The Curation Module and Statistical Analysis on VLO Metadata Quality

Davor Ostojic
ACDH-OEAW
Vienna, Austria
davor.ostojic@oeaw.ac.at

Go Sugimoto
ACDH-OEAW
Vienna, Austria
go.sugimoto@oeaw.ac.at

Matej Ďurčo
ACDH-OEAW
Vienna, Austria
matej.durco@oeaw.ac.at

Abstract

The Curation Module is developed to facilitate the metadata ingestion and curation process of the Virtual Language Observatory (VLO) by providing a systematic method to measure metadata quality and a user-friendly interface to inspect profiles, records, and collections of the Component MetaData Infrastructure (CMDI) used for the VLO. A large amount of useful statistics generate a comprehensive data matrix including information about the quality score, publication status, facet coverage, and metadata header, as well as the number of records and concepts. The module helps various stakeholders to automatically and systematically identify the metadata problems. Whilst metadata modellers can evaluate the quality of shared profiles, data creators assess the validity of newly created records. Data providers can use it for the improvement of their metadata for better discoverability and accessibility of valuable linguistic contents, whereas working groups could examine the actual use of profiles and records to define the next version of CMDI and VLO. Thus, the Curation Module supports all stages of metadata management and fosters the analysis and improvement of metadata quality to enhance the CLARIN services. In this article, we present a selection of statistical information on the metadata quality made possible by the Curation Module.

1 Background

Metadata quality is central to resource discovery. It determines the discoverability and accessibility of resources for the users and metadata curation plays an essential role to control the quality. CLARIN is not an exception. Its main metadata catalogue of language resources, Virtual Language Observatory (VLO)[1] suffers from a backlash of the flexibility of Component MetaData Infrastructure (CMDI)[2], which is a standardised metadata framework underlying VLO. In fact, metadata curation has been a long standing issue in CLARIN, hence the Metadata Curation Task Force was founded to tackle it. Most recently, we investigated the variability issues of metadata in VLO (King et al., 2016) and the idea of a Curation Module was formalised to provide a solution to assess the quality of the ingested metadata. By now we know how many CLARIN centres are registered (centre registry[3]), some of which are the data providers of VLO, how many records are ingested into VLO (its home page), how many collections we

[1] https://vlo.clarin.eu
[2] https://www.clarin.eu/content/component-metadata
[3] https://centres.clarin.eu/

Davor Ostojic, Go Sugimoto and Matej Ďurčo 2017. The curation module and statistical analysis on VLO metadata quality. *Selected papers from the CLARIN Annual Conference 2016.* Linköping Electronic Conference Proceedings 136: 90–101.

have received (CMDI harvester web view[4]), and how many metadata concepts (CLARIN Concept Registry[5] hereafter CCR) and profiles (Component Registry[6]) are created to define and semantically bind different types of resource descriptions. In addition, extra efforts brought us such valuable information as to the structure of CMD profiles and the reuse of CMD components and concepts (SMC Browser[7])(Ďurčo 2013) and what percentage of VLO facets are covered (King et al., 2016; Odijk, 2014). However, it was not possible to systematically and automatically collect statistics about the quality of the CMDI metadata. In 2015, we presented the general functional concept of the Curation Module in the context of overall VLO data ingestion workflow (King et al., 2016) in accordance with some previous works (Kemps-Snijders, 2014; Trippel et al., 2014). The Curation Module then became one of the deliverables of CLARIN-PLUS project[8]. This paper will outline the ongoing development of the module and demonstrate the first findings on the metadata quality made possible by this module, as well as other relevant statistics.

2 The Curation Module

2.1 Overview

The Curation Module is a software tool developed as a component of the CLARIN metadata infrastructure for curation and quality assessment / benchmarking of CMD records, collections and profiles. It is intended as technical support for human curation work to monitor and improve the metadata quality. The design of the module was guided by the following four use cases[9]:

1. The metadata editor checks (on-the-fly) the quality and validity of a newly created record.
2. The metadata modeller evaluates the quality of profiles (especially facet coverage), when selecting an existing profile or creating a new profile for new resources.
3. The data provider, repository administrator, or collection manager checks the overall quality of metadata in his/her repository, including the facet coverage.
4. All records ingested into the VLO undergo a systematic process of curation, validation, normalisation and quality assessment (benchmarking).

The Curation Module consists of two parts: a core Java application that works standalone or can be used in other software as library, and a web application which provides a web-based interface as well as a RESTful API. The module can process web resources via URL as well as locally stored CMD records and collections. In addition to the interface for assessing own data, the user can explore pre-processed assessments of public profiles[10] (figure 1) and the collections harvested by the CLARIN aggregator. The Curation Module heavily depends on other CLARIN infrastructure services such as the Component Registry from where it fetches the XSD schema files of the CMD profiles and the Concept Registry from where it retrieves information about concepts.

