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

Go Sugimoto
ACDH-OEAW
Vienna, Austria
go.sugimoto@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 on the interaction between human curators and the automatic processes.

1 Introduction

CLARIN runs a mature well-established metadata infrastructure, harvesting metadata from more than sixty providers on a weekly basis using the standardised OAI-PMH[1] protocol. Up to a million records[2] are being collected and provided via the main metadata catalogue, 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; Broeder et al., 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 catalogue, 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 on this issue in other institutional contexts and within the CLARIN community until now, we concentrate on the issues underlying the problem of variant values within the facets in the VLO, exemplified primarily by the Resource Type facet, and propose a strategy for the implementation of a metadata curation workflow that could rectify (some of) the described problems.

2 State of research

The general problem of the curation, harmonisation and normalisation of metadata has been central to libraries, academic and cultural institutions for many years. Thus, before looking at CLARIN's approach to the curation of metadata and normalising facet values within the VLO, we reflect on other institutions' treatment of this issue.

2.1 Cases of other communities

Within the library community several approaches have been implemented in dealing with similar issues. Calarco et al. (2014) elaborate at a theoretical level on the role of metadata normalisation in resource discovery. They outline three principles as a basis of a good metadata curation strategy: 1. rigid standards, 2. cooperation with data providers, and 3. technological enhancements. Rigid standards reduce ambiguity and variation, facilitating user discovery. Communicating and including the data providers in

[1] https://www.openarchives.org/pmh/

[2] With considerable fluctuations

the process is the most effective way to avoid future problems. Planning for ongoing technological enrichment of the metadata allows for future development and technical sustainability.

Huffman (2015) shares a case study on normalising variant subject and name values in EAD files. Using OpenRefine[3] and some XSLT processing, He quickly analyses the variability of the values (applying OpenRefine's "cluster and edit" feature) and combines manual inspection and automatic application to reduce the number of unique values by 7%. This procedure resembles in some ways the approach to be proposed later in the paper for the VLO (manual inspection and normalisation of distinct values, followed by automatic application of a normalisation map, see Section 5.4). While OpenRefine in general can be geared toward large-scale "messy" data, being a standalone application which incorporates database-like spreadsheets and on-the-fly facet construction, it relies too much on its own structures, which would make it difficult to integrate it into the VLO workflow which requires an on-going, collaborative process (issues like how to sustain the process over multiple iterations, how to reach agreement on mapping in a large group, etc. would be difficult to resolve).

On the other side of the spectrum, Europeana deals with massive amounts of data in a broad range of data types and formats from many cultural sectors, libraries, museums, archives etc., from many European countries, requiring a robust processing infrastructure. In line with the principles by Calarco et al., Europeana produced comprehensive documentation of the native metadata schema (originally ESE Europeana Semantic Elements (Europeana, 2009), and its successor EDM, the Europeana Data Model (Europeana, 2014)), including definitions of all classes and properties, as well as a number of case studies for mapping from existing formats to EDM[4]. While we can only focus on some aspects relevant to the topic of our paper, we would like to emphasise that the well-defined data model as well as the extensive user and data provider oriented documentation represent best practices that the CLARIN community should use as inspiration for the work on CMDI and metadata curation, especially in the light of the planned harvesting of parts of Europeana data into the VLO.

Europeana's normalisation workflow as it is represented in the guidelines provides several clear applications to the problem of normalising the VLO's variant values. First they designate certain elements as mandatory (and another few as recommended). Starting with required elements is indispensable to ensure a minimal common denominator for describing the resources. The short list of required elements as well as the allowed alternatives is a savvy and pragmatic strategy to balance the data provider-specific situations and the need for a minimal, consistent, descriptive information set.

As for restricting the allowed values of the metadata elements, EDM – in accordance with the Semantic Web principles – instructs data providers to use URL references wherever possible, with some elements allowing both literals and references and others allowing only references. It utilises restrictions of the usable references where possible (e.g. the important required field *edm:rights* that has to take a reference to one of the rights statements endorsed by Europeana[5]). Thus Europeana provides controlled vocabularies in a way that is compatible with the Semantic Web. In one case there is an explicit list of allowed values given. The only element with an explicit list of allowed values is the *edm:type* with the vocabulary: `TEXT, VIDEO, SOUND, IMAGE, 3D`.

Since *edm:type* is semantically closely related to the VLO facet *ResourceType* which serves as the primary example in this paper and is being dealt with intensively by the Curation taskforce, it will be dealt with in more detail here. The type element is innately prone to varying interpretations. Europeana pre-empted this problem by only allowing the five values mentioned above. Initially (in the prototyping phase) Europeana had their providers submit a spreadsheet with mapping for each object to one of the (then) four values (3D was only added in EDM) and applied these centrally. This process evolved toward the content providers supplying the correct term and contacting Europeana only in difficult cases. Meanwhile most of the content of Europeana is provided via the intermediate country- or domain-specific aggregators (The European Library, OpenUp, CARARE, etc.[6]), who take over certain curation tasks. An example of the professionalisation of the aggregation and curation process is also the Europeana's MINT (Metadata Interoperability) platform[7] that aims to "facilitate aggregation initiatives for cultural heritage content and metadata in Europe". Note that next to *edm:type* EDM features also the widely

[3] http://openrefine.org/

[4] http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies

[5] http://pro.europeana.eu/web/available-rights-statements

[6] http://www.europeana.eu/portal/browse/sources

[7] http://labs.europeana.eu/apps/mint

used *dc:type*. While *edm:type* occurs exactly once with one of the predefined values, *dc:type* is optional and repeatable and can be used much more broadly, drawing values from any custom vocabulary.

Interestingly the *europeana:unstored* element defined in ESE and recommended to be used for any information that that does not fit into any of the predefined elements, the *edm:unstored* element has been removed as of version 2.2 (Europeana, 2014, p. 46), without any indication of a substitute.

Next to *edm:type* and *dc:type,* EDM also adopts SKOS[8] for classifying resources and to represent controlled vocabularies. This allows for standard-conforming normalisation and enrichment on the part of the content provider allowing them to clean their values and align their metadata to any given thesaurus. This is demonstrated for example in the curation of the new Europeana Sounds collection, or enriching Europeana data with AAT (Art and Architecture Thesaurus)[9].

Some insights can be drawn from this survey of other institutions' approaches to value normalisation with regard to CLARIN's quest to improve the VLO's metadata and especially its facet values. From a smaller case study to value normalisation we gleaned guidelines from another commonly used technology for curation of messy data, OpenRefine, and from a theoretical library study on clean metadata values, we echo the three key ingredients namely, strict adherence to standards, cooperation with data providers, and technological enhancements, finally by an in depth investigation of Europeana's approach, we saw a way to implement these principles on a large scale.

2.2 Case of CLARIN

The CLARIN community is acutely aware of the problems concerning metadata quality, especially the variation of metadata values. It has discussed the question of how to curate metadata and especially normalise 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 try to remedy some of the encountered problems. It has not yet been able to sustain a sufficient level of concerted 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 have brought about 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.[10] 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 such definitions are only of use when it is integrated into the infrastructure and taken up by data 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 identifies a number of concrete issues and some proposed 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 underscores 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 of achieving clean facets. Based on the assumption that "the providers in general case cannot improve their metadata" (Odijk, 2015) the main actor in the curation process is the curation task force operating on the harvested metadata. The main reason why the metadata in CLARIN domain cannot be improved on the side of the data providers seems to be the lack of resources available for improving legacy data. CMDI in its complexity may also pose a steep challenge to data providers with limited resources. It is perhaps an unreasonable expectation for data providers to select the right CMD profile without guidance. Finally, in the provider's own realm the metadata may be perfectly consistent and homogeneous, it is just through aggregation that inconsistencies arise.

[8] http://www.w3.org/2004/02/skos/

[9] http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies/europeana-aat

10 as indicated in informal talks with members of the taskforce

3 VLO metadata: a closer look

Thus the mission of the CLARIN metadata curation task force in (in normalising the variant facet values) is twofold. In the first place it must analyse the different problems of variation and its effect on discoverability. The second practical aim is to create and implement a strategy for curation within the framework of CLARIN's social structures.

3.1 Variation of values

We can identify different types of variation. These vary from trivial ones like case or whitespaces ("WrittenCorpus" vs. "Written Corpus"), to a combination of multiple values in one field with arbitrary (or even no) delimiters (e.g. "AddressesAnthologiesLinguistic corporaCorpus"), similar concepts ("text" vs "written") and, most problematically, complex (confusing) values that carry information that should be assigned to another facet (e,g. "bioscoop" (cinema in Dutch) and "bible" as *ResourceType*).

Odijk points to the data provider's 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 assign the value "text" to "Tacitus' Annals" while others choose 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 organisation the values in facets mostly "clear up" and appear as a consistent set.

