Curation Face to face meeting January 2018

https://trac.clarin.eu/wiki/Taskforces/Curation/Meetings/2018-01-30

- What?
  - Progress around metadata curation, controlled vocabularies and value normalisation for selected facets
- Who?
  - TF Curation, VLO
  - Attendance sheet
- When?
  - 30 January - 1 February 2018
- Where?
  - organized by: ACDH-OEAW, Vienna, Sonnenfelsgasse 19
  - meeting venue: ÖAW Main building, Dr.-Ignaz-Seipel-Platz 2, 1010 Vienna, 2. Stock Museumszimmer

Agenda

30. 1. Overview

13:00 - 15:00 **State of the affairs**

- VLO
- Value Mapping workflow
- VLO-curation
- Curation module

15:30 - 17:00 **Use Cases**

- Daria & Jakob: Searching in VLO
- Jan: Tools & Services
- ...?

18:30 Dinner & Wine
Singerstraße 28, 1010 Wien

31. 1. Digging into specific facets

Whole day dedicated to hands-on work with controlled vocabularies and value normalisation
Optionally partly divided in (two) smaller groups
Main candidate facets:

- Resource Type
- Availability/Accessibility/Usability/Licensing

If we have capacity we can investigate others like subject or genre

09:00 - 10:30 **Review controlled vocabularies**
review and discuss existing, potential candidates

11:00 - 12:30 **Inspect existing data**

12:30 - 13:30 Sandwiches lunch

13:30 - 17:00 **Do the mapping**
try to map existing data to agreed upon controlled vocabulary (or multiple candidate vocabularies)
(Overnight - apply mappings on the ingested CMD records (in curation VLO)

18:30 Dinner - http://www.xpedit.at/ Wiesingerstraße 6
https://www.google.at/maps/place/Wiesingerstra%C3%9Fe+6,+1010+Wien

1. 2. Review & Conclusions

09:00 - 10:30 **In the aftermath - Review the results**
Investigate the results of the mapping / value normalisation work from yesterday

11:00 - 12:30 **Next steps & tasks**

Minutes

[The following does not strictly follow the order of discussion during the meeting, where we jumped there and back between issues, the notes were reorganized by individual topics.
Specific cases encountered during the meeting are collected in a separate section Curation cases]
There is also a nice summary by Jakob Leonardic as blog-article on CLARIN-EU website.

State of the affairs

VLO

There was a VLO development meeting in Nijmegen/Utrecht on 2018-01-16 - 19 VLO-dev meeting notes

With new harvester viewer there will be integration with curation module.
Working towards a dashboard for the whole harvesting, curation and ingestion process,
Especially also allowing to investigate development over time (remember old results)

VLO-curation - instance with extra features (explicit missing value, more facets, however somewhat outdated data at the moment)

Viewer for the facet-concept mapping (features a profile checker):
https://cmdi.clarin.eu/mapping/index.html#check

Value Mapping workflow

Proposed new workflow for VLO-mapping (See diagram little outdated, or Updated sketch from VLO-dev meeting (open at own risk))

Proposal for VLO value mapping workflow.gdoc
Involves CLAVAS, github, VLO
github: Value mapping scenarios

https://github.com/clarin-eric/VLO-mapping-creator
Simple java-tool converting csv-mapping file to XML format

Curation module

https://clarin.oeaw.ac.at/curate/
=> Weighting of the scoring
=> To used by the Centre Assessment committee

Vocabulary Management

Repositories for publishing: CLAVAS (OpenSKOS) &|vs. acdh vocabs (SKOSMOS)
Management / collaborative editing yet to be defined.
Overview of vocabulary management solutions
One low-tech possibility would be via github - complete provenance/versioning control out of the box. Would be pushed “manually” to CLAVAS upon release.

