November 2014

Metadata quality assurance for CLARIN

Marc Kemps-Snijders

[Type the abstract of the document here. The abstract is typically a short summary of the contents of the document.]

Table of Contents

Introduction

Bruce and Hillman provocatively opened their article ‘THE CONTINUUM OF METADATA QUALITY: DEFINING, EXPRESSING, EXPLOITING’ by stating ‘Like pornography, metadata quality is difficult to define.’ Indeed quality has many dimensions, assessing the quality across all possible dimensions may become a difficult and cumbersome task. However, limiting the number of quality dimensions and selecting dimensions suitable for automated processing quality assurance and assessment procedures may provide useful tools to deliver metadata records with a high level of conformance to specifications.

This document contains an overview of experiences with metadata quality by Dutch CLARIN centers and highlights some of the most problematic areas in CLARIN metadata creation process. In particular knowledge exchange and early feedback on the metadata quality are of interest here. Based on a literature review it proposes additional quality assurance steps to be incorporated as part of the metadata production process. It also proposed a number of quality metrics across a number of accepted dimensions that can be gathered as part of the metadata creation process. This not only helps to determine quality of metadata records eventually listed in CLARIN’s Virtual Language Observatory, but also provides indicators that can be used during the creation process itself and help to decide whether a set of metadata records should continue to the next phase of the production process.

Quality and the metadata life cycle

Three types of metadata are usually distinguished:

- Descriptive metadata, describes a resource for purposes of discovery and identification
- Structural metadata, describes the structural and relational aspects of the resource
- Administrative metadata, provides information to help manage a resource

The metadata life cycle handles the process of creating, maintaining, updating, storing and publishing metadata as well as handling deletions. Although metadata and data life cycles are strongly interdependent, the metadata life cycle may extend that of the resource. Sometimes metadata is created before a resource becomes available, e.g. to indicate its future availability, and metadata may remain available even after the resource has been removed. Several metadata life cycle models have been proposed to assist digital repositories in defining their (meta)data management processes, e.g. OAI, DDC DDI-I. As part of the Dasish project [DASISH 2014] an alternative extended life cycle has been proposed based closely on familiar life cycle models to support more dynamic metadata issues. These life cycle models also serve as a basis to assess the quality of these business processes. In the CLARIN community the Data Seal of Approval(DSA), taking the OAI model as a starting point, is a minimal requirement for organizations aspiring to become CLARIN centers. The DSA evaluation focuses on assessing the quality of general digital preservation procedures to be able to designate a repository as a Trusted Digital Repository. While assessment of these business processes is an important step towards overall quality assessment they are by itself not enough to ensure that the outcome of these processes, i.e. metadata and data, are sufficient for the community. Issues of semantic interoperability and quality assessment methods for individual metadata records are simply not within the scope of digital repository assessments.

To address the issues of semantic interoperability CLARIN poses additional requirements such as use of ISOcat's Data Category Registry for (metadata) concept definitions and use of the Component Registry for specifying metadata profiles/schemas. There appears to be a growing awareness outside of the CLARIN community that these could also be of great relevance to other research infrastructures. From the afore mentioned Dasish report:

- *One recommendation would be that the three infrastructures could agree to define a common list of metadata elements that - crossing the different communities and standards – can be used as compatible between the different communities.*
- *Furthermore, easily accessible definitions of these elements and mappings across the different metadata standards should be available*

Explicit semantics provide great benefits for automated processes, such as indexing of metadata records [Zhang 2012]. But use of the Data Category Registry and the Component Registry within CLARIN has certainly not been unproblematic. Concerns over the proliferation of both data categories and components have led to development of additional systems such as RelCat(α-version only), a relation registry, and CLAVAS, a vocabulary service. These are technological answers to some of the problems that have appeared in this area but do not provide a solution on *prevention* of proliferation and other quality related issues, such as poor or ambiguous definitions. It is worthwhile considering how these problems may be addressed before they appear in the infrastructure, rather than providing solutions to amend these afterwards.

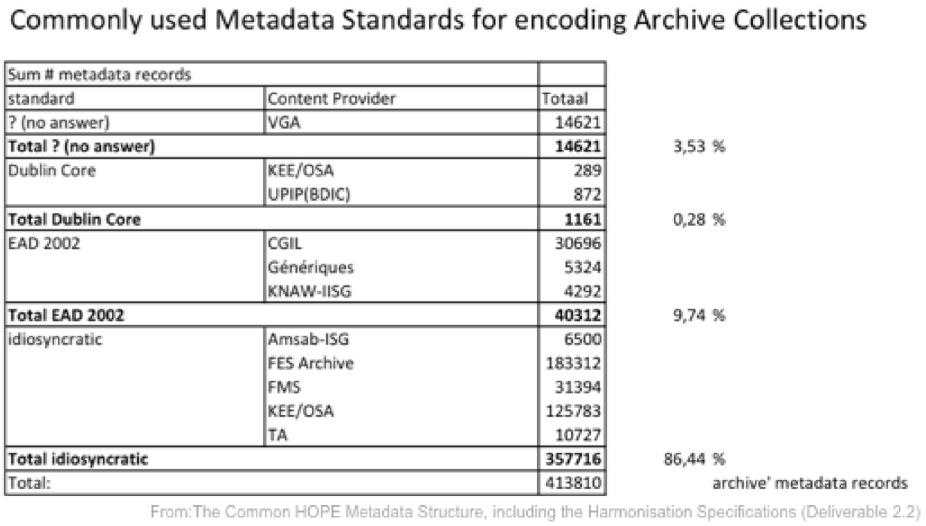
Even with a semantic interoperability issues with metadata quality still remain. Questions about accuracy, completeness, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility of metadata records can only be answered by looking at the individual record level or by placing the record in its repository context. The quality of a metadata record is not only determined by the amount of information (number of elements) in a record, but also by how easy it is to find that record in a repository through its distinctive features. Manual inspection of randomly selected record samples is currently the only way to achieve this.

The remainder of this document will address some of the common quality issues encountered during the metadata creation process and proposed a number of measures to improve the process. Also a set of quantitative metrics, in combination with structure human evaluation, is proposed to gain a better insight in the overall metadata quality.

Metadata quality issues in CLARIN

From its early beginnings, CLARIN has employed a number of strategies that facilitate use and reuse of metadata and metadata schemas (fragments) and semantic interoperability across metadata schemas. While these are considered to be preconditions to improve the overall metadata quality experience have shown that proliferation of both the Data Category Registry and the Component Registry is a rising concern.

The design of the CMDI framework acknowledges that metadata for research data often requires use of custom metadata schemas rather than reusing existing standard metadata schema such Dublin Core, ESD or METS. Results from other projects support these findings. For example, the HOPE project found that over 86% of metadata schemas in use were reported to be idiosyncratic in nature. Also, it is recognized that the profiles registered in CLARIN’s Component Registry find low levels of reuse.



Commonly used Metadata Standards for encoding Archive Collections

Sum # metadata records			
standard	Content Provider	Totaal	
? (no answer)	VGA	14621	
Total ? (no answer)		**14621**	3,53 %
Dublin Core	KEE/OSA	289	
	UPIP(BDIC)	872	
Total Dublin Core		**1161**	0,28 %
EAD 2002	CGIL	30696	
	Génériques	5324	
	KNAW-IISG	4292	
Total EAD 2002		**40312**	9,74 %
idiosyncratic	Amsab-ISG	6500	
	FES Archive	183312	
	FMS	31394	
	KEE/OSA	125783	
	TA	10727	
Total idiosyncratic		**357716**	86,44 %
Total:		413810	archive' metadata records

From:The Common HOPE Metadata Structure, including the Harmonisation Specifications (Deliverable 2.2)

Data Category Registry

Broeder et all[Broeder 2014] have identified several problems related to the use of data categories in CLARIN. These relate to the standardization procedures associated with the current ISOcat implementation, problems encountered in the data model itself and usability of the ISOcat tool itself. Although any effort in raising the quality of the of metadata specifications within the CLARIN domain should also address the standardization and tool usability issues these are largely ignored in this document[1].

Data model related problems are reported concerning several aspects: proliferation due to type, distinction between data category types (open, closed, constrained, container) and the demands for a rare blend of expertise combining linguistic and technical expertise. In addition, the same authors have also expressed concerns related to the quality of data categories at other occasions: (huge differences in quality, ambiguous definitions), number of data categories(proliferation) and semantic consistency of data categories[Schuurman 2013].

[1] The current ISOcat implementation will be discontinued at the end of this year and will be replaced by an OpenSKOS implementation. ISO TC 37 and CLARIN will be decoupled removing some of the standardization working procedure issues reported earlier. Also, any new implementation may also address usability issues currently associated with ISOcat.

Some of these concerns may be alleviated once a new Data Category Registry is in place(see footnote at bottom of page) such as the distinction between data category types . Other problems may be lessened if the process of registering profiles and data categories are more closely monitored and evaluation takes place as standard practice during the initial metadata creation process. Currently, data categories are submitted to the national Data Category Registry coordinator as one of the final deliverables of CLARIN-NL projects or are not submitted to the national coordinator at all. While registration of data categories is mandatory when submitting the final results of CLARIN projects in the Netherlands submission and review by the national DCR coordinator is not. One of the problems related to this is described as: “Users are to go over lots of existing definitions to check to see whether these are reusable → time consuming/boring” By involving the national DCR coordinator at much earlier stages of the process he/she might be able to provide useful alternatives and check the quality of definitions and provide suggestions for improvements of the data category specification.

Actual usage of data categories within CLARIN metadata may be monitored through the Component Registry. It is recommended practice in CLARIN that all metadata elements contain references to these data categories. This also separates metadata categories used within the CLARIN by those used within other communities. An easy to use tool for evaluating the contents of the Component Registry is the SMC browser[2].

Using the *reports* option of the SMC browser it is fairly simple to create a list of the most heavily used data categories in the Component Registry. The table below list the top most data categories referenced by elements in the Component Registry[3].

key	Definition	used in Profiles	referenced by Elements
description [isocat:DC-2520]	String	130	2335
url [isocat:DC-2546]	string	78	671
address [isocat:DC-2505]	address [isocat:DC-2505]	address [isocat:DC-2505]	address [isocat:DC-2505]
email [isocat:DC-2521]	string	114	488

[2] http://clarin.oeaw.ac.at/exist/apps/smc-browser/index.html

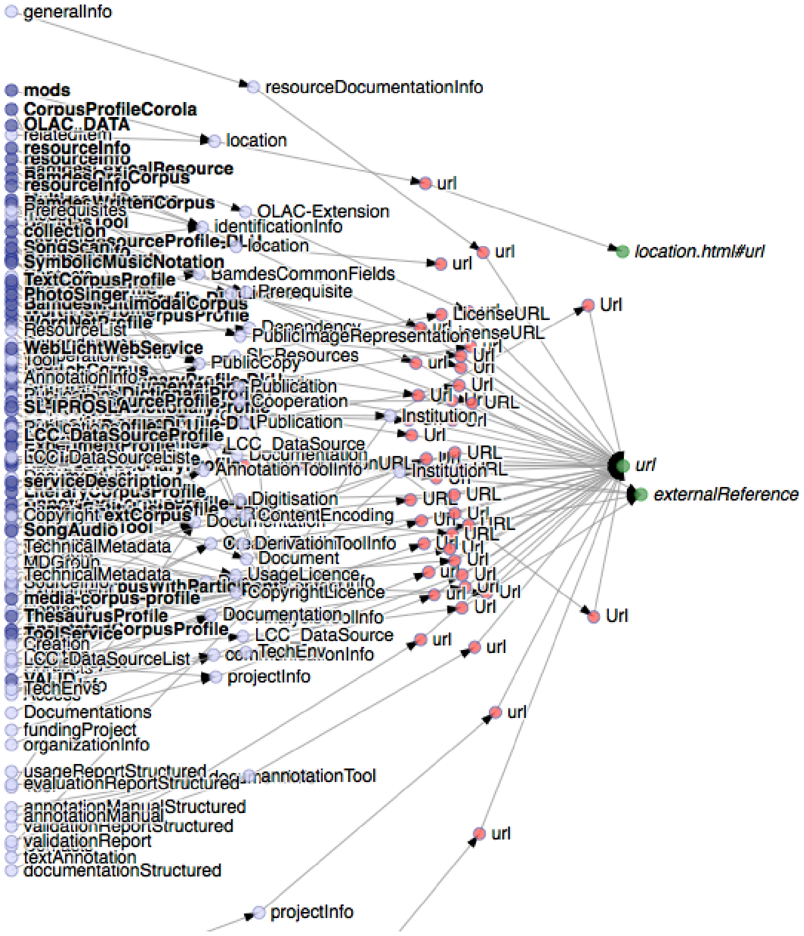
[3] http://clarin.oeaw.ac.at/exist/apps/smc-browser/data/smc_stats_datcat1.html

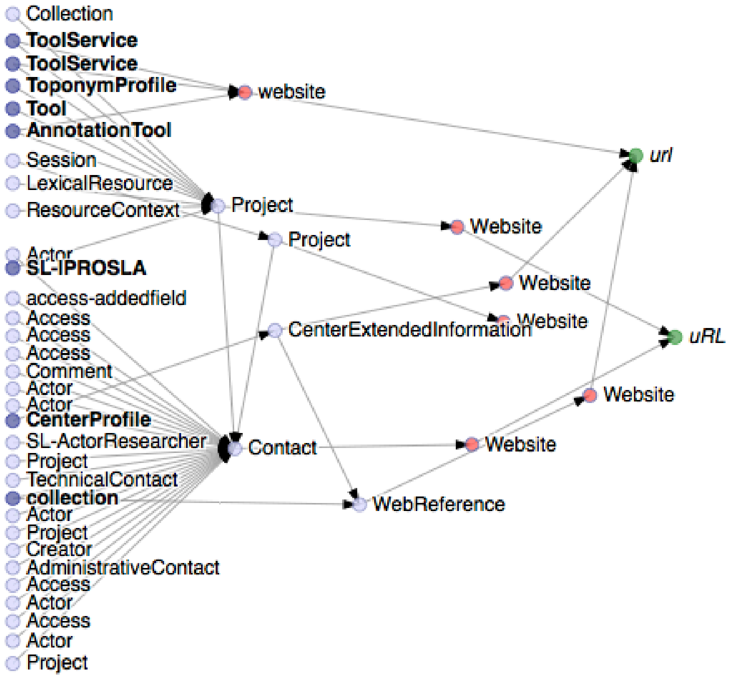
telephoneNumber [isocat:DC-2461]	string	102	467
Organisation [isocat:DC-2979]	the name of an organisation	98	445
size [isocat:DC-2580]	string	100	373
languageID [isocat:DC-2482]	string XML Schema regular expression [a-z]{3}	118	369
Person [isocat:DC-2978]	the name of a person	97	352

By far the most widely used data category is description, both in terms of the number of elements that refer to it as well as the number of profiles making use of it. The (full) list of data categories also contains several examples of data categories that are also found to have alternative representations. These represent excellent candidates for attempting to harmonize the definition of these data categories in the Data Category Registry.

The *description* data category has been defined several times, although name and definition may vary slightly. While data categories, such as *description* [isocat:DC-2520], *Description* [dct:description], *Description* [dce:description], [*http://www.loc.gov/standards/mods/userguide/name.html#description*],[*http://www.loc.gov/standards/mods/userguide/physicaldescription.html*], *msDesc* [isocat:DC-6211], *physDesc* [isocat:DC-6246], *typeDesc* [isocat:DC-6247]

Data categories with ‘url’ references are encountered as *url* , *uRL* or *[http://www.loc.gov/standards/mods/userguide/location.html#url]* in de Data Category Registry. Also, *externalReference* is used as an alternative data category in metadata elements (46 elements found). It is noted that in the figure shown below the *aURL* data category is not referenced by any metadata fields containing ‘url’, but appears to be used exclusively to indicate website references.


generalInfo
resourceDocumentationInfo
mods
CorpusProfileCorola
OLAC_DATA
relatedItem
location
resourceInfo
resourceInfo
BamdesLexicalResource
BamdesOralCorpus
resourceInfo
url
BamdesWrittenCorpus
OLAC-Extension
Prerequisites
identificationInfo
collection
SongScan
location
SymbolicMusicNotation
url
url
location.html#url
TextCorpusProfile
BamdesCommonFields
PhotoSinger
Prerequisite
BamdesMultimodalCorpus
LicenseURL
Url
WordNetProfile
ResourceList
Dependency
PublicImageRepresentation
LicenseURL
WebLichtWebService
SL-Resources
PublicCopy
url
Url
Url
AnnotationInfo
Publication
Url
Cooperation
URL
SL-IPROSLA
Institution
Url
Publication
LCC_DataSourceProfile
Url
LCC_DataSource
LCC_DataSourceList
Documentation
URL
AnnotationToolInfo
Institution
URL
url
serviceDescription
externalReference
LiteraryCorpusProfile
Digitisation
URL
URL
Copyright
extCorpus
ContentEncoding
Url
URL
SongAudio
Tool
DerivationToolInfo
Url
URL
TechnicalMetadata
MDGroup
Document
Url
TechnicalMetadata
UsageLicence
Url
Url
media-corpus-profile
CopyrightLicence
url
Url
Url
ThesaurusProfile
Documentation
ToolService
CorpusProfile
Url
Creation
LCC_DataSource
url
LCC_DataSourceList
TechEnv
url
VALID
projectInfo
TechEnvs
Documentations
url
fundingProject
organizationInfo
usageReportStructured
annotationTool
evaluationReportStructured
annotationManualStructured
annotationManual
validationReportStructured
url
validationReport
textAnnotation
documentationStructured
projectInfo



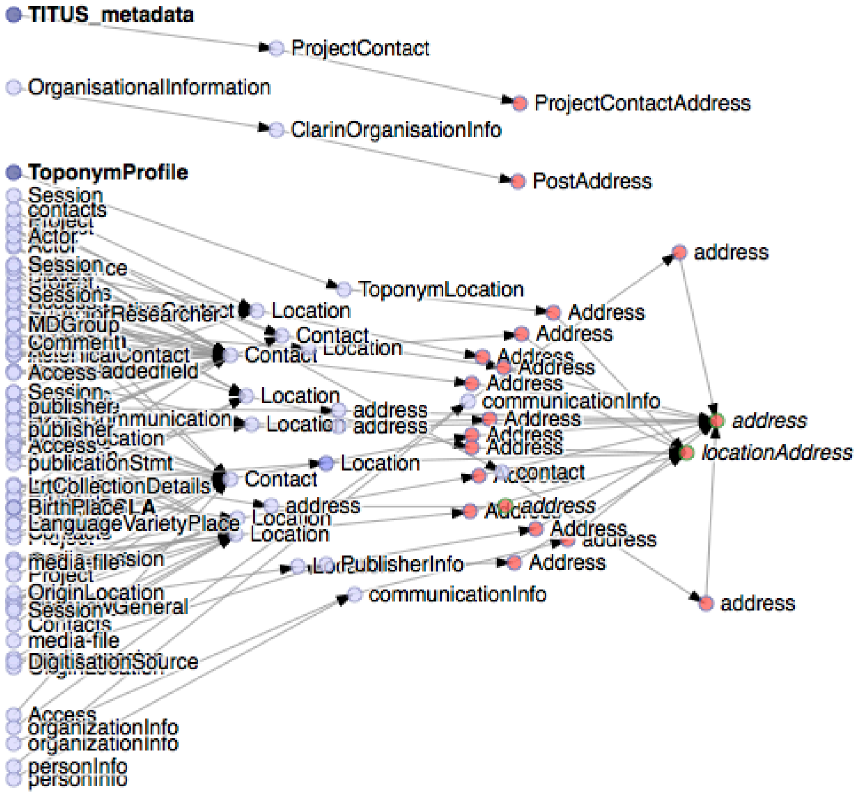
The table below shows the different variants for address.