The obvious solution to the problem 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 is provided in the Recommendations by the VLO Taskforce. Even these vocabularies are rarely used. In the *ResourceType* facet only 15,000 records use one of the 25 recommended values. All in all round 250 different values are used in the *ResourceType* facet. The most common reason for variation is the inclusion of extra information (unrelated to *ResourceType* but to some other facet). For example Shakespeare's King Lear is described by the *ResourceType* "poem", a value which would belong in the Genre facet while the most suitable value for the *ResourceType* is "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 certain 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 *ResourceType* one reason for omitting it may be that it is implicitly provided in the name of the underlying CMD profile (e.g. *TextCorpusProfile, LexicalResourceProfile*).

Whatever the reasons, the extent of the problem is alarming. Some facets cover only about one third of the records, so that of 631000 records found in the VLO at the time of writing, typically around five hundred thousand are not visible and findable in each facet (except for the automatic/obligatory ones: *Collection*, *Data Provider,* as well as well-recorded *Format* and *National Project*). Figure 1 lists the number of null values for each facet. Given these alarming figures, it is clear that facet browsing, one of the most significant functionalities of the VLO, is not effective for resource discovery, calling for urgent action.

A minimal remedy (or rather a "patch") to the problem of facets without specified values is to make this information explicit to the end-users (e.g. with a default value "[missing value]"). A more advanced solution is to borrow values for certain facets from other facets or metadata fields, for example filling Continent facet based on values from Country facet. We aim for complete coverage, i.e. every record should be represented at least once.

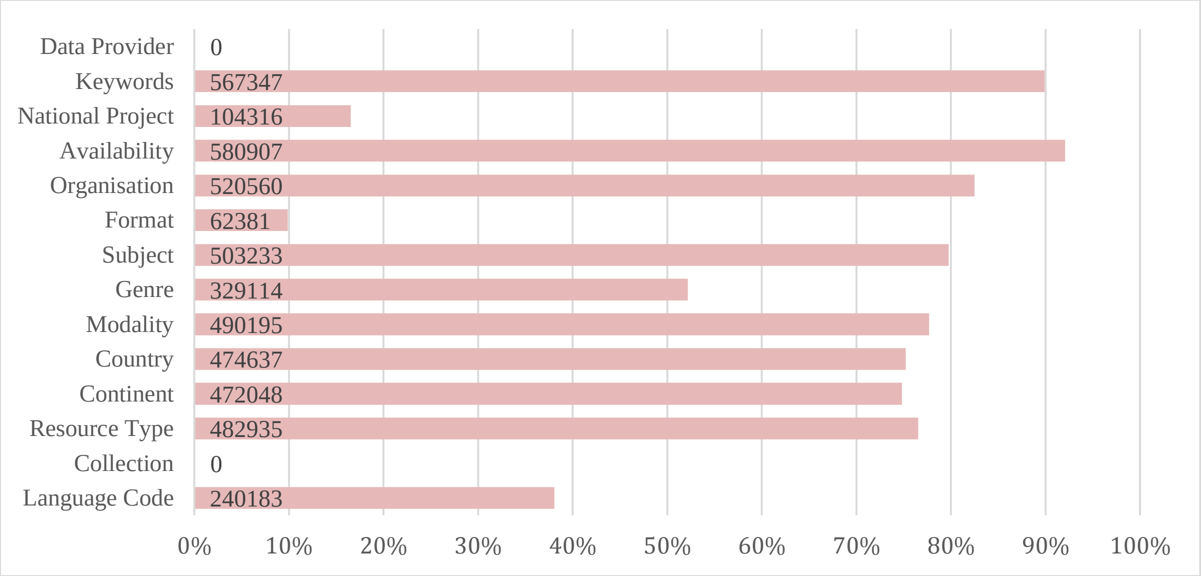


Figure 1 Number of records not covered within given facet in the VLO (on a sample of 631 000 records)

3.3 Missing facets

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

3.4 Missing guidance

A recurring theme in the CMDI/VLO analysis is the lack of coherent guidance for the data providers. Some of the problems with VLO facets may be caused by the data providers not being aware enough of complex mappings taking place in the VLO ingestion process (see Section 4), especially the concepts-to-facets mapping. Their main concern (and a requirement to submit data to CLARIN) is to comply with CMDI and not with VLO facets, CMDI is not specifically designed for the VLO and its scope is broader, but it is indeed a precondition of the latter. Therefore, it is absolutely essential to provide clear guidance for the two aspects. In this respect the facet map checking tool provided by Windhouwer[11] is very useful, but it is not widely circulated nor is it integrated into a coherent set of guidelines for the data providers. The issue of guidance and proposed solutions are further discussed in Section 6.1.

3.5 Need for an efficient curation workflow

As mentioned above much effort has been done to establish the types of problems that exist in the area of facet value normalisation, most notably in Odijk (2014). While some of the trivial problems in value variation can be solved programmatically (case folding, whitespace normalisation), all of the more complex issues like synonyms and complex values require human input – a mapping of variant values to recommended ones. A few tentative mappings have been created as a result of the analysis done by Odijk or the team of authors. Besides the question of the reliability of and broader agreement about 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.

[11] https://lux17.mpi.nl/isocat/clarin/vlo/mapping/index.html

Some automatic curation steps have been applied during the ingestion of the metadata into the indexer for some time (the so-called "post-processing"). Initially, this was limited to simple programmatic corrections of values. Gradually mappings between actual and normalised values were applied to individual facets (*Organisation, Availability, Language, nationalProject*). 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 that the curation process has access to the most current version of the mappings. Meanwhile a more elaborate process is being implemented which is described in, Section 5.4.

4 The mapping and normalisation mechanism

The previous chapter introduced and analysed the issues of metadata quality based on the observation and statistics of the records. This chapter focusses on the underlying mapping and normalisation mechanisms and their impact on the present problems of metadata variability. This section presents the current setup followed by three approaches which will be evaluated by the ability to achieve an improvement in data integrity, the discoverability of resources, and the usability of the VLO.

Before discussing the details of the mapping and normalisation, it is important to touch upon the underlying mechanisms delivered by the CMD framework. The principal interoperability mechanism devised by the CMD framework is the linking of individual CMD elements defined in the schema (or CMD profile) to well-defined concepts (Broeder et al. 2010). This delivers sound semantic grounding of the defined elements independent of the structural aspects of the schemas. Moreover by reusing the same concepts in multiple schemas they can serve as crosswalks. Moving from traditional pair-wise crosswalks (between each pair of schemas) to a conceptual pivotal layer to map individual schemas is a far-reaching paradigm shift.

Even though the CLARIN community and this paper concentrate on the problems of the CMDI framework, some 200 defined CMD profiles, a number of concepts linked from dozens or even hundreds of profiles, and the VLO (and other exploitation applications) relying on this semantic interoperability layer, are a solid demonstration that this approach indeed works. Thus when exploring the alternative mappings in the following sections, we need to bear in mind that CMD, through the use of concept links, already delivers a first (substantial) reduction of variability (on the schema level) using a many-to-one mapping between metadata elements and concepts (see part 1 in Figure 2).

When presenting the mapping scenarios we distinguish three factors or degrees of freedom which have an effect on the display of CMD records in the VLO: 1. *schema mapping* (elements in different schemas referring the same concept as described above), 2. *facet mapping* (multiple concepts mapped to one facet in the VLO), 3. *value mapping* (or "normalisation", values encountered in the metadata replaced with values from a controlled vocabulary according to a normalisation map).

4.1 Scenario 1 (current configuration)

The current procedure of concept mapping and value normalisation is illustrated with an example. A given CMD record contains *ms:OrganisationName* "Summer Institute of Lingustics" and *olac:Type* "Diaries". Like all other CMD records, the format/structure of this record is defined in a XMLSchema derived from a specific CMD profile. A record from another data set defined by another CMD profile includes ex:AgencyName "RADIO ORANJE" and ex:ItemType "plainText". The abovementioned mechanism of concept links (schema mapping) ensures that these elements are also semantically grounded. For instance, *ms:OrganisationName* is linked to the CCR concept *ccr:C-2459*, whereas *olac:Type* element corresponds to a concept *dc:Type*[12]. Similarly, *ex:AgencyName* is mapped to *ccr:C-2979*, whereas *ex:ItemType* is defined as an equivalent to *ccr:C-5424*.

In the ingestion process the framework uses a facet mapping file which defines the mapping between concepts and VLO facets (facet mapping). In our case, *ccr:C-2459* as well as *ccr:C-2979* are mapped to the VLO Organisation facet. Likewise, *ccr:C-5424* and *dc:type* are mapped to VLO *ResourceType* facet[13]. In parallel, normalisation of the values takes place for selected facets (value mapping). Values

[12] CCR is (by design) not the only conceptual reference used for VLO. Especially DCMI was since the beginning handled as equally valid source of concepts.

