this forth concept to the list in facetsConcepts or discuss it at the next meeting or however things are handled right now?

Ticket is issued: https://trac.clarin.eu/ticket/931

Email from Hanna Hedeland sent on 09/05/2016:

I'm afraid at least some of the issues I described after the launch of the latest VLO release are still there, I think they could be defined as mapping problems?

- *Even if http://hdl.handle.net/11459/CCR_C-5439_98bb103d-476a-7f62-54b4-bf9de24d2229 is provided, there are still entries with double values for distribution type (which should not be possible?), e.g. for https://vlo.clarin.eu/record?q=hzsk&docId=http_58__47__47_hdl.handle.net_47_11022_47_0000-0000-534B-F.*
- *It is not clear how to achieve that VLO entries get the laundry tag "PLAN", and entries with seemingly identical CMDI "legally relevant" metadata don't seem to receive the identical set of laundry tags (https://vlo.clarin.eu/record?q=hzsk&docId=http_58__47__47_hdl.handle.net_47_11022_47_0000-0000-631F-F and https://vlo.clarin.eu/record?q=hzsk&docId=http_58__47__47_hdl.handle.net_47_11022_47_0000-0000-5C5F-0).*

Menzo: we need to check mappings and importer.

Pavel: superprofile should be consider, a profile containing mandatory fields like author, title, licence, resourceType, from which all others are derived.

Normalisation workflow

Davor / Henk: How to discover values that haven't been normalised?

Menzo: Developers can query SOLR and obtain values.

Davor: collect them during the import process.

Davor: How to maintain controlled vocabularies, updating them and ensure that VLO always uses the latest version?

Menzo: github might be a proper solution. It gives nice preview for csv files. Push unmapped values to the github file (maintainer will be notified automatically).

Francesca: Github provides users with in-browser editor

ResourceType

Davor: Normalisation map created by Jan Odijk is still not used. Should we use it?

No agreement on this (again).

Davor: how about to create a concept and to link it with controlled vocabulary and recommend it for resType

Menzo: CCR TF was thinking about doing something similar for all facets. Another option is to create a component.

Henk: too many values without a hierarchy are senseless but creating a hierarchy could possibly lead to display issues (maybe they can't fit the screen).