Proposed maintenance of the vocabulary via github:
https://github.com/acdh-oeaw/VLO-mapping/blob/master/vocabs/resourcetype.csv

Common API for SKOSMOS and OpenSKOS?
=> compare the APIs new OpenSKOS
Or take ELDA on top SPARQL (Menzo tried with SISSvoc)

Use cases

Slovenian

Exploring VLO searching different types of resources: Parallel, Newspaper, Parliamentary, CMC, blog corpora - number of issus encountered Detailed report by Darja & Jakob

Summary of discussion:

- Case: search: “parliament*” vs. “parliament* corpus”
  Different order of results (better results/ranking with broader query) - Why?:
  - Big (verbose) records get penalty due to ranking-algorithm
  - more restrictive query may discard records which don’t use the right keyword (“corpus”)
  - Granularity issue (distinction between individual items and big collections/corpora

- Many findings relate to bad quality of the facets (bad coverage + variability),
  E.g. many corpora are not typed as “corpus”
  => this we are trying to tackle in general, and specifically in this meeting
- Definition of the facets
  - current definitions are available via the tooltips
  - more specific or more broad (esp. country)
- Genre/Subject are a mess, but eventually clear up in the context of a specific collection
  => proposition to make them conditional facets, to only show after first filter (on collection or national project)
- Pavel: Identify and collect Non-english values and let them be translated (integrate translation into the mapping files)
- Pavel: distinguish languages mappable to ISO from the non-standard one
- Pavel: Keyword vs. subject are so mixed (value and definition-wise). Wouldn’t it make sense to collapse them into one facet?

Jan Odijk - Tools/Services/Software

In an email prior to meeting Jan Odijk raised the issue with tools and services not being duly represented in the VLO and shared the work done in CLARIN.NL on a specialized search(mode) with a separate set of facets for this type of resource.

> *An implementation that we can consider a prototype for a solution that can be incorporated (somehow, partially or completely, concretely or conceptually) into the VLO will be available in March. For now we will focus on mapping to a resource type for software/service.*

See also current overview of resource on CLARIN.NL: http://portal.clarin.nl/clarin-resource-list-fs

=> This topic will be partly tackled in the discussion on resource type, but otherwise, agreement was to wait to see, what colleagues in CLARIN.NL will come up with.

Facets / Vocabularies

Discussion on individual facets and the (potentially underlying) controlled vocabularies

Overview of facet values for the usual suspects [as of 2018-01-31]:

facet	count facet values	count uncovered records
resourceClass	**346**	**533289**
genre	1679	1272431
subject	51869	701614
collection	627	0
modality	158	1437841
format	122	38091

In the meeting we concentrate on resourceClass, but it is interesting to see what is the situation with other facets, especially those with semantic or value-domain overlap.

Interesting for comparison statistics on usage of metadata in META-SHARE

- at CELR: https://metashare.ut.ee/stats/top/%7B%%20url%20metashare.views.frontpage%20%20%20%20%20%%7Dstats/usage/#fieldvalues
- The same at ELDA meta-share node: http://metashare.elda.org/stats/usage/

Organisation

Currently applied mapping (old mapping-file format) for organisation: https://github.com/clarin-eric/VLO-mapping/blob/master/uniform-maps/OrganisationControlledVocabulary.xml

=> just for info, not further explored in the meeting, but will be adapted to the new value-mapping approach as well.

Resource Type

Current state of discussion on the vocabulary for RT: trac-wiki: Vocabulary for Resource Type

Some important aspects, questions:

- Decomposition
  `AnnotatedTextCorpus => collection text annotation`
- Is *Corpus* a *Collection*?
- Distinguish *Tool/Software/(Web) Service/(Web) Application*
- Is *speech recording* synonymous to *audio*?

Information about resourceClass facet (in alph-vlo):
solr-query@alpha-vlo: all values of resourceClass facet (json) |346|
solr-query@alpha-vlo: all records that don't have resourceClass field populated (json, vlo) |533289|!
solr-query@curation-vlo: all values of resourceClass facet (json)

Tentative mapping file resourceclass.csv
(features all |346| values encountered in the resourceClass facet, some of them tentatively mapped to the vocabulary)
Tentative mapping file profileName2resourceClass.csv
(features all |100| profiles found in vlo-curation, some of them tentatively mapped to the vocabulary)

HZSK and BAS (and may be others) are using conceptlink on the profile to indicate the resource type: Ccr: recording session concept
=> This could be exploited for filling *resourceClass* facet.

