Variability of the Facet Values in the VLO – a Case for Metadata Curation

Margaret King
ACDH-OEAW
Vienna, Austria
margaret.king
@oeaw.ac.at

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

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

Abstract

In this paper we propose a strategy for metadata curation especially with respect to the variability of the values encountered in the metadata records and hence in the facets of the main CLARIN metadata catalogue, the VLO. The approach concentrates on measures on the side of the infrastructure and the interaction between human curators and the automatic processes.

1 Introduction

CLARIN runs a mature well-established metadata infrastructure, harvesting metadata from more than 60 providers on a weekly basis using the standardized OAI-PMH protocol. Some seven hundred thousand records are collected and provided via the main metadata catalog, the Virtual Language Observatory or VLO (Van Uytvanck et al, 2010). It aims to provide access to a broad range of linguistic resources from many disciplines and countries based on the flexible metadata framework CMDI (Broeder et al., 2010, 2012). After a few years of intensive use by the community and continuous growth of the body of data made available via this service a number of issues have been identified (Broeder et al., 2014) concerning the functionality of the catalog, but mainly the quality of the metadata provided by the data providers such as the variation in metadata values. These irregularities seriously hamper the discoverability of resources.

After reviewing the work done within the CLARIN community until now, this paper concentrates on the issue of variant values within the facets in the VLO, exemplifying primarily by the Resource Type facet, and proposes a strategy for the implementation of a metadata curation workflow that could rectify (some of) the described problems.

2 State of Research

The CLARIN community is acutely aware of the issue at hand, and has discussed the question of how to curate metadata and especially normalize the VLO's facet values on multiple occasions. A Metadata Curation Taskforce was established in 2013 by the Centre's Committee (SCCTC) with delegates from member countries, however this taskforce until now could only collect ideas, describe the situation and tried to remedy some of the encountered problems. It wasn't able to sustain a concerted level of activity to systematically approach this problem.

CLARIN-D established a separate VLO Taskforce in October 2013 (Haaf et al., 2014) which worked out recommendations for the VLO facets in an attempt to provide more guidance and clarity regarding the usage and meaning of the facets to the data providers. The VLO Taskforce meetings throughout 2014 and 2015 provided small steps towards a solution. However the Taskforce has concentrated on recommendations and sound definitions, the actual implementation is not seen as one of its tasks.[1] A sound definition of the facets and recommended values for the facets is certainly a necessary condition and a good starting point towards answering the problem under consideration. However it is of little use if it is not integrated in the infrastructure nor taken up by resource providers.

In 2014, Odijk conducted an in depth survey of the VLO from the point of view of discoverability of linguistic resources (Odijk, 2014). The comprehensive report identified a number of concrete issues and

[1] as indicated in informal talks with members of the taskforce

proposed possible solutions. These identified problems pertain both to the schema level (e.g. crucial elements not obligatory), to the instance level of the data (fields not filled, variation of the values), and also to the functionality provided by the VLO (missing facets, multi-selection). He also underscored the aspect of granularity, a related point currently much discussed throughout CLARIN but one which falls outside the scope of this paper.

In an unpublished follow-up internal CLARIN report in 2015, Odijk lays out a strategy for metadata curation, concentrating on the main goal to achieve clean facets. Based on the assumption that "the providers in general case cannot improve their metadata" the main actor in the curation process is the curation task force operating on the harvested metadata (Odijk, 2015). The main reason why the metadata cannot be improved on the side of the data providers is the lack of resources to invest in improving legacy data. CMDI in its complexity may pose a steep challenge to data provider with limited resources, it seems not trivial for data providers to select the right CMDI profile without guidance. Finally, in provider's own realm the metadata may be perfectly consistent and homogeneous, it is just through aggregation that inconsistencies arise.

3 VLO Metadata: a closer look

Thus the mission of the CLARIN metadata curation task force in (in normalizing the variant facets) is twofold. In the first place it must analyze the different problems of variation and its effect on discoverability. The second practical aim is that of creating and implementing a strategy for curation within the framework of CLARIN's social structures.