[13] The labels Resource Type and Resource Class have been used inconsistently and seemingly synonymously for the facet and the corresponding concept. We strongly encourage the standardisation of the labelling convention. We maintain "ResourceType" throughout this paper.

in the metadata elements are mapped into values from controlled vocabularies according to a manually maintained normalisation file. For instance, “Summer Institute of Linguistics” and others including “SIL of University of Oklahoma” in the Organisation facet will be normalised as “SIL”. The values such as “Fictions” and “Diaries” are mapped to “plainText” in the Resource Type facet. We will not discuss whether those mappings are correct or not, but we merely present here how data is curated in the VLO ingestion pipeline.

At first glance it is not problematic, however considering that the semantics of the VLO facets are not yet clearly defined and agreed upon, we face some semantic problems for resource discovery. Cases may arise where the Organisation facet includes not only organisations providing language resource, but also some auxiliary organisations involved in text processing, or as sponsor, depositor or license holder. The same holds true for other facets, like *Country* and *Date*.

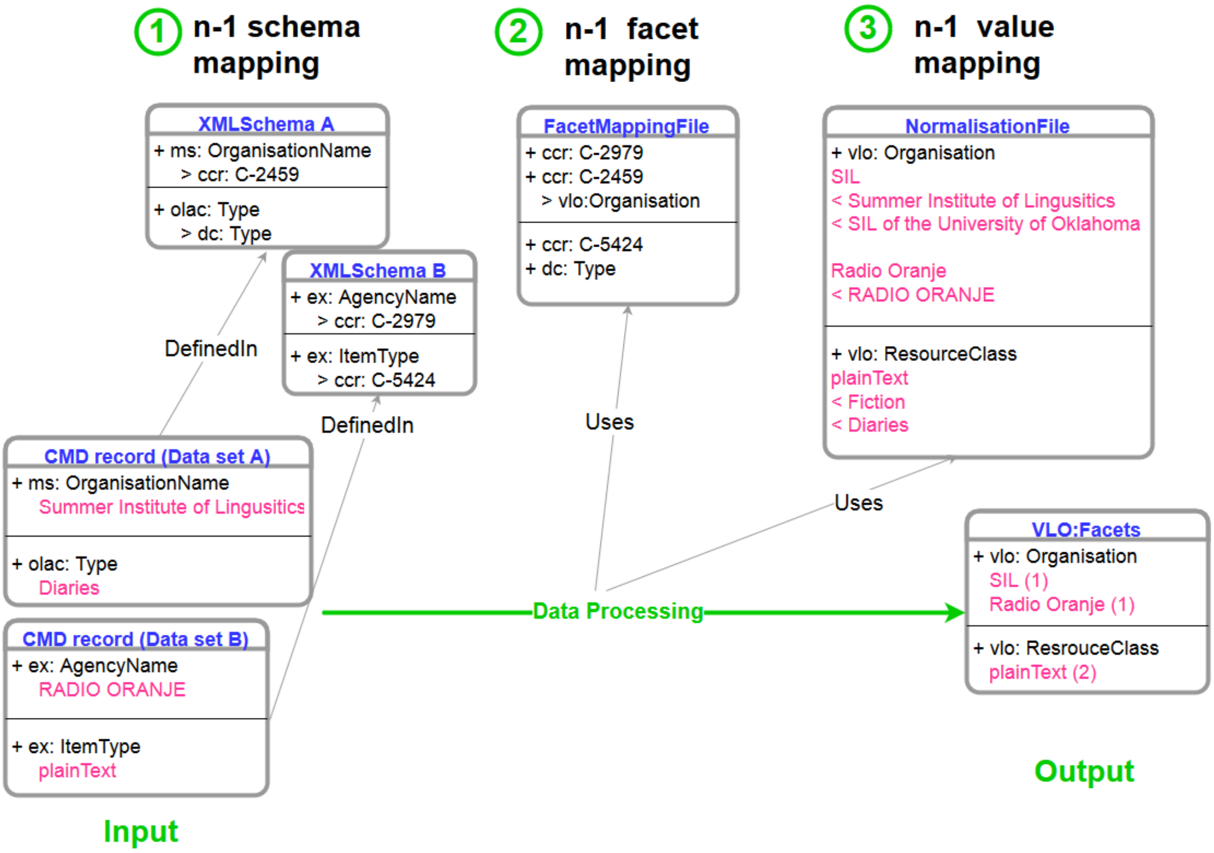


Figure 2. Current scenario with two example records.

Principally, it is acceptable to have a very broad definition of the facets, as long as this definition is made clear to the providers and users. The CCR working group is currently formulating such solid definitions based on their previous work. However, there is a strong sentiment in the CCR and metadata task forces that the facets’ definitions should be narrower and stricter, in order to provide clarity to the users and offer them the (supposedly) most popular semantics of the facets. There are good arguments for this direction, but it further undermines the issue of (already poor) facet coverage. In addition, as long as we apply a many-to-one concepts-to-facet mapping in the VLO configuration, the definition of the latter is necessarily broader than the former. By enlarging the scope of the definitions, we deliver a wide spectrum of concepts defined by the data providers and at least sustain the coverage of records for the VLO facets. In any case it is necessary to assess the impact of restricting the semantic scope of the VLO facets on the facet coverage.

4.2 Scenario 2 (one-to-one facet mapping)

In scenario two, the mapping between the concepts and VLO facets will be one-to-one (Figure 3). In the *FacetMappingFile*, one concept corresponds to exactly one counterpart facet. (In all probability in this scenario the *FacetMappingFile* will be obsolete and the CCR concepts will be used directly to define the indices/facets in the VLO.) For instance, the first CMD record of the previous scenario has the same XML schema, but this time *FacetMappingFile* defines that *ccr:C-2459* will be mapped to *vlo:Organisation* and *ccr:C-5424* will be mapped to *vlo:ResourceType*. The slightly different second CMD record includes "University X" in *ex:PublicationInstitution* field and "plainText" in *ex:Category* field. It will use its own XMLSchema to define a mapping to *ccr:C-xxxx* and *ccr:C-yyyy* concepts. These concepts will further reach *vlo:PubOrganisation* and *vlo:ResourceCategory* facets.

The one-to-one mapping relieves the semantic mismatch problem – every facet carries the semantics of the underlying concept. The problem is that hundreds of concepts are linked from the many defined CMD profiles/schemas, many of them semantically similar (which is exactly the reason why a facet mapping is being applied in the first place), so this option will pass the ambiguity and semantic proximity problems to the user. Furthermore, a user interface with dozens of facets would not be user-friendly. Usually, faceted search interfaces do not employ more than 10 facets (e.g. Europeana six, Gallica nine), and 100% or high data coverage is expected.

With this in mind, a joint effort of the curation and CMDI task forces agreed in October 2015 to explore alternative display methods and user interface layouts, in order to accommodate a substantially higher number of facets/indices, especially concentrating on dynamic or conditional and hierarchical facets. Conditional facets are displayed only under certain circumstances, e.g. bound to a certain resource type, or collection, a given coverage ratio or explicit user selection. This would allow (advanced) users to customize which facets should be displayed.

Hierarchical facets have a potential to resolve some of the limitations of the semantic mapping. For example the user will be able to search in a broad facet *vlo:Organisation*, but will have the option to narrow down to *vlo:PubOrganisation*, as illustrated in Figure 3. Although the implementation would be very challenging and would require substantial development resources, the feature has a potential to become an innovative practice. The development will definitely need to be accompanied by an extensive analysis of the usability of the hierarchical facets, because it is not a very widespread functionality, thus it may cause a reverse effect, confusing the end users.

Also, the facet hierarchy (the hierarchical relations between the concepts) as well as the other dependencies (conditions) between the facets still needs to be defined somewhere, basically moving the facet mapping challenge to another level. Nevertheless this approach may be useful as it makes the concept-to-facet mapping more explicit and transparent to the users. Indeed it would allow to move the mapping from indexing time to query time, yielding a much more flexible exploration interface. The faceted browser developed by the Meertens Institute[14] already adopts this strategy to a certain extent.

We also need to take into account, that the one-to-one facet mapping would have a strong impact on the value mapping, as the normalisation maps are currently defined per facet. When the number of facets would grow considerably, so would potentially also the number of needed normalisation mappings.

4.3 Scenario 3 (dumb-down schema mapping)

The third option features the same one-to-one facet mapping, but it differs from the previous one in that the number of relevant concepts used are reduced to the minimum by ensuring that the relevant concepts used in the schemas exactly match the VLO facets (see part 1 of Figure 4) The *FacetMappingFile* will become obsolete, given the one-to-one facet mapping. The disadvantage is obvious. During the mapping from original metadata elements to the CCR concepts, some semantics are lost. For example, *ms:Organisation* and *ex:PublicationOrganisation* are dumbed down to the general *vlo:Organisation* facet. With regard to value mapping, this scenario would result in even higher variability and ambiguity of the values in each facet. We consider this scenario a purely theoretical option, as it would require the change of the definitions of most of the existing CMD profiles and would introduce a significant loss of semantic precision in the schema definitions contrary to the basic principles of CMDI.