Data category specification for address concepts

Key	Definition	used in Profiles	referenced by Elements
Address [isocat:DC-2505]	address [isocat:DC-2505]	address [isocat:DC-2505]	address [isocat:DC-2505]
locationAddress [isocat:DC-2528]	String	52	60
Address [isocat:DC-6207]	contains a postal address, for example of a publisher, an organization, or an individual.	1	1
addrLine [isocat:DC-6208]	contains one line of a postal address.	1	1

Street [isocat:DC-6209]	contains a full street address including any name or number identifying a building as well as the name of the street or route on which it is located.	1	1

Use of address in data categories and elements



Information gathered through the SMC browser may thus provide useful feedback for attempting to harmonize the use of similar data categories in the CLARIN metadata domain. A simple strategy towards harmonization would be to define appropriate relations, for example '*SameAs*', in the Relation Registry. A more future proof approach would be to attempt to reduce the number of similar data categories in the Data Category Registry. This requires a coordinated effort of the CLARIN community to harmonize these definitions in the Data category Registry, phase out obsolete data categories and modify the existing element links in the Component Registry. Care must be taken here as software modules developed by participating CLARIN centers may be affected. Hence, not only the metadata creation process must be evaluated by each data provider in the community, but also related software modules.

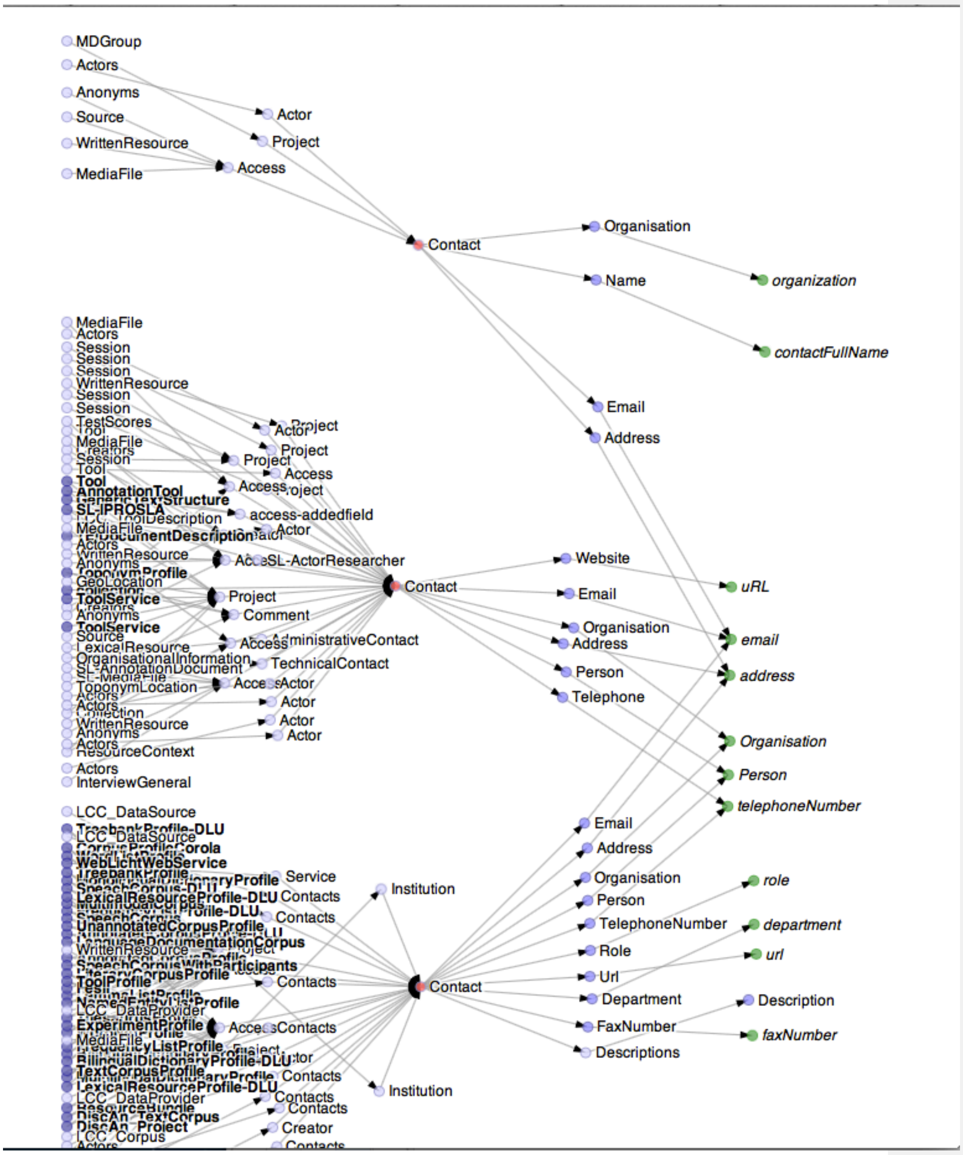
Component Registry

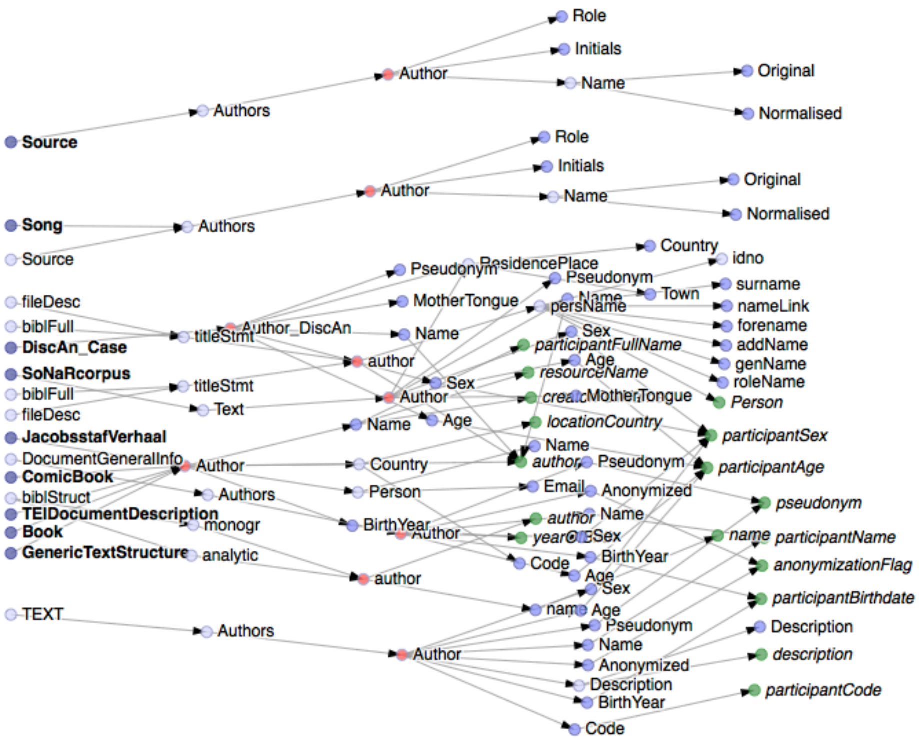
Use of the Component Registry is considered mandatory in CLARIN. Metadata profiles are commonly[4] designed in the Component Registry.

Qualitative evaluation of the Component Registry through the SMC browser[5] shows that some of the most frequently used components, such as *Contact*, are defined multiple times, with a large overlap in data categories. These components appear to be excellent candidates for harmonization. Harmonization efforts will need to take place at the European level.

[4] Some examples may be found in the Virtual Language Observatory where the schema is located outside of the Component Registry, e.g. http://media.dwds.de/dta/media/schema/cmdi-header.xsd . These can be easily retrieved by evaluating the xsi:schemalocation attribute of CMDI records.

[5] http://clarin.oeaw.ac.at/exist/apps/smc-browser/index.html


MDGroup
Actors
Anonyms
Source
WrittenResource
MediaFile
Actor
Project
Access
Contact
Organisation
Name
organization
contactFullName
Email
Address
MediaFile
Actors
Session
WrittenResource
TestScores
Tool
AnnotationTool
SL-IPROSLA
ToponymProfile
GeoLocation
ToolService
LexicalResource
OrganisationalInformation
SL-AnnotationDocument
ToponymLocation
ResourceContext
InterviewGeneral
Project
Actor
Access
access-addedfield
SL-ActorResearcher
Comment
AdministrativeContact
TechnicalContact
Website
Email
Organisation
Address
Person
Telephone
uRL
email
address
Organisation
Person
telephoneNumber
LCC_DataSource
TreebankProfile-DLU
CorpusProfileCorola
WebLichtWebService
LexicalResourceProfile-DLU
UnannotatedCorpusProfile
LanguageDocumentationCorpus
SpeechCorpusWithParticipants
ToolProfile
ExperimentProfile
FrequencyListProfile
BilingualDictionaryProfile-DLU
TextCorpusProfile
LexicalResourceProfile-DLU
LCC_DataProvider
ResourceBundle
DiscAn_TextCorpus
DiscAn_Project
LCC_Corpus
Service
Contacts
Institution
Creator
Email
Address
Organisation
Person
TelephoneNumber
Role
Url
Department
FaxNumber
Descriptions
role
department
url
Description
faxNumber



Current center experiences

A series of interviews were conducted to get feedback from the Dutch CLARIN centers on their experiences with CMDI metadata and curation processes and their stance towards metadata quality.

All centers indicate that primary ownership of the metadata records lies with the researchers who are involved in a project. From this, researchers are also considered to be primarily responsible for the resulting quality of the metadata records. Particularly for older records this presents a problem as researchers tend to move on to other projects or even organizations and it no longer becomes possible to involve the original owners. Efforts to curate these older resources therefore prove problematic, as they often require background knowledge (and decisions) by the original owner). Lack of funding and rising curation costs are also mentioned as obstacles for raising the overall metadata quality level. The costs of correcting or extending rise quickly if modifications of the metadata records are required after the funding period of a project has ended. Only in cases were data is reused in other projects new opportunities arise for reevaluating the earlier work.

Technical personnel who are responsible for storing or processing the metadata for further purposes generally do evaluation, and sometimes correction, of the metadata. Metadata quality is often assessed by manual inspecting individual samples. Sometimes automated scripts are used to assess general characteristics, such as availability of resources.

With respect to CLARIN's CMDI approach all centers have a strategy were CMDI is mostly generated as a byproduct. Metadata is mostly stored in non-CMDI proprietary or standard formats and CMDI records are made available to the CLARIN infrastructure as requested or required through CLARIN projects. Centers have been experimenting with using CMDI directly in newly funded CLARIN-NL projects as part of the metadata production process but this has yet to reach the scale of full implementation at the organizational level. One reason for this is that full adoption of CMDI requires organizations to invest in CMDI support in all stages of the data management life cycle. For one, it requires repository systems capable of handling the diversity of CMDI profiles. But also at the level of making resources available to the intended end user audience support for CMDI metadata records needs to be built in. While some organizations have opted for a advanced approach in this direction others take a more conservative approach by extending their current systems at the boundaries with the CLARIN infrastructure.

Experiences with CMDI are mixed. Some centers indicate that they have no problems with the CMDI model while others indicate to have problems deciding upon which profiles and data categories to use or create. One center has indicated that they yet have to grasp the concepts behind CMDI. There appears to be a clear need for a more hands on exchange of experiences with CMDI profiles and data categories during the metadata creation process. Training sessions on the use of data categories and profiles have been a mandatory component of all CLARIN-NL projects. However, these are perceived to provide a mainly technically oriented approach focusing on the main principles rather than providing practical solutions for the projects at hand. Here it is felt that closer collaboration with CLAIRN centers might help. Also, proliferation of the data category registry and the component registry are regarded as obstacles, along with the perceived complexity from a user perspective of both systems.

Adapting the Metadata Quality Assessment Certification Process for CLARIN

The Metadata Quality Assurance Certification Process (MQACP) presented by N. Palivitisinis[] consists of a number of consecutive phases (Metadata design, Testing, Calibration, Building Critical Mass and Regular Operation) and can be modified to reflect the CMDI related work processes to improve quality from initial CMDI profile creation right up to usage in the CLARIN. The MQACP process distinguishes a number of consecutive phases (Metadata design, Testing, Calibration, Building Critical Mass and Regular operation) involving metadata experts, domain experts, content annotators and content users/consumers. Within CLARIN, scientific or technical project team members, representatives from CLARIN centers, CLARIN metadata experts, reviewers and the user community take on these roles at various stages of the CMDI production process. There is currently no clear distinction between the proposed MQACP phases and CLARIN's CMDI production process and it seems worthwhile evaluating how the MQACP phases could be integrated in the CLARIN process and which quality measures can be taken to assure well defined quality levels at the end of each phase. This separates the CMDI production process in clearly defined stages for which the desired outcomes can be specified but also introduces a number of quality assessment steps in which CLARIN is actively involved. It is recommended that outcomes of these quality assessment steps be listed as part of the projects final reports.

Evaluation of past CLARIN-NL projects show that there is no clear distinction between MQACP phases and quality assurance generally only takes place during the general operation phase. Although the urge for high quality metadata is felt throughout the CLARIN community quality assurance is largely focused on evaluating the end results of metadata, profile and data category creation. This usually takes place after projects have been submitted for final approval to CLARIN. Modifications to profiles and metadata content are thus received after project funding has run out requiring CLARIN centers and project members to invest additional time and resources to meet these requirements. Also, these requirements are often not explicitly clear at the start of these projects leading to discussions afterwards on which data categories should be present to make the provided information useful to the CLARIN community. Quality assessment within CLARIN currently is largely qualitative in nature. A reviewer selects random examples from the Virtual Language Observatory and provides feedback the CLARIN centers and project participants during the final project evaluation process. The review results are not publically available. Quantitative quality assessment has gained growing attention in within the CLARIN community. Trippel et all [Trippel 2014] have described a scoring method taking into account several quality metrics, some of which are also encountered in other literature. Incidentally, this article also provides an insight into the core data category fields that are considered relevant by the authors.

Within CLARIN the metadata production process requires interaction between different groups during the process. (Quantitative) feedback on the metadata quality should be gathered as an integral part of the production process to be able to identify problems at an early stage. Also interaction between different stakeholders such as researchers, CLARIN center representatives and Virtual Language Observatory administrators should be part of the process to accommodate for knowledge exchange. Metadata profiles and records are produced within the project teams covering the Metadata design, Testing, and Calibration phases. Upon completion, metadata records are submitted to the allocated CLARIN center responsible for publishing the metadata to the CLARIN infrastructure (Building Critical Mass phase). Finally, in the Regular Operation phase, metadata is harvested by CLARIN and made available to the end user community through the Virtual Language Observatory. The full representation of the MQACP process in the CLARIN context is provided in the figure below.

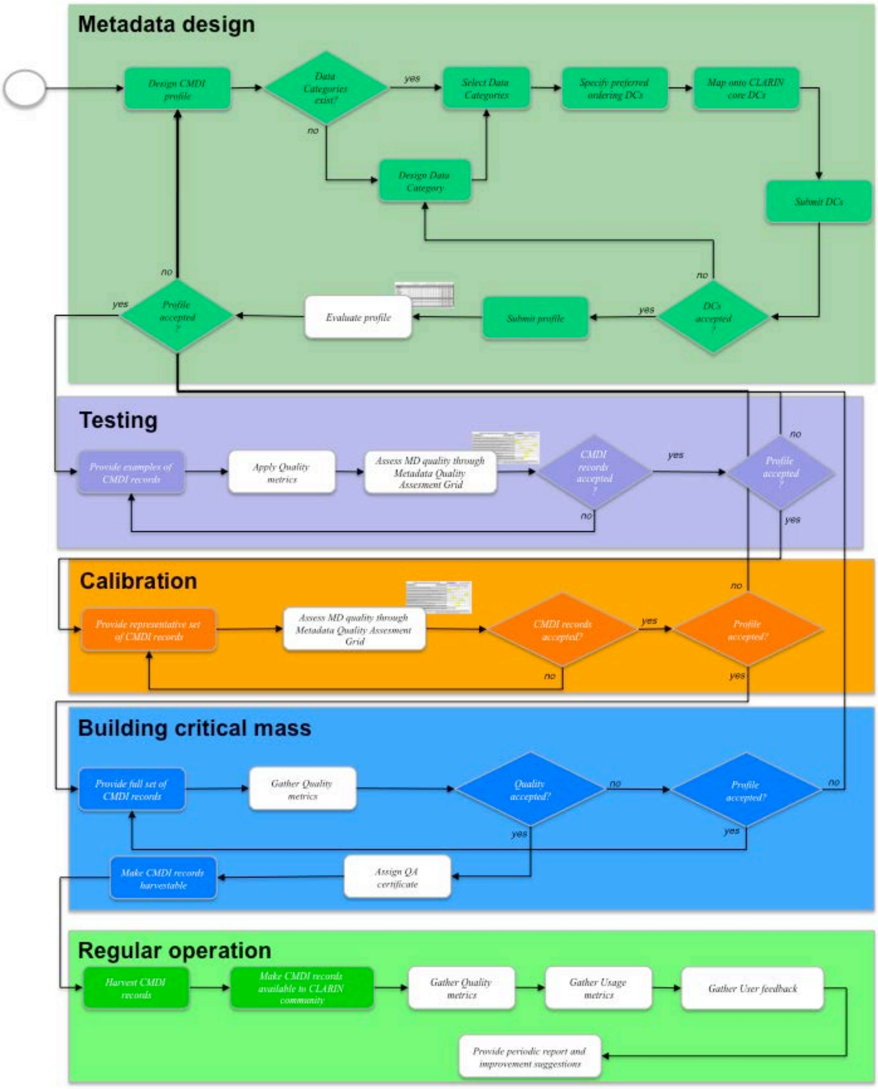


Figure 1: Proposed MQACP process for CLARIN

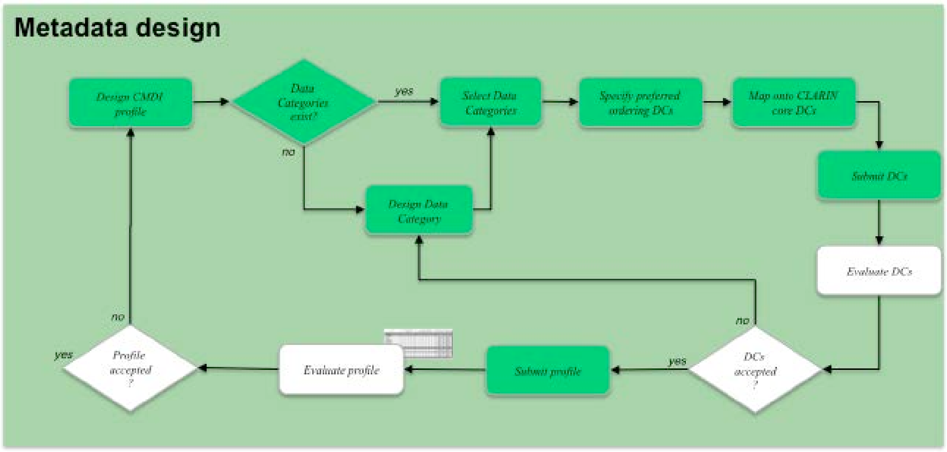
Metadata design phase.