[4] https://vlo.clarin.eu/data/

[5] https://openskos.meertens.knaw.nl/ccr/browser/

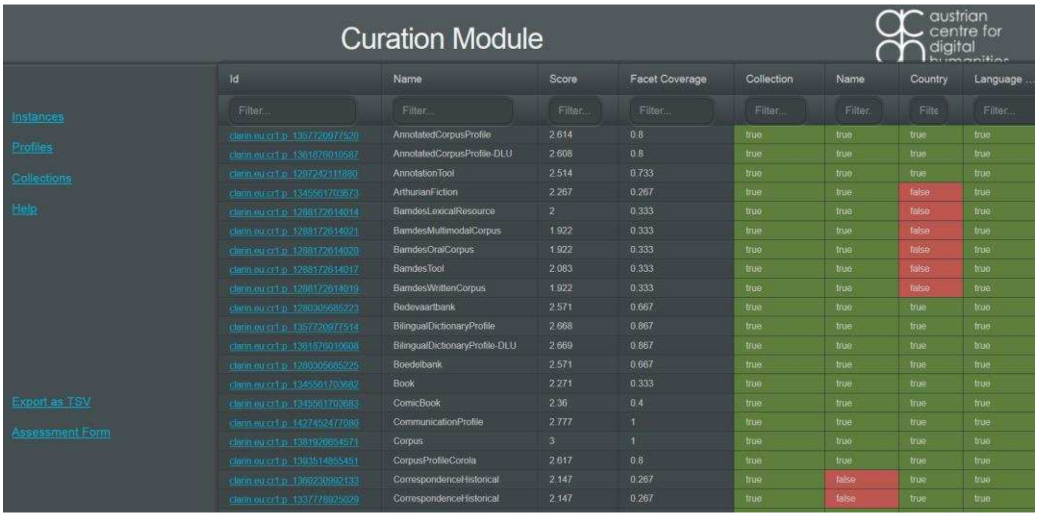
[6] http://catalog.clarin.eu/ds/ComponentRegistry/

[7] https://clarin.oeaw.ac.at/ /smc-browser/

[8] https://www.clarin.eu/node/4213

[9] https://office.clarin.eu/v/CE-2016-0742-CLARINPLUS-D2_1.pdf

[10] Public profiles are the profiles publicly shared in the Component Registry, as opposed to the private profiles (or non-public profiles) which are only visible to the creator.



Id	Name	Score	Facet Coverage	Collection	Name	Country	Language ...
Filter...	Filter...	Filter...	Filter...	Filter...	Filter..	Filte	Filter...
clarin.eu:cr1:p_1357720977520	AnnotatedCorpusProfile	2.614	0.8	true	true	true	true
clarin.eu:cr1:p_1361876010587	AnnotatedCorpusProfile-DLU	2.608	0.8	true	true	true	true
clarin.eu:cr1:p_1297242111880	AnnotationTool	2.514	0.733	true	true	true	true
clarin.eu:cr1:p_1345561703673	ArthurianFiction	2.267	0.267	true	true	false	true
clarin.eu:cr1:p_1288172614014	BamdesLexicalResource	2	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614021	BamdesMultimodalCorpus	1.922	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614020	BamdesOralCorpus	1.922	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614017	BamdesTool	2.083	0.333	true	true	false	true
clarin.eu:cr1:p_1288172614019	BamdesWrittenCorpus	1.922	0.333	true	true	false	true
clarin.eu:cr1:p_1280305685223	Bedevaartbank	2.571	0.667	true	true	true	true
clarin.eu:cr1:p_1357720977514	BilingualDictionaryProfile	2.668	0.867	true	true	true	true
clarin.eu:cr1:p_1361876010608	BilingualDictionaryProfile-DLU	2.669	0.867	true	true	true	true
clarin.eu:cr1:p_1280305685225	Boedelbank	2.571	0.667	true	true	true	true
clarin.eu:cr1:p_1345561703682	Book	2.271	0.333	true	true	true	true
clarin.eu:cr1:p_1345561703683	ComicBook	2.36	0.4	true	true	true	true
clarin.eu:cr1:p_1427452477080	CommunicationProfile	2.777	1	true	true	true	true
clarin.eu:cr1:p_1381926654571	Corpus	3	1	true	true	true	true
clarin.eu:cr1:p_1393514855451	CorpusProfileCorola	2.617	0.8	true	true	true	true
clarin.eu:cr1:p_1360230992133	CorrespondenceHistorical	2.147	0.267	true	false	true	true
clarin.eu:cr1:p_1337778925029	CorrespondenceHistorical	2.147	0.267	true	false	true	true

Figure 1 Curation Module lists the assessment of public profiles

For each input type (CMD record, profile, or collection) the Curation Module defines a distinct workflow of phases each of which collects statistics from different VLO components (the Component Registry and CCR) and generates a corresponding XML report (described in the following subsections). There is also a special type of report which is generated in case of non-recoverable errors during the data assembling workflow. This type of report contains only error messages. If the algorithm of a phase runs successfully, statistics are generated out of the gathered information and the quality assessment score for that phase is calculated. The overall score is calculated by summing up individual scores from each phase in the workflow. Finally, a report is created by combining these statistics, scores, and eventual issues. Each issue has information about the phase in which it occurred, the severity (warnings and errors), and a verbose message. The primary output format for this assessment report is XML, but it is also rendered in HTML for a user-friendly view in the web application. In the next subsections we describe the main features of the three different types of reports.

2.2 Profile report

Profile assessment workflow is divided into three phases: header, components/concepts and facet mapping assessment. In terms of profile scoring (table 1), points are added for the publication status of the profile (public or private), the percentage of elements annotated with concepts, and the VLO facet coverage. The maximum score is 3.0.

Criteria	**Score**
Publication status	0 or 1
The percentage of elements annotated with concepts	[0.0 .. 1.0]
Facet coverage	[0.0 .. 1.0]
Total	**0.0 .. 3.0**

Table 1. Scoring criteria for profiles

A report is structured with the following sections (matching the different phases of the workflow):

1. Meta information about the profile with name, id, description, link to schema, CMDI version and lifecycle status of the profile. The score of this phase is based on the publication status: if the profile is public, one point is given.

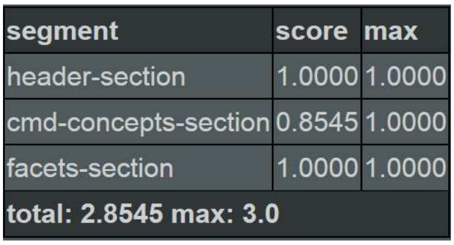
2. Score section presents the summary of the scores from each phase and total score in form of a table (figure 2).

3. Facet mapping section provides information about covered facets. The score represents the facet coverage of the profile.

4. Component section lists the components used in the profile in a table with the following columns: component name, id and count.

