ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

Curation module in action - its preliminary findings on VLO metadata quality

Davor Ostojić, Go Sugimoto, Matej Ďurčo
(Austrian Centre for Digital Humanities)
CLARIN Annual Conference 2016, Aix-en-Provence, France
2016-10-28

Outline

1. Why Curation Module?
2. What is Curation Module?
3. How does it work?
4. What can we find out?
5. Into the future

1. Why Curation Module?

- It's all about VLO (and VLO is about resource discovery)
- CMDI complexity and flexibility
- Impact on Metadata Quality
- VLO problems for resource discovery
- Trippel, et al. (2014), Kemps-Snijders (2014), ACDH team (2015), King et al. (2015), Odijk (2014, 2015)

What we know by now:

- How many CLARIN centers? (Center Registry) - **38**
- How many records in VLO? (VLO search) - **913629**
- How many records harvested? (CMDI harvester) - **881268**
- How many (public) profiles / components (CompReg) - **194 / 1207**
- How many concepts (CCR) - **3160**
- What structure and stats of CMDI metadata? (SMC browser)
- What percentage of facets covered? (Odijk 2014, King et al 2015)

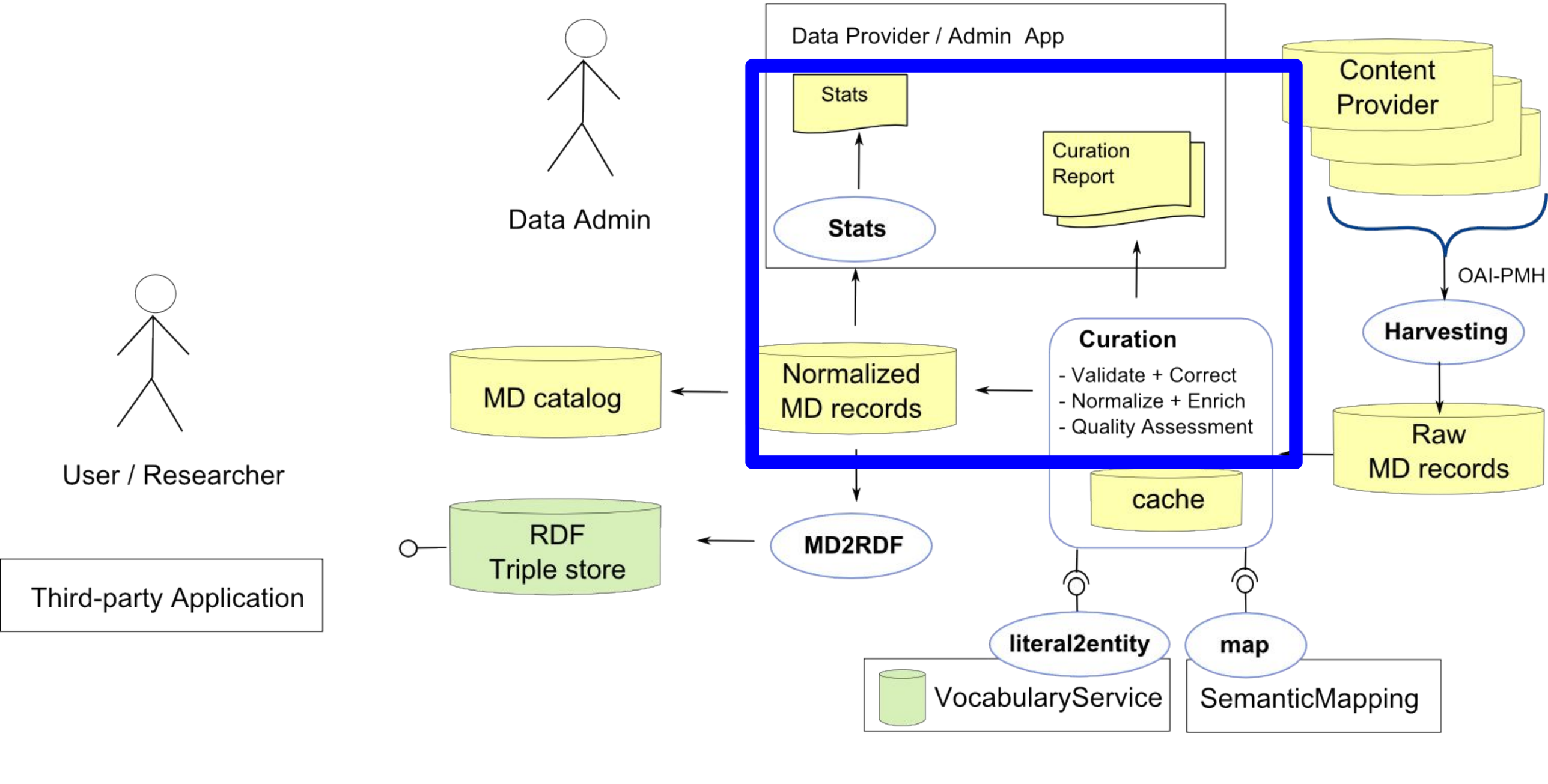
2. What is Curation Module?

