CLARIN
Common Language Resources and Technology Infrastructure

CMDI Usage, profiles & components

Dieter Van Uytvanck
Max Planck Institute for Psycholinguistics
Dieter.VanUytvanck@mpi.nl
CMDI compatibility workshop, Utrecht
2013-06-04

CMDI providers

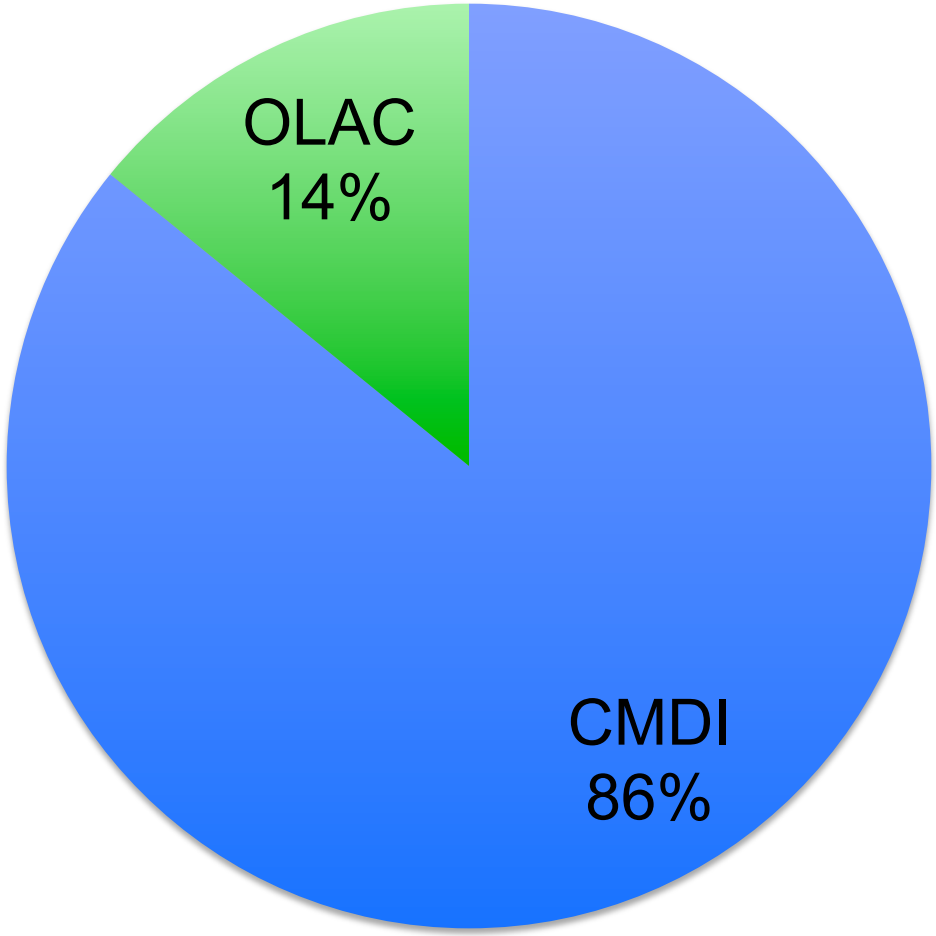
- Currently, about 14 CMDI-providers: see Center Registry:
  - https://centerregistry-clarin.esc.rzg.mpg.de/
- All CLARIN B-centres (currently about 20 candidates) will provide it
- Up to now:
  - 124 profiles
  - 696 components

Conversion workflows

- www.clarin.eu/faq-page/274

Overview (2013-06-03)