5. Concept section firstly shows statistics about elements. Following table lists information about the concept names which links to the corresponding CCR page, CCR status and count. The percentage of annotated elements with the CCR concepts represents the score from this section.

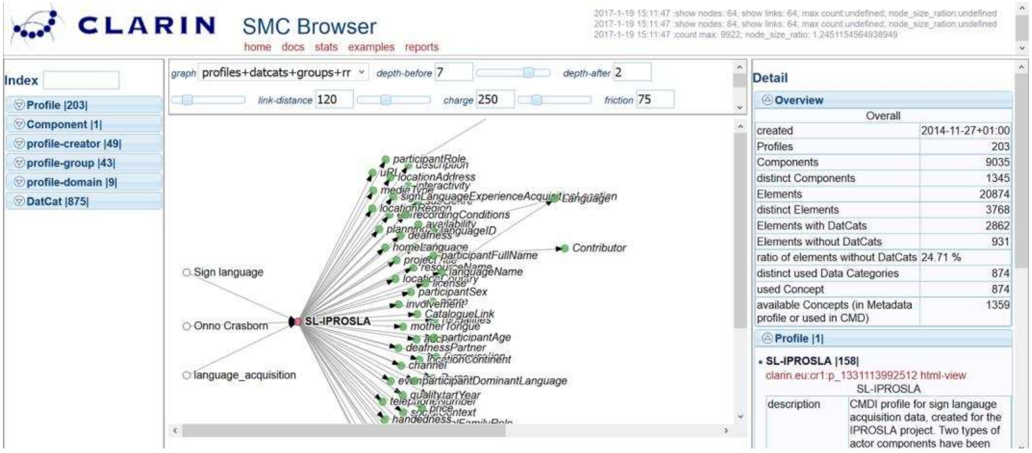
6. Last part of the report shows the issues encountered during assessment workflow. The user can see in which phase they occurred, severity level, and message about the problem.



segment	score	max
header-section	1.0000	1.0000
cmd-concepts-section	0.8545	1.0000
facets-section	1.0000	1.0000
total: 2.8545 max: 3.0		

Figure 2. Score summary for profile assessment

In the pre-processed profile assessment, a direct link is given to the SMC Browser, where the users can explore the complex network of the CMD profile within an interactive application for exploring graph data (figure 3).



CLARIN SMC Browser

home docs stats examples reports

2017-1-19 15:11:47 :show nodes: 64; show links: 64; max count:undefined; node_size_ration:undefined
2017-1-19 15:11:47 :show nodes: 64; show links: 64; max count:undefined; node_size_ration:undefined
2017-1-19 15:11:47 :count max: 9922; node_size_ratio: 1.2451154564938949

Index

- Profile |203|
- Component |1|
- profile-creator |49|
- profile-group |43|
- profile-domain |9|
- DatCat |875|

graph profiles+datcats+groups+rr depth-before 7 depth-after 2

link-distance 120 charge 250 friction 75

Detail

Overview

Overall	
created	2014-11-27+01:00
Profiles	203
Components	9035
distinct Components	1345
Elements	20874
distinct Elements	3768
Elements with DatCats	2862
Elements without DatCats	931
ratio of elements without DatCats	24.71 %
distinct used Data Categories	874
used Concept	874
available Concepts (in Metadata profile or used in CMD)	1359

Profile |1|

- SL-IPROSLA |158|
  clarin.eu:cr1:p_1331113992512 html-view

SL-IPROSLA

description	CMDI profile for sign langauge acquisition data, created for the IPROSLA project. Two types of actor components have been

Figure 3. A profile in Curation Module directly links to the profile view in the SMC Browser

2.3 Instance report

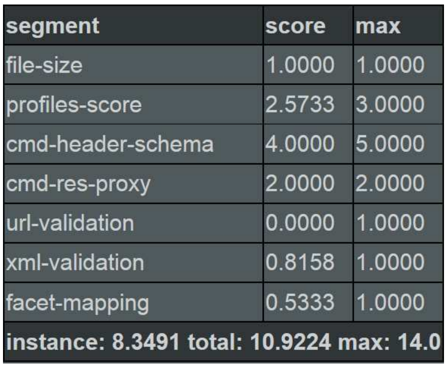
The workflow for instance assessment consists of the following seven processing phases: file, header, profile, resource references, XML validation, facet mapping and URL validation. The score for an instance is based on the score of the used profile, the record size, the completeness of the CMD header, the presence of references to resources, XML syntax validation, the resolvability of links, and facet coverage (table 2). The maximum score is 14.0.

Criteria	**Score**
Profile's score	[0.0 .. 3.0]
File size is less than 10 Mb	0 or 1
Schema location is present	0 or 1
Schema resides in Component Registry	0 or 1
MdProfile element contains a valid value	0 or 1
MdCollectionDisplayName is not empty	0 or 1
MdSelfLink is not empty	0 or 1
Rate of resources with MIME type	[0.0 .. 1.0]
Rate of resource proxy elements with references	[0.0 .. 1.0]
Rate of non-empty XML elements	[0.0 .. 1.0]
Rate of accessible URLs	[0.0 .. 1.0]
Facet coverage	[0.0 .. 1.0]
Total	**0.0 .. 14.0**