- tool for curation, normalisation and quality assessment / benchmarking of CMD records, collections and profiles.
- automatically and systematically collects statistics about metadata quantity and quality
- Web Application, RESTful API, and Java library

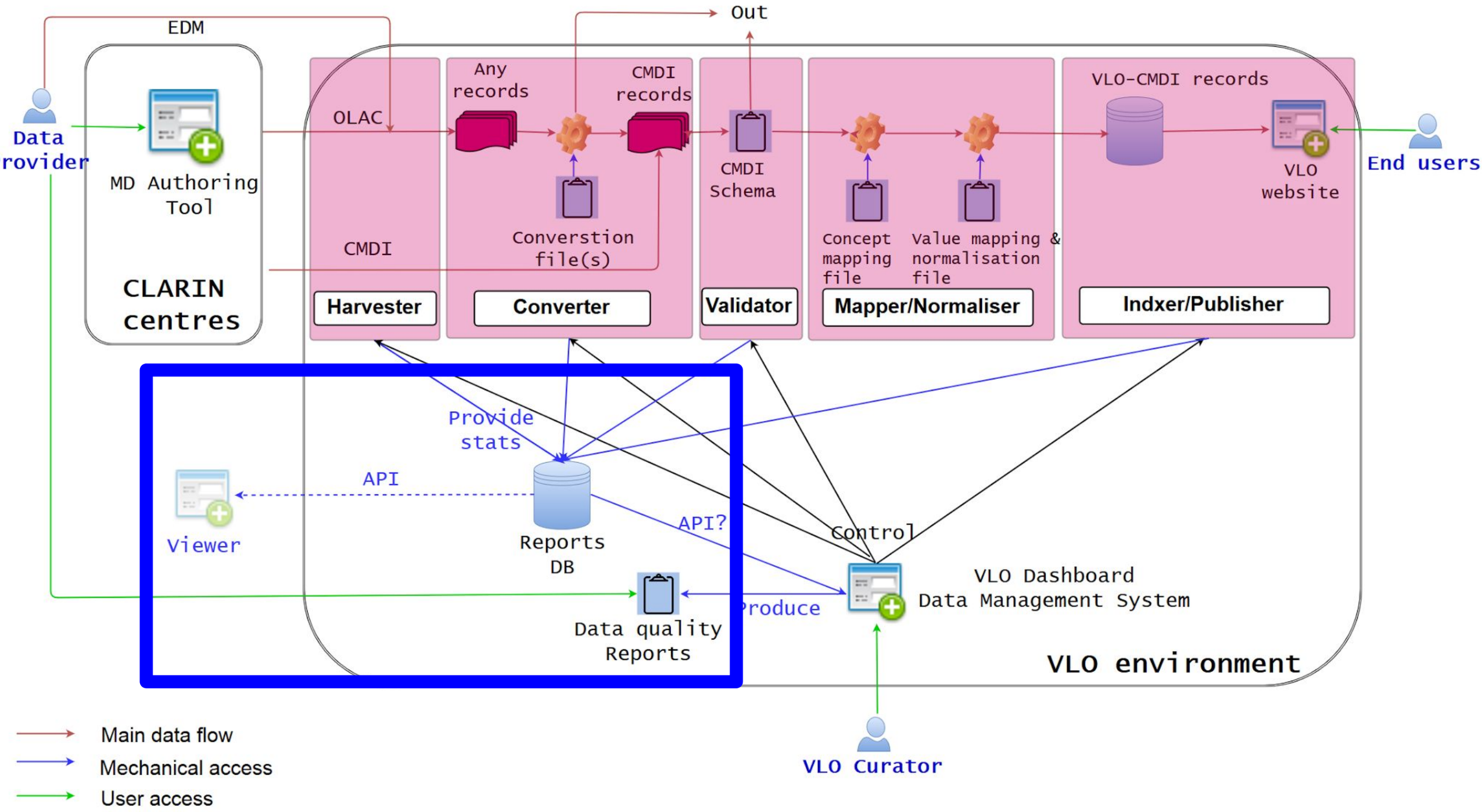
https://clarin.oeaw.ac.at/curate
(https://clarin.oeaw.ac.at/vlo)

Use cases

1. **Metadata author** can validate and assess the quality of a new record
2. **Metadata modeler** can assess the quality of profiles
3. **Repository administrator** can assess the quality of his repository
4. **Metadata curators** can investigate additional aspects of CMDI (beyond VLO facets)
5. Integration in VLO workflow



Initial Curation Module concept:
Durco, Matej, and Karlheinz Mörth. 2014. "Towards a DH Knowledge Hub-Step 1: Vocabularies." Clarin Conference 2014.