Availability/Licensing

CLARINO-licences overview
2017-06-28 - Curation Meeting on licenses
LicenseAvailibilityMap.gsheet
Current license - availability map
Europeana + DPLA: rightstatements.org
Alternative approach suggestion: Can I access it/Can I use it? (github:vlo#104)

ResourceType vocabulary

trac-wiki: Vocabulary for Resource Type
Vocabulary csv
Mapping csv
Mapping rules and examples: see https://github.com/clarin-eric/VLO-mapping-creator
Solr: http://alpha-vlo.clarin.eu/solr/#/vlo-index

Discussion on the definition of ResourceType

Note: There was agreement in the group that it is not possible to devise one globally valid definition, but rather we try/need to come to a **working definition and make it explicit**, what we mean both by the facet resourceClass and by the individual values. Make this “definition” available to the users via tooltips and accompanying documentation and also discuss (with examples) the implications and possible remedies (e.g. with combination with facet:Format)

We concluded that we are mixing two different meanings in the resourceClass facet. Trying to discern these we came to following two “definitions” (and respective values):

a) the nature of the resource /how the information is encoded depending on the way it is perceived, consumed (abstraction of format)
   (audio, video, text, image, collection, dataset*, tool/service, physicalObject)

b) Kind of digital object resulting from or intended for a language-based research
Kind of information carried and intended use
(corpus, lexical resource, grammar, session, annotation)

Still apply decomposition where sensible.
However the decomposition can have different semantics:

a) A resource features the different aspects
Facsimile of a text: image;text
b) Multiple resources each covering different aspects
Audio file + Text transcription => audio;annotation

How does it relate to:

- format : there is some overlap (esp. to definition part a), but it's much more technical and fine-grained, based on mime/type
=> it would be very interesting and useful to compare values in facets resourceClass and format
- Genre : there are resourcetype like value in genre and vice versa, but this is in error and should be cleaned up, though we did not even try to come up with a definition
- Meta-Share typology
- DataCite definition of Text

Oftentimes the mapping to a normalized value is not valid in general, but only in the context of a specific collection, or currently encountered data. To make this explicit, we decided to introduce field **TF-applicability** that allows to indicate this distinction. It can take three different values:

- *General* : generally applicable mapping, clean normalisation, like spelling variations or unambiguous semantic synonyms
- *Specific* : mapping only applicable based on the currently encountered data; in the long-term there should be a mechanism indicating if new data is contributing to given value and reevaluate if the mapping still holds.
In practice, we have oftentimes very specific values contributed only by one provider, where the danger of shifted semantics among collections is minimal.
- *Not-applicable* : when no mapping is possible. If the target value stays empty, the original value will be ignored by mapping machinery anyhow, "not-applicable" value is only to indicate for the human curators, that the value was looked at and we already decided that it should not be processed.

"Corpus" definitions

http://www.dictionary.com/browse/corpus:

Linguistics. a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis.

=> not very useful for our purposes

Alternative definitions [contributed by Hanna from similar discussion in CLARIN-D(?)]:

Scherer, Korpuslinguistik, Heidelberg 2006, S. 3f.:

*„Ein Korpus ist eine **Sammlung von Texten oder Textteilen, die bewusst nach bestimmten sprachwissenschaftlichen Kriterien ausgewählt und geordnet werden**. Unter Text sind in*

diesem Zusammenhang nicht nur Produkte der Schriftsprache wie Zeitungsartikel, Romane, Kochbücher, E-Mails, Briefe oder Tagebücher zu verstehen, sondern auch mündliche Äußerungen, sei es in Form von Vorträgen, Radiosendungen, Telefongesprächen oder dem zwanglosen Gespräch am Mittagstisch. Die Texte, die in einem Korpus enthalten sind, werden als Primärdaten bezeichnet.
Das Korpus hat den ***Zweck, als Ausschnitt der Sprache zu dienen, die untersucht werden soll****. Dabei ist es wichtig, sich klarzumachen, ob man eine Sprache ganz allgemein untersuchen will, also das Deutsche in seiner Gesamtheit, oder nur eine bestimmte Varietät. […]"*

→ a collection of texts or text parts which have deliberately been gathered and organized with regard to linguistic criteria
→ aims at illustrating the language that is planned to be analyzed