Table 2. Scoring criteria for collections

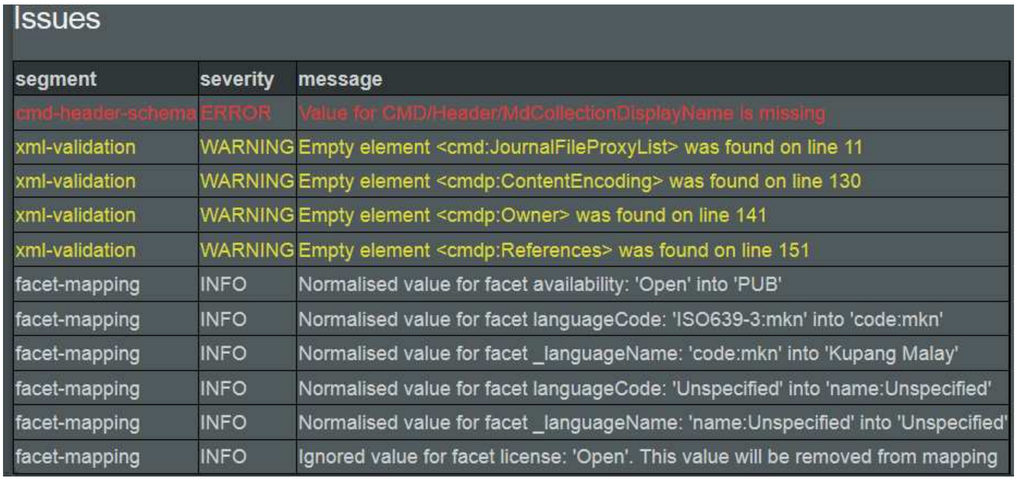
An instance report presents information in the following order:

1. The first section shows combined information from the first three phases of assessment. The user can see information about the file name, profile id linked to the respective profile report, and the size of the record in bytes.
2. Score section (figure 4) displays the summary of individual scores and the overall score with and without the underlying profile score.
3. Facet section tells how the VLO "sees" the record. Beside facets and normalised values, XPath and concept information are available. Eventual missing facets will be listed at the bottom of the table. The score from this phase is used for facet coverage.
4. Resource proxy section presents statistics about the resources described by the record. Score is the summary of percentage of the resource with MIME type and with references.
5. XML validation section shows statistics, gathered during XML validation against the schema, about the elements in the record. The rate of populated elements gives the score from this phase.
6. URL validation section delivers statistics about HTTP links from the record and their resolvability (or persistency). The score is represented with the rate of resolvable links.
7. The last part of the report shows human-readable messages about the errors, warnings and other potentially useful information issued in each assessment phase (figure 5).



segment	score	max
file-size	1.0000	1.0000
profiles-score	2.5733	3.0000
cmd-header-schema	4.0000	5.0000
cmd-res-proxy	2.0000	2.0000
url-validation	0.0000	1.0000
xml-validation	0.8158	1.0000
facet-mapping	0.5333	1.0000
instance: 8.3491 total: 10.9224 max: 14.0		

Figure 4. Example of the score section in the report



Issues

segment	severity	message
cmd-header-schema	ERROR	Value for CMD/Header/MdCollectionDisplayName is missing
xml-validation	WARNING	Empty element <cmd:JournalFileProxyList> was found on line 11
xml-validation	WARNING	Empty element <cmdp:ContentEncoding> was found on line 130
xml-validation	WARNING	Empty element <cmdp:Owner> was found on line 141
xml-validation	WARNING	Empty element <cmdp:References> was found on line 151
facet-mapping	INFO	Normalised value for facet availability: 'Open' into 'PUB'
facet-mapping	INFO	Normalised value for facet languageCode: 'ISO639-3:mkn' into 'code:mkn'
facet-mapping	INFO	Normalised value for facet _languageName: 'code:mkn' into 'Kupang Malay'
facet-mapping	INFO	Normalised value for facet languageCode: 'Unspecified' into 'name:Unspecified'
facet-mapping	INFO	Normalised value for facet _languageName: 'name:Unspecified' into 'Unspecified'
facet-mapping	INFO	Ignored value for facet license: 'Open'. This value will be removed from mapping

Figure 5. Example of the issue section for instance assessment

2.4 Collection report

Workflow for collection assessment consists of assessment of all the contained records and aggregation of collected statistics from each individual record. Currently collections can be only processed in command line mode with a file system path as an input. A collection report contains the following sections:

1. Overview with the name of the collection, the total score (the sum of scores from all instances within), the average score, the minimal and maximal score in the collection.
2. File section supplies statistics about the number of records as well as the total, minimal, maximal and average size of them.
3. Header section lists the profiles referenced by the records in the collection, with the respective records count, score, and links to the corresponding reports.
4. Facet section lists average facet coverage for each individual facet.
5. Resource Proxy and XML validation sections display aggregated statistics as the total and average figures that come from the corresponding sections of the instance reports.

3 Preliminary quality analysis of existing data[11]

In this section we present the main results of the analyses of the two pre-processed datasets, namely profiles and collections, giving some idea about the status quo of the CMDI data as available in the VLO. In addition to the reports of the Curation Module, a few other statistics about the VLO data were collected, visualised, analysed, and interpreted to complement our analyses.

[11] All the numbers are updated as of January 2017.

3.1 Profile analysis

Examining all the profiles referenced by harvested CMD records, the highest score is 2.87 out of 3.0, and 0.79 is the lowest (figure 6). The score distribution is generally good at the moment, suggesting a fairly good initial setup of the score calculation. There is a clear gap between public profiles and private ones, because all the private profiles are below average, implying an easy improvement potential for the private profiles, when being converted to public. This scoring system, thus, attempts to give incentive to make profiles public, because the essence of CMDI is the collective effort of interoperability by publicly sharing common profiles to be reused.