4.4 Scenario 4 (many-to-many value mapping – multi-facet decomposition)

This scenario further develops the idea of value mapping introduced above, introducing the "multi-facet decomposition", i.e. one value can be mapped to multiple values in different facets.

[14] http://www.meertens.knaw.nl/cmdi/search/

For example, a metadata element with values “diaries” or “Bibles” in the *olac:Type* field that is mapped assigned to the *vlo:ResourceType* facet (using the facet mapping). Normally, the *vlo:ResourceType* facet should contain similar values as DCMI type vocabulary[15]. This suggests a problem with the semantics. A proposed remedy is to normalise the values as well as to re-map them to other facets such as subject and genre. The result for the “diaries” would be, for instance, “plainText” in *vlo:ResourceType* facet, plus “diary” in *vlo:Genre* facet, and, for the “Bibles”, “plainText” in *vlo:ResourceType*, while “Bible” in *vlo:Subject* (or *vlo:ResourceTitle*) and “religious text” in *vlo:Genre*. Regarding schema and facet mapping, this scenario is the same as the current configuration (Figure 2).

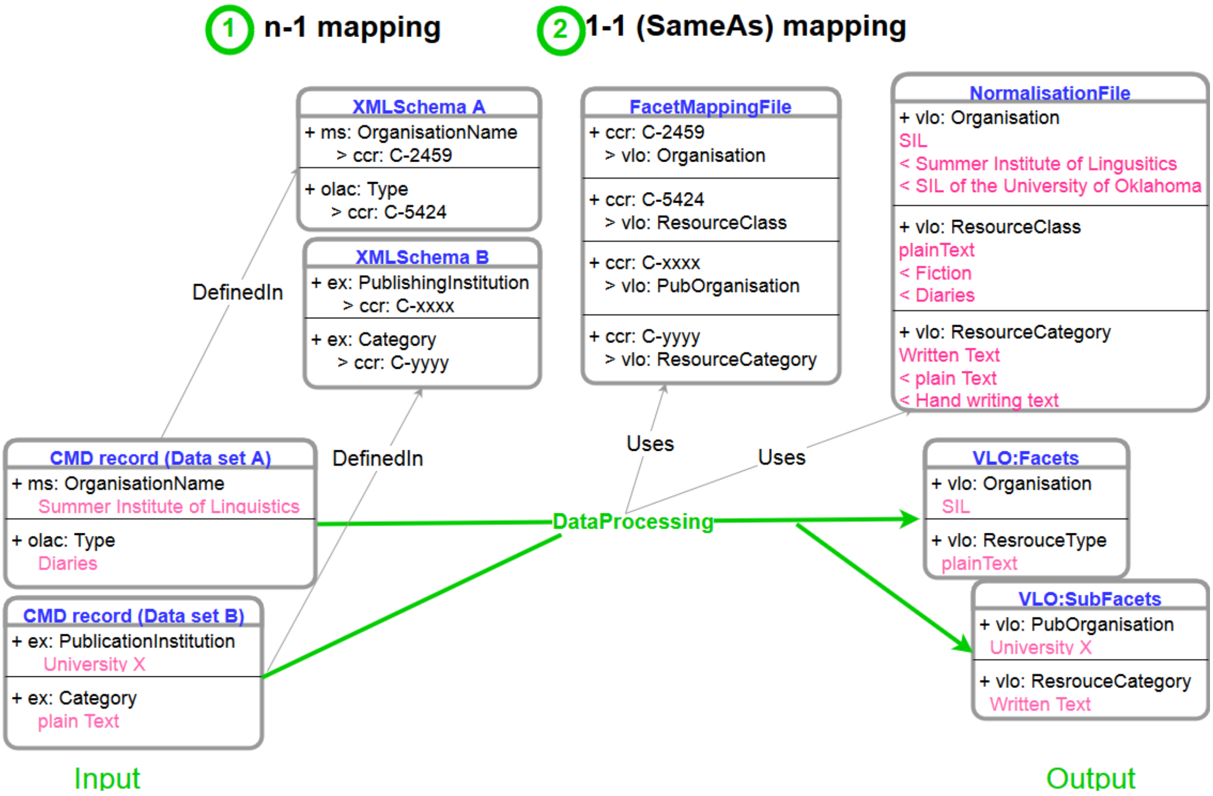


Figure 3. Scenario 2. One-to-one facet mapping.