Lemnitzer/Zinsmeister: Korpuslinguistik. Eine Einführung, Tübingen 2010 (2nd ed.), S. 8:
„Ein Korpus ist eine ***Sammlung schriftlicher oder gesprochener Äußerungen****. Die Daten des Korpus sind typischerweise* ***digitalisiert****, d.h. auf Rechnern gesichert und* ***maschinenlesbar****. Die* ***Bestandteile des Korpus*** *bestehen aus den* ***Daten*** *selber sowie möglicherweise aus* ***Metadaten****, die diese Daten beschreiben, und aus linguistischen* ***Annotationen****, die diesen Daten zugeordnet sind."*

→ collection of written or spoken utterances
→ digitized, machine-readable
→ typical components: primary data, metadata, annotations

Glück (Hg.), Metzler-Lexikon Sprache, Stuttgart/Weimar 2000 (2nd ed.):
„Korpus n., pl. Korpora (lat. corpus ‚Körper')
1. ***Sprachl[iche] Daten, die einer sprachwiss[enschaftlichen] Analyse als Grundlage dienen****. K[orpora] spielen u.a. eine zentrale Rolle bei der Erfassung relativ unerforschter Spr[achen] […], aber auch bei der Bearbeitung ausgewählter sprachl[icher] Phänomene.*
2. I. e. S. ***Sammlung einer möglichst hohen, notwendigerweise aber immer begrenzten Anzahl möglichst zusammenhängender sprachl[icher] Äußerungen (gesprochen oder/und geschrieben)*** *aus möglichst natürl[ichen] Kommunikationssituationen. Aufbereitet sind solche K[orpora] wichtige Hilfsmittel z.B. in der Lexikographie, in der maschinellen bzw. maschinengestützten Übersetzung, aber auch bei der Erstellung von Sprachlehrmaterialien."*

→ language data which can be used as basis of a linguistic analysis
→ collection of a huge (though always limited) amount of continuous language utterances (spoken or written)

"Collection" vs "Corpus"
"Corpus" is not equivalent, not even subtype of "collection".
"Collection" is considered any set of items of any type, where the items are principally accessible individually as well.
For a "corpus" there are numerous definitions, but we tend towards a narrow linguistic definition, where a body of text (collected according to some guiding principles) is accessible as one dataset via (ideally linguistic) search means (via a corpus search engine).
There may be a certain overlap, where all the texts of a corpus are accessible individually as well, but this it not the general case.
Pavel (and others): Term "corpus" is oftentimes used in a broad meaning (e.g. also a corpus of images), thus if we aim for higher recall we should adopt the broader definition and consider

“corpus” and “collection” as synonymous.
Hanna, Susanne, Matej: But this would on the other hand confuse and frustrate people who search specifically for a corpus in the narrow meaning. In any case we need to be clear and explicit about our understanding of the term.
Hanna: We miss a clear definition what we mean by *resource type*. There seems to be a discrepancy between generic item level types: text, image, audio, video and language resource specific types like corpus of lexical resource. [=> this has been addressed later, see above section on ResourcType vocabulary]

Observations from the VLO:

- Items that are collections based on their structure (ResourceProxy=Metadata) https://vlo.clarin.eu/search?fqType=_hasPartCount:not_empty
- Items **named** collection that are not a CMDI collection: https://vlo.clarin.eu/search?q=-_hasPartCount:*+name:*collection*
- Items based on the *‘collection’* profile: https://vlo.clarin.eu/search?q=_componentProfile:collection (note: all but 3 of these are CMDI collections)

Tool/Software/Service/Application

Another problematic complex is that of tools, services, software, applications etc.
Observations:

- Matej: There is a clear strong distinction between software and service
  The former is a digital object, structured information about how to algorithmically process data, the other is (according e.g. to CIDOC-CRM:) willingness and ability of an actor to execute specific actions on demand of a client. Which more specifically for an “online” or “E-Service” means some (server) process that is running and can be called and executed without having to install anything, or even having to have access to the underlying software. (Even though every service clearly is “running” on an underlying software)
- One could introduce a general distinction:
  - Web service
  - Web application
  - Desktop application
  - Command-line application
- However we agreed that this would be a skewed to detailed distinction in the resource type categorisation and should be moved to a dedicated facet (as proposed by Jan Odijk)
- Also while this may be a conceptually more sound distinction it does not necessarily mean the definitions are adopted by wide audience.
- Thus for purposes of the general resource type classification, for pragmatic reasons, and also favouring primarily recall in a discovery service, we decided to stay with one generic category “**tool service**” to cover all of the above. (with more detailed distinction left for a dedicated facet)

Value mapping hands-on

In a hands-on session three groups of three people worked on mapping files according to new mapping workflow:

- one group on mapping profileName → resourceClass |100|
- one group on resourceClass → resourceClass (from A → Z) |346|
  one group on resourceClass → resourceClass (from Z → A) |346|

Next to providing mappings agreed upon by a group, one goal of the exercise was also to test and validate the proposed procedure:

1. There are mapping files (created beforehand based on the actual data in the VLO) committed to a dedicated git-repo (VLO-mapping/value-maps)
2. Curator(s) fork given repo and propose their changes (new normalized values) in separate forks in a separate branch
3. When they are finished, they issue a pull request from their fork/branch to the main git-repo
4. This gets picked up by the VLO-maintainers (later probably automatically by the vlo-environment) and is converted from CSV to map.xml according to defined schema by a simple java conversion tool (VLO-mapping-creator)
5. These mapping files in xml format are picked up by the VLO-importer
   (Currently they are injected by hand as part of the configuration, later the conversion to xml and use can be automated)

Following collects observation identified by the curator groups during the exercise:

Profile 2 resourceclass

- The list contains all profiles for which instances where encountered in the (curation) vlo
  Some of these profiles may actually have a valid and functioning mapping to resourceClass facet. This can be checked either via a solr-query or in the curation module, where in the precomputed list of profiles the facet coverage is listed for each profile and each facet.
  => Only profiles actually used and with not facet-mapping should be in this list and attempted to be mapped manually.
  [Matej, 2018-03-04: this was meanwhil implemented, by retrieving from alph-vlo only profiles, where resoureClass facet is empty, query: -resourceClass:* . the corresponding mapping file is: profileName2resourceClass_tf-extended_noResourceClassProfiles.csv]
- Profile vs. instance
  → information allowed by the profile might be missing in the instance
  => This is a well-known issue, also covered by the curation module (separate score for profile and instance facet coverage
- “Song” profile not found in Component Registry, though lots of records in the VLO
  => “Song” profile is not yet published by the author (Meertens Institute)
  How to map it to resourceClass?
  => map to “image;song”, solved?
- “lat-session” allows for video, audio and annotation; but not all records in the VLO include instances of all resource types
  => map to “session”, solved?
- talkbank-license-session: not in the list, not in curation module; ~~“audio;annotation”~~
  => add it to the mapping list; map to “session;audio;annotation”
- “OLAC-DcmiTerms”: multiple resource types possible (audio, text, corpus…)
  => ignore in mapping (in general, generic profiles allowing multiple resource types cannot be mapped to a resourceClass on the profile level; There has to be a field indicating the resource type
- Profile OralHistoryInterviewDANS (p_1369752611610) has element “media type” with value “audio” (concept link: http://hdl.handle.net/11459/CCR_C-2570_6f596a6d-4d5c-f336-d971-e61a310e2c8c)
  → question was: is the concept “media type” used for the resource type facet, and if not, should it? In case of the named profile it would help extracting the resource type from a record

- teiHeader: 3 profiles; ID: 1380106710826 contains “images” as well; ID: 1345180279115 and 1381926654438 is “text” only; ID: 1282306194508 ??

