Discovering Resources in CLARIN: Problems and Suggestions for Solutions

Jan Odijk
UiL-OTS, Utrecht University
j.odijk@uu.nl

1 Introduction

Discovery of research resources (data and software) relevant to a researcher is an important functionality that CLARIN aims to offer. To that end, it has set up the Component-based Metadata Infrastructure (CMDI, Broeder *et al.* (2010)) so that metadata can be made for all resources in the CLARIN infrastructure. It also makes available the Virtual Language Observatory (VLO, Van Uytvanck (2014)), which enables a user to search for resources through their metadata.

Unfortunately, finding resources[1] by searching in their metadata through the VLO is not as easy as it should be. I will illustrate this in this paper. I will also analyze what causes this and make suggestions for solutions. I am a linguist, and focus on searching resources that are interesting to me as a linguist. Similar analyses should be done by researchers from other disciplines, to analyze whether the metadata and the VLO are useful to them and to specify how these can be improved to serve their needs.

The major causes for the problems can be found in (1) the nature of the metadata, and (2) limitations of the VLO.

The paper is structured as follows. I will describe what metadata are and what purposes they must serve in section 2. An important factor is the wildly varying granularity of the metadata descriptions: section 3 is dedicated to this topic. Some basic characteristics of the VLO are described in section 4. Section 5 will describe the facets currently offered by the VLO, as well as some suggestions for new facets. I will then use some test queries to show how difficult it is to use the VLO, given the metadata it operates on, and the limited number of facets it can use, for discovery of resources (section 6). I summarize the problems encountered in section 7, and make several recommendations for improving the situation in section 8. I end with conclusions in section 9.

[1] *Resources* include both *data* and *software*, However, there are currently very little metadata for software, and no facets in the VLO specific for software at all, so this report will focus on data.

2 Purpose of resource descriptions and the VLO

Metadata[2] is a description of a resource and serves as (more or less formalized) documentation of the resource. One of the main purposes of metadata is to enable discovery of a resource.

CMDI offers a very flexible framework for making metadata: it provides a model and a format for metadata, but does not in any way prescribe the contents of the metadata. This flexibility is needed, since the world of resources in the humanities is too varied for fixed metadata schemes. But in my view, there is currently too much flexibility, so much that CMDI cannot properly serve its purpose (discovery of resources) anymore.

Most metadata are made in isolation by a specific research group or data provider.[3] This often leads to (often unnecessary) differences between metadata. CMDI offers a registry of metadata profiles and components, but these are basically flat lists of profiles and components, and without search facilities it is very difficult for a user to determine whether there are profiles and components that can be re-used, and which ones these are.[4] A proliferation of profiles and components results. Often crucial information is lacking in the metadata (properties that are 'obvious' for a researcher are often not included in the metadata). CMDI gives too little guidance to metadata providers to assist them in making 'useful' metadata.

Too many metadata elements allow open vocabularies, which leads to many different terms for the same data category or concept, and to many different interpretations of the same term. Setting up closed vocabularies is possible in CLARIN, but most researchers are hesitant to do so. First, researchers often do not agree what values should be included. Second, nobody has a complete overview of a domain, so there is the danger of missing values. Third, there will undoubtedly be new developments in the future that nobody can foresee, which will require adaptation of the closed vocabulary. But adapting a closed vocabulary in the CLARIN-supported data category registry is not possible once it has been made public.[5] If an adaptation is needed, one must make a new version of the closed data category, but there no good facilities for making a new version, marking it as a new version of an existing data category, marking the original one as obsolete etc.[6]

3 Granularity

The granularity of metadata, i.e. the size of the resource they describe, differs wildly. In many cases, the granularity of the metadata is rather small. This can be useful, because it enables much more detailed descriptions of resources, but it is only useful if one can search for properties of these more detailed descriptions.

In many cases a collection of fine-grained resources should be viewed as a resource itself

[2] I actually prefer the term *resource description* over the term *metadata*, but will use the term *metadata* in this paper.

[3] At least in the Netherlands CLARIN-NL project.

[4] Ďurčo and Windhouwer (2014) have developed a tool, the SMC-browser, to investigate the relations between profiles and components.