The original MQACP process describes the Metadata design phase as:

In this phase, the metadata standard or specification to be used in the envisaged LOR[6] is selected and the necessary modifications are made to "profile" it to meet the application context. More specifically, a metadata

[6] Learning Object Repository

standard is chosen to fit the generic needs of the application domain and it's profiled and adapted based on the limitations and requirements of the field it's applied to.



Here, the metadata characteristics are selected and organized to meet the needs of the application domain or end user group. The purpose is to produces an initial CMDI profile and list and register all relevant data categories. To provide an initial estimate for the relative importance factor α_i of the ***Completeness*** measure data categories are ordered. It might also be feasible to ask participants to assign relative importance directly to circumvent the problem of having to map the ordering onto a relative importance. In addition, project participants should be asked to map the proposed set of data categories onto the CLARIN core set of data categories. The latter are highly recommended by CLARIN and are, for example, used in the Virtual Language Observatory to provide different facets to the end user community. The resulting profile is also evaluated by representatives of CLARIN centers (other than the center publishing the resources) to provide feedback on the understandability and usefulness of the proposed profile fields.

Metadata experts and domain experts contributing at this stage from within the project should consist of at least be a representative from the CLARIN centre responsible for storing and publishing the final CMDI records and data resources. The domain experts are (often) scientific staff members from the organizations the data originates from and who have considerable experience with the data made available to CLARIN. CLARIN contributes to the quality assessment through the national Data Category Registry coordinator who evaluates the proposed DCs and representatives from other CLARIN centers who jointly evaluate the proposed profile for its usefulness and other parts of the CLARIN community. It is recommend to include the results of these quality assessments in the final reports of the project.

- Design CMDI profile

  A CMDI profile describes the metadata characteristics from the perspective of the intended application domain or end user community. CMDI profiles are designed in CLARIN's Component Registry[7]. It provides support for creating and reusing fully fledged profiles or smaller readily usable building blocks ('components'). Examples of existing metadata schemas, such as IMDI, OLAC and TEI, are present and reuse of profiles (and components) is strongly recommended and encouraged.
  - **Actors**: Metadata experts and domain experts (project members).
  - **Comments**: Within CLARIN the metadata experts should be a representative from the CLARIN centre responsible for storing and publishing the final CMDI records and data resources. The domain experts are (often) scientific staff members from the organizations the data originates from and who have considerable experience and insight into the data made available to CLARIN.
- Data Categories exist?
  - At the data element level of both CMDI profiles and components links to data categories are to be present. A data category is an elementary descriptor in a linguistic structure or an annotation scheme and plays a key role in semantic interoperability in the CLARIN infrastructure.
  - **Actors**: Metadata experts and domain experts
  - **Comments**:
- Select Data Categories
  - It is recommended to reuse Data Categories whenever possible. Persistent identifier links may be inserted into the profile at the data element level
  - **Actors**: Metadata experts and domain experts
- Design Data Categories
  - If no appropriate data category is present a new one may be created. Data Categories may be created in the Data Category Registry[8].
  - **Actors**: Metadata experts and domain experts
- Specify preferred ordering DCs

  The **completeness** measure provides a metric for the amount of information that is present in a metadata record. The weight assigned to each metadata field is expressed in a relative importance factor α. The purpose of ordering the

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

[8] At the time of writing ISOcat (http://www.isocat.org) acts as CLARIN's Data Category Registry. This may change in the near future as ISOcat is currently being phased out.

Data Categories is to provide input for the relative importance factor of each field. Also it provides a guidelines for metadata creators for prioritizing metadata fields when producing metadata records

    - **Actors**:Domain experts and metadata experts.
- Map onto CLARIN core DCs

    Within the CLARIN infrastructure several end user applications make use of specific data categories. The Virtual Language Observatory, for example, uses data categories associated with fields such as language, continent and genre to provide different facets to the end user community. It is strongly recommended to include these data categories in the CMDI profile or provide an approximation to one of the core data categories.

    - **Actors**:Metadata experts and domain experts.
- Submit DCs

    The list of proposed (new) data categories is submitted to CLARIN's national DCR coordinator for evaluation.

    - **Actors**: Metadata experts and national DCR coordinator
- Evaluate DCs

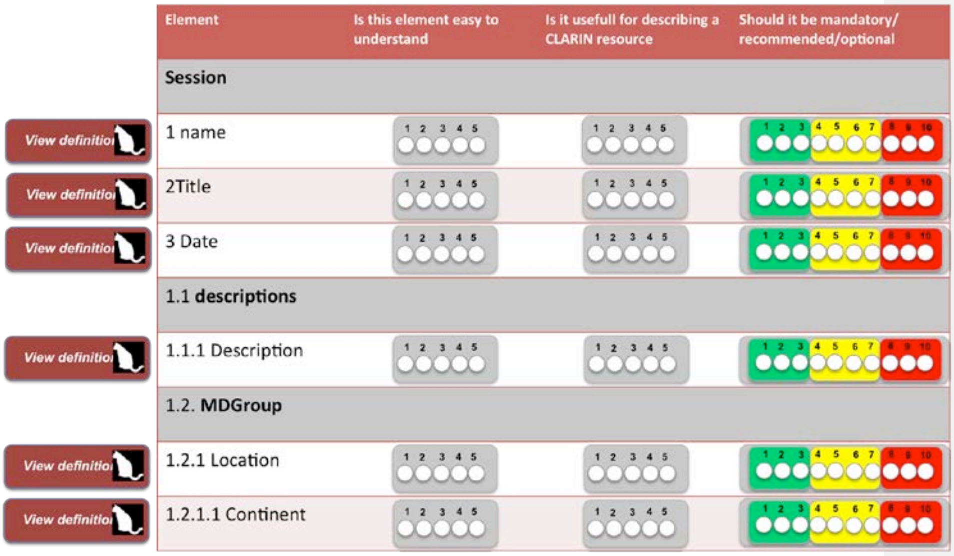
    CLARIN's national DCR coordinator evaluated the list of proposed new or modified data categories. The purpose here is to ensure that the specification of the data categories is formally correct and semantic interoperability across projects is stimulated. Suggestions for modifications, such as improved definition or example usage, may be proposed to raise the quality level of the data category specifications or alternative data categories may be suggested for use in the application profile.

    - **Actors**: National DCR coordinator
- DCs accepted?

    If the data categories are not accepted by the national DCR coordinator the metadata expert may be requested to modify the data category specification, use an alternative, similar data category or provide additional/alternative mappings onto CLARIN's core data categories.

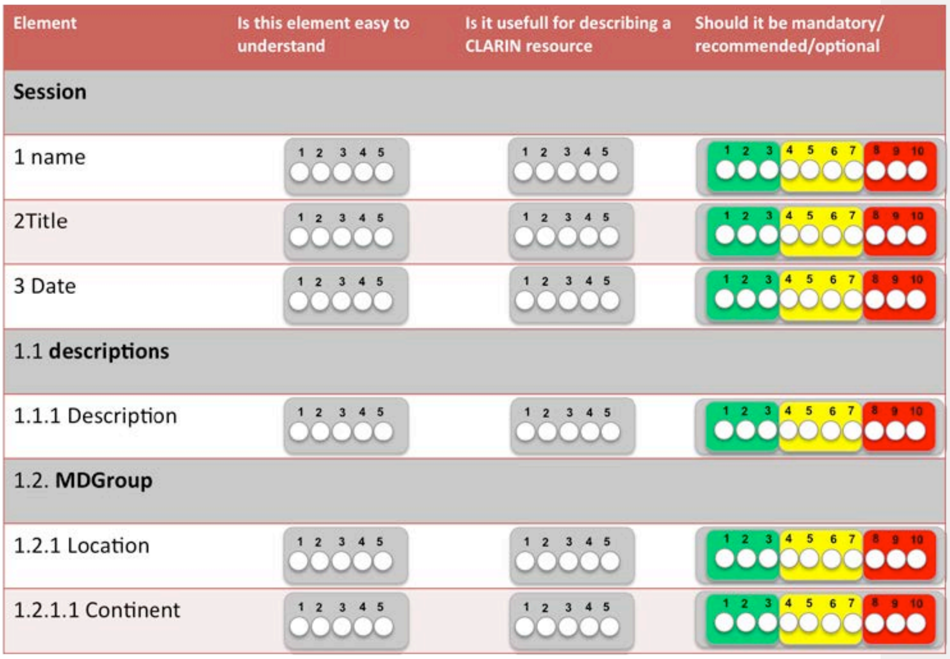
    - **Actors**: National DCR coordinator
- Submit profile
    - Once data category specifications have been accepted the profile may be submitted for review.
- Evaluate profile
    - Representatives from other CLARIN centers evaluate the profile. They represent a broad perspective on the CLARIN domain and provide feedback on the understandability, usefulness and whether elements should be considered to be mandatory, recommended or optional. For this evaluation process a Application Profile Design tool may used similar to the one described in the MQACP process. All

necessary information such as component and data category information can be automatically generated from the submitted profile using the Component Registry and Data Category Registry. The results are gathered and used for further evaluation.

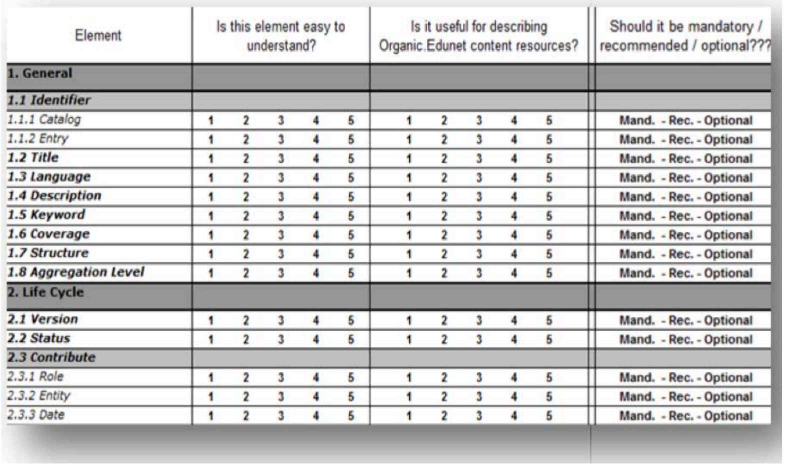


Element	Is this element easy to understand	Is it usefull for describing a CLARIN resource	Should it be mandatory/ recommended/optional
Session			
1 name	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
2Title	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
3 Date	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.1 **descriptions**			
1.1.1 Description	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.2. **MDGroup**			
1.2.1 Location	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.2.1.1 Continent	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○

Figure 2: Application Profile Design Tool, CLARIN style

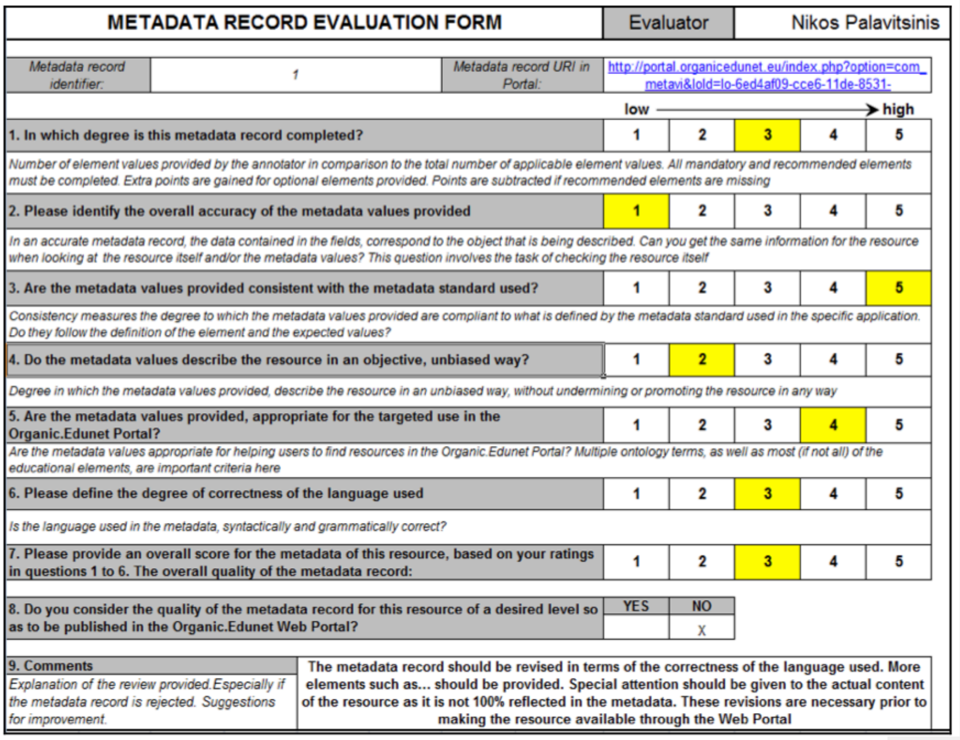


Element	Is this element easy to understand	Is it usefull for describing a CLARIN resource	Should it be mandatory/ recommended/optional
Session			
1 name	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
2Title	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
3 Date	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.1 **descriptions**			
1.1.1 Description	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.2. **MDGroup**			
1.2.1 Location	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.2.1.1 Continent	1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○



Element	Is this element easy to understand?	Is it useful for describing Organic.Edunet content resources?	Should it be mandatory / recommended / optional???
1. General			
1.1 Identifier			
1.1.1 Catalog	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.1.2 Entry	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.2 Title	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.3 Language	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.4 Description	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.5 Keyword	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.6 Coverage	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.7 Structure	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.8 Aggregation Level	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2. Life Cycle			
2.1 Version	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.2 Status	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.3 Contribute			
2.3.1 Role	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.3.2 Entity	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.3.3 Date	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional

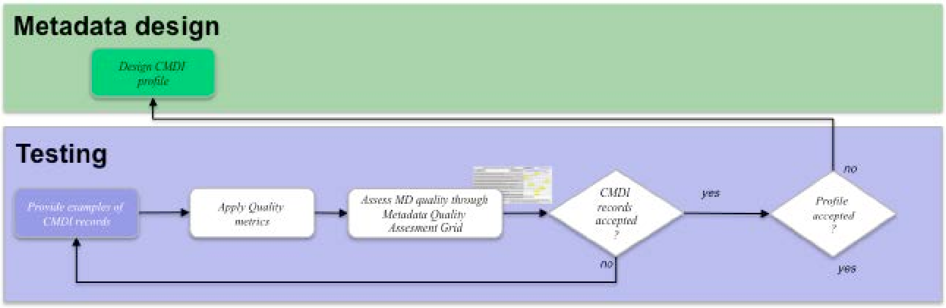
- Profile accepted?
    - If there is no general consensus on the practical applicability of the profile metadata designers and domain experts may decide to modify the profile, structure or add/remove data categories before resumission.



METADATA RECORD EVALUATION FORM	Evaluator	Nikos Palavitsinis			
Metadata record identifier: 1	Metadata record URI in Portal:	http://portal.organicedunet.eu/index.php?option=com_metavi&loId=lo-6ed4af09-cce6-11de-8531-			
	low → high				
1. In which degree is this metadata record completed?	1	2	[✓ 3]	4	5
Number of element values provided by the annotator in comparison to the total number of applicable element values. All mandatory and recommended elements must be completed. Extra points are gained for optional elements provided. Points are subtracted if recommended elements are missing					
2. Please identify the overall accuracy of the metadata values provided	[✓ 1]	2	3	4	5
In an accurate metadata record, the data contained in the fields, correspond to the object that is being described. Can you get the same information for the resource when looking at the resource itself and/or the metadata values? This question involves the task of checking the resource itself					
3. Are the metadata values provided consistent with the metadata standard used?	1	2	3	4	[✓ 5]
Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata standard used in the specific application. Do they follow the definition of the element and the expected values?					
4. Do the metadata values describe the resource in an objective, unbiased way?	1	[✓ 2]	3	4	5
Degree in which the metadata values provided, describe the resource in an unbiased way, without undermining or promoting the resource in any way					
5. Are the metadata values provided, appropriate for the targeted use in the Organic.Edunet Portal?	1	2	3	[✓ 4]	5
Are the metadata values appropriate for helping users to find resources in the Organic.Edunet Portal? Multiple ontology terms, as well as most (if not all) of the educational elements, are important criteria here					
6. Please define the degree of correctness of the language used	1	2	[✓ 3]	4	5
Is the language used in the metadata, syntactically and grammatically correct?					
7. Please provide an overall score for the metadata of this resource, based on your ratings in questions 1 to 6. The overall quality of the metadata record:	1	2	[✓ 3]	4	5
8. Do you consider the quality of the metadata record for this resource of a desired level so as to be published in the Organic.Edunet Web Portal?	YES	NO			
		X			
9. Comments *Explanation of the review provided.Especially if the metadata record is rejected. Suggestions for improvement.*	**The metadata record should be revised in terms of the correctness of the language used. More elements such as... should be provided. Special attention should be given to the actual content of the resource as it is not 100% reflected in the metadata. These revisions are necessary prior to making the resource available through the Web Portal**				

Testing phase

In this phase, a test implementation of the content repository management system an be used for hands-on experience with metadata....Domain experts provide metadata for a limited set of resources, using the application profile....This process allows the domain experts to get accustomed to the application profile and the metadata experts to get some preliminary feedback on the use of metadata....



