Curation module in action

-preliminary findings on VLO metadata quality

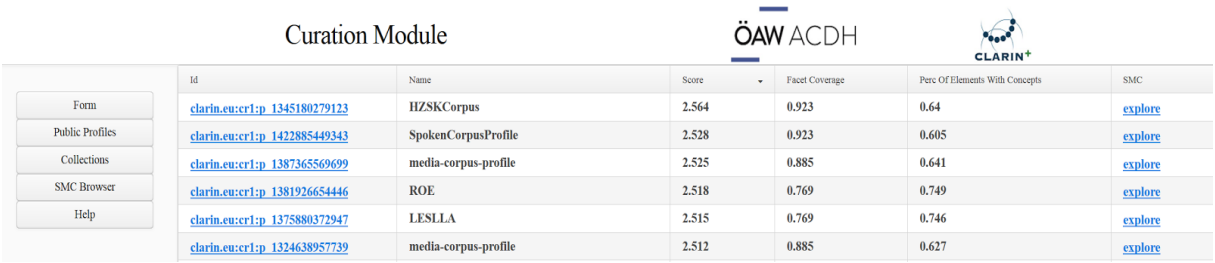
Davor Ostojic	Go Sugimoto	Matej Ďurčo
ACDH-OEAW	ACDH-OEAW	ACDH-OEAW
Vienna, Austria	Vienna, Austria	Vienna, Austria
davor.ostojic	go.sugimoto	matej.durco
@oeaw.ac.at	@oeaw.ac.at	@oeaw.ac.at

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)[i] suffers from a backlash of the flexibility of Component MetaData Infrastructure (CMDI)[ii], which is a standardised metadata framework underlying VLO. In fact, metadata curation has been a long standing issue in CLARIN, hence Metadata Curation Task Force was founded to tackle it. Most recently, we have investigated the variability issues of metadata in VLO (King et al., 2015) and the idea of 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[iii]), some of which are data providers of VLO, how many records are ingested into VLO (its home page), how many collections we have received (CMDI harvester web view[iv]), and how many metadata concepts (CLARIN Concept Registry[v]) and profiles (Component Registry[vi]) 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 the CMD profiles and the reuse of CMD components and concepts (SMC Browser[vii]) and what percentage of VLO facets are covered (Odijk, 2014 and King et al., 2015). 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., 2015) in accordance with some previous works (Trippel et al., 2014; Kemps-Snijders, 2014). This paper will outline the on-going development of the module in CLARIN-PLUS project[viii] and demonstrate the first findings on the metadata quality.

2 Curation Module

The Curation module[ix] is a software tool developed as a component of the CLARIN metadata infrastructure for curation, normalisation 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 output of the tool is a report in XML format containing various statistics, quality assessment scores, and information about issues encountered during the validation and curation according to an array of quality criteria. The curation module consists of two parts: a core 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 resources on the web via URL and profile ID as well as local resources of the CMD records and collections. In addition to the interface for assessing own data, the user can explore pre-processed assessments of public profiles (Figure 1) and collections harvested by the CLARIN aggregator. The curation module heavily depends on the Component Registry from where it fetches XSD schema files of the CMD profiles.



Id	Name	Score	Facet Coverage	Perc Of Elements With Concepts	SMC
clarin.eu:cr1:p_1345180279123	HZSKCorpus	2.564	0.923	0.64	explore
clarin.eu:cr1:p_1422885449343	SpokenCorpusProfile	2.528	0.923	0.605	explore
clarin.eu:cr1:p_1387365569699	media-corpus-profile	2.525	0.885	0.641	explore
clarin.eu:cr1:p_1381926654446	ROE	2.518	0.769	0.749	explore
clarin.eu:cr1:p_1375880372947	LESLLA	2.515	0.769	0.746	explore
clarin.eu:cr1:p_1324638957739	media-corpus-profile	2.512	0.885	0.627	explore

Figure 1 Curation module lists the assessment of public profiles

For each input type (CMD instance, profile or collection) the module generates a specific XML report. Every report provides statistics about the input and a quality assessment score computed from a number of criteria. In case of the profile, points are added for the VLO facet coverage, the percentage of elements annotated with concepts, and the publication status of the profile (public or private). The maximum score is 3.0. The score for the instance is based on the score of the used profile, record size, completeness of the CMD header, presence of references to resources, XML syntax and link validation, and facet coverage. The maximal score is 14.0. The collection score is the average of the scores of all instances within. In addition to the statistics and scores, the report also contains information useful for metadata curators to identify and solve issues. If any recoverable errors occurred during assessment process, they will be presented in a human readable form with information about the severity, the phase when issued, and the details of the errors. The CMD instance report also describes each VLO facet value with provenance (XPath and concept), which is invaluable information for detecting facet coverage issues.