[5] For this reason, some people (including myself) do use closed data categories but postpone publishing them so that they can keep adapting it.

[6] As an alternative, *half-open* data categories have been proposed, i.e existing values cannot be modified but new ones can be added, but these have been rejected by the ISOCAT committee.

(because it has properties that hold at this collection level), but that is mostly not the case in the current set of metadata. One can classify a collection of fine-grained resources in the metadata as a collection, but that is just a property: a collection is not itself associated with properties and has no associated resource description (metadata record). For example, FESLI[7] consists of 55 resources, each with data from one session (and with one speaker). FESLI has been made for the purpose of research into specific language impairment, in combination with bilinguality. There is a metadata record for each of the sessions, and each of them is classified as belonging to the collection FESLI. But this collection has properties of its own, which are either not described at all, or repeated 55 times in the metadata record for each session:

- the description of the project (project name, organisation, location, project description, etc.) that created the original data consist of properties at the collection level. (it is currently repeated 55 times): e.g. the project descriptions: *BISLI is a project that aims to disentangle effects of bilingualism en specific language impairment in the domain of inflection*
- contact person, organisation etc is a property at the collection level
- usually a license is linked to the corpus as a whole
- documentation is usually a property of the corpus as a whole, not of individual sessions. It usually also describes design decisions, such as the rationale for having 55 sessions (rather than 5 or 550), a justification for the proportion between monolingual and bilingual speakers, a justification of the material used, etc. etc. Many properties related to this are usually also documented here:
  - the number of participants
  - the proportion of monolingual v. bilingual participants
  - the proportion of bilingual speakers with Dutch as first language v. with Turkish as first language

  True, these could be computed automatically from the session metadata, by not by the current VLO![8]
- many users will be interested in using the collection as a whole, not just some individual sessions: a user currently has to download 2*55 files to obtain the corpus, which is not really user-friendly!

It therefore makes sense to organize resources in a hierarchy (with explicit relations between resource descriptions in the same hierarchy), and make a resource description (metadata) for each distinguished resource. It may also make sense to present resources initially, or as the default option, at a less granular level, i.e., present initially only the metadata that

[7] In part on the basis of the description of the FESLI collection given here, the producers have adapted the metadata for this collection, so what is stated here is not fully true anymore of the FESLI collection.

[8] And since computers are slow, especially when they have to work across the net, it makes sense to compute such aggregates only once and then store them. They are usually static, so updates are seldomly needed.

describe large resources (e.g., the highest one in the hierarchy), and go to a smaller granular level only after initial selections have been made. The facets offered can adapt dynamically in function of the level of granularity. (See below under *dynamic facets*)

4 VLO

The VLO enables faceted search and string search. Faceted search is limited to a small number of facets, and therefore cannot make use of the fine-grained metadata descriptions. String search has its inherent limitations: it searches just for the strings entered, it searches in the whole record without taking context into account, etc. It should be offered but should be a fall-back option rather than being presented as the most prominent search option. In addition, the string search offered in the VLO is, even for string search, very primitive (e.g. no Boolean operators, no or very limited query expansion). Also it is not possible to search for strings inside a particular metadata element. An option to search specifically for the *title* or *name* of a resource is lacking, let alone that one can do fuzzy search for title or name (fuzzy search is required for this to be useful).

I will discuss the facets for search currently offered by the VLO in section 5, and will make some suggestions for new facets there as well.

5 Facets

The VLO offers the following facets for search: *language, subject, collection, format, resource type, organisation, continent, national project, country, keyword, modality, data provider, genre.* I will describe each of them in section 5.1.

The values for the facets of the VLO are derived from metadata elements according to the scheme in `https://lux17.mpi.nl/isocat/clarin/vlo/mapping/index.html` and is based on ISOCAT DCs, specific CMD paths, and exclusion of specific paths.

One can speak of *discovery of a resource* by a user if the user finds a resource of which the user did not know in advance that it existed. In my view, only a few of the facets currently offered are interesting for this purpose: *language, subject, format, resource type, keyword, modality, genre, continent.* The other facets are more interesting when a user knows about the existence of a resource, but tries to find it or its exact properties.