Curation Module in the long-term **Dashboard** and its workflow (King et al, 2016)

3. How does it work

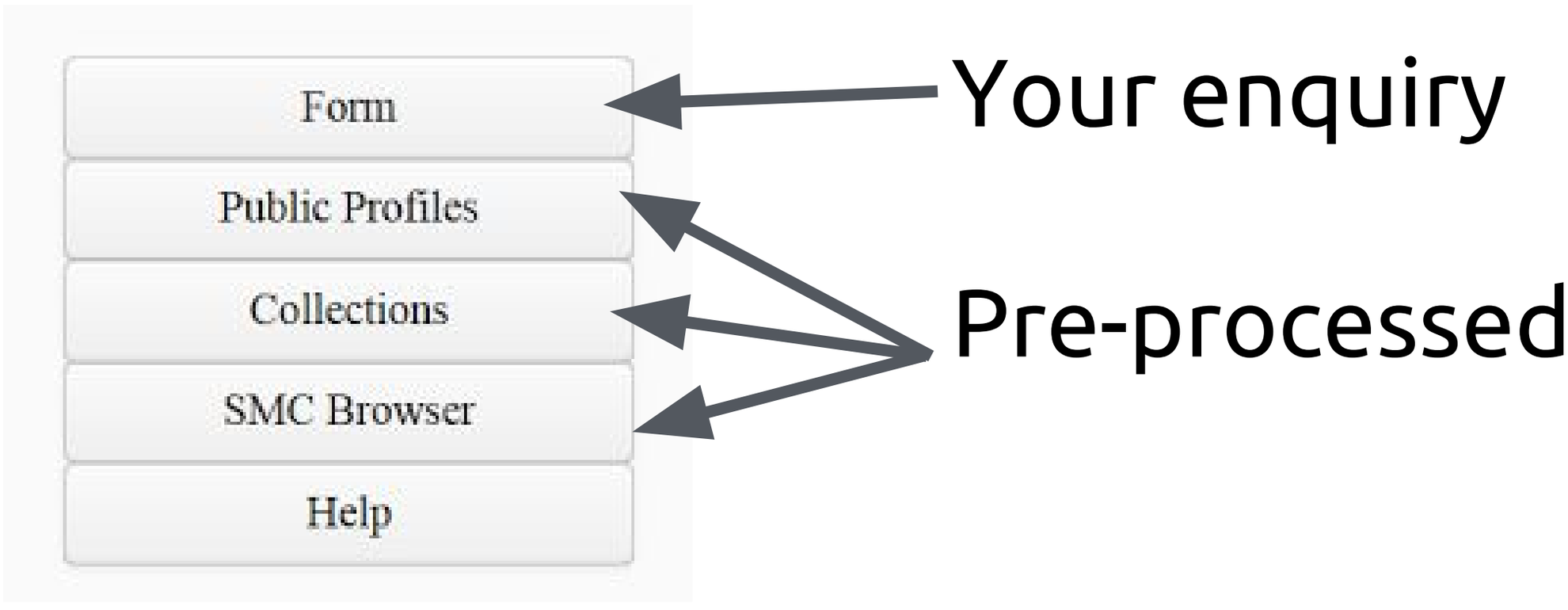
Curation Module

ÖAW ACDH

CLARIN+

Form
Public Profiles
Collections
SMC Browser
Help

Id	Name	Score	Facet Coverage	Perc Of Elements With Concepts	SMC
clarin.eu:cr1:p_1345180279123	HZSKCorpus	2.564	0.923	0.64	explore
clarin.eu:cr1:p_1422885449343	SpokenCorpusProfile	2.528	0.923	0.605	explore
clarin.eu:cr1:p_1387365569699	media-corpus-profile	2.525	0.885	0.641	explore
clarin.eu:cr1:p_1381926654446	ROE	2.518	0.769	0.749	explore
clarin.eu:cr1:p_1375880372947	LESLLA	2.515	0.769	0.746	explore
clarin.eu:cr1:p_1324638957739	media-corpus-profile	2.512	0.885	0.627	explore



Record 37 of 17,260

< previous next >

mona034

HUMANITIES

Record details | Resources (1) | Availability | All metadata | Technical details

Self link	http://hdl.handle.net/11312/c-00030129-1
ID	hdl_58_11312_47_c-00030129-1
Data provider	CLARIN Centres
Last seen	10/18/16
Profile name	talkbank-session
Metadata source	http://vlo.clarin.eu/data/clarin/results/cmdi/CHILDES/oai_childes_talkbank_org_childes_French_York_Max_mona034.xml
Hierarchy level	0

Curation Module

ÖAW ACDH

CLARIN+

Form
Public Profiles
Collections
SMC Browser
Help

http://vlo.clarin.eu/data/clarin/results/cmdi/CHILDES/oai_childes_talkbank_org_childes_French_York_Max_mona034.xml