3 Preliminary quality analysis

This section brings some overall statistics about the state of CMD metadata processing, generated by the curation module. It gives some initial ideas of what it can do and what the status quo of the CMDI and VLO is. For example, examining all public profiles, the highest score is 2.564 out of 3.0, and 1.231 is the lowest (figure 2). As we give penalties for private profiles, the lowest score of all profiles is 0.4045. This scoring system, thus, attempts to give incentive to publish profiles publicly. The facet coverage ranges from 23.1% to 92.3%, implying that there is a variety and there is no perfect profile in terms of VLO facet. Among the top 10 public profiles, the most uncovered facets are distributionType, rightsHolder, and lifeCycleStatus.

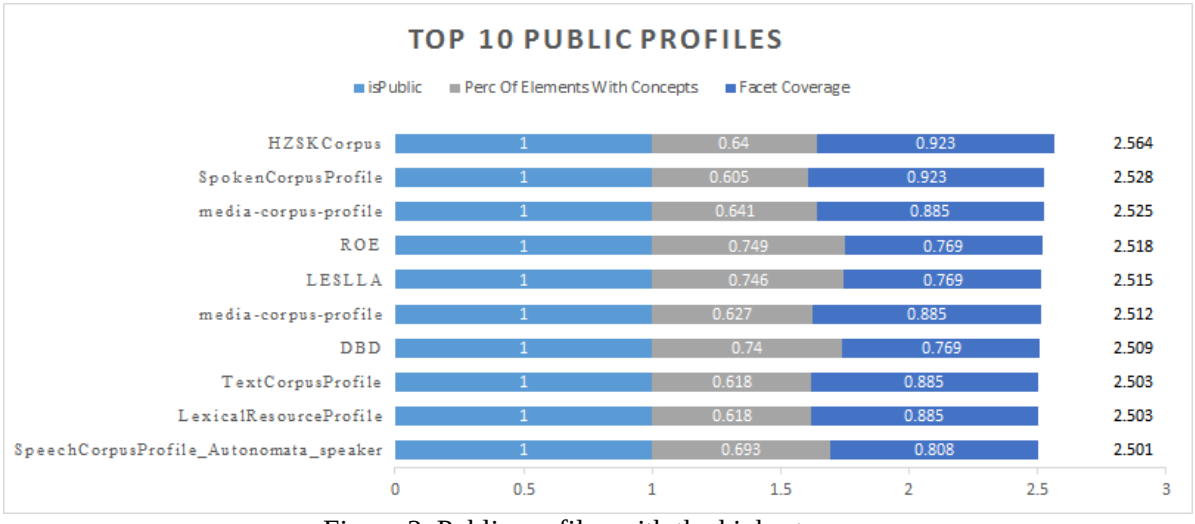


Figure 2. Public profiles with the highest score

Figure 3 shows the 10 profiles referenced by most CMD instances. Four out of ten are not public. The first one, the Song profile, is associated with over 155.000 instances, albeit relatively low facet coverage (38.5% or 10 of 26). It contains following facets: id, selfLink, collection, name, languageCode, genre, format, nationalProject, text, and componentProfile.