[15] http://dublincore.org/documents/2003/11/19/dcmi-type-vocabulary/

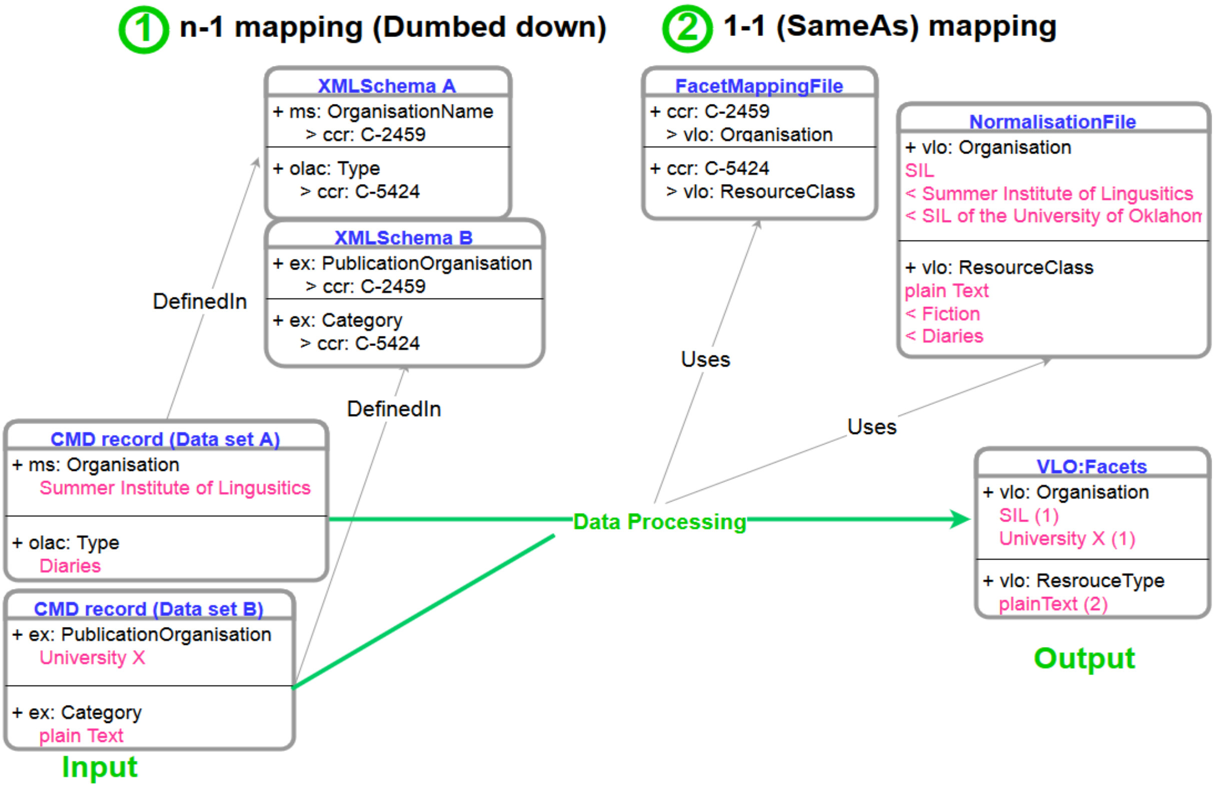


Figure 4. Scenario 3 Dumbed-down schema mapping

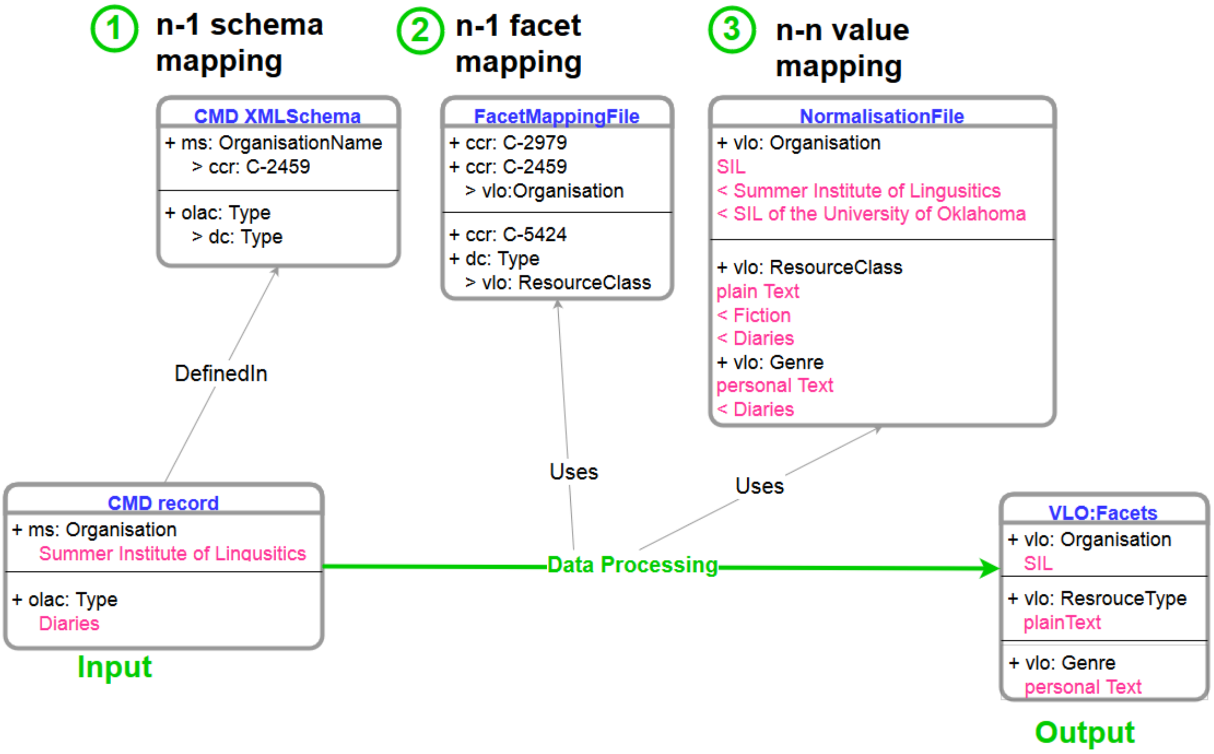


Figure 5. Scenario 4. Many-to-many value mapping

This strategy is motivated by the view that is unlikely that the normalisation would happen on the side of the data providers and that the central curation team has to take charge. On the surface (in the VLO), the approach is effective, because the facets appear clean and more understandable for the end-users. However, this approach implies a significant interpretative addition on the side of the curation. The modification may be far from the data provider's intention and the end user's expectation and could negatively affect the resource discoverability. This can serve as a short-term solution for maintaining data consistency in VLO but extensive steps must be put into place for the curators to ensure that resource discoverability is not compromised. See Section 5.5 for some details on the case study mapping values for the *vlo:ResourceType* facet.

4.5 Summary of the mapping scenarios

In this section, we have elaborated on different scenarios for concept mapping and value normalisation. The focus of the discussion is the second part of the data processing (mapping between concepts and VLO facets). It is currently a many-to-one model. If we do not change this relationship (scenario 1), broader definitions for the VLO facets are recommended in order to accommodate various types of similar concepts in one facet. If we move to a one-to-one relationship, there are theoretically two paths to be followed. Scenario 2 maintains (and maybe also fine-tunes) all the existing concepts. As it is not possible to guarantee the same number of facets as concepts, an extensive investigation is necessary on how to accommodate this complexity in the user interface without hampering the usability. Dynamic and hierarchical faceting is a promising strategy here. Scenario 3 also aims at one-to-one facet mapping, but shifts the semantic reduction towards the schema mapping, which is both not feasible (all the profiles must be changed) and unacceptable (contrary to core CMDI principles). The last scenario introduces the idea of multi-facet decomposition on the value mapping level, inspired by the “messy” (overloaded) values of the metadata in some facets. However, this mechanism may lead to an unintended manipulation and interpretation of data, which would distort resource discovery. Thus, it can only be applied with great caution and in a conservative manner. The above analysis makes explicit the complexity of the process of ingesting and mapping CMD records into the VLO. Even if we disregard the problems coming from the data providers – there are three levels of semantic engineering in this process, which makes it a very demanding task to trace back to the source the different problems with quality of metadata and in the VLO.

5 The data processing workflow/pipeline

After investigating qualitative and quantitative aspects of the question at hand as well as mapping and normalisation mechanisms in detail, in this section we focus on the actual implementation of the ingestion and curation workflow and propose optimisations aiming mainly at a more integrated, more ergonomic setup and better communication with the data providers.

5.1 Current setup

Figure 6 illustrates a simplified view of the current workflow. It is a well-established chain of actions starting from a data submission through data processing to indexing and publishing on the VLO website. The starting point is when a data provider accesses a metadata authoring tool often hosted at a CLARIN national centre to design and create their records. Typical examples are ARBIL[16] in the Netherlands, COMEDI[17] developed in Norway, and the custom submission form of DSpace as implemented in Czech Republic[18] or in Poland[19]. Most of the tools are tightly integrated with the underlying data repository, where the metadata is stored together with the digital resources. Some allow for the use of any (or multiple) CMD profiles, some are tailored towards one specific profile. The metadata is exposed via an OAI-PMH endpoint from where it is fetched by VLO harvester on a regular basis. OLAC[20] and CMDI are the two major metadata formats that can be imported into the VLO environment, and the former is converted to CMDI by a predefined mapping. When CMDI is ready, it is being ingested into the Solr/Lucene index, governed by a set of configuration files: a facet mapping file and value mapping and normalisation files described in Section 4. The processed data is indexed and published on the VLO website, where the end users can browse and search the data.

While the authoring tools try to provide a local control over the quality of the metadata, offering a custom auto-complete functionality based on local controlled vocabularies and various consistency

[16] https://tla.mpi.nl/tools/tla-tools/arbil/
[17] http://clarino.uib.no/comedi/page
[18] https://github.com/ufal/lindat-dspace
[19] https://clarin-pl.eu/dspace/
[20] http://www.language-archives.org/OLAC/metadata.html

checks, and there are also already individual infrastructural services available for data providers to check their metadata (OAI-endpoint validator, schema validator), a coherent, formal and rigorous mechanism for VLO data ingestion is lacking.

The CLAVAS service especially dedicated to shared management of controlled vocabularies, although advocated on several occasions, is not yet fully functional as an authoritative source of controlled vocabularies in the CMD infrastructure, mainly due to lack of well-defined organisational procedures. An integrated user interface for editing the facet mapping file and the value normalisation maps is missing (see Section 5.4 for details of current usage) There is also no automatic (and very little manual) feedback from the VLO team after data ingestion, thus the data providers are required to exert a significant amount of effort to improve the metadata quality by individual consultation. Both the VLO curators and the data providers can examine the quality/integrity of the metadata only on the public website.

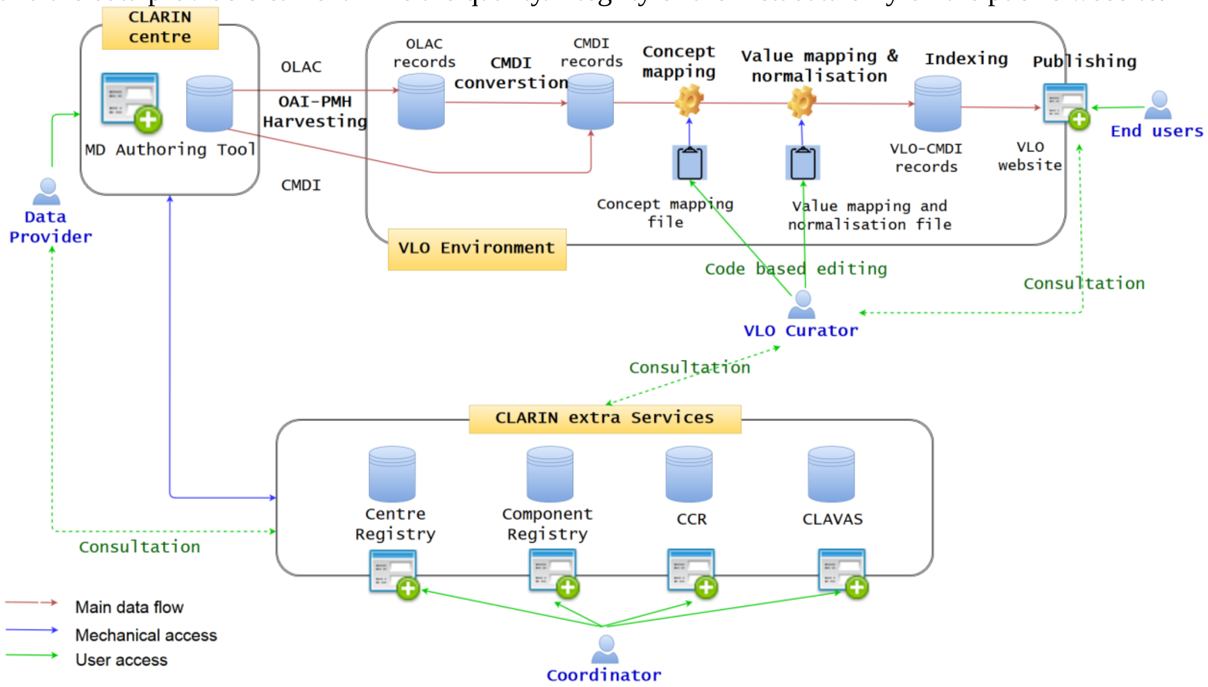


Figure 6. Current workflow.