3.1 Variation of Values

We can identify different types of variation. From trivial ones like case or whitespaces ("WrittenCorpus" vs. "Written Corpus"), to combination of multiple values in one field with arbitrary (or even no) delimiters (e.g. "AddressesAnthologiesLinguistic corporaCorpus"), synonyms ("spoken" vs. "audio", "text" vs "written") and, most problematically, complex (confusing) values that carry too much information and need to be decomposed to multiple values possibly in multiple facets.

Odijk points to the data provider isolation as a main cause for the variation of values (Odijk, 2014). Indeed, it is clear that different people describe things in different ways. Some providers assigned the value "text" to Tacitus' Annals while someone else chose to create a new value called "Annals". This assumption is also supported by the fact that once the data is restricted to a single collection or organization the values in facets mostly "clear up" and appear as a consistent set.

The obvious answer from the infrastructure point of view is to reach better coordination between the data providers, basically applying shared controlled vocabularies (Durco and Moerth, 2014). Presently the only guidance regarding recommended vocabularies for individual facets was provided in the Recommendations by the VLO-Taskforce. Even these vocabularies are rarely used. In the Resource Type facet only 15,000 records use one of the 25 recommended values. All in all round 250 different values are used in the Resource Type facet, the most common reason for variation is the inclusion of extra information (unrelated to Resource Type but to some other facet). For example Shakespeare's King Lear is described by the Resource Type "poem" which belongs in the Genre facet with the Resource Type "text". A controlled vocabulary could help data providers to assign the details to the correct facet.

3.2 Missing values

Even worse than the variation of the values is the fact, that many records do not provide any value for some of the facets. Odijk attributes this mainly to the lack of obligatory metadata elements in CMDI and the fact that the metadata authors are often 'blind' to the 'obvious' aspects of their resources, like language or type. For the special case of the Resource Type the main reason may be the fact that the type is implicit in the underlying CMDI-Profile (e.g. TextCorpusProfile, LexicalResourceProfile).

Whatever the reasons, the extent of the problem is alarming. Most of the facets cover only about ⅓ of the records, so from the some 700 thousand records around 500 thousand are not visible and findable in each facet (except for the automatic/obligatory ones: Collection, Data Provider). Table 1 lists the number of null values for each facet.

A minimal remedy to deal with facets without specified values, would be to collect all records without appropriate value facets should have default value (e.g. "unspecified" or "unknown")[2]. More advanced solution would be to evaluate values for certain facets from other facets or metadata fields, like "continent" from "country." We aim for complete coverage, i.e. every record needs to be represented at once.

Facet	null count	Facet	null count
Language Code	240 183	Subject	503 233
Collection	0	Format	62 381
Resource Type	482 935	Organisation	520 560
Continent	472 048	Availability	580 907
Country	474 637	National Project	104 316
Modality	490 195	Keywords	567 347
Genre	329 114	Data Provider	0

Table 1 Number of records not covered within given facet in the VLO (on a sample of 631 000 records)

3.3 Missing facets

One source of the problem with confusing values may be the lack of appropriate facets. When trying to normalize the values of the Resource Type facet it was sometimes unclear in dealing with an overloaded value exactly where the information should go. For example, mediums of information such as radio, internet, mobile phone as well as more technical entries did not have a clear value among the recommendations for this facet. This lack of facets was also identified by Odijk (2014), who suggests adding a dedicated facet for Linguistic Annotation, as well as by the VLO task force, proposing new facets for Lifecycle Status, Rights Holder and License. However adding more facets also raises the complexity of the user interface, and the mapping, so the impact of such additions would need to be carefully examined.

3.4 Need for an Efficient Curation Workflow

As mentioned in the State of Research section a great deal of inquiry has been spent on establishing exactly what types of problems exist in the area of facet value normalization most notably in Odijk (2014). While some of the trivial problems in value variation can be solved programmatically (case folding, whitespace normalization), all the more complex issues like synonyms and complex values require human input - a mapping of variant values to recommended ones. There exist some first, tentative mappings available as a result of the analysis done by Odijk or the team of authors. Besides the question of the reliability of such mappings, the next challenge is how to integrate such a mapping into the established harvesting and ingestion workflow, especially how to ensure a sustainable and consistent process over time.