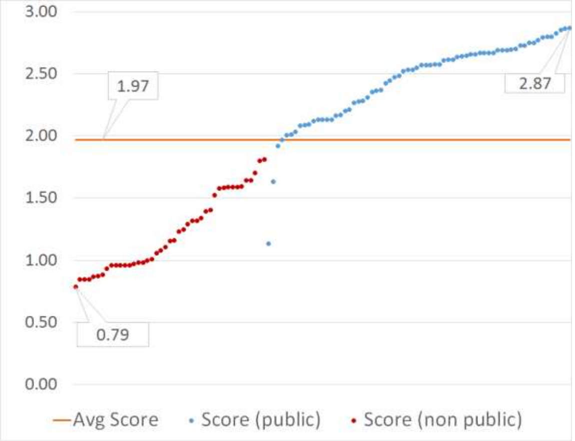


Figure 6. Profile score distribution

Figure 7 demonstrates the top 10 profiles with the highest score, which is a valuable source of information for the CMDI guidance. Data providers (or their data modellers) can go down this list ordered by the score, to find the most suitable profile to use to describe their resources, ensuring the best possible discoverability of their resources. It is also possible that the CMDI working group can learn the usage of CMDI profiles to discuss and develop the future version of the metadata model/framework. In addition, it is our recommendation that CMDI and VLO data ingestion guidelines can be produced and describe the recommendations of profiles according to the outcome of the Curation Module.

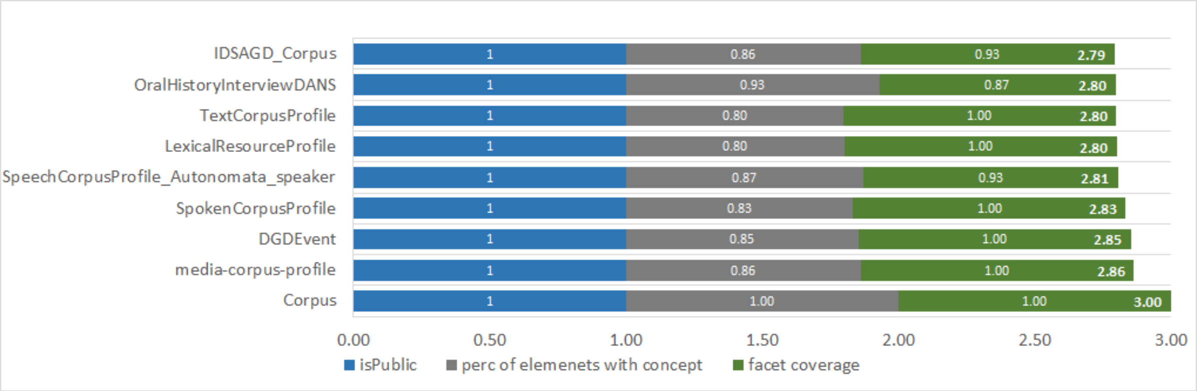


Figure 7. Public profiles with the highest score

Figure 8 shows the ten most referenced profiles by CMD instances/records. Interestingly, four out of ten are not public. In fact, as also discovered above, if the private profiles are changed to public, the scores would jump substantially with very little effort. For example, the Song profile, is associated with over 155.000 instances, albeit relatively low facet coverage (40% or 6 of 15). It contains the following facets: format, name, genre, nationalProject, collection, and languageCode.

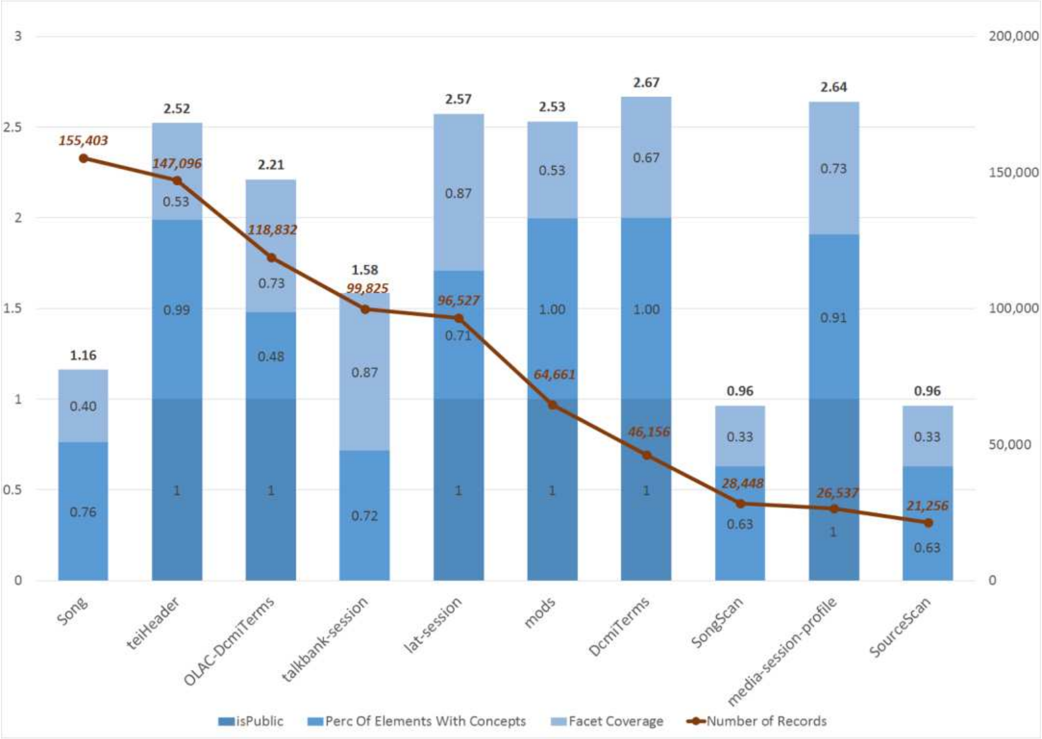


Figure 8. Profiles referenced in VLO with the highest number of records

In terms of the proportion of profiles, figures 9 suggests that approximately two thirds of the profiles are public. In addition, we counted the number of public and used profiles (figure 10). It turned out that about two thirds of public profiles are unused (or, more precisely, there are no records published via VLO). It is likely that when a (published) profile needs to be updated, the profile modeller/creator would abandon the old profile and create a new one, leading to a large number of unused public profiles. However, with CMDI 1.2 this issue has been tackled by establishing a lifecycle for profiles and components, including deprecation that should prevent further proliferation.