Values of facets are not normalized in any way. Even if two values differ just by capitalization, they appear separately in the list (e.g. *Text* and *text* are considered distinct values). There are many such small differences. In addition, the VLO allows a user to select only one value.[9] As a consequence, it is almost impossible to select the exact set of resources that a user is interested in.

I believe that, from a linguistic perspective, some facets are lacking and should be added.

I will first discuss the existing facets (section 5.1), and then the new facets proposed by me (section 5.2).

[9] That is odd since even garden variety software such as Excel offers multiple selections, so why does the major CLARIN metadata search interface not offer this functionality?

5.1 Existing Facets

Language Filtering by language gives a list of the 10 languages with most resource descriptions. Unfortunately, for Dutch not only *Dutch* occurs but also *nl_NL*, but one can select only one of these (not both at the same time).

Selecting "Dutch" yields (2014-08-21) 148,076 results[10]

Subject This facet offers a wild variety of things: how on earth can anyone find anything? There is a long list with just an alphabetic ordering:

- some are subject codes followed by a German description
- some are completely incomprehensible codes: e.g. *!181570343*
- some belong more to genre or resource type: 'language resources', 'lexical conceptual resource', 'multilingual lexicon'.
- some are multiple values in a single metadata element
- some are language names which have been stated twice in the OLAC record (under language and under subject)
- some clearly belong to language, e.g. certain olac-based records have things like

```
<subject olac-language="">Greek</subject>
```

I believe that a mapping to a small range of fields with closed vocabularies is possible, but did not try it for this facet yet.

Collection OK, but only useful if you already know that a specific collection exists. "Collection" in the VLO currently is just a value of an attribute of metadata descriptions. However, many "collections" are better described as resources in their own right, see section 3.

Format OK, mime type values, though some are not particularly useful (*application/octet-stream* not incorrect for .doc, .docx, .xls, and .xslx formats but not specific enough.

Is *text/plain-bas* a valid mime type, and what does it mean?

Multiple selection is highly desirable here.

Resource type is a complete mess: it mixes values for resource type and genre, is not normalized, has occasionally multiple values in one field (even as one word).

My suggestion: Make a mapping from values of this field to a combination of values of the following fields:

- resource type, restricted to values of DCMI Type or a small extension of it
- Genre

[10] By the way, if you now go back with the back button (that is what it is for!) you lost your language selection and have start all over again. This is a problem of almost every web application I have seen so it is all a mess, and the VLO is no exception to this.

- Subject
- Linguistic Annotation (to be added, closed vocabulary, see section 5.2)

Organisation Only useful if you know or suspect that a resources originates / is at a particular organisation

No standardized names, so it's a mess, e.g. the following are just a small selection of the different variants.[11]

- *Max Planck Gesellschaft* and *Max-Planck-Gesellschaft*
- *Max Planck Institut; Max Planck Institute,*
- *Max Planck Institute for Evolutionary Anthropology; Max-Planck-Institute for Evolutionary Anthropology; Max Planck Institute for Evolutionary Anthropology, Department of Linguistics*
- *Max Planck Insitute for Psycholinguistics; Max Planck Institute for Psycho-Linguistics; Max Planck Institute for Psycholinguistics; Max-Planck Institute for Psycholinguistics; Max-Planck-Institute for Psycholinguistics; Max-Planck-Institute-for Psycholinguistics; Max Planck Institute for Psycholinguistics, Nijmegen; Max Planck Institute for Psycholinguistics, Nijmegen Netherlands; Max Planck Institute for Psycholinguistics, Nijmegen, NL; Max Planck Institute for Psycholinguistics, Nijmegen, Nl*

and one cannot select multiple values. Work for normalizing this has been done in CLAVAS but is not used in the VLO (yet?).

Continent OK but it is of limited value. Some records have explicitly *unspecified* marked for this element, but many are not marked for continent at all. Though it often can be derived by a simple table from the *country* field.

National project OK but limited value. Of course, not all are marked as such. For example, the FESLI collection has not been marked for national project, so cannot be found in this way.

Country OK but limited value.