At the moment any automatic curation steps happen during the ingestion of the metadata into the indexer (the so-called "post-processing"). However this is currently limited to simple programmatic corrections of values, a mapping between actual and normalized values is only applied for the "Organization" facet. What is especially missing is a procedure to ensure that the mappings are kept up to date (new previously unseen values are added and mapped) and the curation process has access to the most current version of the mappings.

We will concentrate in the following section on the general framework that needs to be in place to ensure collaborative maintenance of vocabularies and mappings, their integration in the automatic curation process, and their wider adoption by the data providers. It is crucial to ensure that all changes are transparent to the data provider and to the user of the VLO. Another requirement is to make the workflow more modular, especially allow for the curation module to be encapsulated enough so as to be reusable in other contexts.

4 Proposed solution for normalization of facet values

The first step in coming to a solution will be to ensure that the recommended values are sufficient for the current state of the VLO. Once there is a (relatively) stable version of these recommendations, a manual, case by case mapping must be completed for the individual facets. Very importantly, these

[2] As we actually did on our test instance of VLO when evaluating data for this work.

mappings must be constantly modified (new values/mappings added) as new metadata is added to the VLO) In practice, the work on the recommendations and the mappings will go hand in hand. Also given the sheer size of the value lists of the individual facets, we need to set up a priority list, and process the facets one by one. Only the combination of human and automatic curations can lead to an efficient and workable solution. Only a human can unpack the kinds of variations that exist but only an automatic procedure can implement the corrections consistently
This strategy was applied by the team of authors to the Resource Type facet in a testing instance of VLO, and has proven workable and lead to a substantial consolidation of the facet values. The wider adoption of the process, especially inclusion of colleagues from other national consortia still needs to be implemented.

4.1 Integration into Workflow

There are multiple strategies where changes to the processed data can be introduced in the workflow:

a) The most radical approach is to define a separate profile guided solely by the goal of better discoverability, with elements reflecting one to one the facets from VLO. While this would make the life of the VLO developers very easy, it would move the burden of mapping the existing profiles to this one profile either to the data providers or to the curation task force.

b) The other extreme is to normalize values only at the moment of indexing the records, which is the approach currently already adopted in some facets within the VLO ("post-processing").

c) Next option is to create amended copies of the metadata records by the curation module while staying within the confines of the original schema/profile.

d) A variant of the previous option is to keep the original records and only indicate proposed changes by means of annotations.

4.2 Management of Vocabularies and Mapping

A relatively simple (and partly already implemented) approach to the management of the mappings is to maintain a vocabulary in the vocabulary repository CLAVAS, where, based on the SKOS data model, every entity or concept is maintained as a separate record/item (skos:Concept), with a skos:prefLabel as the normalized name/label for given concepts and all variants encountered in the actual metadata stored as skos:altLabel (or skos:hiddenLabel). This information can be easily retrieved from CLAVAS and injected in the harvesting/curation workflow of the VLO. This is currently being done for the Organization names. The change introduced in CMDI 1.2 (Goosen et al., 2014) allowing to indicate a controlled vocabulary for given element in the CMDI-profile should in the mid-term also help with the handling of vocabularies with relation to the metadata elements.

What is still missing is an automatic procedure to add new previously unseen values to CLAVAS. The application underlying CLAVAS, OpenSKOS exposes a rich RESTful API that allows us not only to query but also to manipulate the data. So technically it would be possible for the curation module to add new candidate concepts. Human interaction is crucial here. These candidate concepts need to be clearly marked and kept in "quarantine" until they are checked and approved by the curators.

However, even if this whole process is set up it does not offer a solution to the more complex problem, when a value in one facet needs to be decomposed to multiple values in multiple facets. The ad-hoc experiments until now showed that a natural data structure would be a simple table with the encountered values in first column, a separate column for the other facets, allowing the curators to decompose the facet value intuitively/ergonomically into the appropriate facets.

If this file is stored as text/csv file and maintained under version control in the CLARIN's code repository, it can be easily edited by a team of curators, seeing who has done what when and can equally easily be retrieved and processed by any application, most notably the curation module.

A final technical issue is the testing phase. In order to prove that metadata quality and VLO discoverability are improved by curation module, test cases have to be designed by experts. Each class of identified problems should be covered and generated reports should be used by metadata curators and software developers for further improvements.