Asses Instance

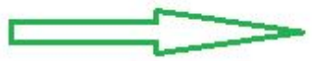
URL or ID of a profile

Asses Profile

CMD Record Report

CMD Record: http://vlo.clarin.eu/data/clarin/results/cmdi/CHILDES/oai_childes_talkbank_org_childes_French_York_Max_mona034.xml

profileID: clarin.eu:cr1:p_1393514855466



Assess profile

file size: 10657 B

score-section

segment	score	max
file-size	1.0000	1.0000
profiles-score	1.6075	3.0000
cmd-header-schema	5.0000	5.0000
cmd-res-proxy	2.0000	2.0000
xml-validation	0.7483	1.0000
url-validation	0.0000	1.0000
facet-mapping	0.7308	1.0000
instance: 9.4791 total: 11.0865 max: 14.0		

resProxy-section

total number ResourceProxies: 1

number of ResourceProxies having specified MIME type: 1

percent of ResourceProxies having specified MIME type: 1.0000

number of ResourceProxies having reference: 1

percent of ResourceProxies having reference: 1.0000

resource type	count
Resource	1

xml-validation-section

number of XML elements: 191

number of simple XML elements: 147

number of empty XML elements: 37

percentage of populated XML elements: 0.7483

url-validation-section

number of links: 1

number of unique links: 1

number of links in resourceProxy reference 0

number of broken links 1

percentage of valid links XML elements: 0.0000

Facet section

* - facet is derived

Facet	Value	Normalised Value	Concept	XPath
continent	Europe		http://hdl.handle.net/11459/CCR_C-2531_e0427265-2fc4-d23e-0a7c-a21981ec3734	/CMD/Components/talkbank-session/Session/MDGroup/Location/Continent/text()
country	France		http://hdl.handle.net/11459/CCR_C-2532_d004b0a6-fd1d-3ca3-abf1-1e6aeb3e37b2	/CMD/Components/talkbank-session/Session/MDGroup/Location/Country/text()
modality	speech		http://hdl.handle.net/11459/CCR_C-2490_44bc38a3-1799-4149-c791-40ac0176f0ff	/CMD/Components/talkbank-session/Session/MDGroup/Content/Modalities/text()
availability	open access	PUB	http://hdl.handle.net/11459/CCR_C-2453_1f0c3ea5-7966-ae11-d3c6-448424d4e6e8	/CMD/Components/talkbank-session/Session/Resources/WrittenResource/Access/Availability/text()
languageCode	ISO639-3:fra	code:fra	http://hdl.handle.net/11459/CCR_C-2482_08eded24-4086-7e3f-88e5-e0807fb01e17	/CMD/Components/talkbank-session/Session/MDGroup/Content/Content_Languages/Content_Language/Id/text()
license	missing value			
name	mona034		http://hdl.handle.net/11459/CCR_C-2544_3626545e-a21d-058c-ebfd-241c0464e7e5	/CMD/Components/talkbank-session/Session/Name/text()
projectName	Child Language Data Exchange System		http://hdl.handle.net/11459/CCR_C-2537_fa206273-223a-f4fa-dde3-ba59b965701f	/CMD/Components/talkbank-session/Session/MDGroup/Pro/Title/text()
	CHILDES		http://hdl.handle.net/11459/CCR_C-2536_13fc5f10-c14a-1f64-a669-32736f6d3ef5	/CMD/Components/talkbank-session/Session/MDGroup/Pro/Name/text()
nationalProject	TalkBank			/CMD/Header/MdCollectionDisplayName/text()
_languageName*	code:fra	French		
format	missing value			
rightsHolder	not covered by profile			

total: 26 coveredByInstance: 19 instanceCoverage: 0.7308 coveredByProfile: 20 profileCoverage: 0.7692

Issues

segment	severity	message
xml-validation	WARNING	Empty element <cmd:JournalFileProxyList> was found on line 6
xml-validation	WARNING	Empty element <cmd:ResourceRelationList> was found on line 6
xml-validation	WARNING	Empty element <cmdp:Region> was found on line 16
xml-validation	WARNING	Empty element <cmdp:Address> was found on line 17
xml-validation	WARNING	Empty element <cmdp:Name> was found on line 158
xml-validation	WARNING	Empty element <cmdp:Address> was found on line 159
xml-validation	WARNING	Empty element <cmdp:Email> was found on line 160
xml-validation	WARNING	Empty element <cmdp:Organisation> was found on line 161
xml-validation	WARNING	Empty element <cmdp:Keys> was found on line 163
xml-validation	WARNING	Empty element <cmdp:MediaResourceLink> was found on line 178
xml-validation	WARNING	Empty element <cmdp:ContentEncoding> was found on line 186
xml-validation	WARNING	Empty element <cmdp:LanguageId> was found on line 187
url-validation	ERROR	URL: http://childes.talkbank.org/data-orig/French/York/Max/mona034.cha STATUS:java.net.SocketTimeoutEx
facet-mapping	INFO	Normalised value for facet availability: 'open access' into 'PUB'
facet-mapping	INFO	Normalised value for facet languageCode: 'ISO639-3:fra' into 'code:fra'
facet-mapping	INFO	Normalised value for facet _languageName: 'code:fra' into 'French'
facet-mapping	INFO	Ignored value for facet license: 'open access'. This value will be removed from mapping