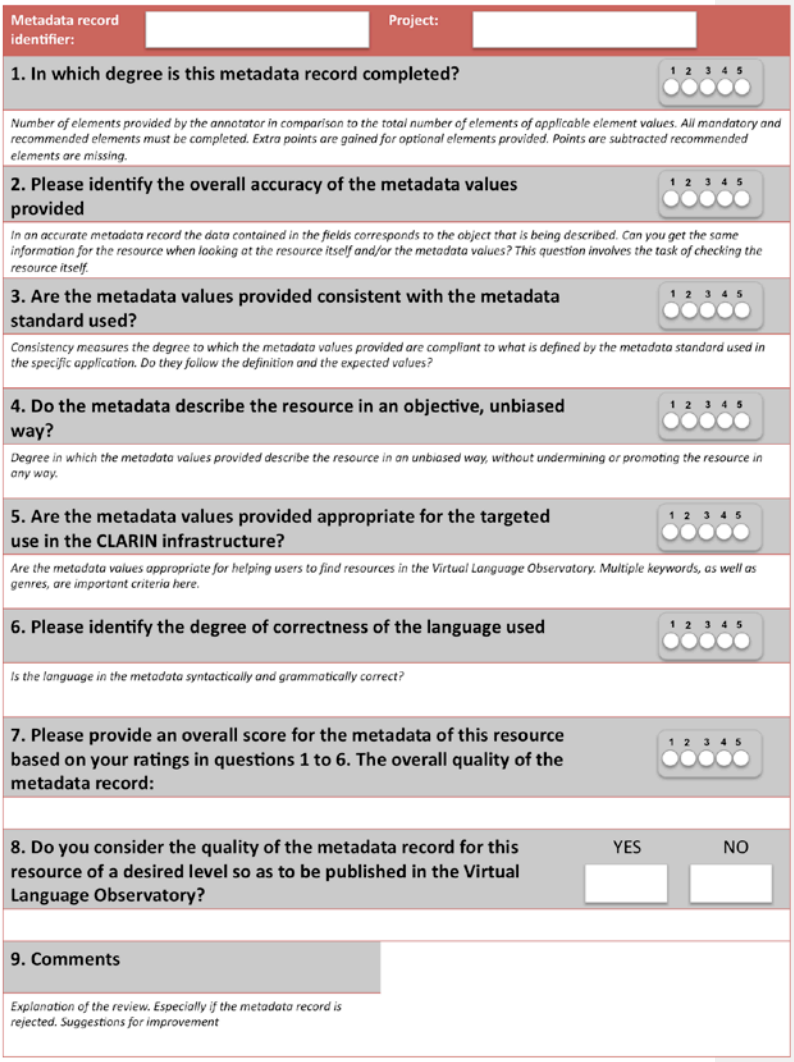
Within most of the CLARIN projects the testing phase is currently absent.

Question: Are examples of CMDI records currently evaluated within the CLARIN-NL projects?

- Provide examples of CMDI records

    - Bases on the CMDI profile created in the Metadata design phase a small representative set of CMDI records is produced by domain and metadata experts associated with the project Upon completion these are submitted to CLARIN for initial quality assessment.
    - **Actors**: Metadata experts and domain experts (project members).
- Apply quality metrics
    - To gain a first impression of the quality of the metadata records quality metrics are determined across several metadata quality dimensions. Metadata quality dimensions and associated metrics considered feasible in the CLARIN context are discussed elsewhere in this document.
    - **Actors**: Metadata experts and domain experts
- Asses MD quality through Metadata Quality Assessment Grid
    - The Metadata Quality Assessment Grid has also been described in the original MQACP model proposed byPalavitsinis [Palavitsinis 2013]. The assessment grid is a questionnaire assessing the same metadata quality dimensions as used in the previous step providing the same view but from a human perspective. Representatives from CLARIN centers are asked to perform this assessment. This will contribute to a wider community perspective on the quality of the metadata and can also help to build a common understanding of good metadata practices[9]. In the previous CLARIN-NL round this could be considered to be task within the
    - **Actors**: Metadata experts (CLARIN center representatives)

[9] Expressed in terms of the current CLARIN-NL projects this could be implemented as a shared task in the IIP (Infrastructure Implementation Plan) project.



Metadata record identifier: ______ **Project:** ______

1. In which degree is this metadata record completed? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Number of elements provided by the annotator in comparison to the total number of elements of applicable element values. All mandatory and recommended elements must be completed. Extra points are gained for optional elements provided. Points are subtracted recommended elements are missing.

2. Please identify the overall accuracy of the metadata values provided 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

In an accurate metadata record the data contained in the fields corresponds to the object that is being described. Can you get the same information for the resource when looking at the resource itself and/or the metadata values? This question involves the task of checking the resource itself.

3. Are the metadata values provided consistent with the metadata standard used? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata standard used in the specific application. Do they follow the definition and the expected values?

4. Do the metadata describe the resource in an objective, unbiased way? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Degree in which the metadata values provided describe the resource in an unbiased way, without undermining or promoting the resource in any way.

5. Are the metadata values provided appropriate for the targeted use in the CLARIN infrastructure? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Are the metadata values appropriate for helping users to find resources in the Virtual Language Observatory. Multiple keywords, as well as genres, are important criteria here.

6. Please identify the degree of correctness of the language used 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Is the language in the metadata syntactically and grammatically correct?

7. Please provide an overall score for the metadata of this resource based on your ratings in questions 1 to 6. The overall quality of the metadata record: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

8. Do you consider the quality of the metadata record for this resource of a desired level so as to be published in the Virtual Language Observatory? YES □ NO □

9. Comments

Explanation of the review. Especially if the metadata record is rejected. Suggestions for improvement

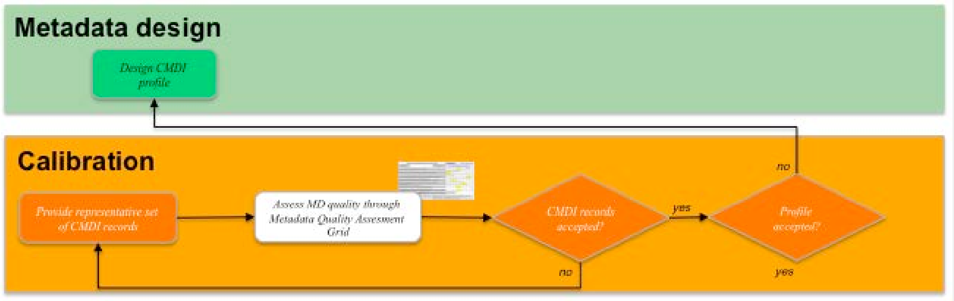
- CMDI records accepted?
    - This step represents a decision moment where the quality of the provided sample records is considered to be sufficient by CLARIN. If the quality of the metadata records are not considered to be sufficient, based on either the quality metrics of the view of the reviewers, the project participants may be requested to provide a new sample with suggested improvements. Suggested improvements relate to the content of the metadata records. Suggestions for structural changes are covered in the next step.
    - **Actors**: Metadata experts (CLARIN center representatives) and project lead??
- Profile accepted?
    - If the problems observed suggestions for improvement relate to the structure of the metadata profile a redesign of the CMDI profile should be considered. This often involves

restructuring of the data categories and adding/removing data categories. At this stage domain experts have started to get accustomed to the metadata profile and new insights may have been gained that were not discovered during the Metadata design phase. It is quite common to iterate over a metadata profile a number of times in order to collect all relevant information fields. It is recommend that the results of the previous two steps are recorded and the resulting decision are recorded as part of the standard project's deliverables. If the metadata records are not accepted

Comment [MKS1]: Hoe te voorkomen dat de metadata experts biuj de centra overladen worden met kleine aanpassingen en wijzigingen vanuit allerlei projecten?

Calibration phase

During this phase, the various technical components (web front-end, content management system, etc) are put together and part of the content is available online. Content providers are still involved in the process and more specifically continue to annotate resources using the tool(s) deployed. A larger body of resources is now uploaded on the tool and a metadata peer review exercise takes place on a representative sample of resources.


Metadata design
Design CMDI profile
Calibration
Provide representative set of CMDI records
Assess MD quality through Metadata Quality Assesment Grid
CMDI records accepted?
yes
no
Profile accepted?
no
yes

- Provide representative set of CMDI records

  Bases on the CMDI profile created in the Metadata design phase a small representative set of CMDI records is produced by domain and metadata experts associated with the project Upon completion these are submitted to CLARIN for initial quality assessment.

  - **Actors**: Metadata experts and domain experts (project members).
- Assess MD quality through Metadata Quality Assessment Grid
  - To determine the usability of the metadata records from a wider perspective the quality of the metadata records is manually scored by other CMDI experts.
  - **Actors**: Metadata experts (CLARIN center representatives).
- CMDI records accepted?

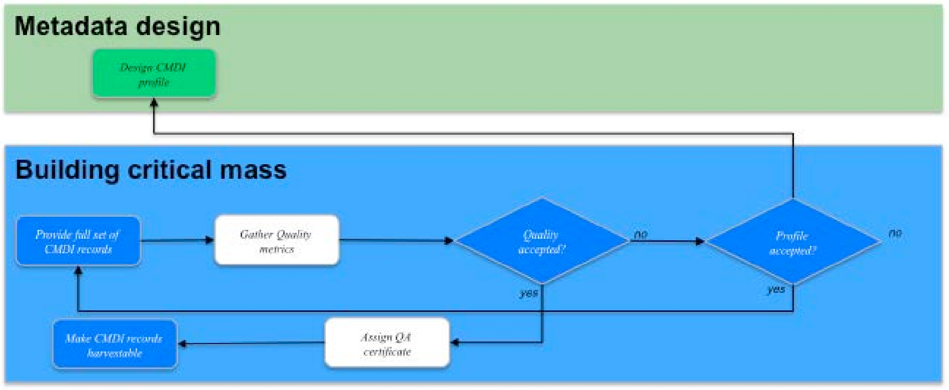
If problems exist within the sample set of metadata records it may be decided to try to add further metadata information to records.

- Profile accepted?

  If problems with the metadata records relate to the underlying profile specification the profile is redirected back into the design phase

Building critical mass phase

Critical mass is the phase during which the tool(s) have reached a high maturity level and are ready to accept large numbers of content with their respective metadata. The application profile used is now completed and final, so not a great deal of changes can take place and in addition a significant number of metadata records are available for the metadata experts to review and analyze.



The Building critical mass phase consists of the following steps:

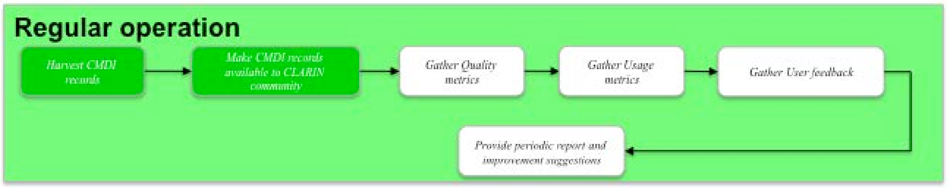
- Provide full set of CMDI records
  - The project has completed the full set of metadata descriptions and is ready to hand them over to the responsible CLARIN center.
  - **Actors**: Metadata experts and domain experts (project members).
- Gather quality metrics
  - The quality of the metadata descriptions is assessed automatically before metadata records are to be accepted by the CLARIN center.

    Quality metrics at this stage only take the local context into account at this stage:
    - Completeness: Q_{comp} and Q_{wcomp}
    - Accuracy: AccR(y) and Q_{accu}
    - Logical Consistency: ConR(y) and Q_{cons}

        - Accessibility: Q_{read}
    - **Actors**: Metadata experts and domain experts (project members).
- Quality accepted?
    - A decision is made whether the quality of the metadata records is acceptable. For projects funded through CLARIN a minimal set of quality criteria may apply. If the quality not accepted this may indicate a problem with the metadata profile.
    - **Actors**: Metadata experts(CLARIN centers) and CLARIN
- Profile accepted?
    - If there are problems with the metadata records and these relate to the profile then the profile should be redirected to the design phase. It should be noted that rejection of a profile at this stage has a significant impact on the whole metadata production process.
    - **Actors**: Metadata experts(CLARIN centers) and CLARIN
- Assign QA certificate
    - As an incentive a Quality assurance certificate may be attached to the metadata document indicating that it has passed the MQACP process.
    - **Actors**: CLARIN
- Make CMDI records harvestable
    - Here, MCDI records are made available thought the center's OAI-PMH server.
    - **Actors**: Metadata experts(CLARIN centers)

Regular operation phase

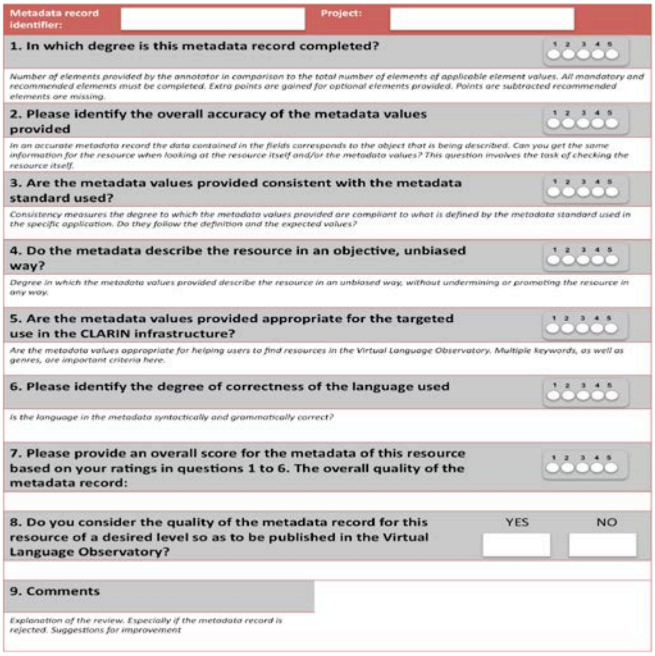
During regular operation, the metadata elements used in the tool(s) are considered to be final. The tools themselves and the content providers are now annotating resources regularly but not necessarily intensively like in the previous phase. This period covers the remainder of the LOR lifecycle.



The regular operation phase consists of the following steps:

- Harvest CMDI records
    - The CMDI records are harvested through the center's OAI-PMH end points
    - **Actors**: Metadata experts.

- Make CMDI records available to CLARIN community
    - The CMDI records are indexed and published in the Virtual Language Observatory
    - **Actors**: Metadata experts.
- Gather quality metrics
    - Quality metrics for all metadata records are gathered, both at the individual level as well as the aggregated level
    - **Actors**: Metadata experts.
- Gather usage metrics
    - Part of the operational phase should be gathering of usage statistics. This will provide useful indications of the user behavior.
    - **Actors**: Metadata experts.
- Gather user feedback
    - To be able to match the user's expectations with the previously collected quality metrics it is necessary that the user is able to provide feedback on the usability of the metadata records. The previously introduced feedback form may be used here.
    - **Actors**: Metadata experts.



Metadata record identifier: □ **Project:** □

1. In which degree is this metadata record completed? 1 2 3 4 5 ○○○○○

Number of elements provided by the annotator in comparison to the total number of elements of applicable element values. All mandatory and recommended elements must be completed. Extra points are gained for optional elements provided. Points are subtracted recommended elements are missing.

2. Please identify the overall accuracy of the metadata values provided 1 2 3 4 5 ○○○○○

In an accurate metadata record the data contained in the fields corresponds to the object that is being described. Can you get the same information for the resource when looking at the resource itself and/or the metadata values? This question involves the task of checking the resource itself.

3. Are the metadata values provided consistent with the metadata standard used? 1 2 3 4 5 ○○○○○

Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata standard used in the specific application. Do they follow the definition and the expected values?

4. Do the metadata describe the resource in an objective, unbiased way? 1 2 3 4 5 ○○○○○

Degree in which the metadata values provided describe the resource in an unbiased way, without undermining or promoting the resource in any way.

5. Are the metadata values provided appropriate for the targeted use in the CLARIN infrastructure? 1 2 3 4 5 ○○○○○

Are the metadata values appropriate for helping users to find resources in the Virtual Language Observatory. Multiple keywords, as well as genres, are important criteria here.

6. Please identify the degree of correctness of the language used 1 2 3 4 5 ○○○○○

Is the language in the metadata syntactically and grammatically correct?

7. Please provide an overall score for the metadata of this resource based on your ratings in questions 1 to 6. The overall quality of the metadata record: 1 2 3 4 5 ○○○○○

8. Do you consider the quality of the metadata record for this resource of a desired level so as to be published in the Virtual Language Observatory? YES □ NO □

9. Comments

Explanation of the review. Especially if the metadata record is rejected. Suggestions for improvement

- Provide periodic report and improvement suggestions.
    - Communication of the results, i.e. quality metrics and user results should be communicated back to the participating CLARIN centers so they can compare CLARIN results with their own results.
    - **Actors**: Metadata experts.

Questionnaires in the MQACP process

The CLARIN MQACP process contains three quality assessment forms designed to quality assessment information at various stages of the quality assurance process; the Profile Assessment Grid, the Metadata Quality Assessment Grid and the User Assessment Grid.

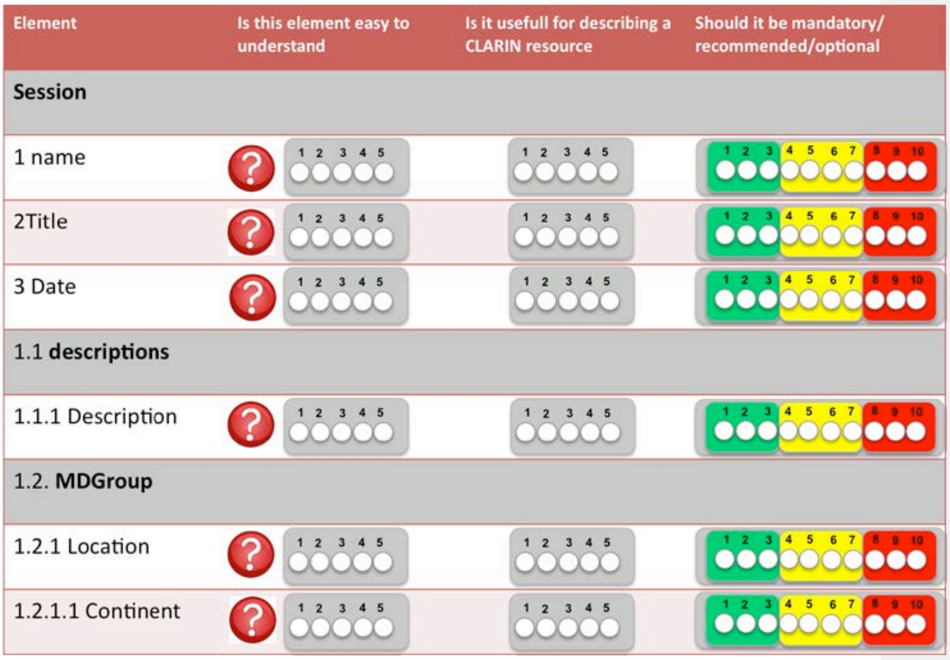
Profile Assessment Grid

The Profile Assessment Grid is used to provide feedback on the CMDI profile created in the *Metadata design* phase. The intended end users are CLARIN center representatives acting as metadata experts. The grid consists of three questions for each of the data categories represented in the CMDI profile:

1. Is the element easy to understand?
   a. An element should be easy to understand by both metadata annotators and content users/consumers. It should be placed logically in the profile's structure and the element's definition should be clear and consice.
2. Is the element useful for describing CLARIN resources?
   a. The element should be relevant to the use within the CLARIN infrastructure. Since the profile is evaluated by multiple CLARIN center representatives serving different domains of CLARIN's intended end user audience the idea is to balance these different perspectives.
3. Should the element be mandatory, recommended or optional?
   a. By indicating whether an element is expected to be mandatory, recommended or optional designers of the metadata profile are made aware of the relevance of the fields. Also, the scores obtained from this may be used as input for the relative importance parameters in the Q_{wcomp},, *AccR(y)*, and *ConR(y)* measures.