Keyword Probably useful, though of course the keywords are in multiple languages and not normalized:

- Biologie (1698)
- Biology (17)
- biology (3)

but we can select only one of these three at a time.

[11] All the problems one expects when one uses natural language for communication, which is hopeless (Odijk (1993))

Modality is a complete mess:

- there is *speech* and *spoken* (but you cannot select both)
- there[12] is *speech;gestures* ||*speech/gestures* ||*co-speech;gesture* ||*cospeech gestures* ||*speech, gestures* ||*speech, pointing gestures* ||*speech, pointing gestures, gestures* ||*gestures; speech* ||*gestures, speech* ||*speech, gesture* ||*speech and gestures* ||*speech gestures* ||*speech,pointing gestures* and *speech,pointing gestures,gestures* (but you can select only one at a time).

There is a data category in ISOCAT, part of Athens Core, called *modalities*, but

- it is an open category (i.e. any string is a valid value)
- its name (plural) suggests that multiple values are allowed, which is of course asking for disasters (as proven by the speech / gestures example), because these multiple values are not represented formally and cannot be recognized as such by software (such as the VLO)
- it provides *one example*[13], which represents multiple categories, reinforcing the suggestions that multiple categories are allowed as a single value: *Unknown; Unspecified; speech; writing; gestures; pointing-gestures; signs; eye-gaze; facial-expressions; emotional-state; haptic; song; instrumental music;* Actually, for many of these strings simple categories exist in ISOCAT, but they are not used here.
- it contains information pertaining to other properties than modality, such as *genre* and *subject*.

Clearly, this is a prototypical case where an attribute should be used (and be obligatory in CMDI resource descriptions) that allows multiple occurrences but selects it values from a closed (or half-open) list of possible values.

Yes, it is true that scientists will never agree on what the values for such an attribute should be, and yes it is true that nobody can list all possible values (simply because each of us has only a limited view of the world), and yes new values will be needed in the future that we cannot foresee now: but only having an open category attribute that also allows informally represented multiple values is giving so much flexibility that the attribute becomes completely useless. Every scientist will agree that *speech* and *spoken* in (5.1) refer to the same, and every scientist will agree that the wild variety in (B) would better be represented by two or three occurrences of a *modality* attribute with values *speech*, *gestures*, and *pointing gestures*[14]

Having a closed category is the best way to guarantee that only valid values are used. However, it is very likely that this closed category is too restricted very soon. It can be used therefore only when (1) it is easy to make a new version of the data category (marked as a new version) with additional values, and (2) official endorsement of this new version

[12] We use ||here to separate the different examples, since most interpunction symbols alrerady occur in the values.

[13] i.e. in terms of the ISOCAT structure for representing examples

[14] I assume that *co-speech* is not really different from *speech*, but just speech which is used in combination with some other modality. If that is not correct, we should add *co-speech* as a possible value for the attribute *modality*.

within CLARIN[15] is quickly decided upon, so that new users of the category will use the most recent version.[16]

If this is not feasible, a half-open category is desirable. It allows users to pick from existing values where these can be used, but allows them to add new values if none of the existing values is suited. Such half open lists of course have big disadvantages: one is dependent on the discipline of users to select an existing value whenever that is needed, and it remains to be seen whether that is the case. It will also have to be ensured that new values really contribute to a partitioning if the category: adding a new value that covers a subpart of an existing value should not be allowed (if that is necessary, other means should be used). This will require close monitoring by a metadata quality team[17]

I strongly recommend to semi-automatically clean up this attribute. I have added a possible mapping into a combination of attribute-value pairs (*modality, genre, subject*), some of which have a closed vocabulary, in appendix B

Data provider Distinguishes only CLARIN centres from other data providers. Not very interesting for discovery of resources.

Genre Also a complete mess:

- sometimes describes the genres of subparts of a collection
- sometimes multiple values
- sometimes more appropriate to the title/name field or the *subject* field

A mapping to a more normalized set of values, possibly distributed over multiple fields is desirable, but not done by me here yet.

A small hierarchical taxonomy is desirable, e.g. putting all different kinds of songs under *song*.