Id	Name	Score	Facet Coverage	Perc Of Elements With Concepts	SMC
clarin.eu:cr1:p_1455633534543	DGDCorpus	2.978	1	0.978	explore
clarin.eu:cr1:p_1456409483189	DGDEvent	2.934	0.962	0.972	explore
clarin.eu:cr1:p_1290431694580	TextCorpusProfile	2.919	0.923	0.995	explore
clarin.eu:cr1:p_1290431694579	LexicalResourceProfile	2.918	0.923	0.995	explore
clarin.eu:cr1:p_1422885449343	SpokenCorpusProfile	2.916	0.962	0.955	explore
clarin.eu:cr1:p_1290431694581	ToolProfile	2.909	0.923	0.986	explore
clarin.eu:cr1:p_1387365569699	media-corpus-profile	2.888	0.923	0.965	explore
clarin.eu:cr1:p_1302702320451	ExperimentProfile	2.881	0.885	0.996	explore
clarin.eu:cr1:p_1324638957739	media-corpus-profile	2.881	0.923	0.957	explore
clarin.eu:cr1:p_1320657629649	ResourceBundle	2.873	0.885	0.989	explore
clarin.eu:cr1:p_1392642184799	SpeechCorpusWithParticipants	2.842	0.846	0.996	explore

Public profiles assessment

- Profile ID
- Profile Name
- Score (0.0 - 3.0)
- Facet coverage (0.0 - 1.0)
- Percentage of elements with concepts (0.0-1.0)
- Link to SMC Browser

Name	Avg S...	Num Of Recor...	Size In Bytes	Avg Size	Num Of Profiles	Avg Nu
HZSK Repository	12.521	2,516	37,618,882	14,951	6	6.878
BAS Repository	12.092	26,593	6,912,735,518	259,9...	6	96.171
DSpace at Utrecht University	11.757	64,655	434,576,201	6,721	1	13.934
CLARIN DK UCPH Repository	11.735	147,114	672,461,802	4,571	3	2.363
IDS Repository	11.709	31,086	201,171,434	6,471	13	4.093
LINDAT CLARIN digital library at the Institute of Formal and Applied...	11.703	1,173	6,649,539	5,668	3	1.723
Instituut voor Nederlandse Lexicologie INL Metadata Repository	11.638	21	1,455,958	69,331	6	313.04
IMS Repository	11.594	69	629,290	9,120	6	1.014

Collection Assessment

- Avg. Score (0.0-14.0)
- Number of records
- Overall size (bytes)
- Avg. size per record (bytes)
- Number of used profiles
- Avg. number of Resource Proxy
- Avg. number of XML elements
- Avg. number of empty XML elements
- Avg. rate of XML population (0.0-1.0)
- Avg. facet coverage (0.0-1.0)

ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

Basic integration of the SMC Browser

CLARIN

SMC Browser

home docs stats examples reports

2016-10-12 12:18:9 :show nodes: 48; show links: 51; max count:undefined
node_size_ration:undefined
2016-10-12 12:18:9 :show nodes: 48; show links: 51; max count:undefined