Comment [MKS2]: This might be split into two questions, since the question pertains to 2 different aspects.

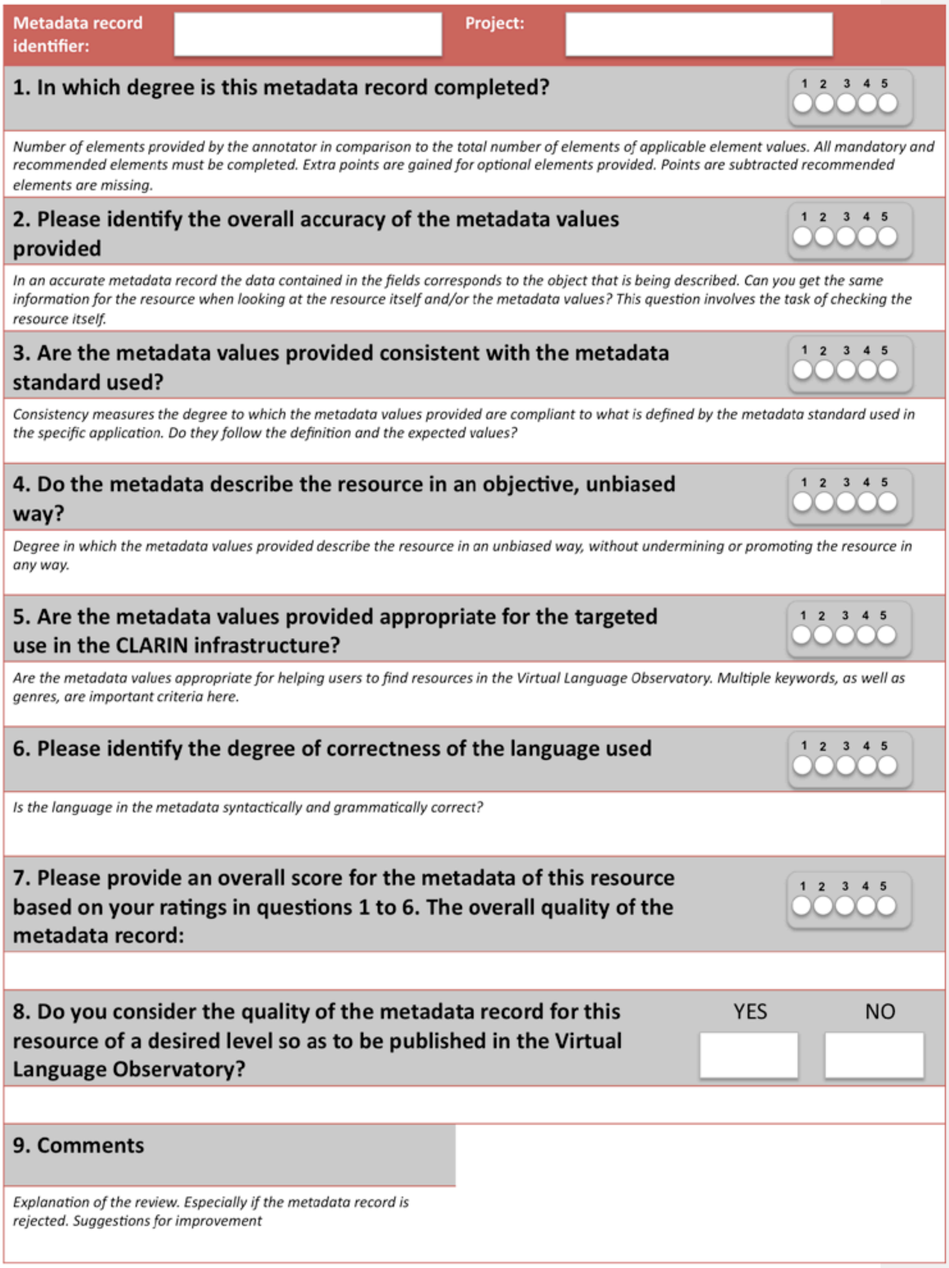
Maintainers of the Data Category Registry and Component Registry may also use the information from the Profile Assessment grid as quality indicators for data categories and CMDI profiles .



Element	Is this element easy to understand	Is it usefull for describing a CLARIN resource	Should it be mandatory/ recommended/optional
Session			
1 name	? 1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
2Title	? 1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
3 Date	? 1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.1 **descriptions**			
1.1.1 Description	? 1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.2. **MDGroup**			
1.2.1 Location	? 1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○
1.2.1.1 Continent	? 1 2 3 4 5 ○○○○○	1 2 3 4 5 ○○○○○	1 2 3 4 5 6 7 8 9 10 ○○○○○○○○○○

Metadata Quality Assessment Grid

The Metadata Quality Assessment Grid is intended to provide human evaluator feedback on the all of the intended quality assurance dimensions, i.e. completeness, accuracy, conformance to expectations and coherence and provide a human assessment of the overall quality. It is a slightly modified version of the original Metadata Quality Assessment Grid as proposed by Palavitsinis. The intended end users are CLARIN center representatives who are requested to provide this feedback during the *Testing* and *Calibration* phases of the metadata creation process. A simple to use overview of the profile specification should accompany the form shown below to assist the evaluator in the feedback process. Mandatory, recommended and optional elements should be indicated appropriately and shortlists of admissible values wherever appropriate. These lists may be collected form the Data Category Registry. One option here is to present the contents of the metadata record directly in this context.



Metadata record identifier: ☐ Project: ☐

1. In which degree is this metadata record completed? 1 2 3 4 5 ○○○○○

Number of elements provided by the annotator in comparison to the total number of elements of applicable element values. All mandatory and recommended elements must be completed. Extra points are gained for optional elements provided. Points are subtracted recommended elements are missing.

2. Please identify the overall accuracy of the metadata values provided 1 2 3 4 5 ○○○○○

In an accurate metadata record the data contained in the fields corresponds to the object that is being described. Can you get the same information for the resource when looking at the resource itself and/or the metadata values? This question involves the task of checking the resource itself.

3. Are the metadata values provided consistent with the metadata standard used? 1 2 3 4 5 ○○○○○

Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata standard used in the specific application. Do they follow the definition and the expected values?

4. Do the metadata describe the resource in an objective, unbiased way? 1 2 3 4 5 ○○○○○

Degree in which the metadata values provided describe the resource in an unbiased way, without undermining or promoting the resource in any way.

5. Are the metadata values provided appropriate for the targeted use in the CLARIN infrastructure? 1 2 3 4 5 ○○○○○

Are the metadata values appropriate for helping users to find resources in the Virtual Language Observatory. Multiple keywords, as well as genres, are important criteria here.

6. Please identify the degree of correctness of the language used 1 2 3 4 5 ○○○○○

Is the language in the metadata syntactically and grammatically correct?

7. Please provide an overall score for the metadata of this resource based on your ratings in questions 1 to 6. The overall quality of the metadata record: 1 2 3 4 5 ○○○○○

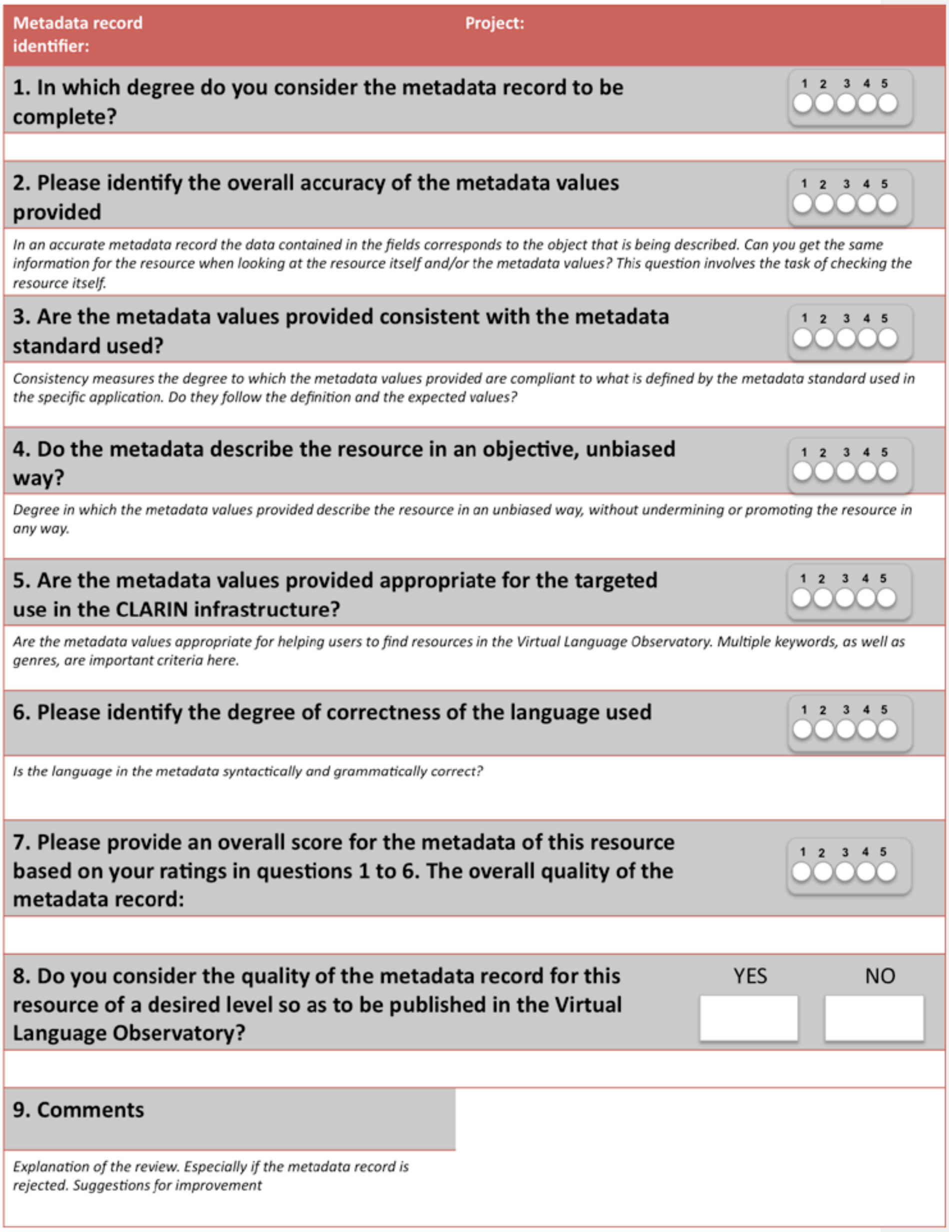
8. Do you consider the quality of the metadata record for this resource of a desired level so as to be published in the Virtual Language Observatory? YES ☐ NO ☐

9. Comments

Explanation of the review. Especially if the metadata record is rejected. Suggestions for improvement

Metadata Quality Assessment Grid (User evaluation)

End users are requested to provide feedback in the quality of metadata records in the *Regular Operation* phase. The purpose of these evaluations is to provide CLARIN centers and resource owners feedback on how to improve metadata quality of their records in future rounds and provide the possibility of comparing end user experiences with automated test results as described in the next section. The form used is essentially the same as the previously presented Metadata Quality Assessment Grid. It should at least cover the same topics, but look and feel and formulation of questions may be adapted to the end user community.



Metadata record identifier: **Project:**

1. In which degree do you consider the metadata record to be complete? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

2. Please identify the overall accuracy of the metadata values provided 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

In an accurate metadata record the data contained in the fields corresponds to the object that is being described. Can you get the same information for the resource when looking at the resource itself and/or the metadata values? This question involves the task of checking the resource itself.

3. Are the metadata values provided consistent with the metadata standard used? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata standard used in the specific application. Do they follow the definition and the expected values?

4. Do the metadata describe the resource in an objective, unbiased way? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Degree in which the metadata values provided describe the resource in an unbiased way, without undermining or promoting the resource in any way.

5. Are the metadata values provided appropriate for the targeted use in the CLARIN infrastructure? 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Are the metadata values appropriate for helping users to find resources in the Virtual Language Observatory. Multiple keywords, as well as genres, are important criteria here.

6. Please identify the degree of correctness of the language used 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Is the language in the metadata syntactically and grammatically correct?

7. Please provide an overall score for the metadata of this resource based on your ratings in questions 1 to 6. The overall quality of the metadata record: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

8. Do you consider the quality of the metadata record for this resource of a desired level so as to be published in the Virtual Language Observatory? YES □ NO □

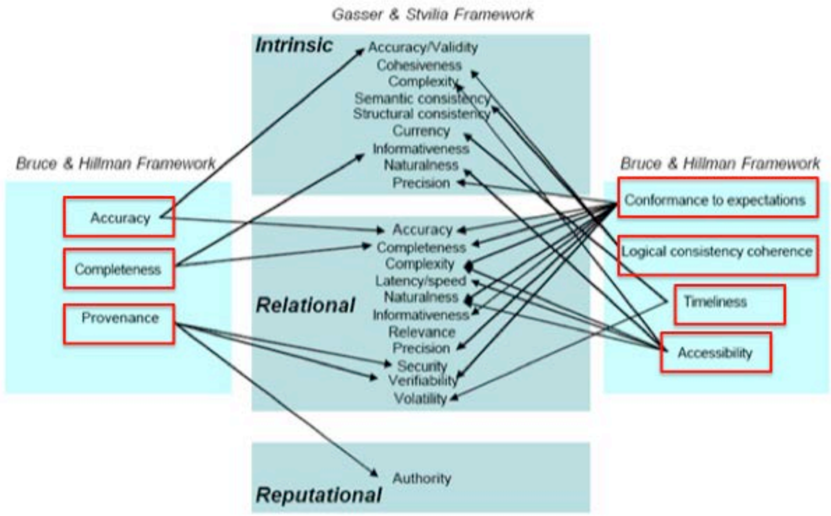
9. Comments

Explanation of the review. Especially if the metadata record is rejected. Suggestions for improvement

Defining quality metrics for CLARIN

Quality measurement categories

Several frameworks have been proposed to categorize metadata quality measures, such as by Moen et all, Stvilia et all or Bruce and Hillman. At least a partial mapping between these frameworks appears to be acceptable. An example of the Gasser&Stvilia framework and Bruce&Hilman Framework is shown below. Ochoa et all have used the Bruce&Hillman framework as a basis for a proposed set of metrics. Similar metrics have been proposed by other authors as well. Section [] provides an overview of these metrics and discusses their relevance for the CLARIN metadata domain.



Metadata quality metrics

Several attempts have been made to quantify the Bruce & Hillman metrics. These are described below as part of the individual metrics section. Ochoa et all[Ochoa 2009] have not only attempted to set up an elaborate set of these metrics, but also attempted to evaluate these by comparing them to agreement by human evaluators. Interestingly, *completeness* and *conformance to expectation* metrics show highest level of agreement here and could thus provide the most reliable quality metrics in a quality assessment process.

Metric Section Bruce&Hillman	Metric	Required scope	Reported human evaluator agreement	Comment
Completeness	Q_{comp}	Local	90%	
Completeness	Q_{Wcomp}	Local	70%	Requires relevance weights
Accuracy	$AccR(y)$	Local	?	Requires relevance weights and spell checkers
Accuracy	Q_{accu}	Local	30%	Requires textual resource analysis
Conformance to expectation	Q_{cinfo}	Repository	20%	
Conformance to expectation	Q_{Tinfo}	Repository	80%	
Logical consistency	Q_{cons}	Local	-	
Logical consistency	Q_{coh}	Repository	40%	
Logical consistency	$ConR(y)$	Local	?	Requires relative importance
Accessibility	Q_{link}	Repository	-	
Accessibility	Q_{read}	Local	30%	
Timeliness	Q_{time}	Repository	-	Requires consecutive updates
Provenance	Q_{prov}	Repository	-	
Total	Q_{avg}		90%	

Completeness

A metadata instance should describe the resources as fully as possible. Also, the metadata fields should be filled in for the majority of the resource population in order to make them useful for any kind of service. The completeness metric can be used to measure how much information is available about a resource.

The most direct approach is to measure whether metadata fields have been filled in or not. The resulting score is expressed as a completeness measure:

$$Q_{comp} = \frac{\sum_{i=1}^{N} P(i)}{N}$$

Where $P(i)$ is 1 if the field has a no-null value and 0 otherwise. N is the number of fields defined in the metadata profile. There appears to be large reported agreement(90%) with human evaluators on the metric.

A more advanced version of this metric, the weighed completeness measure [Ochoa 2009][Bellini 2013], , takes the relative importance of each of the metadata fields into account:

Error! Bookmark not defined. $Q_{wcomp} = \frac{\sum_{i=1}^{N} \alpha_i * P(i)}{\sum_{i=1}^{N} \alpha_i}$

Where $P(i)$ is 1 if the field has a no-null value and 0 otherwise. N is the number of fields defined in the metadata profile. α_i is the relative importance of the *i*-th field. Reported agreement with human evaluators is slightly lower(70%) than for the Q_{com} measure.

Determining the relative importance of metadata fields.

The Q_{wcomp} measure requires a relative weight to be assigned to each of the metadata fields. The question thus becomes how to assign relative weights to metadata fields, and more specifically, in the CLARIN context.

The relative importance has been estimated for DC-fields[Bellini 2013] by averaging the weights assigned by members of the Open Access community The results are shown below:

Fields	Weights
Creator	0.95
Title	0.95
Data	0.86
Identifier	0.8
Description	0.78
Subject	0.73
Type	0.72
Rights	0.7
Contributor	0.68
Format	0.66
Language	0.66

Although this overview provides a first impression of some of the important metadata fields, these fields do not necessarily reflect the views of the CLARIN community. The Virtual Language Observatory[10] uses the following set of categories/fields in its user interface to narrow down the search results. The order in which they are presented on the user interface might suggests an ordering preference here. Noteworthy is that not all facets of the Virtual Language Observatory, such as Data Provider and National Project are not reflected in the metadata record or data category fields, but appear to have been added by the maintainers of the site.

These fields largely overlap with the 14 core CLARIN fields described in [Trippel 2014]. The list of core categories includes mappings onto ISOcat data categories and of Dublin Core elements. The table is provided below.