5.2 Dashboard – an integrated data management system

In order to better manage the full VLO workflow, we propose a dashboard component (Figure 7) as a central interface or single entry point that will integrate all the procedural information from the individual steps of the data processing pipeline into one user-friendly GUI web interface with which the VLO curators, administrators and data providers can work on data management much more efficiently and coherently in a uniform manner, fostering the metadata quality and the VLO user experience. It should offer an intuitive monitoring view, illustrating each stage of the entire process, starting from harvesting, converting, and validating, to indexing and distributing. Note, that the dashboard would not perform any of the tasks in the process itself, but rather interact with the individual components of the VLO framework – harvester, converter, validator, mapper/normaliser, and indexer/publisher. The functionalities of the dashboard should include (but are not limited to):

F1. List of the datasets (OAI-PMH sets), optionally grouped per data provider and per CLARIN centres/countries (MUST)[21]

F2. Status and statistics of the datasets within the ingestion pipeline (errors, progress indicator) (MUST) (export as PDF, XML, CSV etc. (SHOULD))

F3. Simple visualisation of the statistics in F2, including pie charts, bar charts etc. (COULD)

[21] Suggestions of priorities are made using MoSCoW method.

F4. Browse the data quality reports per set (MUST) (export as PDF, XML, CSV etc (SHOULD)) including a link checker which lists broken links (COULD)

F5. Deliver the data quality report to the data provider/CLARIN centre (via automatic email, and/or via a web interface) (SHOULD)

F6. Edit the concept to facet mapping (MUST)

F7. Edit the value mapping and normalisation (MUST)

F8. Manual data management (deactivate indexing of the sets, delete the data sets, invoke harvesting of data sets etc) (MUST)

F9. Browse the log files of the VLO systems (COULD)

F10. Browse the Piwik web traffic monitoring (COULD) (do it per data set (COULD))

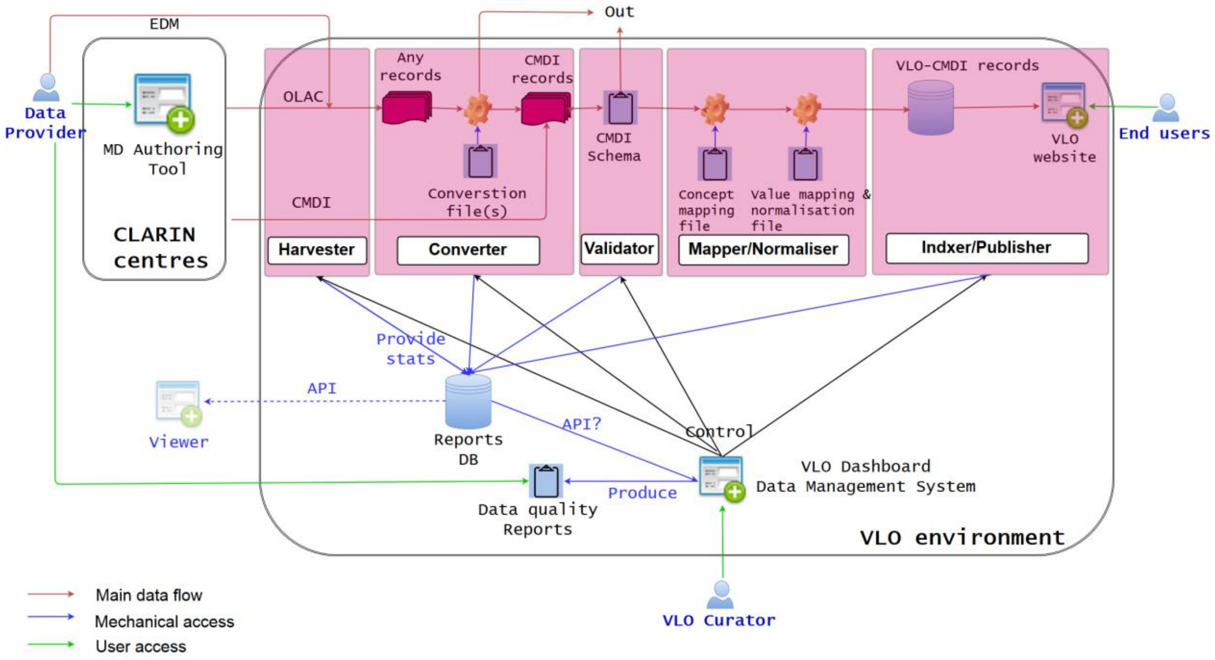
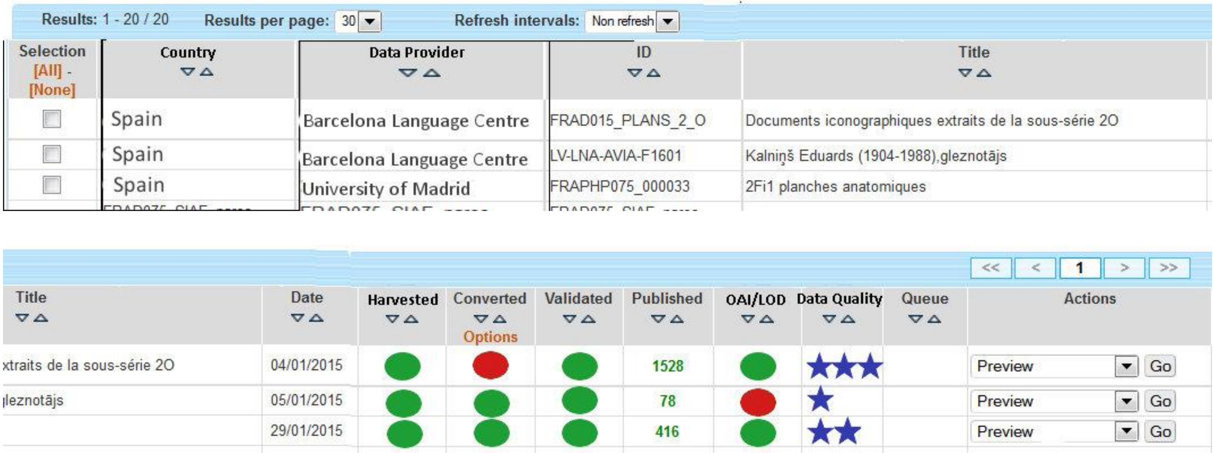


Figure 7. Proposed workflow – integrated dashboard.

With regard to the metadata problems, the dashboard is not supposed to resolve them per se. Rather it helps to identify them and keep track of them by means of a user-friendly interface (F1, 2, 3). In this sense the metadata quality report (F4) is probably the most prominent function. It compiles information from the harvesting, validation and curation step and makes it available to both VLO curators and data providers (F5). It is crucial that the users are able to edit/configure the complex mapping and normalisation within the integrated environment in sync with the core infrastructural registries Component Registry, CCR and CLAVAS. Additionally, the dashboard environment could allow for alternative instantiations of the facet mapping, so that the curator can experiment with different versions, approaches etc. before publishing the metadata records. The dashboard would also allow us to invoke actions on datasets directly in the interface (F8):

- (Re-)harvesting of the data set
- Disable indexing
- Delete the data set
- Show the data quality report (see Section 5.3) (download them as XML, PDF, etc.)
- Show the error messages (download them as PDF etc.)
- Show the metadata records
- Show the schema/profile (with the link to Component Registry, CLAVAS, and CCR)
- Send an email to the data provider (e.g. data quality report)

Figure 8 visualises the idea of the dashboard user interface in which the VLO curators can monitor the whole data processing workflow. In the mock-up, OAI-PMH data sets are listed as rows, and can easily be sorted per country and data provider. Following the ID and title, there is a date of latest update (e.g. harvesting date or latest actions). The status of data processing is clearly visible with green and red circles (harvested, converted to CMD, validated against CMD). When the dataset is indexed and published, the number of records is shown. Should the data be offered for further distribution (e.g. OAI-PMH, Linked Open Data etc.) in the future, the status will also be indicated here. Finally, an indication of the data quality could be provided with stars (or the metric delivered by quality assessment). Different actions will be selectable per set, according to the status of the data (F8).