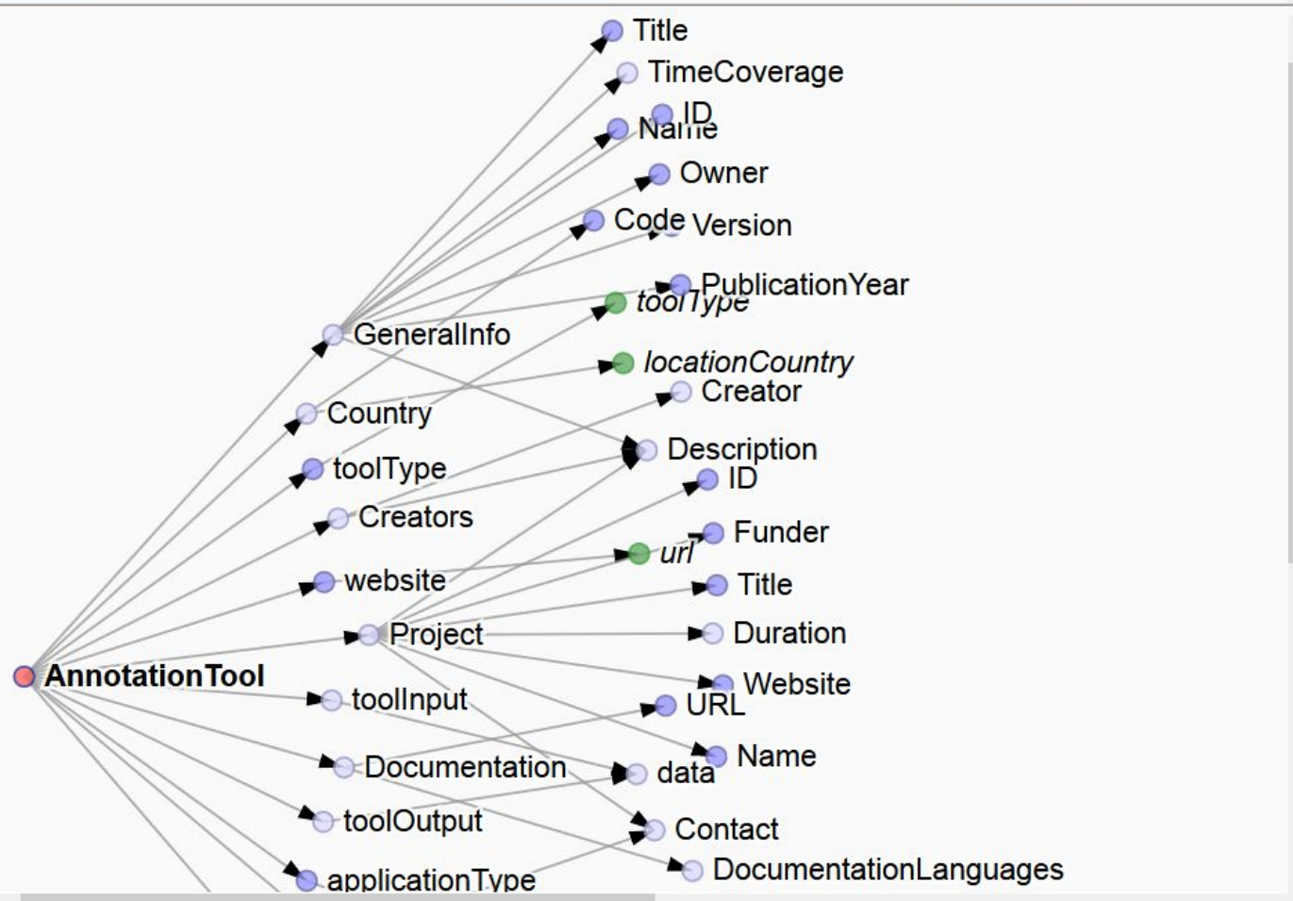
Index

Profile |203|

- AnnotatedCorpusProfile |165|
- AnnotatedCorpusProfile-DLU |174|
- **AnnotationTool |110|**
- ArthurianFiction |43|
- BASWebService |97|
- BASWebService |105|
- BamdesLexicalResource |15|
- BamdesMultimodalCorpus |17|
- BamdesOralCorpus |17|
- BamdesTool |13|
- BamdesWrittenCorpus |17|
- Bedevaartbank |112|
- BilingualDictionaryProfile |266|
- BilingualDictionaryProfile-DLU |269|
- Boedelbank |112|
- Book |23|
- CLARINWebService |79|
- CRM |33|
- CRMCollection |26|
- CenterProfile |56|
- Chromosome_Example |6|
- ComicBook |35|
- Communication_Recording |22|
- Communication_Transcript |152|
- Corpus |22|

graph SMC graph basic depth-before 2 depth-after

2 link-distance 120 charge 250



Detail

Overview

	Overall
created	
Profiles	
Components	
distinct Components	
Elements	
distinct Elements	
Elements with DatCats	
Elements without DatCats	
ratio of elements without DatCats	
distinct used Data Categories	
used Concept	
available Concepts (in Metadata profil used in CMD)	

Profile |1|

- **AnnotationTool |110|**

clarin.eu:cr1:p_1297242111880 htm

	AnnotationT
description	Description adept for in
registrationDate	2011-02-09
creatorName	Eric Sander
domainName	

How we calculate the score - Profile

profile is public	*1.0*
rate of elements annotated with concept	*[0 .. 1]*
facets coverage	*[0 .. 1]*
Max	***3.0***