5.2 New Facets

For successful linguistic searches, facets for *linguistic annotation*, *period covered* and *language variety* information are crucial. I discuss each of them here.

Linguistic Annotation Metadata should describe formally what kind of annotations the data it describes has. Though this is true for annotation in general, I will limit myself here to *linguistic annotation.* A initial proposal for values of such a metadata element, in a small taxonomy, has been included in appendix A.

[15] Within CLARIN is enough, endorsement by official standardisation organisations is not needed, actually irrelevant, and will take too much time anyway.

[16] Of course, the old version will continue to exist, but be marked as obsolete, with a reference to the newer version.

[17] I do not believe that there are formal objections to half-open lists. As far as I can see, the semantics of the individual values remain constant, as well as the semantics of the half-open category itself (which has the meaning: the category is partitoned in the values distinguished plus an unknown and unclassified other part).

Period The existing TimeCoverage component probably will do, provided that dates need not be precise to the day (by century, decade, year, or by month should also be allowed), and not both the begin date and the end date need be present. Maybe some extensions are needed, because sometimes we can characterize a period only negatively (e.g. *not before/after some date*).

LanguageVarietyInfo The existing LanguageVarietyInfo component might suffice, though the value 'standard language' should be added there for the *languageVarietyType* metadata element.

6 Test Cases

I describe a number of examples of search for resources. They have a strong linguistic bias, so it would be interesting to see how researchers from other disciplines experience the VLO.

6.1 Searching by language

I have tested a search for the value *Dutch*. As described above, for Dutch not only *Dutch* but also *nl_NL* occurs as a value, and one cannot select both. Apart from that, the results for *Dutch* appear reasonable. Most data sets of which I know that they are visible via the VLO and contain Dutch are in the result set.[18] Some are there more by accident than by design: e.g. the FESLI records are not marked for language. 31 of them occur here for reasons that I do not understand at all.[19] Some records are missing, e.g. the PILNAR data have not been marked for language at all! And the same is true for the Academia Collection provided by the Netherlands Institute for Sound & Vision. There are several data for which no records are available in the VLO at all (e.g. GrNe (Classical Greek-Dutch dictionary), Corpus of Modern Dutch).

An indication of the resources that have not been marked for language at all would be very useful.

6.2 Searching by resource type and linguistic annotation

There is no facet for *linguistic annotation*. Searches for

- text corpora that are Pos-tagged
- text corpora that are enriched with discourse annotation
- text corpora in which each sentence is assigned a syntactic structure (treebanks)

[18] DuELME, Cornetto, Corpus Gysseling, FESLI, DiDDD, Dynasand, GTRP, LESLLA, Discan, NEHOL, UBU data (some however with code *nl_NL*), VU-DNC, the resources accessible via the Integrated Language Database (GTB).

[19] since apparently language values for the metadata element *FirstLanguage* (of a speaker) ends up in this facet. The metadata element *SecondLanguage*, however, does not (e.g. one will not find any FESLI records under *Turkish* via the language facet, even though several records contain this information. However, I do not find any FESLI records under Turkish, though I believe that some speakers have Turkish as their first language.

- lexicons with phonetic transcriptions
- speech corpora with orthographic transcriptions

are therefore not easy. One is obliged to use string search.

Text corpora that are Pos-tagged One has to use string search: *pos tag* yields 15 results, *part of speech tag* yields 9 (mostly different) results, *part of speech* yields 12999 results![20] An OR-operator is not available for string search: One cannot use a query such as *pos tag OR part of speech* to get both results together.

Text corpora that have discourse annotation Searching for *discourse annotation* yields zero results. Searching for discourse yields 86467 results, but these are mainly results where the word *discourse* (which, as any decent natural language word, is highly ambiguous) occurs in the metadata. It is not clear that there are any resources with discourse *annotation*. Unfortunately, the metadata for resources from a collection that are annotated for discourse properties (TLA:Discan) do not describe that the data have these annotations at all, so they are not among the ones found by the string "discourse"![21]