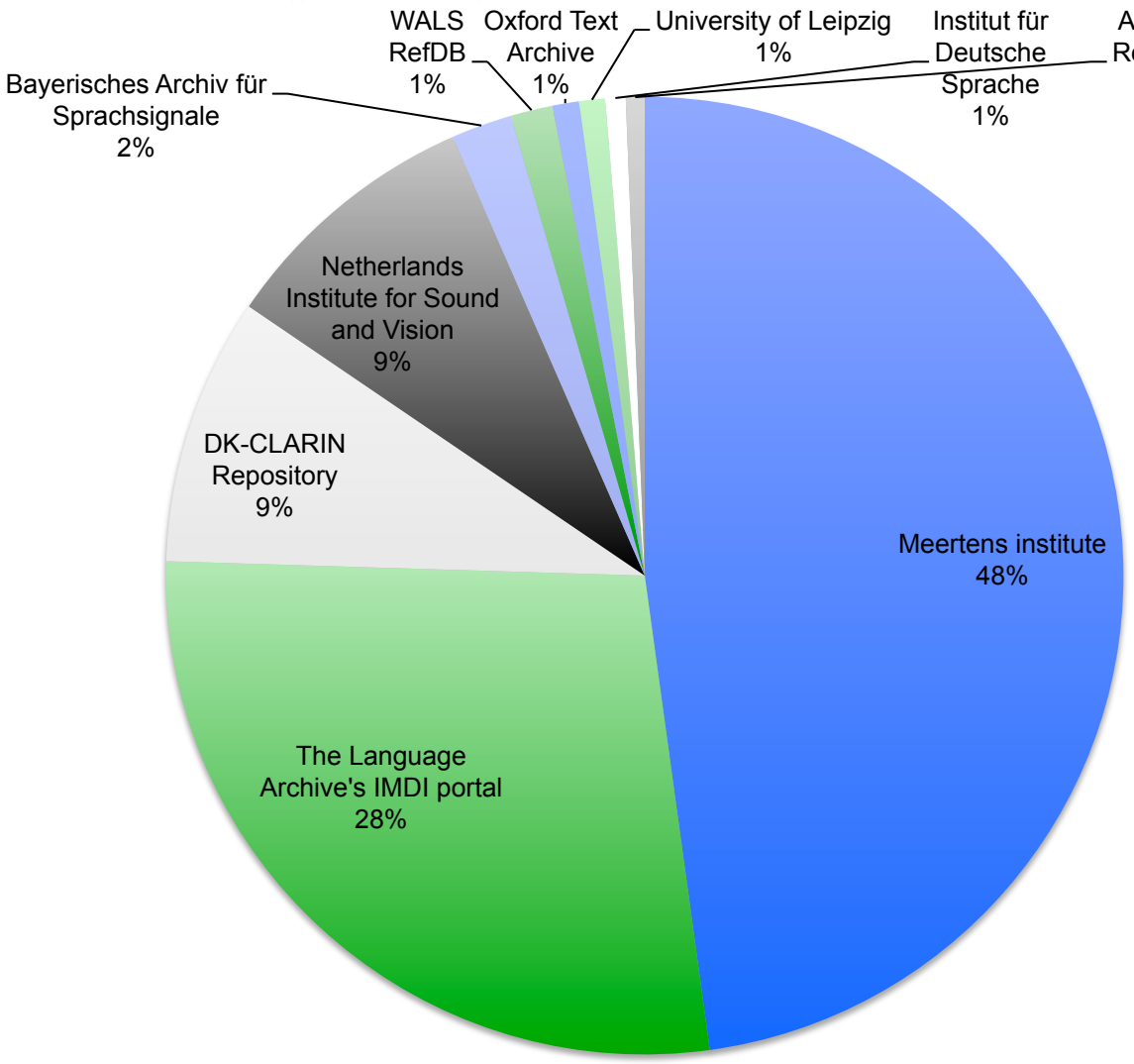
- CMDI (445.000) vs OLAC records (74.000)



Profiles used (OLAC)