VLO facet	Core Category	Data category identifier in ISOcat (DC) or Dublin Core

[10] http://catalog.clarin.eu/vlo/

	Project name	DC-2536 DC-2537 DC-5414
	Resource name	DC-5428 DC-5127 DC-4160
		DC-4114 DC-2544 DC-2545
		DC-6119
		Dublin Core: title
	Date indication	DC-2509 DC-2510 DC-2538
		DC-6176
		Dublin Core: created, date, issued
CONTINENT	*Continent*	DC-2531 DC-3791
COUNTRY	*Country*	DC-2532 DC-3792 DC-2092
LANGUAGE	*Language*	DC-2482 DC-2484 DC-5361
		DC-5358
ORGANISATION	*Organization*	DC-2459 DC-2979 DC-6134 DublinCore:
		Publisher
GENRE	*Genre*	DC-2470 DC-3899
MODALITY	*Modality*	DC-2490
SUBJECT	*Subject*	DC-2591 DC-6147 DC-5316
		Dublin Core: subject

	Description	DC-2520 DC-6124
		Dublin Core: description
	Resource class	DC-5424 DC-3806
		Dublin Core: type
FORMAT	*Format*	DC-2571
KEYWORD	*Keywords*	DC-5436
COLLECTION		
RESOURCE TYPE		
NATIONAL PROJECT		
DATA PROVIDER		

Figure 3: VLO core facet fields

Two questions thus become relevant here:

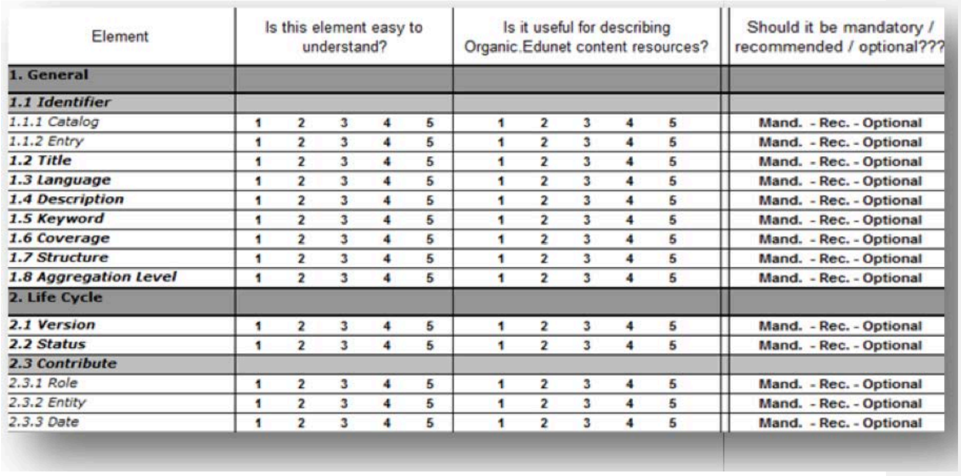
- Which data categories are of broad relevance? These data categories should be recommended for inclusion to ALL metadata experts for inclusion in their profiles.
- Which data categories are relevant to a specific domain? Mostly metadata experts will create specific profiles for describing their resources. Although CLARIN promotes reuse of profiles and components across communities it appears that generally only components are reused.

The first question may be answered by consulting CLARIN experts who deal with metadata aggregation and searchability at the CLARIN EU level.

To get an insight into the perceived importance of data categories at the individual profile level, a questionnaire type approach as was used in the Metadata Understanding Session as part of the MQACP[11] process may be

[11] Nikos Palavitsinis, Metadata Quality issues in learning repositories, Doctoral Thesis November 2013

used.



Element	Is this element easy to understand?	Is it useful for describing Organic.Edunet content resources?	Should it be mandatory / recommended / optional???
1. General			
1.1 Identifier			
1.1.1 Catalog	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.1.2 Entry	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.2 Title	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.3 Language	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.4 Description	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.5 Keyword	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.6 Coverage	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.7 Structure	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
1.8 Aggregation Level	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2. Life Cycle			
2.1 Version	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.2 Status	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.3 Contribute			
2.3.1 Role	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.3.2 Entity	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional
2.3.3 Date	1 2 3 4 5	1 2 3 4 5	Mand. - Rec. - Optional

To set up such a questionnaire one needs to extract all relevant data category fields for each profile. Since the Component Registry only exposes public profiles through their web service API this process will require some help from the Component Registry's administrators as some organizations keep their profiles private[12]. According to the Questionnaire is supported by a focus group meeting for metadata design and is most likely targeted at Metadata Experts.

provides a checklist across all(?) relevant quality dimensions(Completeness, Accuracy, Consistency, Objectiveness, Appropriateness, Correctness and overal

The relevance of this all for metadata creators and editors is that the CMDI profiles that are being used/created should fit the available information as tightly as possible to minimize the number of open fields in the metadata instances. Also, in the CMDI profile creation stage a set of commonly used or recommended data categories should be used that are shared by the wider community.

CMDI creators

[12] In the current setup of the Component Registry public profiles can no longer be updated. To retain control over the profiles and allow for subsequent updates of the profiles some organizations opt to keep their profiles private.

Accuracy metrics

Accuracy describes the extent to which the information in the in a record provides correct and factual information on the resource being described.

Semantic difference approach

Ochoa et all [Ochoa 2009] propose an accuracy metrics approach calculating the semantic difference between the metadata instance and the resources that contain *textual* information. The method uses Vector Space[13] modeling techniques taken from the domain of Information Retrieval.

$$Q_{accu} = \frac{\sum_{i=1}^{N} tfresource_i * tfmetadata_i}{\sqrt{\sum_{i=1}^{N} tfresource_i^2 * \sum_{i=1}^{N} tfmetadata_i^2}}$$

As an extension to this method the authors propose a Latent Semantic Analysis[14] (LSA) approach to detect words with close semantic relations. This extension assumes a sizeable corpus of text documents to compare against is available. While the Q_{accu} can be directly calculated for a given metadata document and its resource applying the LSA method requires a broader scope.

The Q_{accu} method assumes that there is indeed a direct relation between the resource and that useful text information can be directly extracted from the associated resource, both of which may not always be the case.

Concerning the relation between the metadata and the related resource it is not always clear that the text of a metadata record has a clear relation to the resource content. Consider for example a child language data experiment where children are provided with a number of elicitation tasks. Content words appearing in utterances are not described as part of the metadata

Also, extraction of text can become problematic in cases. For binary data, such as images or audiovisual material no Qaccu can be calculated. Other resources, such as binary tables prove to be problematic. Here only the (column) headers may contain some useful information provided these are not encoded in an idiosyncratic form(e.g. LNG for 'language'). Within CLARIN it is envisaged that these will be associated with data categories so the textual information from the data category's specification can be

[13] http://en.wikipedia.org/wiki/Vector_space_model

[14] http://en.wikipedia.org/wiki/Latent_semantic_analysis

included in the measure. Association of content header fields and data categories is currently not common practice within CLARIN. On the practical side, this also implies that the language in which the metadata is specified is part of the metadata description to ensure that the relevant language section for the data category's definitions, explanations and examples are used. The same consideration applies to files containing part of speech or other linguistic phenomena that are textual by nature but where data content is not reflected in the metadata. Again, only consistent use of the terminology used to describe the phenomena recorded in the data and described in the metadata can be calculated through this metric. Even fully textual resources such as questionnaires do not lend themselves for this type of analysis.

Another type of problem using this metric arises from CLARIN's authorization methods. Since resources may be protected automatic retrieval of resources is not always possible and hence this metric cannot be applied.

For textual resources that do lend themselves for this type of analysis a number of text extraction tools are needed to extract text from document formats such as PDF or Word. Here the CLARIN tool suite may prove to be helpful.

Editing quality

High-quality editing is another aspect mentioned in this context ; the absence of spelling or formatting mistakes in the record.

Bellini/Nesi [Bellini 2013] consider a metadata accurate when:

- There are no typographical error in the free text fields,
- The values in the fields are in the format expected

The derived accuracy metrics is expressed as:

0, if an accuracy issue is detected

Field accuracy g(x)= 1, no problem found

The record accuracy becomes

$$AccR(y) = \frac{\sum_{i=1}^{nField_{Acc}(y)} g(x_i(y)) * w_i}{\sum_{j=1}^{nField_{Acc}(y)} w_j}$$

where *w* is a weighing factor for the i-th field. *AccR(y)* thus produces a weighed average across all fields. This approach requires relevance weights to be determined for each of the metadata fields.

Practical applicability of this metric (*AccR(y)*) proves problematic.

Typographical errors in free text fields

Typographical errors originate from spelling mistakes in (open) data category fields. Typing in a wrong value where a closed vocabulary is intended (closed data categories) is considered to be a consistency problem.

Suspected spelling mistakes may be detected by evaluating the entered text using a spell checker, but this requires the language in which the metadata was written to be known and spell checkers for each of these languages to be available. Even so not all words may be present in the spell checker's dictionary. The word *CLARIN*, for example, is most likely not represented in any standard spell-checking module. A practical implication for CLARIN is that the working language[15] should to be specified for each metadata document and that good spell checkers should become available for each of these languages.

Formatting of metadata fields

Wrong formatting of metadata fields is another reason for accuracy failures in metadata records. Formatting errors in dates or years provide notorious difficulties in attempting to interpret the field's contents. Often these arise from historical encoding principles, e.g. [1650 ca.] to encode that the year 1650 is an approximation. While datetime values should, ideally, be encoded as such at the schema level it is found that they are often represented as strings. A more coherent approach would be to encode approximate datetimes in CLARIN in a uniform manner using the Extended Data/Time Format (EDTF)[16]. Incidentally, if datetime values are encoded as xsd:dateTime at the schema level these errors would show up as consistency problems since the schema would fail to validate against the metadata record.

More problems here are found in non-vocabulary values such as people names or addresses. First, last and full names or street, city and zip codes are often found mixed here. Good practice is needed here since these fields are very error prone and that are potentially repeated across multiple metadata records or data providers.

[15] Working language is the language used to describe objects while object language refers to the language being described.

[16] http://www.loc.gov/standards/datetime/pre-submission.html

Conformance to expectation metrics

Ochoa et all [Ochao 2009] relate conformance to expectation to the ability to find, identify and select a given metadata record. The more discriminating features a metadata record possesses, the easier it will be to find the record. One could challenge the usefulness of having, for example, a single Spanish resource in an otherwise Dutch repository from a repository point of view. However, this resource will clearly stand out if facetted search across languages is provided. As Ocha at all, have demonstrated, there appears to be high agreement between human evaluators and proposed metrics for this aspect. They propose three types of metrics depending upon whether a field is a categorical field or a free text field.

For categorical fields the information content is defined as:

$$\inf ocontent(cat\_field) = 1 - \frac{\log(times(value))}{\log(n)}$$

This (normalized) version counting the number of times the field category value is encountered compared to the full range of possible category values.

For numerical fields Ochoa et all[Ochoa 2006] suggest a similar approach as provided above for category fields under the condition that the values follow a normal distribution.

For free text fields the related to the TFIDF values of the words appearing in the text[17]. To get the information content of a free text field the TFIDF score of each word is added:

$$\inf ocontent(free\_text) = \sum_{i=1}^{N} tf(word_i)\log\frac{N}{df(word_i)}$$

where tf(wordi) is the term frequency of the ith word, df(wordi) is the document frequency of the ith word and N is the number of word in the field.

The total information content of a metadata record is then determined by adding the individual fields' *infocontent* values (the logarithm is applied to reduce the range of the Information Content value):

[17] Some indexing frameworks provide direct support for determining TFIDF scores. One example is SOLR where a TermVector component(http://wiki.apache.org/solr/TermVectorComponent) may provide this information.

$$Q_{t\inf o} = \log\sum_{i=1}^{N}\inf ocontent(field_i)$$

Comment [MKS3]: DIT MAG DUS NIET!!!

One of the problems associated with evaluating the Qtinfo value is that within CLARIN it is fairly difficult to determine whether a field is to be considered a category field or a free text field. Formally, data categories are marked as either open or closed but practice shows that values are found outside the intended value domain of a closed data category or that open data categories behave as a category field. An example of this are *genre* data categories(DC-2470 and DC-6791). This One practical approach is to determine the full range of values encountered in a field and establish a cutoff point beyond which fields are evaluated as free text fields.

Comment [MKS4]: What nu al ser ALLEEN van de tfidf wordt uitgegaan?

Consistency metrics

Ochoa et all [Ochoa 2009] describe three reasons for consistency problems in metadata records:

1. Instances include fields not defined in the standard or do not include fields that the community sets as mandatory
2. Categorical fields, that should only contain values from a fixed list are filled with non-sanctioned values
3. The combination of values in a categorical field is not recommended by the standard definition, i.e. values in different fields show an interdependency relation(*If value in field X is Y then value in field Z must be A*).

Bellini et all[Bellini 2013] provide additional examples such as publication before creation dates, language of the title being different from the object being described or links to digital objects being broken. The latter are considered as a separate Accessibility metric by Ochoa et all (and in this document).

They propose a simple consistency metric to capture these types of problems:

$$brokeRule_i = \begin{matrix} 0; & if\ \ instance\ complies\ with\ ith\ rule \\ 1; & otherwise \end{matrix}$$

$$Q_{cons} = 1 - \frac{\sum_{i=1}^{N} brokeRule_i}{N}$$

Here N is the number of rules that have been applied. Bellini et all[Bellini 2013] further propose to include the relative importance of metadata fields into this equation:

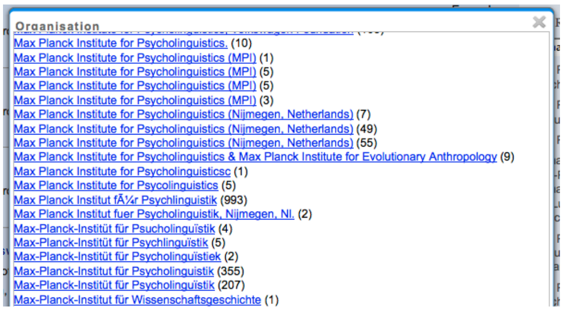
$$ConR(y) = \frac{\sum_{i=1}^{nField_{cons}(y)} h(x_i(y)) * w_i}{\sum_{j=1}^{nField_{cons}(y)} w_j}$$

where *w_i* is the relative importance of the i-th field and *$h(x_i(y))$* is the same as the *brokerule* variable. Again this requires the relative importance of metadata fields to be known. Also, the Q_{cons} version and *ConR(y)* scores move in opposite directions. A high Q_{cons} score indicates good quality, while a high *ConR(y)* scoreindicates lower quality.

Bruce&Hillman[Bruce 2004] propose three tiers of quality indicators, some of which can be interpreted as consistency rules:

- First tier:
    - The ability to validate against a schema
    - The use of appropriate namespace declarations
    - The presence of an administrative wrapper
- Second tier:
    - The presence of controlled vocabularies
    - The definition of elements by a designated community with a publically available application profile
    - Provenance data at a more detailed level
- Third tier:
    - Information on conformance, trust and full provenance information

Usage of wrong values or values outside an intended vocabulary may be a result of handling legacy data or simply bad tool design (offering an open text field rather than a pick list). Organization names are a notorious example of this, as illustrated in the screen shot from the Virtual Language Observatory below(multiple spelling variations of the Max Planck Institute for Psycholinguistics).



Organisation

Max Planck Institute for Psycholinguistics. (10)
Max Planck Institute for Psycholinguistics (MPI) (1)
Max Planck Institute for Psycholinguistics (MPI) (5)
Max Planck Institute for Psycholinguistics (MPI) (5)
Max Planck Institute for Psycholinguistics (MPI) (3)
Max Planck Institute for Psycholinguistics (Nijmegen, Netherlands) (7)
Max Planck Institute for Psycholinguistics (Nijmegen, Netherlands) (49)
Max Planck Institute for Psycholinguistics (Nijmegen, Netherlands) (55)
Max Planck Institute for Psycholinguistics & Max Planck Institute for Evolutionary Anthropology (9)
Max Planck Institute for Psycholinguisticsc (1)
Max Planck Institute for Psycolinguistics (5)
Max Planck Institut fÃ¼r Psychlinguistik (993)
Max Planck Institut fuer Psycholinguistik, Nijmegen, Nl. (2)
Max-Planck-Institüt für Psucholinguïstik (4)
Max-Planck-Institüt für Psychlinguïstik (5)
Max-Planck-Institüt für Psycholinguïstiek (2)
Max-Planck-Institut für Psycholinguistik (355)
Max-Planck-Institüt für Psycholinguïstik (207)
Max-Planck-Institut für Wissenschaftsgeschichte (1)

Another reason for values appearing outside the intended range of a vocabulary stems from situations where the vocabulary does not contain the value entered. While the value itself may be correct is not contained in the vocabulary. One reason for this is that the vocabulary user does not have permission to update the vocabulary itself therefore leaving him/her no option to enter the value manually and hope for the best. Again, organization names may serve as an example here since organization names have been specified as a vocabulary list in CLAVAS as part of an effort by CLARIN to clean up the wide variety of organization names. Organizations not yet represented in that list cannot simply add themselves to CLARIN's organization vocabulary but will have to do this via the vocabulary's owner. One practical way around this, employed by several organizations, is by simple specifying a new, similar data category they control themselves. An example of this in the CLARIN infrastructure may be found by looking at the *genre* data category (http://www.isocat.org/rest/dc/2470 : *The conventionalized discourse or text types of the content of the resource, based on extra-linguistic and internal linguistic criteria.*) describing a closed data category owned by Athens core. An alternative genre specification may be found (http://www.isocat.org/rest/dc/6791 : *A particular style, form or kind of content.*) which is an open data category. Given that the latter data category was created much later than the first it may be assumed that the owner has values outside this proposed list. Judging from the full list of genres presented in the Virtual Language Observatory (screen shot shown below) it appears that other organizations have opted for the strategy to ignore the specified vocabulary and to simply include their own values. An example of this can be found in the VLO under 'HES01-AK' where 'Ritual texts' and 'Religious texts' have been specified as genre values although these do not appear in the data category's value domain.

Coherence

Coherence refers to the degree in which different metadata fields describe the same object in the same way. This is comparable to the Accuracy metrics describing the relation between the metadata fields and the *textual* content of the resource. For the coherence, the semantic distance between the different free text fields is calculated:

$$distance(f1, f2) = \frac{\sum_{i=1}^{N} tfidf_{i,f1} * tfidf_{i,f2}}{\sqrt{\sum_{i=1}^{N} tfidf_{i,f1}^2 * \sum_{i=1}^{N} tfidf_{i,f2}^2}}$$

Where tfidf$_{i,field}$ is the Term Frequncy Inverse Document frequency of the *i*-th word in the textual field f. N is the total number of different words in the field 1 and 2.

Individual semantic distances are then aggregated to yield the coherence measure:

$$Q_{coh} = \frac{\sum_{i}^{N}\sum_{j}^{N}\begin{array}{ll}\sum distance(field_i, field_j); & if\ i < j \\ 0; & otherwise\end{array}}{\frac{N*(N-1)}{2}}$$

Where N is the number of textual fields that describe the object.

Accessibility metrics

One method to determine the accessibility of a metadata record is to record the number of time a record has been retrieved via a search operation.