Note: Even if a profile is not published, its definition (in XML or XSD) can be looked up in the component registry. E.g. for clarin.eu:cr1:p_1282306194508: https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1282306194508/xml

Status of work on this task after the common hands-on session: In http://alpha-vlo.clarin.eu/solr/vlo-index/select?facet.field=_componentProfile&facet=on&q=-resourceClass:[*%20TO%20*]&rows=0 we worked from the top and got till (including) “collection”

ResourceClass correction

- https://vlo.clarin.eu/?fqType=resourceClass:or&fq=resourceClass:Zemljevidi
  Zemljevidi => Maps
  => non-english values
- https://vlo.clarin.eu/?q=resourceClass:Virtual_manuscript
  "Virtual_manuscript" => "image;text;manuscript"
  => introduced “manuscript” as further type

OTA & CoCoON - Two main offenders in resourceClass contamination

- resourceClass in Oxford Text Archive | 154 | http://alpha-vlo.clarin.eu/solr/vlo-index/select?q=collection:%22Oxford+Text+Archive%22&facet=true&facet.field=resourceClass&wt=xml&rows=0&facet.limit=-1&facet.sort=text&facet.mincount=1
- resourceClass in COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO) |64| https://vlo.clarin.eu/search?fqType=collection:or&fq=collection:COllections+de+COrpus+Oraux+Numeriques+%28CoCoON+ex-CRDO%29

=> resolved by skipping plain dc:type[not(@schema=’DCMI-Type’)] in OLAC2DCMI conversion: https://github.com/clarin-eric/metadata-conversion/commit/56426933dc1bfee9a9ca36801986ecb5244f2d80 [Menzo]

Concatenation of multiple values in dc:type
https://vlo.clarin.eu/search?q=resourceClass:%22AddressesAnthologiesLinguistic+corpora%22
https://vlo.clarin.eu/data/clarin/results/cmdi/Oxford_Text_Archive/oai_ota_oucs_1733.xml
[Menzo] this problem already exists in the OLAC/DC we get from OTA:
https://vlo.clarin.eu/data/clarin/oai-pmh/Oxford_Text_Archive/oai_ota_oucs_1733.xml
=> problem disappears (at least in VLO) thanks to above correction (skipping of the plain dc:type)
=> still worth notifying provider

1.2.2018

Progress & Summary

Conceptlink in META-SHARE profile added
concept CCR_C-6582_... added to concept mapping and as a concept link to the *resourceType* element of a number of *resourceInfo* and *LINDAT_CLARIN* profiles (metashare), applied to http://alpha-vlo.clarin.eu.

Removed dc:type[not(@DCMI-Type)] elements from OLAC2CMDI mapping
=> reduced number of facet values **from |346| to |138|** #1041

Reviewed and closed a number of old curation issues
People from curation team assigned to curation tickets

Next steps

Continuously process open Curation issues
(and add there new ones, by marking them: *component: Metadata curation*)

Value normalization:

- Generate new resourceClass list (without OTA, Cocoon) (try to add “offending” collection”)
- Profile 2 resourceClass - reduce the list (just those profiles, which don’t have a resource type-facet covered)
- Update the lists in git,
- divide the lists among: 5 persons: Jakob, Susanne, Hanna, Go, Matej
- Try to fill in by 15.2.

Development

- Debug the implementation of the mapping
  Testcase: _componentProfile:teiHeader should be mapped to resourceType:text
  Test-record in vlo@hephaistos (restricted access)
  Profiles currently in vlo@hephaistos
- Update the diagram on the vlo-mapping workflow
  With Updated sketch from VLO-dev meeting in Nijmegen/Utrecht
- Curation module
  - Integrated with new harvester
  - Link to best practices
    https://www.clarin.eu/content/cmdi-best-practice-guide
  - Highlight empty resource proxy
  - Tighten check for CLARIN-B centre

Licensing / AVailability -
Resource proxy - URL checking => availability

Next meeting: 2018-02-21 10:00 telco

Curation Cases

This section collects specific curation cases encountered over the course of the meeting.
To be filed as curation of the MD curation queue.