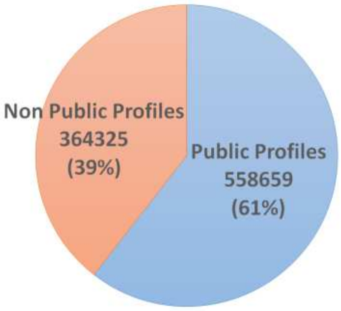


Figure 9. The number of instances using public and non-public profiles

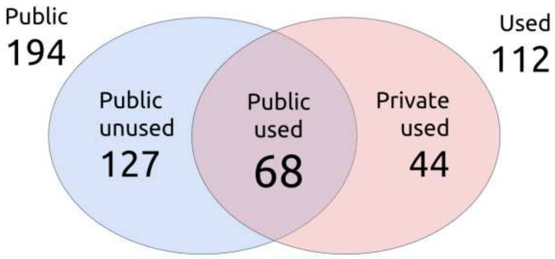


Figure 10. The proportion of public and used profiles

3.2 Collection analysis

The collections, as a set of the CMDI records gathered from the individual data providers, represent a primary discerning principle/dimension for the large body of CMD records harvested and indexed regularly by VLO. The reports for the collections, computed by aggregating the reports of instance assessment, are the most comprehensive report type, primarily relevant for the use case of a repository administrator checking the records that the repository exports to the VLO. Additionally, the contrasting juxtaposition of the metrics of all the individual collections reveals the overview on the statistics of the VLO records. For example, we can easily find that the number of records varies greatly from collection to collection, ranging from 1 to 249,659, and the size in bytes from 1 kB to 5 GB. While most collections use only a single profile, there is a collection using 36 distinct profiles.

The average score for all collections is 10.6 (out of 14.0) suggests a good overall quality (figure 11). However, the overall score distribution is rather concentrated in the area between 10.0 and 13.0, making it hard to differentiate the qualities of each collection. We have to examine the distribution more in detail to know whether the scores are too optimistic, thus not reflecting the intended results of reality, or not. The scoring criteria is always an area of discussion (Kemps-Snijders, 2014; Trippel et al., 2014), requiring a continuous reviewing and calibration. As such, the statistics may be used as rough indicators. The important point is that we now have a tool to automatically measure the quality of all data in a consistent and transparent manner.

Figure 12 illustrates the top 10 highly scored collections with reference to the data volume. It is of particular value to recognise that there are some large collections with low scores. By consulting the data providers and improving the metadata descriptions, we will be able to increase the overall quality of the data in the most efficient way. It is also the advantage of the Curation Module that it can indicate the best practice of (re-)defining profiles. The statistical overview available in the module is especially informative for CLARIN curation team, as they can determine what priorities and strategies should be taken in order to maintain and balance the quality of VLO in a short and long term.

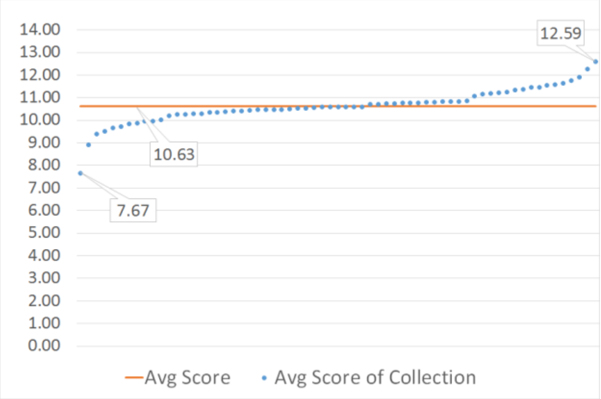


Figure 11. Collections score distribution

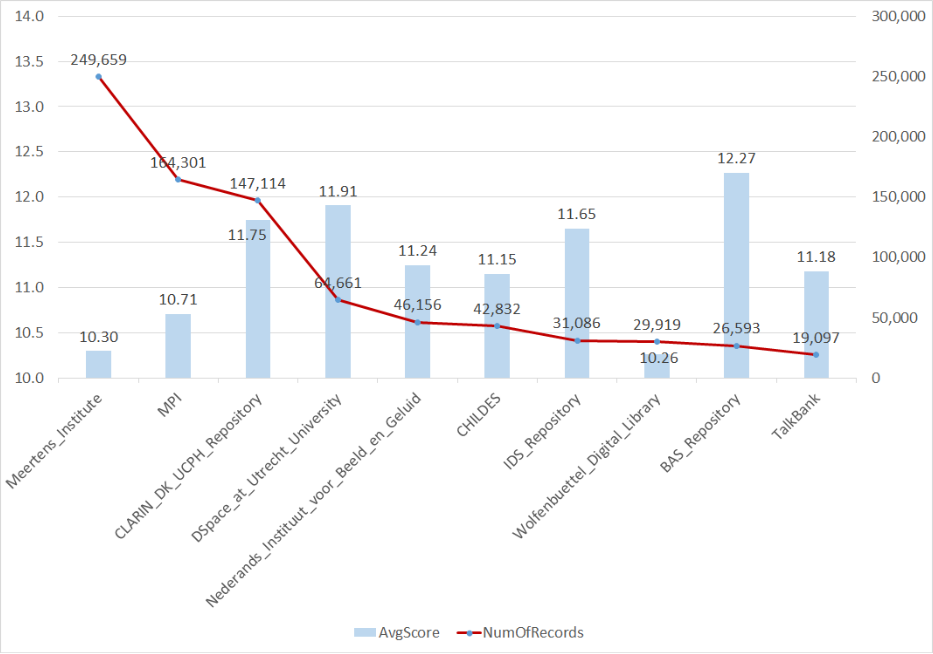


Figure 12. Top 10 highly scored collections with reference to the number of records