Ochoa et all[Ochoa 2009] propose two alternative methods that are independent of the search environment. One is by counting the number of links from the metadata record to other records. Links may be explicit, for example through isPartOf relation, or via shared terms in metadata fields such as keywords, authors or genres. The linkage metric is defined as follows:

$$Q_{link} = \frac{links(instance_k)}{\max_{i=1}^{N}(links(instance_i))}$$

Here, *links(instance)* refers to the number of to or from links from a metadata records and *max(links(instancei$_i$)* is the maximum number of links encountered in the repository.

The second method addresses cognitive accessibility, i.e. the degree to which the metadata is easy to understand by end users. They propose to use the Flesch index for determining the readability of a metadata field. Other possible methods to use here are the Flesch-Kincaid grade level, For scale, SMOG index, Coleman-Liau index, Automated Readability index or Linsear Write Formula[18]. The readability metric for the metadata record is then measured by calculating the normalized average of the individual field Flesch indices.

$$Q_{read} = \frac{\sum_{i=1}^{N} Flesch(\, fieldtext_i)}{100 * N}$$

Where N is the number of textual fields and *Flesch(fieldtext$_i$)* represents the Flesch calculation for the i-th field. The Q$_{\text{read}}$ is reported to show a 30% agreement with human annotators.

The formula of the Flesch readability index is:

$$Flesch = 206.835 - 1.015\left(\frac{total \quad words}{total \quad sentences}\right) - 84.6\left(\frac{total \quad syllables}{total \quad words}\right)$$

The Flesch index exhibits a preference for short sentences and short words. The following table is useful to relate the Flesch score to the perceived readability of a text.

90-100 : Very Easy

80-89 : Easy

70-79 : Fairly Easy

60-69 : Standard

50-59 : Fairly Difficult

30-49 : Difficult

0-29 : Very Confusing

Calculating the average quality.

From the proposed CLARIN MQACP process it becomes clear that not all quality metrics can be calculated at any given stage in the process. Quality

[18] Examples of these can be found at: http://www.readabilityformulas.com/free-readability-formula-tests.php

metrics such as the completeness measure (Q_{comp} and Q_{wcomp}) may already be applied in the *Testing* phase. Information content (Q_{tinfo}) on the other hand requires information that is only available from the repository context in which metadata records are embedded. In the *Building Critical Mass* phase this information may be supplied by the CLARIN centers thus providing information on this quality aspect in their repository. For CLARIN assessing quality metrics are most relevant in the *Regular Operation* phase from a context such as the Virtual Language Observatory.

Timeliness metrics

To measure changes of quality over time Ochoa et all propose a timeliness measure. Quality of a set of metadata records may vary as new collections are brought into the repository or existing collections are upgraded. The timeliness measure may capture these variations over time and provide useful feedback to metadata record owners. Within CLARIN this measure is relevant to CLARIN centers during the *Building Critical Mass* phase to monitor fluctuations in metadata quality in their own institutional repositories. For the CLARIN infrastructure this measure becomes relevant during the *Regular Operation* phase each time a new set of CMDI records is harvested. The timeliness measure depends on assessment of the current quality of a metadata record:

$$Q_{curr} = Q_{avg} = \frac{\displaystyle\sum_{i=1}^{N} \frac{(Q_i - \min Q_i)}{(\max Q_i - \min Q_i)}}{N}$$

Where Q_i is the value of the i-th metric and $\min Q_i$ and $\max Q_i$ are the minimum and maxim values of that metric encountered in the repository context. The Q_{avg} is then the average of the different metrics for a metadata record.

The timeliness measure measures changes of this metric over time:

$$Q_{time} = \frac{Q_{curr_{t2}} - Q_{curr_{t1}}}{Q_{curr_{t1}} * (t2 - t1)}$$

Provenance metrics

TODO

Preliminary test results

A number of quality metrics were evaluated against a sample of the latest OAI-harvest of CLARIN records in the VLO. In total, a sample of 1006 records was taken covering all CMDI OAI-PMH end point providers. For each record some, not all, of the metrics discussed above were applied to determine as a first test to yield initial quantitative results and determine the usability of the metrics discussed above. In particular the Qtinfo metric, which according to literature should be an important indicator, can only be evaluated in a full repository context, such as the VLO.

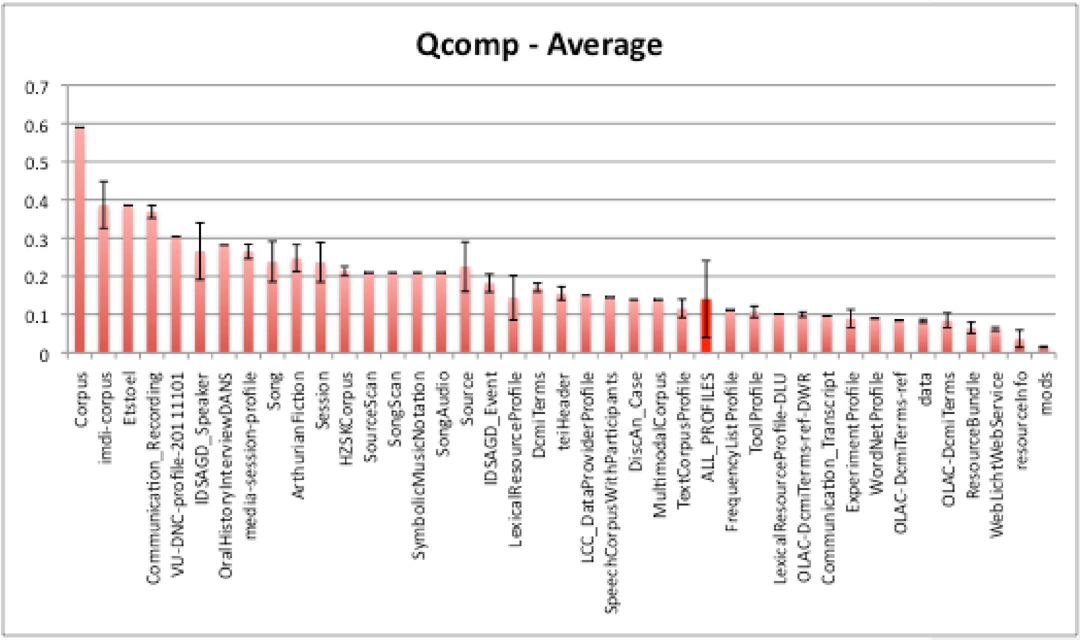
The interested reader displays may find all results in the tables presented below as they are all displayed as in-document Excel objects. Tables are standard sorted in descending order by through the Average column.

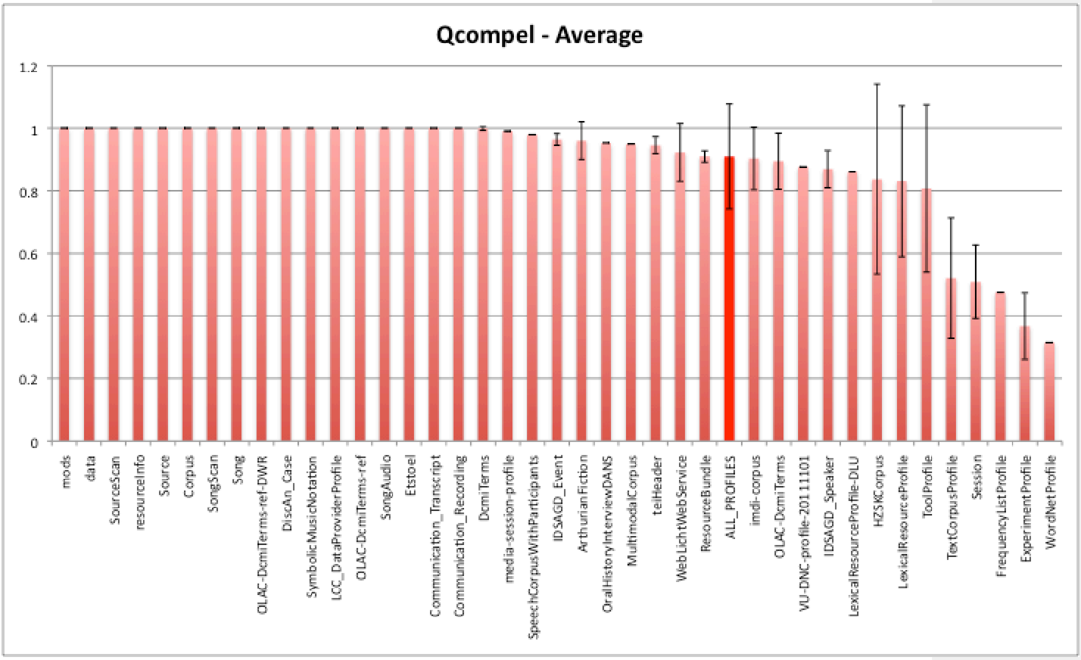
The results of individual CMDI records have been grouped by their profile to obtain an indication of how the different profiles compare to each other.   The standard deviation is also provided. A word of warning here; a standard deviation of 0 may indicate that only a single record from that profile was sampled. In a production environment ***all*** CMDI records would need to be evaluated producing more comprehensive overviews. A large standard deviation however makes the records in a profile set excellent candidates for further inspection as it suggests the overall results of the profile may be raised by addressing the CMDI records with the lower quality indicator values.

Completeness

Completeness checks the amount of fields that have been filled out for a given schema. Schema locations were extracted from each CMDI document's by evaluating the schemaLocation attribute of the document. Most schemas were located at CLARIN's Component Registry, except for 42 CMDI documents. These schemas were located at on a server associated with Das Digitale Wörterbuch der deutschen Sprache. Each schema was evaluated using Meertens' SchemaParser to extract xPath expressions to obtain a reference to the relevant document node.

Since CLARIN promotes the use of elements rather than attributes and since it was expected that quite a large number of unused attributes (such as 'ref') would be present in the CMDI documents two approaches were used. The Qcomp counts both elements and attributes, while the Qcompel only takes elements into account.


Qcomp - Average
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Corpus
imdi-corpus
Etstoel
Communication_Recording
VU-DNC-profile-20111101
IDSAGD_Speaker
OralHistoryInterviewDANS
media-session-profile
Song
ArthurianFiction
Session
HZSKCorpus
SourceScan
SongScan
SymbolicMusicNotation
SongAudio
Source
IDSAGD_Event
LexicalResourceProfile
DcmiTerms
teiHeader
LCC_DataProviderProfile
SpeechCorpusWithParticipants
DiscAn_Case
MultimodalCorpus
TextCorpusProfile
ALL_PROFILES
FrequencyListProfile
ToolProfile
LexicalResourceProfile-DLU
OLAC-DcmiTerms-ref-DWR
Communication_Transcript
ExperimentProfile
WordNetProfile
OLAC-DcmiTerms-ref
data
OLAC-DcmiTerms
ResourceBundle
WebLichtWebService
resourceInfo
mods

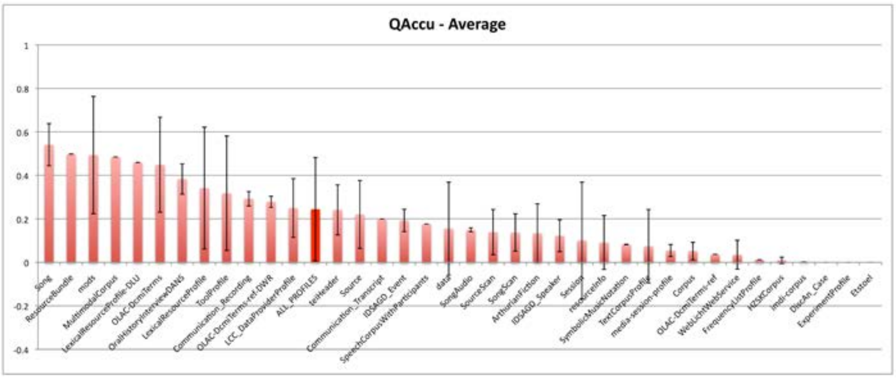

Qcompel - Average
1.2
1
0.8
0.6
0.4
0.2
0
mods
data
SourceScan
resourceInfo
Source
Corpus
SongScan
Song
OLAC-DcmiTerms-ref-DWR
DiscAn_Case
SymbolicMusicNotation
LCC_DataProviderProfile
OLAC-DcmiTerms-ref
SongAudio
Etstoel
Communication_Transcript
Communication_Recording
DcmiTerms
media-session-profile
SpeechCorpusWithParticipants
IDSAGD_Event
ArthurianFiction
OralHistoryInterviewDANS
MultimodalCorpus
teiHeader
WebLichtWebService
ResourceBundle
ALL_PROFILES
imdi-corpus
OLAC-DcmiTerms
VU-DNC-profile-20111101
IDSAGD_Speaker
LexicalResourceProfile-DLU
HZSKCorpus
LexicalResourceProfile
ToolProfile
TextCorpusProfile
Session
FrequencyListProfile
ExperimentProfile
WordNetProfile

As can be seen from figure * the Qcomp completeness of all profiles is around 14%. When only taking element completeness (Qcompel) into account, the situation is much better. A significant portion of the profiles, around 91%, has filled out all available elements. Another observation that can be made here is that more elaborate profiles, such as META-SHARE's *TextCorpusProfile* or *WordNetProfile*, Nalida's *FrequencyListProfile* or IMDI's *Session* profile, tend to be more sparsely filled.

The discrepancy between Qcomp end Qcompel suggests that attributes are seldom used and that maintainers of profiles through the ComponentRegistry could probably do without them to a large extent.

Then figures also suggest that profiles with a large standard deviation such as the *HZSKCorpus* could be subject of further investigation to explain the individual differences between metadata instances.

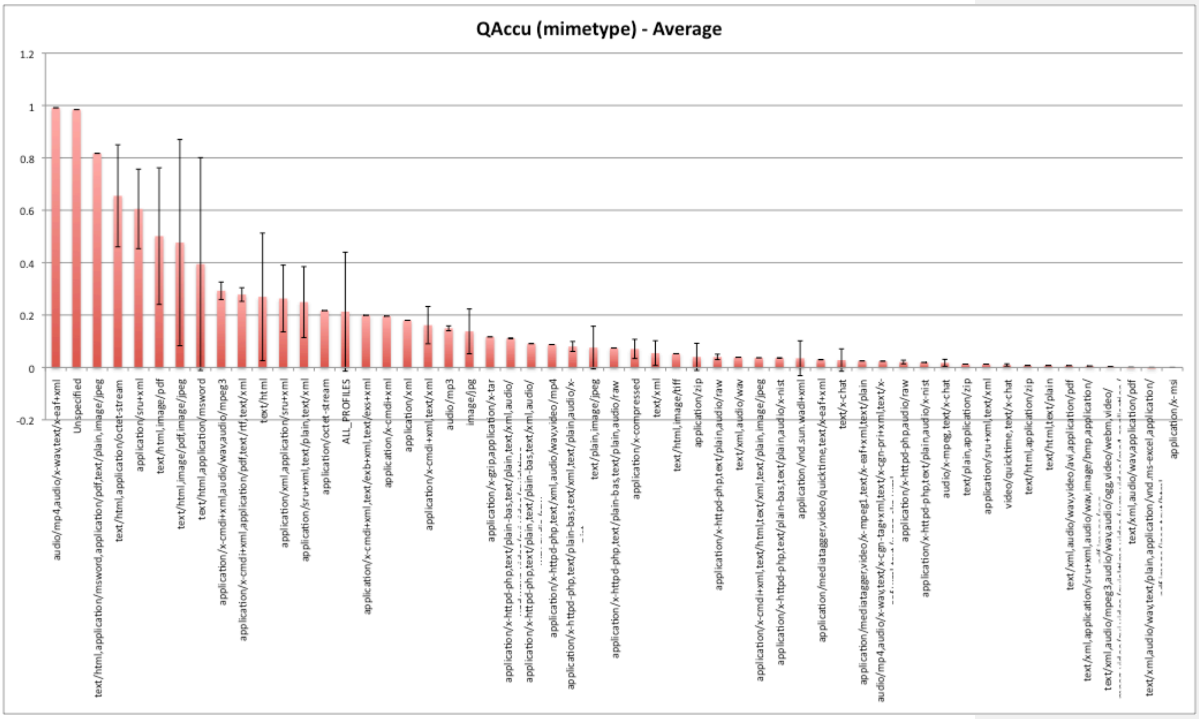
Accuracy



The QAccu measures the semantic distance between the contents of the metadata description overlaps with the contents of the associated resource(s). In the current approach the contents of *all* associated resources is extracted, whenever possible, and combined into a single data block. This includes all other CMDI records, if the metadata record happens to describe a collection.

The overall accuracy across all profiles is around 24%. Preliminary results seem to suggest that the extent to which metadata and content overlap strongly depend strongly the data types of the associated resources as can be seen in the figure below. Care must be taken however when interpreting this figure since these represent combinations of data types and ***not*** individual data types. The number of metadata records found which each of these combinations are often below a statistically relevant threshold as can be observed from the underlying excel sheet. The influence of associated CMDI records and individual data types on QAccu remains to be further investigated.

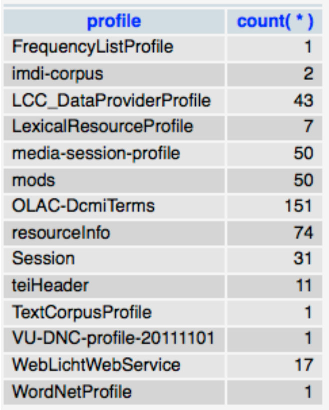
The results suggest that at least some of the expected QAccu behaviour can be observed in the figure shown above. Comparing the two resources at the far end of the spectrum, *Song* and *Etstoel*, the associated *Song* resources contain an HTML page reference largely containing the same information as the CMDI document, while the *Etstoel* metadata records only describing the general characteristics, such as author, time period and region of occurrence of the associated resources.



Logical Consistency

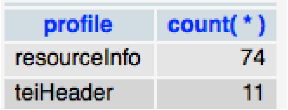
To determine the logical consistency of the CMDI documents each document was validates against its schema and was subsequently evaluated against an example profile. This profile consists of the 14 core CLARIN fields as described in [Trippel 2014]. An additional result from the analysis was that the MdSelfLink was evaluated.