Results: 1 - 20 / 20 Results per page: 30 ▾ Refresh intervals: Non refresh ▾

<< < 1 > >>

Selection [All] - [None]	Country ▽△	Data Provider ▽△	ID ▽△	Title ▽△	Date ▽△	Harvested ▽△	Converted ▽△ Options	Validated ▽△	Published ▽△	OAI/LOD ▽△	Data Quality ▽△	Queue ▽△	Actions
□	Spain	Barcelona Language Centre	FRAD015_PLANS_2_O	Documents iconographiques extraits de la sous-série 2O	04/01/2015	🟢	🔴	🟢	1528	🟢	★★★		Preview ▾ Go
□	Spain	Barcelona Language Centre	LV-LNA-AVIA-F1601	Kalniņš Eduards (1904-1988),gleznotājs	05/01/2015	🟢	🟢	🟢	78	🔴	★		Preview ▾ Go
□	Spain	University of Madrid	FRAPHP075_000033	2Fi1 planches anatomiques	29/01/2015	🟢	🟢	🟢	416	🟢	★★		Preview ▾ Go

Figure 8. Dashboard mock-up (table cut into two parts for readability).

All actions can be applied both to a single data set as well as to multiple data sets in batch mode. The user can search, sort, and filter the information in this table view. In addition, s/he can select multiple data sets by clicking the checkboxes on the left. With this table view, the user can see the overall statistics (and/or selected data sets), including the number of datasets, countries, data providers, the status, the number of records indexed and distributed. These figures are important performance indicators (for CLARIN board, funders, but also the data providers themselves). The user should also be able to export the statistics as PDF and CSV or directly into an online spreadsheet. Equally important is the availability of historic information, i.e. statistics from previous harvests, allowing the curator to spot immediately sudden quality drops, or dramatic changes in the amount of records provided.

In summary, the dashboard will provide manual data processing functions, complementing and governing the automatic data processing. It will allow the VLO curators to monitor the data and manually interact with it without any knowledge of behind-the-scene codes and scripts.

5.3 Curation module

We are aware that the dashboard cannot be built overnight. Thus, it has to be developed block by block. The most imminent block of development will be curation module. Currently, the curation steps are implemented within the VLO ingestion application. The CMDI curation and VLO teams agreed to extract this functionality into a separate module that can be reinserted into the VLO ingestion pipeline, but can be used in other contexts as well. The specification for the Curation Module is based on previous work by Kemps-Snijders (2014), Trippel et al. (2014) and other (technical) documentation created by CLARIN metadata team in the last few years. The curation module will validate and normalize single metadata records, as well as whole collections, assess their quality and produce reports with different information for different actors in the VLO workflow. The module will integrate with the harvester that is being re-implemented and especially will also be tightly integrated with Dashboard application. The following four main use cases were already identified:

a) Metadata creator checks the validity of newly created records

b) Metadata modeller checks the quality of profiles

c) Repository administrator checks quality of metadata in his repository

d) Continuously check all metadata harvested and indexed by VLO

The module is being developed by the ACDH-OEAW implementing the task 2.2.1 of the CLARIN-PLUS project. First version is scheduled for February 2016. Some of the planned features are as follows:

- Schema validation,
- URL inspection,
- Value validation and normalisation against controlled vocabularies and normalisation.
- Assess facet coverage (of the profile and of the record)
- Feedback about errors, per record or per collection.
- Quality metrics
- Provision of instructions on improvement of the metadata optionally accompanied by already amended (normalised) CMD records
- Comparison of the curation results over time

5.4 Management of vocabularies and mapping

We need to take yet a closer look on the handling of the vocabularies in relation to the value mapping. A relatively simple (and partly implemented) approach to the management of the mappings is to maintain the vocabularies in the vocabulary repository CLAVAS, where, based on the SKOS data model, every entity or concept is registered as a separate item (*skos:Concept*), with a *skos:prefLabel* as the normalised label for a given concept and all variants encountered in the actual metadata stored as *skos:altLabel* (or *skos:hiddenLabel*). This information can be easily retrieved from CLAVAS via its REST-API and injected in the harvesting/curation workflow of the VLO. Until now, this has been done for Organisation names. The change introduced in CMDI 1.2 (Goosen et al., 2014) allows the indication of a controlled vocabulary for a given element in the CMD profile which will enable a more consistent handling of vocabularies in 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 not only to query but also to manipulate the data. So technically it is 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 a group of curators.

However, even if this whole process is set up, it does not offer a solution to more complex normalisation scenarios, like the multi-facet decomposition introduced in Section 4.4. Even though this specific scenario, is problematic in a certain respect (potentially severe interpretative intervention), it is clear that more advanced mapping mechanisms will be needed that cannot be served by the simple approach based on *skos:Concept* as proposed above. The current temporary solution for maintaining multi-facet mappings with which we have experimented is to use a simple spreadsheet with the encountered values in first column, and a separate column for the other facets, allowing the curators to assign values in multiple facets for any given value. These files are stored as *text/csv* file and maintained under version control in the CLARIN's code repository[22], so they can be edited by a team of curators, who can see who has done what when, but also retrieved and processed by any application, most notably the curation module. However this is still a very cumbersome and not well integrated process. Ideally, the functionality for value normalisation has to be well integrated into the dashboard (see F7 in Section 5.2). Thus, in addition to the data management view, the Dashboard has to offer a user interface to create and edit the concept mapping and the value mapping and normalisation (F6, 7). This functionality should completely hide the internal mapping mechanism, freeing the VLO curators from the manual editing and tedious syncing of CSV or XML files stored in different places, as is the case currently. However developing an interactive web-based table or spreadsheet application that features at least a minimal set of functionalities is a resource-intensive task itself. The value normalisation interface has to integrate with

[22] https://github.com/clarin-eric/VLO/tree/vlo-3.3-oeaw/vlo-vocabularies/maps/csv

CLAVAS and ideally also other sources of controlled vocabularies, offering the curator normalised values via autocomplete or similar functionality.

5.5 Normalisation example: resource type

Let us take a closer look at the example facet *ResourceType*. Currently, the facet encompasses around 300 different values. There were multiple attempts to define a controlled vocabulary for this facet, among others by the CLARIN-D VLO taskforce and Odijk (2015). These proposals stand next to a number of existing controlled vocabularies from other domains like Europeana (*edm:Type*, see Section 2.1), or DCMI Type. All the controlled vocabularies have some overlapping terms, some omissions and some slight differences in the semantics of the terms.

The curation task force is working from a vocabulary of some ten to twelve terms that tries to accommodate all of the above. The governing aspects are: high-level distinction to keep the number of values low (no more than 15, ideally under 10), and decomposition, i.e. each term signifies a certain “atomic” aspect of the resource and it is allowed to use a combination of values to describe one resource. Example:

```
AnnotatedTextCorpus = collection, text, annotation
Audio recording with transcription = audio, annotation
```

The currently proposed draft vocabulary

- annotation
- audioRecording
- collection
- structuredData
- grammar
- image
- lexicalResource
- physicalObject
- text
- videoRecording
- software
- service/interactiveResource

A crucial aspect of any controlled vocabulary is a sound definition of individual terms. A trial set of definitions is currently being worked out making use of the existing definitions from existing vocabularies. Once a coherent set of definitions is available this vocabulary will be circulated among the relevant CLARIN bodies and colleagues and especially will be discussed with the CLARIN Concept Registry group, in order to achieve a broad agreement for the vocabulary. Next, after adapting the normalisation maps against this authoritative list, a thorough examination of the soundness of the mappings will be undertaken and finally the mappings will be applied in the VLO and the vocabulary will be exposed for public use. We especially plan to use this vocabulary to make *vlo:ResourceType* a primary, prominent facet in the VLO, adorned by appropriate icons, potentially influencing also the customisation of display (different resource type ask for different facets, as proposed in Scenario 2 in Section 4.2).

6 Upcoming work

In this section we place the proposed partial solutions introduced in the previous sections, into a bigger picture. The metadata quality issues cannot be addressed by a single measure, but by a comprehensive set of measures. We point especially to the importance of the social dimension of the solutions. For example, the maintenance of vocabularies and mappings can only be effective if a broad agreement can be reached in a collaborative manner, if they are to be integrated into the automatic curation process, and widely adopted by the data providers.

The technical solutions comprise the following elements: adaptation of the facet mapping, concepts directly available in the VLO accompanied by advanced user interface features like dynamic or hierarchical facets (scenario 2 Section 4.2); (conservative) advanced value normalisation based on shared normalisation maps (scenario 4 Section 4.4); a dashboard as an integrated interface for managing the ingestion and curation workflow (Section 5.2); a curation module assessing various aspects of metadata quality (Section 5.3). It is crucial to ensure that all changes applied during the processing (i.e. the mapping of the records to facets and the value normalisation) are transparent to the data provider and to the user of the VLO. Another requirement is to make the workflow more modular, especially allowing for the curation module to be encapsulated enough to be reusable in other contexts. A final technical issue is the testing phase. In order to ensure that the metadata quality and VLO discoverability are improved by the 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.