- Greek Centre not appearing in the VLO (not registered as Centre)
  => https://trac.clarin.eu/ticket/1046
- https://vlo.clarin.eu/record?docId=http_58__47__47_hdl.handle.net_47_10932_47_00-0332-C0A4-B1CF-7A01-1&q=resourceClass:Verbundprojekt&index=0&count=1
- “Verbundprojekt” - actually a corpus
  => resolved, corrected at source upon request (does not appear anymore)
- Unnamed records - Items without value on *name*-facet:
  Query: -name:* | 62407 |
  => https://trac.clarin.eu/ticket/1045#ticket

Need revisit:

- License coverage: 10% ! -license:*
- https://vlo.clarin.eu/record?2&docId=oai_58_rosettaproject.org_58_rosettaproject_piu_detail-2&fqType=languageCode:or&fq=languageCode:name:1978&index=0&count=1
  language=1978
- https://vlo.clarin.eu/record?1&docId=http_58__47__47_hdl.handle.net_47_11858_47_00-203Z-0000-002E-7B30-C&fqType=languageCode:or&fq=languageCode:name:,+,&fq=languageCode:name:1978&index=0&count=2
  language=”, ,”
- https://vlo.clarin.eu/record?2&docId=europeana_58_aggregation_47_europeana_47_92068_47_URN_NBN_SI_IMG_09NERZUK_&fqType=resourceClass:or&fq=resourceClass:Zemljevidi&index=3&count=50
  Dead link: https://www.dlib.si/streamdb/URN:NBN:SI:IMG-09NERZUK/maxi/edm
  Seems that all records in the collection:92068_Ag_Slovenia_ETravel |697| have their edm links broken
- https://vlo.clarin.eu/record?2&docId=oai_58_clarin-pl.eu_58_11321_47_287&fqType=resourceClass:or&fq=resourceClass:ToolService&fqType=collection:or&fq=collection:CLARIN-PL+digital+repository:+CLARIN-PL&index=1&count=67
  ToolService, but actually a SearchPage for a Corpus should be in the ResourceProxy of the data record
- resourceClass:Translation
  https://vlo.clarin.eu/?fqType=collection:or&fq=collection:Archive+of+the+Indigenous+Languages+of+Latin+America |100|
  https://vlo.clarin.eu/record?docId=oai_58_ailla.utexas.edu_58_94&q=resourceClass:Translation&index=1&count=22
  Dead ResourceProxy- link

Records without any resource proxy:

https://vlo.clarin.eu/search?q=resources:0+-_searchPageRef:*+-_hasPartCount:*+-_contentSearchRef:* |20675|

- From CLARIN centres |7703||
- list of offending collections
- https://trac.clarin.eu/ticket/1038#

ATILF resources - no resource proxies
Example:
https://vlo.clarin.eu/record?1&docId=oai_58_atilf.inalf.fr_58_M277&fqType=resourceClass:or&fq=resourceClass:Poetry&index=1&count=2
https://vlo.clarin.eu/record?docId=81420adb-9b20-4955-981d-8e8857ddb84b&q=resourceClass:Trilogue&index=0&count=1
No resourceProxy!!
No resource proxy link:
https://vlo.clarin.eu/data/clarin/results/cmdi/ORTOLANG_Repository/oai_ortolang_fr_81420adb_9b20_4955_981d_8e8857ddb84b.xml
=> will be removed (not ingested)