4.3 Prevention – fighting the problem at the source

While we pessimistically stated at the beginning that we cannot expect the providers to change their metadata, we cannot give up on them, as it is clearly better to combat the problem at the source. There are indeed a number of measures that can (and need to) be undertaken on the side of the data provider:

a) best practices guides and recommendations (like the CLARIN-D VLO-Taskforce recommendations on the VLO facets), especially a list of recommended profiles (one or two per resource type) needs to be provided, with profiles that have good coverage of the facets and use controlled vocabularies wherever possible

b) provision of detailed curation reports to the providers as separate output of the curation step

c) provision of curated/amended metadata records directly back to the data providers (automatically available with option c) and d))

d) availability of the controlled vocabularies via a simple API (as is provided by the OpenSKOS-API) to be integrated with metadata authoring tools. This functionality has been already planned to be added for at least two metadata editors used by the CLARIN community: Arbil (Withers, 2012) and COMEDI (Lyse et al., 2014)

A crucial ingredient to the proposed strategy is the question of governance, i.e. who is going to steer the process and if not force than still persistently keep reminding data providers of the problems encountered and proposing solutions. CLARIN has well-defined organizational structures and a number of bodies with delegates from all member countries, where decisions can be agreed upon on different levels. In the described case, the primary operative unit is definitely the metadata curation task force with representatives from national consortia, in a tight collaboration with the CMDI task force, both reporting to the SCCTC, which in turn reports to the Board of Directors. Thus both the horizontal coverage over the member countries is ensured, so that national metadata task forces can report shortcomings they have identified, as well as the vertical integration of the decision-making bodies, allowing to integrate small practical technical solutions as well as to propose substantial structural changes, if needed.

5 Conclusion

In this paper we proposed a strategy for curation and normalization of values in the facets of the VLO. We elaborated on the ways how to establish and sustain a workflow that combines systematic, automatic, transparent curation of the metadata with continuous input from human curators providing the mappings from actual values encountered in the metadata to recommended "normalized" values. Integral part of the process must be a suite of test cases that ensure the quality of the mappings and the whole curation process. Finally, all output of the curation (corrections & amended metadata records) must be recycled to the data providers in the hope of preventing further problems and the entire work cycle must repeat as new resources are added. Thus the need for metadata curation is perpetual.

References

[Broeder et al. 2014] Broeder, D., Schuurman, I., & Windhouwer, M.. 2014.Experiences with the ISOcat Data Category Registry. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik*.

[Broeder et al. 2012] Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., & Trippel, T. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme* (p. 1).

[Broeder et al. 2010] Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., & Zinn, C. 2010. A data category registry-and component-based metadata framework. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]* (pp. 43-47). European Language Resources Association (ELRA).

[Durco and Moerth, 2014] Ďurčo, M., & Moerth, K. 2014. Towards a DH Knowledge Hub - Step 1: Vocabularies. Poster at *CLARIN Annual Conference*, Soesterberg, Netherlands.

[Goosen et al., 2014] Goosen, T., Windhouwer, M.A., Ohren, O., Herold, A., Eckart, T., Durco, M., Schonefeld, O. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. At the *CLARIN Annual Conference*. Soesterberg, The Netherlands, October 23 - 25.

[Haaf et al. 2014] Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Eckart, T., Hedeland, H., ... & Van Uytvanck, D.. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres. *CLARIN*. http://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3210/file/Haaf_Fankhauser_CLARINs_virtual_language_observatory_under_scrutiny_2014.pdf

[Lyse et al., 2014] Lyse, G. I., Meurer, P., & De Smedt, K. (n.d.). COMEDI: A New COmponent Metadata EDItor. Presented at the CLARIN Annual Conference 2014, Soesterberg, Netherlands. Retrieved from http://www.clarin.eu/sites/default/files/cac2014_submission_13_0.pdf

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

[Odijk, 2015] Jan Odijk. 2015. Metadata curation strategy. Internal document, unpublished.

[Van Uytvanck, 2010] Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardelleni, M. 2010. Virtual language observatory: The portal to the language resources and technology universe. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]* (pp. 900-903). European Language Resources Association (ELRA).

[Withers, 2012] Withers, P. 2012. Metadata management with Arbil. In *Proceedings of the Workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR at LREC* (pp. 72–75).