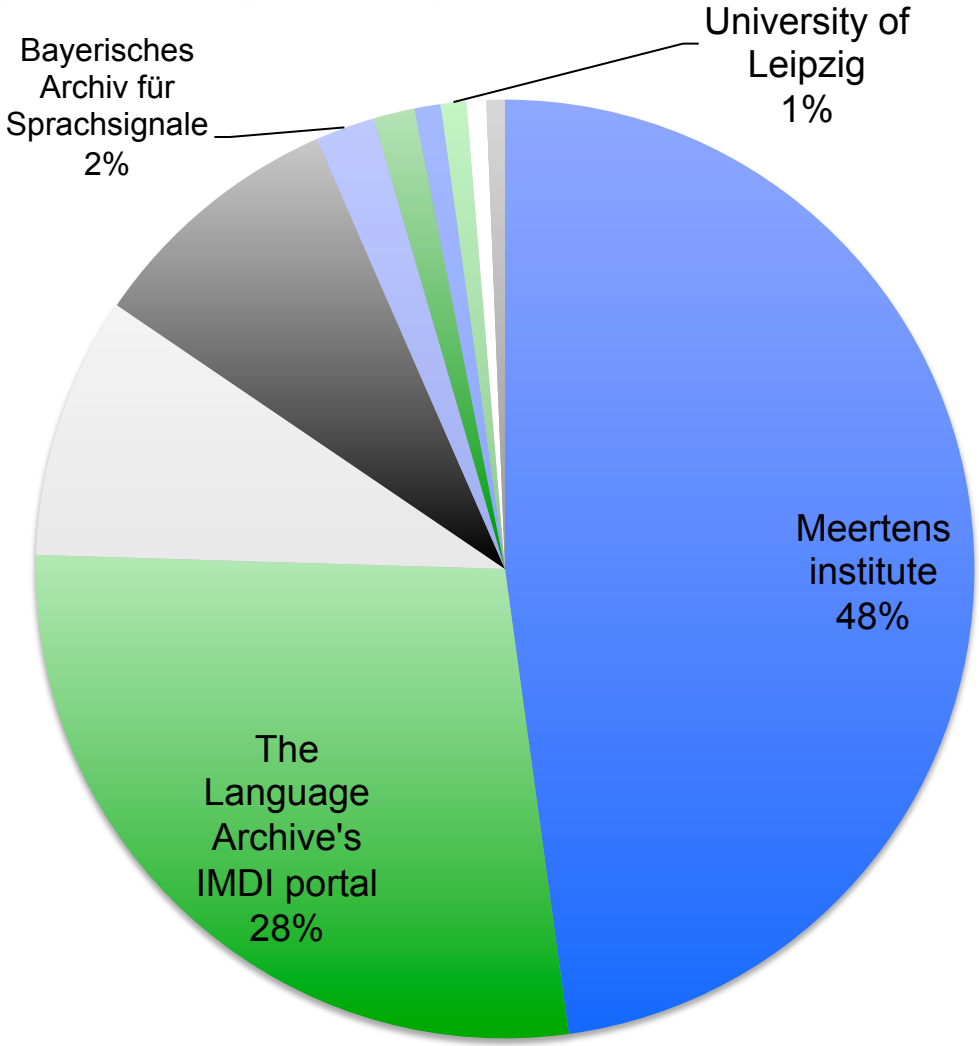
- standard for OLAC/DC records - 1 single profile: OLAC-DcmiTerms
  - http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1288172614026
- Some are using a pure DC profile, DcmiTerms (eg Sound and Vision)
  - http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1288172614023

top 10 MD providers (# recs)



99.6% of all records come from 10 providers

top 10 MD providers (# recs)



- 21% based on DC/OLAC
- 79% of these are based on pure CMDI profiles (no DC derivates)
- Source: http://catalog.clarin.eu/oai-harvester/