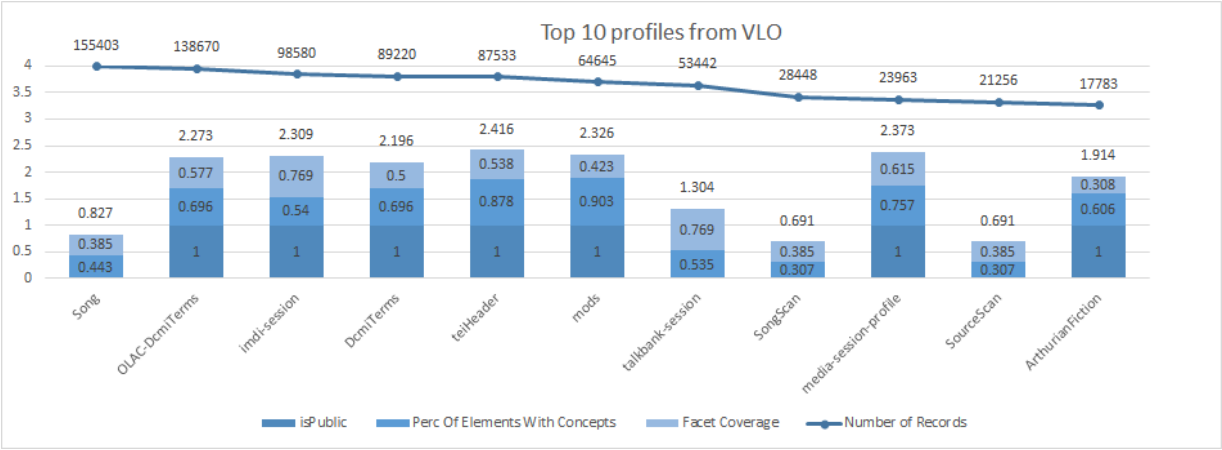


Figure 3. Profiles referenced in VLO with the highest number of references

At instance level, the pre-processed statistics for the harvested collections reveal a new view on the CMD. For example, the number of records varies greatly from collection to collection, ranging from 1 to 249,658, 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. Probably the biggest headache of VLO is the VLO facet coverage. Our statistics show the spread between 23% and 78%, but a closer examination on the distribution of covered and uncovered facets is required in order to identify problematic metadata records.
The average score per collection between 9.0 and 12.0 (out of 14.0) suggests a rather good overall quality. However, we have to examine the score distribution in detail concentrating on the low-end, to identify the problematic records. The statistics may be used as rough indicators, but the scoring criteria is always an area of discussion (Trippel et al., 2014; Kemps-Snijders, 2014), requiring a constant reviewing and calibration. The important point is that we now have a tool to automatically measure the quality of all data in a consistent and transparent manner, ensuring fairness and transparency. The statistical overview available in the module is especially valuable 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.

4 Implication for metadata curation

The curation module is especially effective in the different scenarios of metadata quality control. For example, it can:

- Deliver statistical facts about CMD instances, collections and profiles to any stakeholders
- Enable the curation team to identify (potentially) problematic metadata records, profiles, and collections
- Provide feedback reports to the data providers
- Give positive and/or negative emotion-driven incentive to the data providers to improve their metadata ("let's improve metadata to make VLO better and get encouraging feedback from the users" and "let's improve metadata not to feel ourselves ashamed of presenting low-quality metadata to the users")
- Help a data provider to prepare CMDI datasets and check their quality and problems by themselves before submitting them to VLO
- Help decide metadata modellers and the CLARIN community which profiles should be preferably used

5 Further developments

There is a long wish list for the curation module, and some ideas are highlighted in this section. First of all, there can be more functions to help curation tasks. For instance, comparison of original and normalised values for facets can be presented, and links between values and concepts should be traceable.

Secondly, there is a lot of room for improvement to produce the output report. One of the highest priorities is to organise it with instructive information and more user-friendly graphical interface. SMC Browser can be also integrated to visualise data. Thirdly, simple improvement can be made to add more flexibility to display the statistics. It is of great value if summing, filtering, and sorting are done with different parameters such as country, collection, and data provider. Moreover, as for input method, we are keen to cater for various clients by adding functionalities including a single file and a batch upload to lower the technical threshold. All such small improvement should not be underestimated, because they all contribute to the adaptation of the module and CMDI at large. Furthermore, the curation module is also a part of a long-term vision of VLO backend development. As we suggested (King et al., 2015), we aim for the implementation of Dashboard which manages the whole processing of data within the aggregation infrastructure, ranging from data harvesting, converting, validating, mapping, normalising, to indexing. Lastly, the calibration of the score (Kemps-Snijders, (2014)) should be considered. After careful manual revision of the quality reports, the curation team should be able to suggest a new weighting of the individual criteria to best reflect the reality of the situation.

6 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 is to collectively develop a standard framework to aggregate heterogeneous metadata for language resources. Therefore, the module objectively answers the question of how much CLARIN has achieved in terms of metadata aggregation. As we pointed out that various factors contributed to a number of problems in VLO (King et al., 2015), the module successfully demonstrated to visualise and prove them with detailed statistics. In this sense, our first set of analysis outlined unprecedented views on the quality of CMD metadata. Although several issues and challenges are identified including the user interface, usability, input methods, and the calibration of scoring algorithm, it is our mission to develop the curation module continuously, also in relation to a broader framework, the Dashboard, to reinforce the CMDI. We consider different use cases of the curation module. 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 happens 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. 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. We strongly believe that it fosters the analysis and improvement of metadata quality to support CMDI and VLO.

Reference

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

[Kemps-Snijders 2014] M. Kemps-Snijders. 2014. Metadata quality assurance for CLARIN. Technical report.

[King et al. 2015] M. King, D. Ostojic, M. Ďurčo, and G. Sugimoto. 2015. 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, 2016.

[Odijk 2014] J. Odijk. 2014. Discovering Resources in CLARIN: Problems and Suggestions for Solutions. Utrecht University Repository, Netherlands.

[Trippel et al. 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* [LREC 2014]. Pp. 3851-3856.

[i] https://vlo.clarin.eu
[ii] https://www.clarin.eu/content/component-metadata
[iii] https://centres.clarin.eu/
[iv] https://vlo.clarin.eu/data/
[v] https://openskos.meertens.knaw.nl/ccr/browser/

[vi] http://catalog.clarin.eu/ds/ComponentRegistry/#
[vii] https://clarin.oeaw.ac.at/exist/apps/smc-browser/index.html
[viii] https://office.clarin.eu/v/CE-2016-0742-CLARINPLUS-D2_1.pdf
[ix] https://clarin.oeaw.ac.at/curate/