3.3 Combined analysis (facet coverage)

Probably the biggest headache of VLO is facet coverage. Echoed with Odijk (2014), King et al. (2016) argued that the extremely low percentage of facet coverage severely hampers the discoverability of language resources in the VLO. In this section, a closer examination on the distribution of covered and uncovered facets is executed in order to identify problematic metadata records. Firstly, our basic statistics of the Curation Module indicate the spread of facet coverage between 7.2% and 94.4%. The facets with the lowest coverage are keyword (7.2%), modality (13.6%), and country (13.6 %). Our analysis goes further to compare the current statistics with those in 2016 (figure 13). It has to be noted that there are a number of changes affecting the direct comparison. The data mapping of VLO (i.e. concept to facet, and value normalisation mapping) has been constantly modified and the VLO has received a large amount of records. Despite such changes, it is still useful to track the statistics over time. Good news is that the coverage of many facets is improved. For instance, languageCode facet (116.1%) and national project (90.5%) have improved dramatically, and the other facets such as availability, subject and resourceClass are in the range of 20%. There are only two facets whose coverage were deteriorated. Continent facet became obsolete in the meantime. The number of records without using format facet has increased from 9.9% to 40.6%, while collection facet has risen from 0% to 8.7%. Most likely reasons of those are the ingestion of a number of records which do not comply with the requirements of the VLO facets.

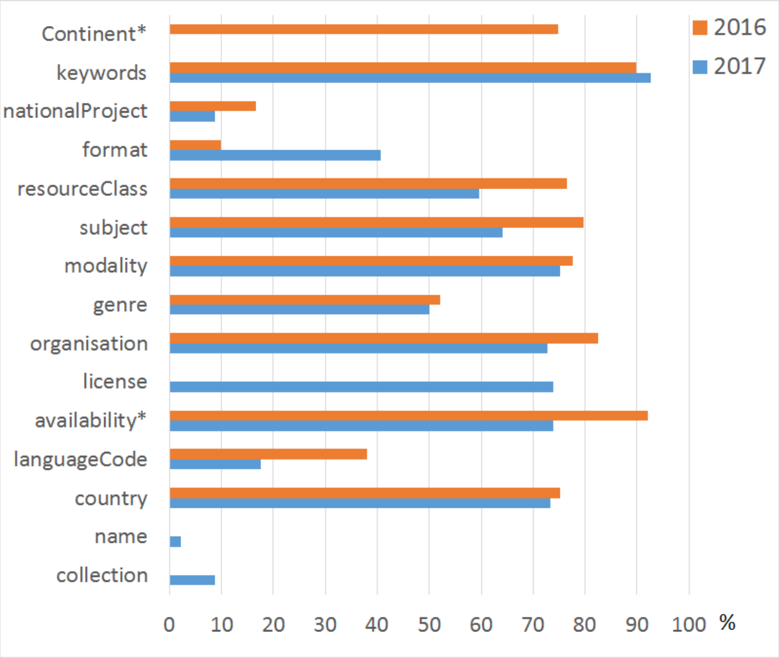


Figure 13. Comparison of uncovered facet of VLO (with data from King et al. (2016))

More interestingly, the collection and profile matrix helps us to generate new types of view on the data. Two figures were produced in order to investigate the relationship between the volume of collections (figure 14) and facets (figure 15) and the average facet coverage. It is clear that there are several collections and profiles which have a significant number of records with relatively low facet coverage (pink areas in the charts). Those can be regarded as the highest priority of improvement with least effort to increase the overall metadata quality of VLO. All in all, we think that the Curation Module is capable of providing a wide spectrum of feedback on the metadata quality.

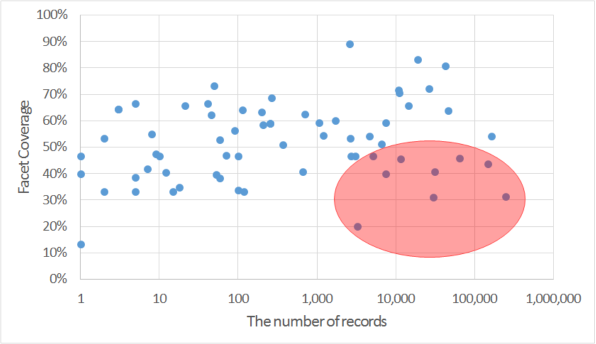


Figure 14. Facet coverage and size of collections

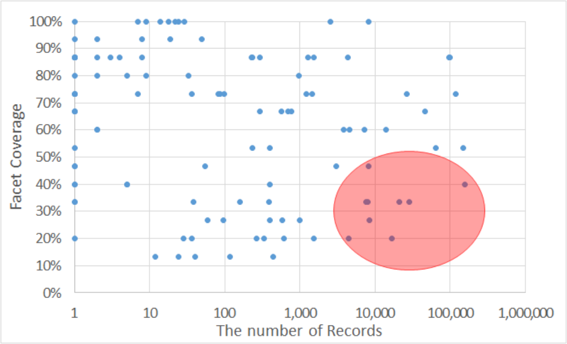


Figure 15. Facet coverage and size of profiles

4 Further developments

We have just finished the second phase of the development of the Curation Module. In the first phase (2015-2016), the Curation Module was provided with baseline functionalities, including the display of statistics with simple tables, the evaluation of a profile or instance by URL, and a simple link to the SMC browser. In the second phase (2016-2017), a series of small improvements were conducted according to internal reviews and feedback from CMDI team and early adopters. First of all, there are more functions to help curation tasks. For instance, comparison of original and normalised values for facets are now included, and links between values and concepts become traceable. Secondly, the output report was adapted with instructive information while web application presents it in a more user-friendly graphical interface. SMC Browser was also more integrated in the level of profiles and concepts. Thirdly, support for filtering and sorting the data in the overview tables for profiles and collections was added, allowing for more efficient exploration of large datasets. Moreover, while URL was the only input method for CMD records before, it becomes possible to upload a single local file. Finally, the table report of profiles and collections is exportable in a tabular format. It enables the users to further process and analyse the statistics with third-party applications. All such small improvement should not be underestimated, because they all contribute to the adaptation of the module and CMDI at large.