It is impressive that CLARIN has developed a unique approach and system for their data aggregation. It has developed CMDI to facilitate the heterogeneity of the metadata for a linguistic domain, accompanied by the impressive automation of the data processing from the harvesting to indexing. It is to some extent effective. However, precisely due to this combination of data ingestion mechanisms, it leaves space for problems. It is evident that satisfactory results cannot be achieved with purely automatic measures. There has to be always a human curation, whether it is by a data provider or central curation team. Considering that the facets are the main selling point of VLO, it is imperative to find a complete solution to the extremely low coverage of records which has not been recognised until recently, and to tackle the issue from a structural point of view.

A crucial ingredient to the proposed strategy is the question of governance, i.e. who is going to steer the process and persistently remind data providers of the problems encountered and propose solutions. CLARIN has well-defined organisational structures and a number of bodies with delegates from all member countries where decisions can be agreed upon at 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 the application of small, practical, technical solutions as well as to propose substantial structural changes, if needed.

6.1 Prevention – fighting the problem at the source

While we pessimistically stated before 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) need to be provided, with profiles that have good coverage of the facets and use controlled vocabularies wherever possible

b) provision of metadata quality reports to the providers

c) provision of curated/amended metadata records directly back to the data providers

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

6.2 Semantics and relations of the concepts

We recognise that CCR is currently undergoing a restructuring, recently replacing the ISOcat in February 2015. However we must not forget that ISOcat had been in intensive use as a semantic layer of the

VLO for several years to build and develop the data aggregation systems. It is now high time to seriously discuss all the concepts in CCR and finalise them in order to ensure the semantic data integrity of concepts used within CLARIN. In particular, this paper has shown the emerging issues of mapping problems between CCR concepts and VLO facets. Broeder et al. (2010, 2014) reiterated that ISOcat forms the basis for semantic interoperability and will enable semantic search over the dataset. However this requires answers to following questions:

- What are the relationships between CCR concepts?
- How do we relate CCR concepts and external concepts such as DCMI?
- What are the needs and requirements for VLO and CCR to implement Semantic Web?

The first and second questions were indeed picked up right from the beginning (Broeder et al, 2010), referring to the need for a Relation Registry (Windhouwer, 2012), accompanying the concept registry. Adopting SKOS, a decision come to on account of the migration to CCR, as data model for the concepts allows us to define relationships such as the hierarchical structure of the concepts (skos:broader, skos:narrower). This at least provides good technical means to help clarify the semantics and relations of CCR concepts and VLO facets, especially given the support for defining the SKOS relations in OpenSKOS[23] the software underlying CLAVAS and CCR.

On the other hand, it remains unclear how these relations relate to the semantics of the defined CMD components and how they can be exploited for resource discovery. It is at least clear that it will add an extra dimension (maybe confusion) to the already complex concepts-facets mapping. It would be worthwhile to explore whether SKOS relations could replace, or serve as basis for the VLO facet mappings. Moreover, SKOS can only define (hierarchical) relations between concepts. For a comprehensive description of properties and relations we need to look to RDFS. This task is already being taken up within the CLARIAH-NL project, building on previous work by Durco and Windhouwer (2014) on expressing the whole of CMD in RDF[24].

7 Conclusion

In this paper we evaluated the VLO metadata quality issues at different levels. First of all, we outlined our observations and statistical analysis of the value variation. Secondly, we pointed out the volatile situation of the current mapping and normalisation mechanisms. Thirdly, the fragmented ingestion workflow was revealed. Different elements of the VLO environment all contribute to the problems. Therefore, our proposal is a comprehensive set of technical and social solutions. We proposed a concrete strategy for curation and normalisation of values in the facets of the VLO. We elaborated on the ways to establish and sustain a data ingestion mechanism and 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 normalised values. An integral part of the process must be a suite of test cases that ensures the quality of the mappings and the whole curation process. Finally, all output of the curation (corrections and amended metadata records) must be recycled to the data providers in the hope of preventing problems in the future and the entire work cycle must repeat as new resources are added. Thus the need for metadata curation is perpetual.

VLO is a living infrastructural service designed to provide a single access point to the European language resources. Thus, it will adopt new ideas and technologies and continue to evolve for the needs of the end-users. In this sense, we always take a heuristic approach. This paper has aimed to deliver a detailed analysis of the current metadata issues in the context of data ingestion mechanism and workflow. We can continue to improve the service accordingly and continuously. In a close collaboration with CLARIN partners in other European countries, we strive to make a regular contribution to the development of high-quality stable and sustainable VLO services.

[23] http://openskos.org

[24] https://github.com/TheLanguageArchive/CMD2RDF

References

[Broeder et al.2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry-and component-based metadata framework. In *Procedings of the Seventh Conference on International Language Resources and Evaluation* [LREC2010]. Pp. 43-47.

[Broeder et al.2012] D. Broeder, M. Windhouwer, D. Van Uytvanck, T. Goosen, and T. Trippel. 2012. CMDI: a component metadata infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 1387-1390.

[Broeder et al.2014] D. Broeder, I. Schuurman, and M. Windhouwer. 2014. Experiences with the ISOcat Data Category Registry. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* [LREC 2014]. Pp. 4565-4568.

[Calarco et al.2014] P. Calarco, L. Conrad, R. Kessler, and M. Vandenburg. 2014. Metadata Challenges in Library Discovery Systems. In *Proceedings of the Charleston Library Conference*. Purdue University e-Pubs. Pp. 533-540.

[Durco and Moerth2014] M. Ďurčo, and K. Mörth. 2014. Towards a DH Knowledge Hub - Step 1: Vocabularies. Presented at *Clarin 2014 Conference* [CAC2014].

[Durco and Windhouwer2014] M. Ďurčo and M. Windhouwer. 2014. From CLARIN Component Metadata to Linked Open Data. In *Proceedings of the Third Workshop on Linked Data in Linguistics* [LDL 2014]. Pp. 13-17.

[Europeana2009] Europeana. 2009. Metadata Mapping & Normalisation Guidelines for the Europeana Prototype: Europeana Version 1.2. Europeana: Think Culture, Den Haag, Netherlands.

[Europeana2014] Europeana. 2014. EDM Mapping Guidelines: Europeana Version 2.2. Europeana: Think Culture, Den Haag, Netherlands.

[Goosen et al.2014] T. Goosen, M. Windhouwer, O. Ohren, A. Herold, T. Eckart, M. Ďurčo and O. Schonefeld. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure [CAC2014]. In *Selected Papers from the CLARIN 2014 Conference* [CAC2014]. Pp. 36-53.

[Haaf et al.2014] S. Haaf, P. Fankhauser, T. Trippel, K. Eckart, T. Eckart, H. Hedeland and D. Van Uytvanck. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres. Presented at *Clarin 2014 Conference* [CAC2014].

[Huffman2015] N. Huffman. 2015. Adventures in metadata hygiene: using Open Refine, XSLT, and Excel to dedup and reconcile name and subject headings in EAD. In *Bitstreams: Notes from the digital projects team.* Duke University Libraries, N.C.

[Kemps-Snijders2014] M. Kemps-Snijders. 2014. Metadata quality assurance for CLARIN. Technical report.

[Lyse et al.2014] G. Lyse, P. Meurer, and K. De Smedt. 2014. COMEDI: A New Component Metadata Editor. In *Papers from the CLARIN 2014 Conference* [CAC2014]. Pp. 82-88.

[Odijk2014] J. Odijk. 2014. Discovering Resources in CLARIN: Problems and Suggestions for Solutions. Utrecht University Repository, Netherlands.

[Odijk2015] J. Odijk. 2015. Metadata curation strategy. Internal document, unpublished.

[Palmer2014] W. Palmer, 2014. Fits metadata normalisation API? Github Repository.

[Sofou and Tzouvaras2015] N. Sofou, and V. Tzouvaras. 2015. MS28: Sounds thesaurus and metadata cleaning and normalization module complete. Europeana Sounds 620591, Den Haag, 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.

[Van Uytvanck2010] D. Van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardelleni. 2010. Virtual Language Observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* [LREC 2010]. Pp. 900-903.

[Van Uytvanck2012] D. Van Uytvanck, H. Stehouwer, and L. Lampen. 2012. Semantic metadata mapping in practice: The Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 1029-1034.

[Windhouwer2012] M. Windhouwer. 2012. RELcat: a Relation Registry for ISOcat data categories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 3661-3664.

[Withers2012] P. Withers. 2012. Metadata management with Arbil. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* [LREC2012]. Pp. 72–75.