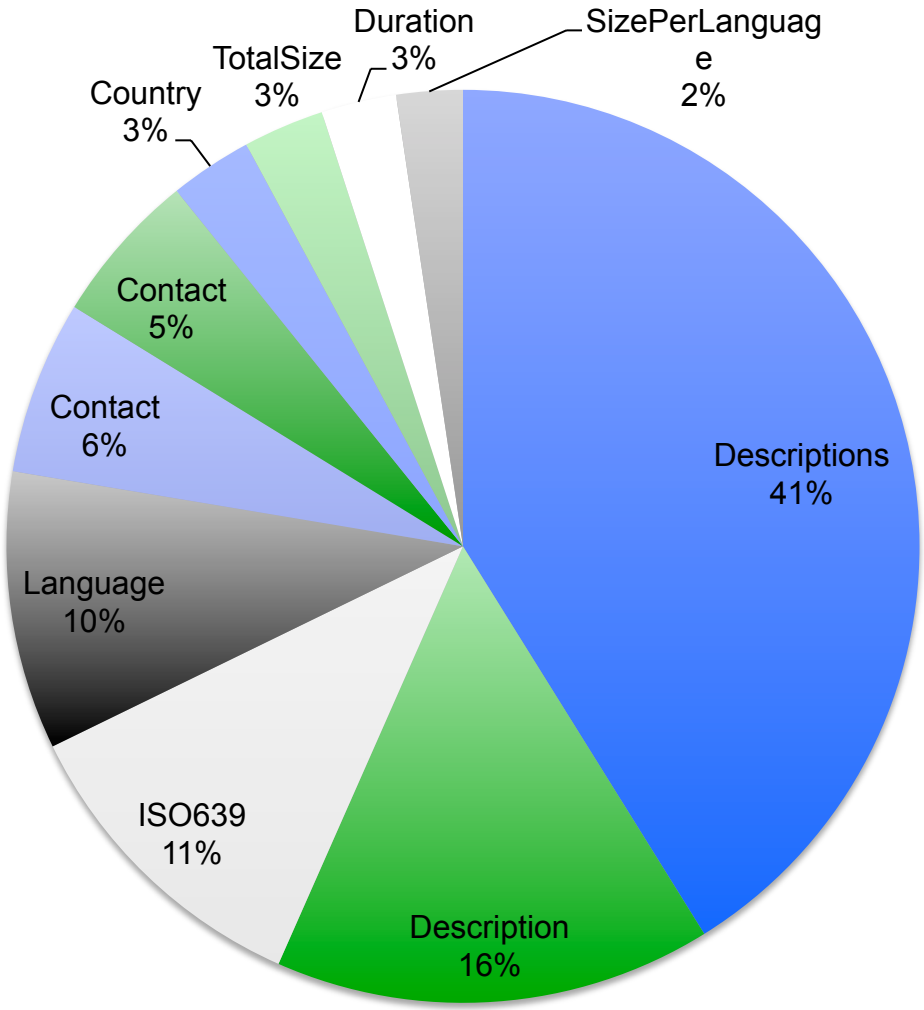
Most-used profiles

- Meertens:
  - Liederenbank (98% of records): SongScan, SourceScan, Source, SongAudio, Text [unpublished]
  - Soundbites (0.8%): Soundbites-recording [unpublished]
- MPI:
  - imdi-session („leaf“ – 90%): http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1271859438204
  - imdi-corpus („collection“– 10%): http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1274880881885
- BAS: media-session-profile:
  - http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1336550377513

Most-used profiles

- Leipzig (some duplicates – older versions?):
  - LCC_CorpusProfile: http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1360931019822
  - LCC_DataProviderProfile: http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1360931019821
  - LCC_DataSourceFile: http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1360931019820
  - LCC_SentenceProfile: http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1360931019819

Top 10 components



- Popularity by usage in profiles
- Source: http://clarin.aac.ac.at/smc-browser/smc_stats.html

Top 20 components

- Descriptions 789
- Description 297
- ISO639 214
- Language 189
- Contact 118
- Contact 104
- Country 56
- TotalSize 55
- Duration 51
- SizePerLanguage 45
- Country 44
- DeploymentToolInfo 43
- Size 41
- CollectionType 39
- Continent 35
- Location 35
- Creator 34
- DocumentationLanguages 32
- Price 32
- Creators 30

Conclusions

- Almost all metadata in the VLO (500k records) is based on less than 20 profiles
- Most used:
  - OLAC-DcmiTerms and DcmiTerms
  - Liederenbank
  - Imdi-session
- Need tools to analyse the combination of profiles/ components and instances