Although the two phases successfully delivered a stable service, there is still a wish list for the future, and some ideas are highlighted here. For example, batch upload for assessment can be developed so that the users can aggregate the statistics for given files and/or compare a set of uploaded files. It is also interesting to add more visualisation such as graphs and charts to the assessment report. In addition, automatic email notification with an attachment of (or link to) curation reports could be sent to the data providers. Such active effort would raise the awareness of the quality issue and hopefully encourage the providers to work on improving the data they delivered. Moreover, the calibration of the score should be considered (Kemps-Snijders, 2014). After careful manual revision of the quality reports, the curation team should be able to suggest a new weighting of the individual criteria to achieve a fair scoring of the metadata quality. Furthermore, the modularisation of the module and other VLO components could streamline the CLARIN service infrastructure for efficient maintainability, as the whole technical infrastructure gets bigger and more complicated.

Another major planned enhancement of the functionality includes the addition of time dimension and a facet-centred view. The former will store history of (primarily collection) assessments and thus enable us to monitor the data quality over time, introducing also a possibility to automatically identify sudden drop in any of the metrics. The latter will basically invert the current overviews of profiles and collections and allow us to explore how well (or badly) specific facets are covered relative to the collections and profiles. It will also feature information about the value variability, delivering much needed systematic input for the value normalisation efforts. If developed, those new features actually transform the Curation Module into a CMDI Analytics, which is similar to the concept of web analytics, to collect, monitor, and analyse the statistics of all stages of metadata management. It naturally allows the users to see a good visualisation with charts and to flexibly and seamlessly (re)generate the data and its matrices by manipulating different parameters of values and dimensions. Indeed, the Curation Module is also a part of a long-term vision of VLO backend development. As we suggested (King et al., 2016), we aim for the implementation of an integrated dashboard application which manages the whole processing of data within the aggregation infrastructure, ranging from data harvesting, converting, validating, mapping, normalising, to indexing. All those developments will be of added value for the VLO development team as well as the user evaluation initiatives and the assessment committee when evaluating new centres.

5 Conclusion

The Curation Module is clearly a big step forward. It does not only inform about the metadata quality, but also the level of collaboration. The idea of CMDI, one of CLARIN's pillar achievements, is to collectively develop a standard framework to aggregate heterogeneous metadata for language resources and tools. Therefore, the module objectively answers the question of how much CLARIN community has achieved together in terms of metadata aggregation. As we pointed out that various factors contribute to a number of problems in VLO (King et al., 2016), the module successfully demonstrates and supports them with detailed statistics and visualisations. In this sense, our first set of analyses outlined unprecedented views on the quality of CMD metadata. Although several issues and challenges are identified over time including the user interface, usability, input methods, data workflow, and the calibration of scoring algorithm, it is our mission to develop and maintain the Curation Module continuously, also in relation to a broader framework, the Dashboard, to reinforce the CMDI.

The Curation Module delivers a myriad of statistical facts about CMD instances, collections and profiles. That means they can be informative for various stakeholders. In the beginning, we considered different use cases. Most notably, the curation team and data providers would benefit from the detailed report on the delivered metadata. They can inform them of exactly what happened with the datasets during the process of metadata ingestion. It has a precaution and treatment function. It, on the one hand, can prevent the data providers from supplying low quality metadata, if they check it beforehand. On the other hand, it can report what went well or wrong after ingesting the metadata. In addition, the Curation Module can be used more widely from the very beginning of metadata creation to the future use of metadata. It helps the metadata modellers to compare different possibilities and select the right profile for the sake of metadata quality and accessibility to their valuable content.

There can be more use cases than initially defined. The module can be used for the research and development of the CMDI framework, because it gives the CMDI community a practical feedback on the actual use of CMDI, creating room for consideration for the future update of the CMDI. Moreover, the CLARIN community may want to have an annual report on the progress of data ingestion. In conclusion, the Curation Module supports all stages of metadata management that CLARIN has worked on, therefore, showing a potential to be transformed into a CMDI Analytics. It should, however, not be forgotten that the Curation Module itself does not do anything to improve the metadata. It has to trigger human actions. Nevertheless, we strongly believe that it fosters the analysis and improvement of metadata quality to support CMDI and VLO.

Reference

[Ďurčo 2013] M. Ďurčo. 2013. *SMC4LRT - Semantic Mapping Component for Language Resources and Technology*. (masters)Technical University, Vienna, Austria. http://permalink.obvsg.at/AC11178534

[Ďurčo and Mörth 2014] M. Ďurčo, and K. Mörth. 2014. Towards a DH Knowledge Hub - Step 1: Vocabularies. In *CLARIN Annual Conference* Soesterberg, Netherlands.

[Kemps-Snijders 2014] Kemps-Snijders, M. 2014. *Metadata quality assurance for CLARIN.* .

[King, Ostojic, Ďurčo, and Sugimoto 2016] M. King, D. Ostojic, M. Ďurčo, and G. Sugimoto. 2016. Variability of the Facet Values in the VLO–a Case for Metadata Curation. In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland* (pp. 25–44) Linköping University Electronic Press. http://www.ep.liu.se/ecp/123/003/ecp15123003.pdf

[Odijk 2014] J. Odijk. 2014. Discovering Resources in CLARIN: Problems and Suggestions for Solutions. http://dspace.library.uu.nl/handle/1874/303788

[Trippel, Broeder, Ďurčo, and Ohren 2014] T. Trippel, D. Broeder, M. Ďurčo, and O. Ohren. 2014. Towards automatic quality assessment of component metadata. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 3851–3856) Reykjavik, Iceland: European Language Resources Association (ELRA). http://lrec2014.lrec-conf.org/en/