How we calculate the score - Instance

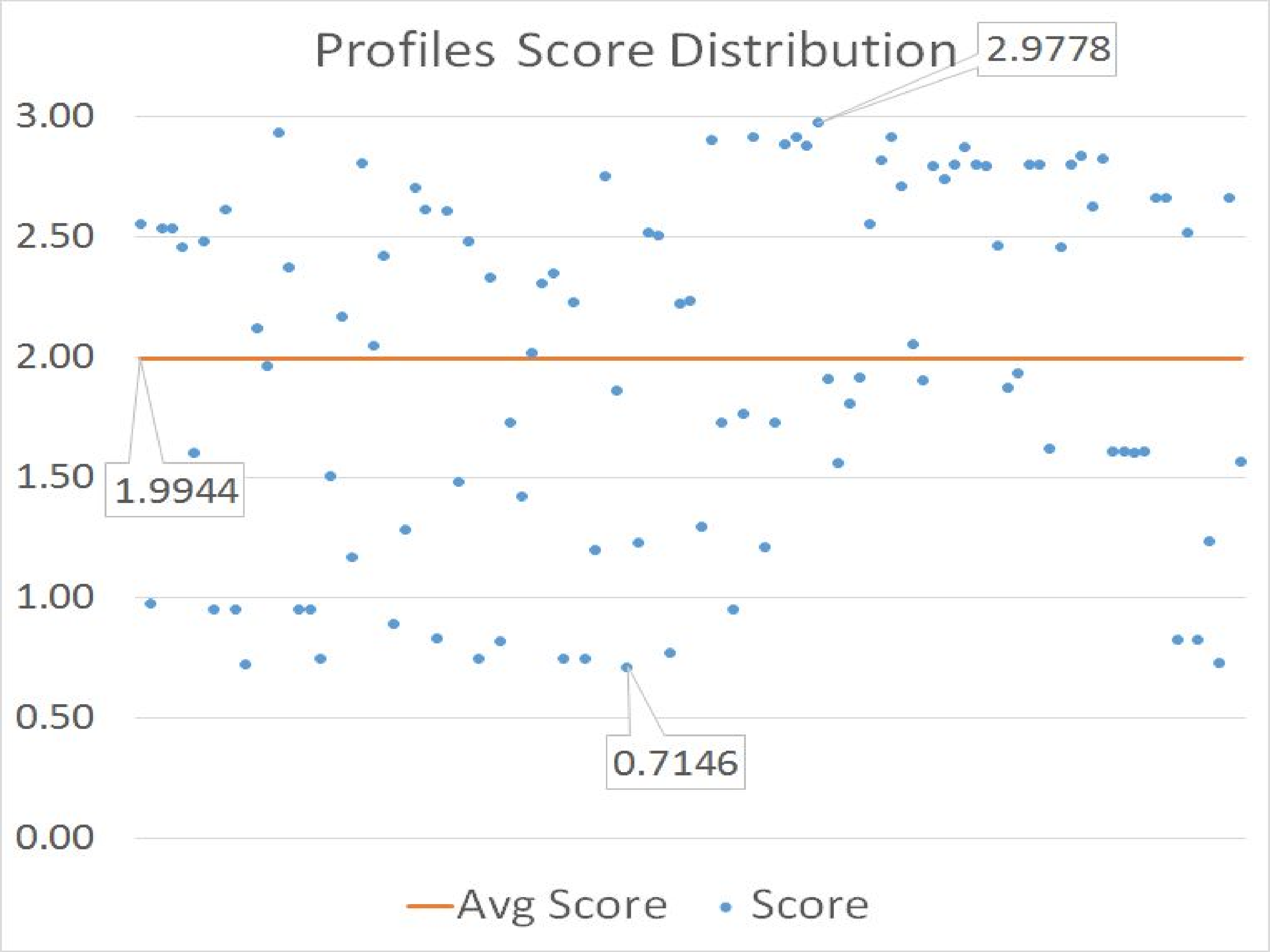
profile's score		*[0 .. 3]*
file size < 10 Mb		*0.0 or 1.0*
Header	*schema is specified*	*0.0 or 1.0*
	schema resides in Component Registry	*0.0 or 1.0*
	MdProfile element contains a valid value	*0.0 or 1.0*
	MdCollectionDisplayName is not empty	*0.0 or 1.0*
	MdSelfLink is not empty	*0.0 or 1.0*
ResProxy	*rate of resources with MIME type*	*[0 .. 1]*
	rate of resources with references	*[0 .. 1]*
	rate of non empty XML elements	*[0 .. 1]*
	rate of accessible URLs	*[0 .. 1]*
	facets coverage	*[0 .. 1]*
	Max	***14.0***

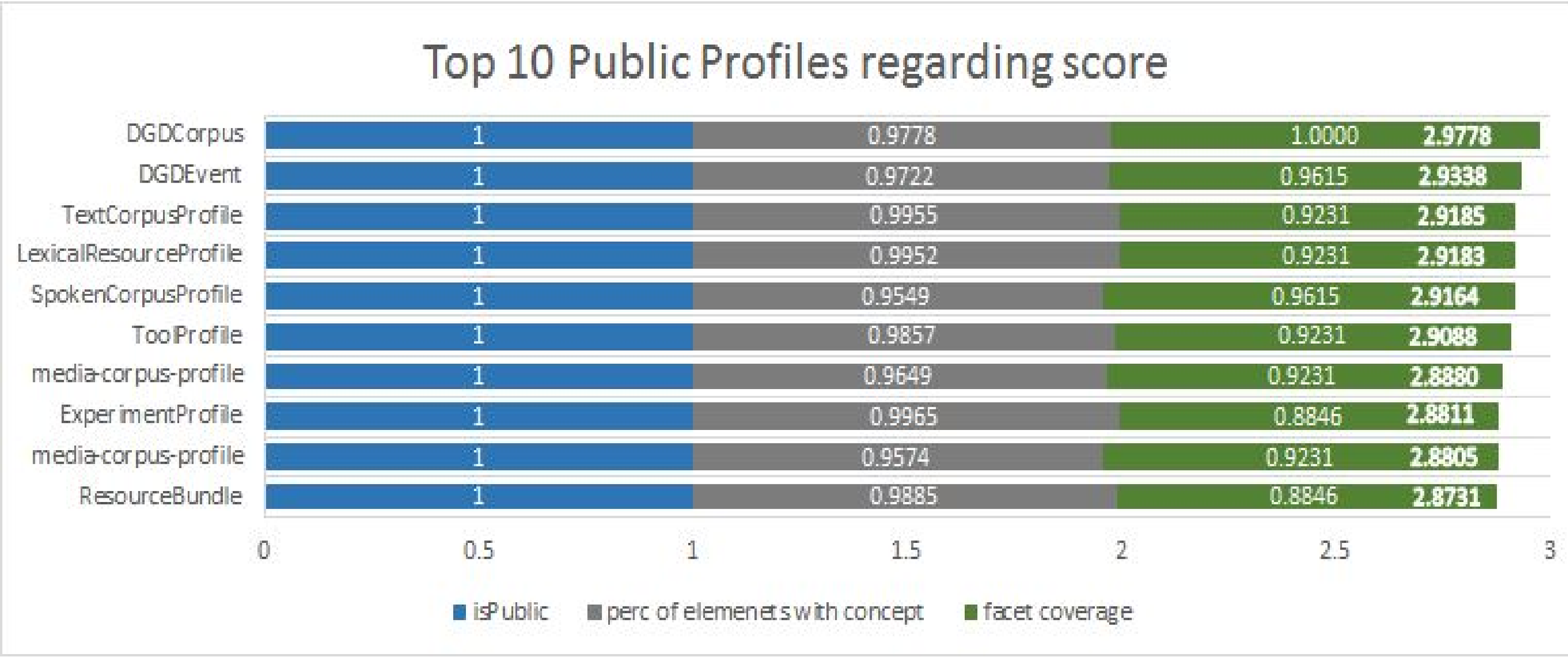
4. What can we find out?