Out of 1006 CMDI documents it was found that 440 did not have an actionable MdSelfLink identifier, 85 of these did not have a MdSelfLink identifier at all. The discussion on whether a PID of a metadata record should be actionable is open for discussion, although the center assessment procedure suggests in its check procedure for persistent identifiers that one should ‘try to resolve a PID for a metadata record’. A missing MdSelfLink however is a clear logical inconsistency.



profile	count(*)
FrequencyListProfile	1
imdi-corpus	2
LCC_DataProviderProfile	43
LexicalResourceProfile	7
media-session-profile	50
mods	50
OLAC-DcmiTerms	151
resourceInfo	74
Session	31
teiHeader	11
TextCorpusProfile	1
VU-DNC-profile-20111101	1
WebLichtWebService	17
WordNetProfile	1

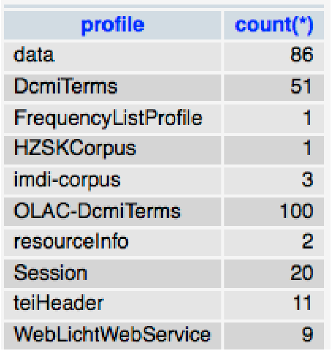
Figure 4: List of profiles with non-resolvable MdSelfLink



profile	count(*)
resourceInfo	74
teiHeader	11

Figure 5: List of profiles with no MdSelfLink

Another consistency check that was performed is whether the CMDI record validates against the schema as expressed in the CMDI document’s *xsi:schemaLocation* attribute. In the processed sample 284 out of 1006 documents could not be validated. Validation of CMDI records against their schema is a CLARIN center requirement so this constitutes a violation of basic CMDI principles.



profile	count(*)
data	86
DcmiTerms	51
FrequencyListProfile	1
HZSKCorpus	1
imdi-corpus	3
OLAC-DcmiTerms	100
resourceInfo	2
Session	20
teiHeader	11
WebLichtWebService	9

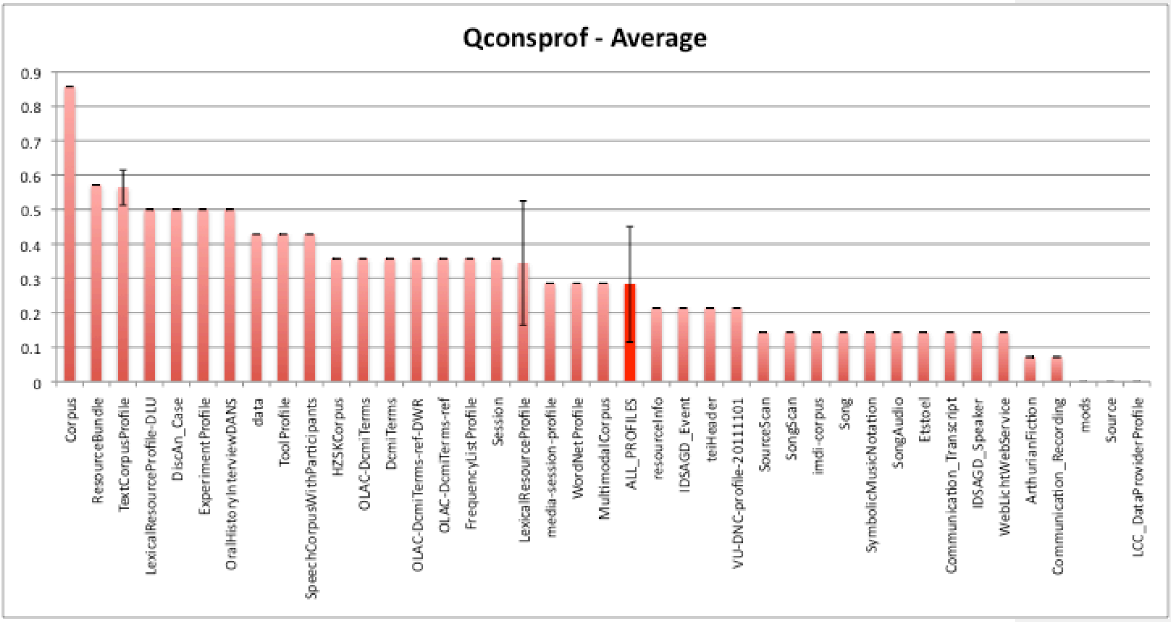
Figure 6: List of profiles with non-validating CMDI records

A further consistency check was performed against the presence of 14 core VLO facet fields as presented by [Trippel 2014]. This provides an insight into the level of conformance to **these** facets. Other facets may be relevant for other communities. For example, from the authors' point of view *modality* is regarded as a core data category. For other communities other data categories may be more relevant. An indication of whether a resource contains annotation layers appears to be relevant for linguistic researchers. Metadata instances can be compared against multiple category lists to determine the suitability of a resource for a specific community.

Core Category	Data category identifier in ISOcat (DC) or Dublin Core
Project name	DC-2536 DC-2537 DC-5414
Resource name	DC-5428 DC-5127 DC-4160
	DC-4114 DC-2544 DC-2545
	DC-6119
	Dublin Core: title
Date indication	DC-2509 DC-2510 DC-2538
	DC-6176
	Dublin Core: created, date, issued
Continent	DC-2531 DC-3791
Country	DC-2532 DC-3792 DC-2092
Language	DC-2482 DC-2484 DC-5361
	DC-5358
Organization	DC-2459 DC-2979 DC-6134 DublinCore:
	Publisher
Genre	DC-2470 DC-3899
Modality	DC-2490

Subject	DC-2591 DC-6147 DC-5316 Dublin Core: subject
Description	DC-2520 DC-6124 Dublin Core: description
Resource class	DC-5424 DC-3806 Dublin Core: type
Format	DC-2571
Keywords	DC-5436

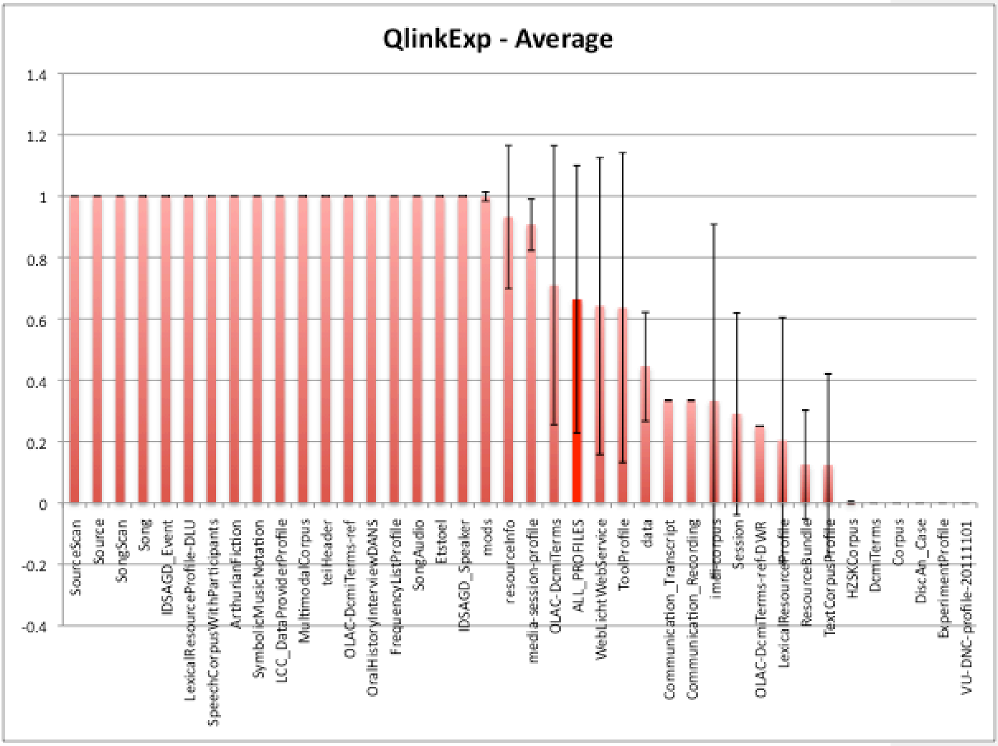
The results for comparing the presence of the data categories presented above are shown below. It appears that the *Corpus* profile is most suitable for the intended purposes of the data category list shown above. *LexicalResourceProfile* and *TextCorpusProfile* appear to have a larger standard deviation compared to the other and show potential for improvement of individual metadata records in this respect.



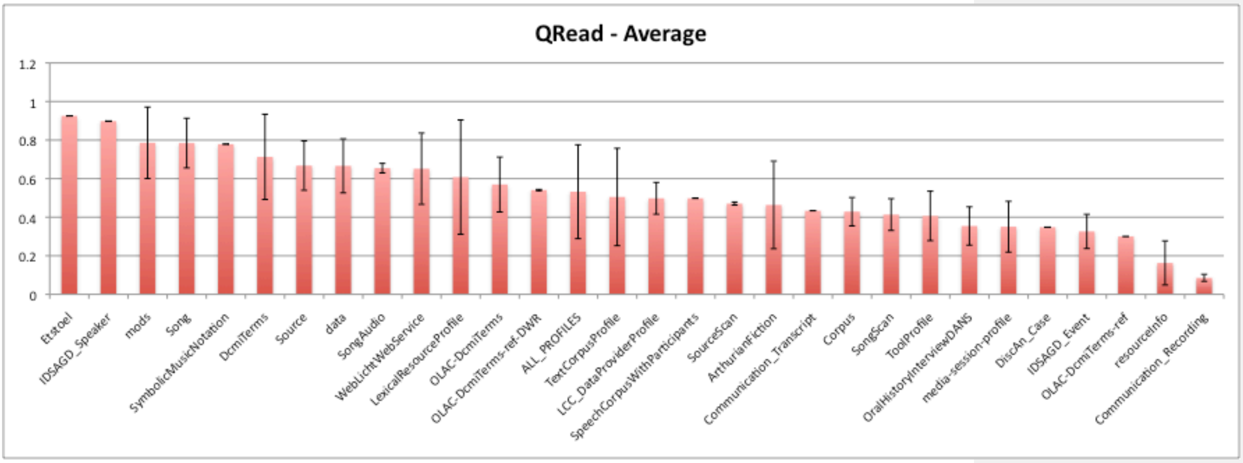
Accessibility

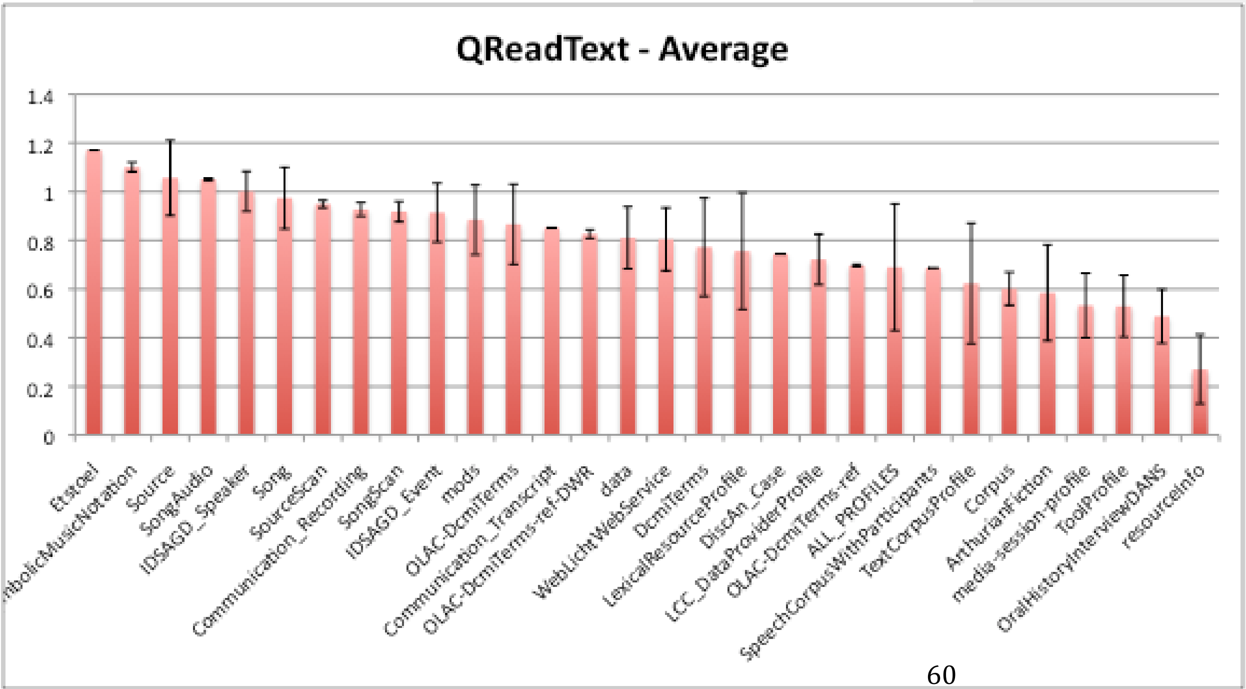
To test the accessibility of the resources the availability of the ResourceRef links was tested (QlinkExp) and the readability of the metadata document was tested in two manners: one by determining the combined Flesch index of all elements (QRead) and one by testing the combined Flesch index of only the text fields.

Preliminary results for QLinkExp are presented below. It appears that around 60% of all links are available. The method for testing the QLinkExp uses a standard HTTP GET` method to try to access the resource. Although the results seem to suggest that not all of the resources are readily available the results may have been influenced by network or latency problems occurring between the test site and data provider site. This remains to be further investigated. Incidentally, availability of the resource links are not currently not listed as a CLARIN center requirement.



The readability of the metadata documents was determined by averaging the Flesch index of each individual metadata field (Qread). Since metadata fields can also contain numerical data types the Flesch index was also determined by evaluating the string type only fields. In general, the latter produce higher values. The results for QRead and QReadText are provided below. The significance of these values their overall effect on perceived metadata quality remains to be further investigated.





Conclusions and recommendations

As a community CLARIN has become more sensitive towards the questions surrounding metadata quality. While the Data Category and Component Registry approaches are seen as vital building blocks in the creation process of high quality metadata, both systems greatly suffer from proliferation of data categories, components and profiles. From the center interviews it has become clear that even for metadata experts working at these centers it becomes increasingly difficult to determine which building blocks to use and to what extent they are shared within a wider community. Here, more knowledge exchange between the metadata experts is definitely welcomed. It has become clear that attempts to correct data after a project has completed is very difficult. Ownership, as in the person responsible, always lies with the participating researcher who most likely will have moved on to other projects. Any attempt to raise the overall quality level of meta records should therefore be incorporated into the project as early as possible. This document proposes a modified MQACP process for CLARIN where different project phases are clearly distinguished and metadata quality checks are performed during the process. These consist of manual feedback from other CMDI experts who are involved at specific steps in during different project phases, but also consist of quantitative analyses of (intermediate) results. Here it is important to stress that any quantitative characteristic that may be extracted from a metadata records or a set of metadata records in itself does not constitute a quality indicator but can only be evaluated with respect to other CMDI records of the same profile or in relation to other profiles. Quality metric results that deviate strongly from previously gathered results for the same profile may serve as a signal. The same holds for large standard deviations of a result set.

As demonstrates with the Qconsprof metric the outcome may predict suitability for a specific end user audience. While the tests in document only focused on one particular set of features reported earlier by CLARIN members several sets may be constructed catering for different end user audiences. As a future possibility it is worth noting that once these metrics have proven their usefulness the results can be incorporated directly into the VLO's indices and be used to guide different end user communities to content that is of more specific interest to them.

To determine the overall quality of a metadata records the end user experience must be taken into account as these metrics have yet to show a clear relation with the end user's verdict. Also, some of the automatically extracted quality indicators can only be assessed as part of the total repository they eventually end up in. For CLARIN this would be the Virtual Language Observatory. In particular this applies to the Qtinfo metric, describing the distinctiveness of the metadata record in a set.

Recommendations for CLARIN:

- The Data Category Registry and Component Registry are the corner stones of the CLARIN infrastructure. It is recognized throughout the CLARIN community that these greatly suffer from proliferation. It is strongly recommended that a cleanup action be carried out on short notice. Relevant sections in this document list some of the data categories, components or profiles that could be evaluated in this context. If cleanup is not feasible on short notice it is recommended that at least the data categories, components and profiles be marked in the Data Category Registry and Component Registry that are considered to be relevant for the CLARIN community as a whole. This will help future users on deciding which profiles, components or data categories to select.
- Invest in early metadata quality assessment methods as part of the metadata creation process.
- Ensure that metadata experts have a direct communication platform with other metadata experts. Involve other metadata experts into the metadata creation process as this helps to reach a next level of convergence for data categories, components and profiles.
- The proposed MQACP process may serve as a template for clarifying tasks and responsibilities and for gathering feedback and quality indicators during the metadata creation process. While such a process may be difficult to enforce at an institutional level it seems feasible to make this part of future CLARIN projects.
- Quantitative quality indicators may provide a complementary set of tools for guiding metadata quality, but have yet to prove their

value. Although it has been possible to extract quantitative information from metadata records it is yet unclear to what extent these match up with end user's opinions on metadata quality. Also, the result of any quantitative indicator should not be interpreted in isolation, but should always be interpreted in relation to the resource it that is being described, the profile that is used and how it compared to similar other metadata descriptions.

References

[Bruce 2004]Bruce, T.R., & Hillmann, D.I. (2004). The Continuum of metadata quality: defining, expressing, exploiting. Metadata in Practice, American Library Association, Chicago. Retrieved from http://www.ecommons.cornell.edu/handle/1813/7895

[Trippel 2014] Trippel, T., Broeder, D., Durco, M., & Ohren, O. Towards automatic quality assessment of component metadata.

[Ochoa 2006] Ochoa, Xavier, and Erik Duval. "Quality Metrics for learning object Metadata." *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006*. 2006.

[Ochoa 2009]Ochoa, Xavier, and Erik Duval. "Automatic evaluation of metadata quality in digital repositories." International Journal on Digital Libraries 10.2-3 (2009): 67-91.

[Bellini 2013] Bellini, Emanuele, and Paolo Nesi. "Metadata Quality assessment tool for Open Access Cultural Heritage institutional repositories." Information Technologies for Performing Arts, Media Access, and Entertainment. Springer Berlin Heidelberg, 2013. 90-103.

[Palavitsinis 2013]Nikos Palavitsinis. "Metadata quality issues in Learning Object Repositories." Doctoral Thesis, November 2013. http://blog.agro-know.com/wp-content/uploads/2014/02/Palavitsinis_Thesis_Final221113.pdf

[Broeder 2014]Broeder, Daan, Ineke Schuurman, and Menzo Windhouwer. "Experiences with the ISOcat Data Category Registry." Ninth International Conference on Language Resources and Evaluation (LREC 2014). 2014

[Schuurman 2013] CLARIN-NL ISOcat experiences. Presentation 2013-12-09 Utrecht. http://www.isocat.org/2013-SR/presentations/clarin-isocat-semreg2013-presented.pdf

[Durco 2013]Durco, Matej, and Menzo Windhouwer. "Semantic Mapping in CLARIN Component Metadata." Metadata and Semantics Research. Springer International Publishing, 2013. 163-168.

[DASISH 2014]Dasish Deliverable D5.2A: Metadata Quality Improvement 1 July 2014

[Zhang 2012] Junte Zhang, Marc Kemps-Snijders, and Hans Bennis. The CMDI MI search engine: Access to language resources and tools using heterogeneous metadata schemas. In P. Zaphiris et al., editor, Proceedings of Theoretic and Practice Digital Li- braries Conference (TPDL 2012), volume 7489, pages 492–495, Berlin / Heidelberg, 2012. Springer.