Text corpora in which each sentence is assigned a syntactic structure (treebanks) Searching for *syntactic structure* yields 34 results, only 2 of which actually are text corpora with utterances annotated with syntactic structures. Searching for treebank yields 135 results, most if which indeed appear to be treebanks. However since the search is just for metadata containing the word *treebank*, the search results also include other types of resources such as a parser trained on the basis of a treebank and a valency lexicon extracted from a treebank. Of course one could use the facet *resource type* to select the desired resources, but here one has to select from *treebank*, *corpus*, *Corpus*, *text* or *Text*. A resource is usually called a *treebank* when the annotation with syntactic structures is the only or main type of annotation: a resource such as the Spoken Dutch Corpus (CGN), which contains all kinds of annotation including syntactic structures for each utterance in a subcorpus is usually not called a treebank. It is therefore not found when searching with the string *treebank*.

Lexicons with phonetic transcriptions Searching for phonetic transcription yields 1571 results. None of the found resources is classified as a lexicon. Searching for phonetic transcriptions yields 43 results, and contains the complete CELEX lexical database and the phonology subset of CELEX but also all other subsets (on orthography, morphology, frequency and syntax) because their descriptions contain the string *phonetic transcriptions*! The complete CELEX lexical database and all its subsets are classified as *resource type = Text*, which is strictly spoken perhaps not false (depending on the interpretation of *resource type* and the value *Text*) but also not very informative, and shows that the interpretation of attributes and their possible values must be defined

[20] Most of these from the Dutch CGN corpus, due to the fine granularity this corpus has been described with.

[21] Minor detail: when selecting one found result, the web page states *Record 1 of 1.215* in which a period is used to separate thousands in the results count. This seems odd for an otherwise English interface: one would expect a comma instead.

clearly (which is only possible with a closed or half-open vocabulary). Searching for *phonetic lexicon* yields 24 results, all of which are classified as *resource type = Sound.* Selecting *resource type = Sound* from the search with *phonetic transcription* yields 8 results, exactly one of which is a phonetic lexicon. Searching for pronunciation lexicon yields 148 results, some of which are a pronunciation lexicon, but most of which are acoustic databases containing a pronunciation lexicon.

speech corpora with orthographic transcriptions Searching for the string *transcriptions* and *resource type = Spoken Corpus* yields 12 results, but then one misses the 256 ones with *Resource Type = sound*, the 20 ones with *Resource Type = corpus*[22], and the single one with *Resource Type = SpeechCorpus*, and the 15 ones with *Resource Type = Dataset*, and surely more.

Searching first for the Language Facet before selecting a resource type reduces this problem significantly in most cases, but of course, that should not be necessary

Summarizing: for all searches illustrated here, it was impossible or very difficult to select exactly the resources that one is interested in.

6.3 Searching by subject

data relevant to language acquisition Searching for language acquisition yields 7257 resources and appear to include all relevant resources. The collections I know are all present in the result set. The results also include texts about language acquisition, and it is not so easy to separate these from real language acquisition data.

data relevant to specific language impairment Searching for specific language impairment yields 186 results, all of which appear relevant. As with *language acquisition*, the results also include texts about specific language impairment.

data relevant to syntax Searching for syntax yields 5491 results, as can be expected from such a single term that is often used. Restricting the selection to language = Dutch reduces the results to 1342. Of course, all CELEX subsets are included here. However, an important resource such as the Alpino Treebank is not included, because it happens not to contain the word *syntax* (but the phrase *syntactically analyzed Dutch sentences*).

I conclude that string search for subjects can yield sensible results if the terminology used for a specific subject is pretty uniform and consists of multiple words. With short terms, the results are less usable, and additional query strings for related words (e.g. *syntactic*, *syntactically* for *syntax*) are needed but cannot be added because no OR operator is allowed.

7 Summary of Problems

Finding data by means of the Virtual Language Observatory, especially data of which it is unknown whether they exist, is currently very difficult and in practice in most cases impossible. I summarize the reasons for this:

[22] of which, unfortunately, some are transcriptions of images.