Collections with no resource type

https://vlo.clarin.eu/?q=-resourceClass:*

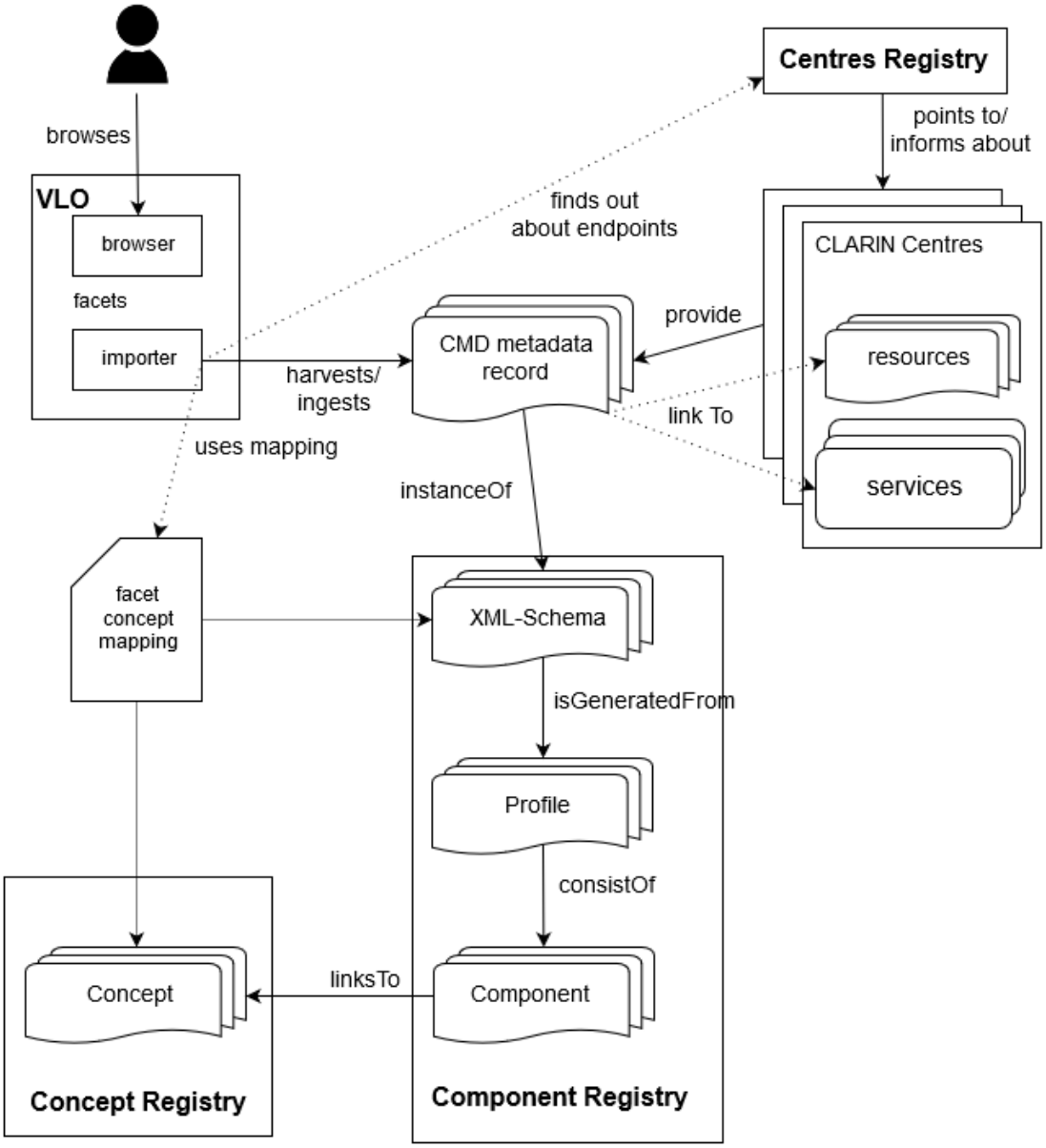
- Meertens collection: Liederenbank (243129)
- TalkBank (65593)
- Bavarian Archive for Speech Signals (BAS) (26654)
- Wolfenbuettel Digital Library: Deutsche Digitale Bibliothek (21435)
- CLARIN Centres (20187)
- Institut für Deutsche Sprache, CLARIN-D Zentrum, Mannheim (17414)
- MPI CGN (12767)

Curation module

- https://lindat.mff.cuni.cz/repository/oai/cite?metadataPrefix=cmdi&handle=11234/1-2605
  Not displaying the mapping in the curation module
- HZSK-record
  Many Validation errors - still not reflected in the score
  xml-validationERRORline: 1, col: 633 - cvc-elt.1: Cannot find the declaration of element 'cmd:CMD'. ??
- 0 REsourceProxy more visible

Appendix - CMDI

http://clarin.eu/cmdi
https://www.clarin.eu/content/cmdi-best-practice-guide
Component Registry used for collaborative definition and publication of CMDI profiles and components:
https://catalog.clarin.eu/ds/ComponentRegistry/


Centres Registry
points to/
informs about
browses
VLO
browser
facets
importer
finds out
about endpoints
CLARIN Centres
provide
CMD metadata
record
resources
harvests/
ingests
link To
services
uses mapping
instanceOf
facet
concept
mapping
XML-Schema
isGeneratedFrom
Profile
consistOf
Concept
linksTo
Component
Concept Registry
Component Registry