1. Several crucial metadata elements are not obligatory.
2. Several important metadata elements that are used in the facets do not have values from a closed vocabulary.
3. The metadata are made too much in isolation so that there are often unnecessary differences.
4. Often crucial information is lacking, especially properties that 'obvious' to the researcher/data provider are often not described at all in the metadata (e.g. language, annotation, time period)
5. The number of attributes covered by the VLO faceted search is small, some important ones are lacking (e.g. annotation).
6. One is never sure when a search result has been obtained, whether it really covers all relevant data. This is caused by the optionality of most metadata elements. The VLO does not yield a list of records that are NOT marked at all for a facet searched for, but it should do so, at least as one of the options.[23]
7. The granularity of the metadata records varies wildly. In many cases it is too small for many searches. In combination with the limited number of facets, the small granularity is totally useless, since one cannot search for the finer distinctions. Collections have been introduced but so far never occur as a result of a search action (only the individual metadata records of the collection do).
8. There are hardly any metadata for software, so finding software via the VLO is almost impossible.

In the next section I will make some recommendations to address these problems.

8 Recommendations

Based on the inspection of the VLO, the metadata it operates on, and the search experiments, many problems were found. Here I make the the following recommendations to address these problems:

Versioning of DCs make special provisions for quickly and easily creating a new version of a Data Category in the supported data category and concept registries (ISOCAT, CLAVAS).

Versioning of Metadata Profiles and Components make special provisions for quickly and easily creating a new version of a metadata profile or component, and for marking existing ones as superseded by new versions

Versioning of Metadata make special provisions for quickly and easily creating a new version of a metadata record, and for marking existing ones as superseded by new versions

[23] This will also be very useful for the Metadata Quality Team.

Granularity Organize resources in a hierarchy (with explicit relations between resource descriptions in the same hierarchy), and make a resource description (metadata) for each distinguished resource.

Granularity & Presentation Present resources initially, or as the default option, at a less granular level, i.e. present initially only the metadata that describe large resources (e.g. the highest one in the hierarchy), and go to a smaller granular level only after initial selections have been made.

Enable selection of multiple metadata in the VLO.

New facets Facets must be added for *linguistic annotation*, *period*, perhaps more

"My Virtual Collections" Add the possibility to select resources and put them in a user-specific repository ("My Virtual Collections")

Obligatory Metadata Elements make certain metadata elements obligatory, in particular *title, name, version, language, annotation, resource type, modality, time period, genre, subject.* Of these, *language, annotation, resource type, modality, genre* must have closed (or half-open) vocabularies, and time period must be a highly constrained value. They may allow multiple values, provided they are formally marked as such.

Formally mark multiple values and NEVER allow multiple values in a single field in any metadata

Dynamic facets Carry out an experiment with dynamic facets: i.e. more facets appear when a subset has been selected with more shared attributes. Initially, this can be done per collection. For example, if you select the FESLI collection, facets become available for *ParticipantCode, ParticipantGender, LanguageImpairmentStatus, FirstLanguage, SecondLanguage, Bilingual, Age*, and *Task*, since all or most metadata records in this collection have these attributes, and these attributes are relevant for searching in this collection. Then use can be made of the fine granularity.

Hierarchy in values Consider making small taxonomies in values, e.g. in *Genre* many different kinds of songs occur: hiding them under the hierarchically higher value *song* and making them visible only when the researcher is interested in the subclasses will make every user happy. Currently almost every user is bothered by the long list of values, most of which are irrelevant to this user.

Clean up and Normalize We want to improve the metadata information. We cannot change the metadata records provided by researchers or data providers ourselves. We do not want to make ourselves dependent on improvements made by these researchers and data providers (because they might never do it). So, what we do is add information the metadata records: for each metadata record, add a record with the attributes used for faceted search. Copy the values from the actual metadata records to the attributes of these records. Do this in a smart way: map currently occurring values to cleaned values, possibly in multiple attributes, and map multiple values in a single field to single values in multiple fields. Set up a Metadata Quality Team for regularly monitoring newly added metadata records for these aspects, and if needed, for manually adapting these records.[24] Use these cleaned facet records for the actual search. Suggest metadata providers to make updated versions of their metadata based on the improvements made by this team. This has several advantages:

- The Metadata Quality Team/software does not modify metadata records provided by others but just adds new information to them
- This additional information can be used for the VLO without being dependent on the metadata record providers: they might not be willing to change, not being able to do it (now), etc.
- A concrete improvement proposal is made to the metadata providers. If the metadata provider agrees to the changes but for some reason cannot make them (e.g. because of lack of technical skills, lack of human resources, etc.), the Metadata Quality Team can make the changes for the metadata provider, if he/she agrees.
- Searching with the VLO will become better and easier

The Metadata Quality Team must ensure semi-automatic and automatic cleanups for (initially) the facet fields where a closed vocabulary is desired:

- apply exceptional string replacements for known problematic cases
- split up a field by separators (;/: etc)
- map known values to standardized values (can be detected automatically)
- regularly adapt for new incoming metadata

See appendix B for a concrete example. In these examples, it is assumed that the information provided is correct but notated in a sloppy manner. If the actual information is incorrect, or completely lacking, more effort will needed to upgrade the metadata.

9 Conclusions

Discovery of research resources (data and software) relevant to a researcher is an important functionality that CLARIN aims to offer. Unfortunately, finding resources by searching in their metadata through the VLO is not as easy as it should be. I have illustrated this in this paper, by discussing the nature of the metadata and the limited capabilities of the VLO, and by carrying out a variety of test queries focusing on searching data that are interesting for a linguist. I have analyzed and summarized the problems, and have made many recommendations to improve the situation.

Acknowledgments

This work was financed by CLARIN-NL.

[24] I.e. writing functions that correct wrong input so that the corrections are re-installed each time a new harvest has been carried out with the same messy data.

References

[Broeder *et al.*, 2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework. In N. Calzolari, B. Maegaard, J. Mariani, J. Odijk, K. Choukri, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 43–47, Valetta, Malta, 2010. European Language Resources Association (ELRA).

[Odijk, 1993] Jan Odijk. Natural language, communication, and man-machine interfaces. November 1993.

[Van Uytvanck, 2014] Dieter Van Uytvanck. How can I find resources using CLARIN? Presentation held at the *Using CLARIN for Digital Research* tutorial workshop at the *2014 Digital Humanities Conference*, Lausanne, Switzerland. `https://www.clarin.eu/sites/default/files/CLARIN-dvu-dh2014_VLO.pdf`, July 2014.

[Ďurčo and Windhouwer, 2014] Matej Ďurčo and Menzo Windhouwer. The CMD cloud. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 687–690, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

A Linguistic Annotation

Initial version of a list of possible values for *linguistic annotation*. Maybe a slightly deeper taxonomy is desirable, e.g. putting *synonymy*, *hyponomy / hyperonomy* etc under l*exico-semantic relations*.

Linguistic Level	Type of Annotation
orthography	orthographic transcription
orthography	normalized orthographic transcription
orthography	hyphenation
orthography	contextual conditions
orthography	frequency
orthography	transliteration
orthography	word separators
orthography	sentence separators
orthography	paragraph separators
phonology/phonetics	phonetic transcription
phonology/phonetics	tone marking
phonology/phonetics	stress marking
phonology/phonetics	prosodic marking
phonology/phonetics	alternative phonetic transcriptions
morphology	compound boundaries
morphology	derivational affix boundaries
morphology	Pos-tags
morphology	inflection class
morphology	lemma
morphology	stem
morphosyntax	inflection
syntax	Disambiguated Pos-tag
syntax	Syntactic structures
syntax	grammatical relations
syntax	dependencies
syntax	phrases
syntax	multi-word expressions
syntax	chunks
Semantics	sense
Semantics	synonyms
Semantics	hyponyms/ hyperonyms
Semantics	meronyms
Semantics	other lexico-semantic relations
Semantics	sense numbers
Semantics	co-reference
Semantics	entailment
Semantics	Topic/Comment/Focus marking
Translation	translation
Discourse	co-reference
Discourse	overlapping reference
Discourse	discourse relations
Discourse	textual units

B Possible Cleanup of Modality

See next page.

The multiple values of the *modality* attribute have been split up into single values. Each single value is mapped to a combination of 3 attributes: *modality*, *genre* and *subject*. Modality and Genre will have a closed vocabulary as their possible values.