Lot of numbers

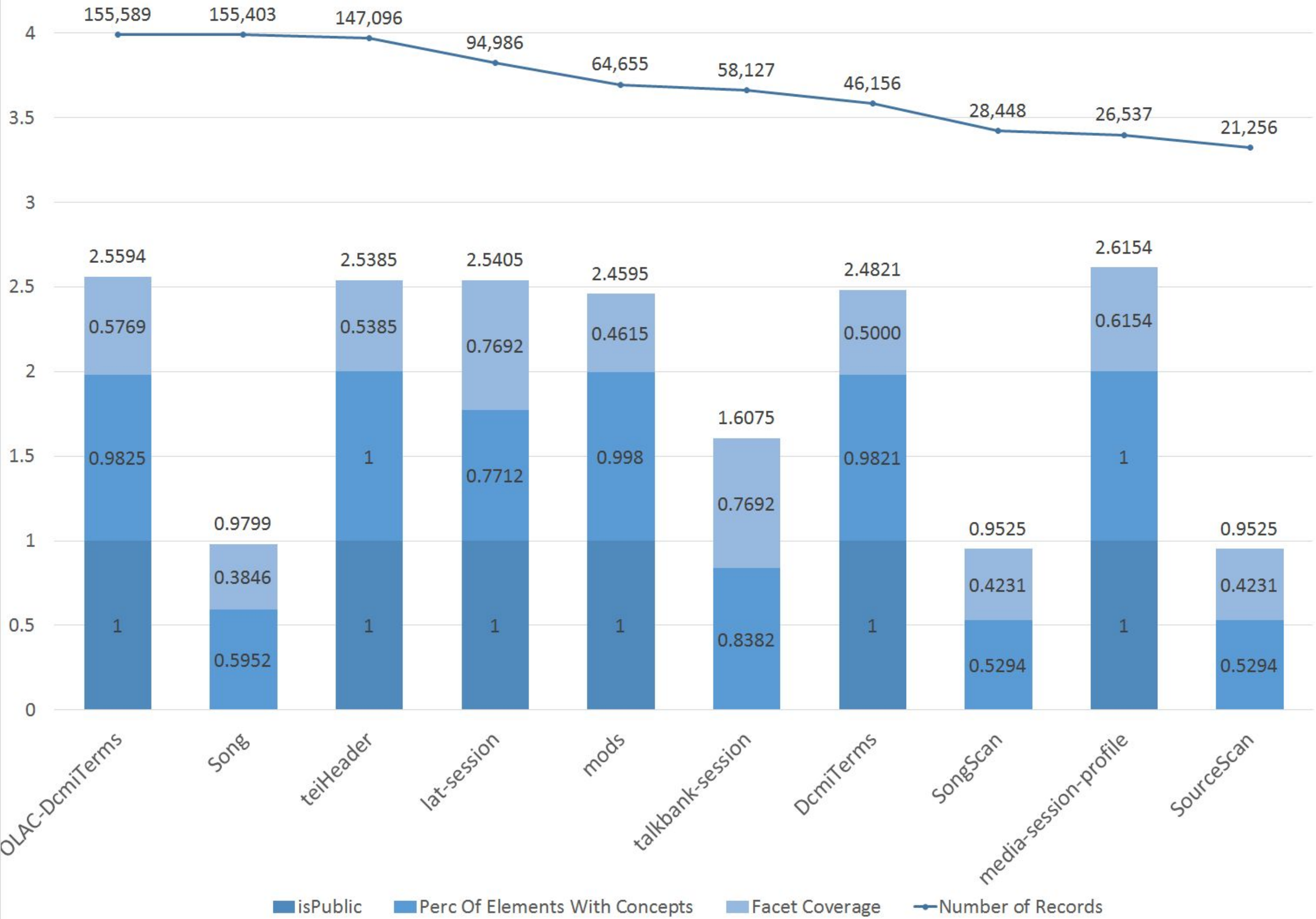
- Assessment criteria -> Scores for profiles, instances & collections
- esp. Facet coverage of profiles, instances & collections
- Concept coverage of profiles
- Statistics on size of collections

- Problems with ResourceProxies (link checks)
- Detailed info messages on issues

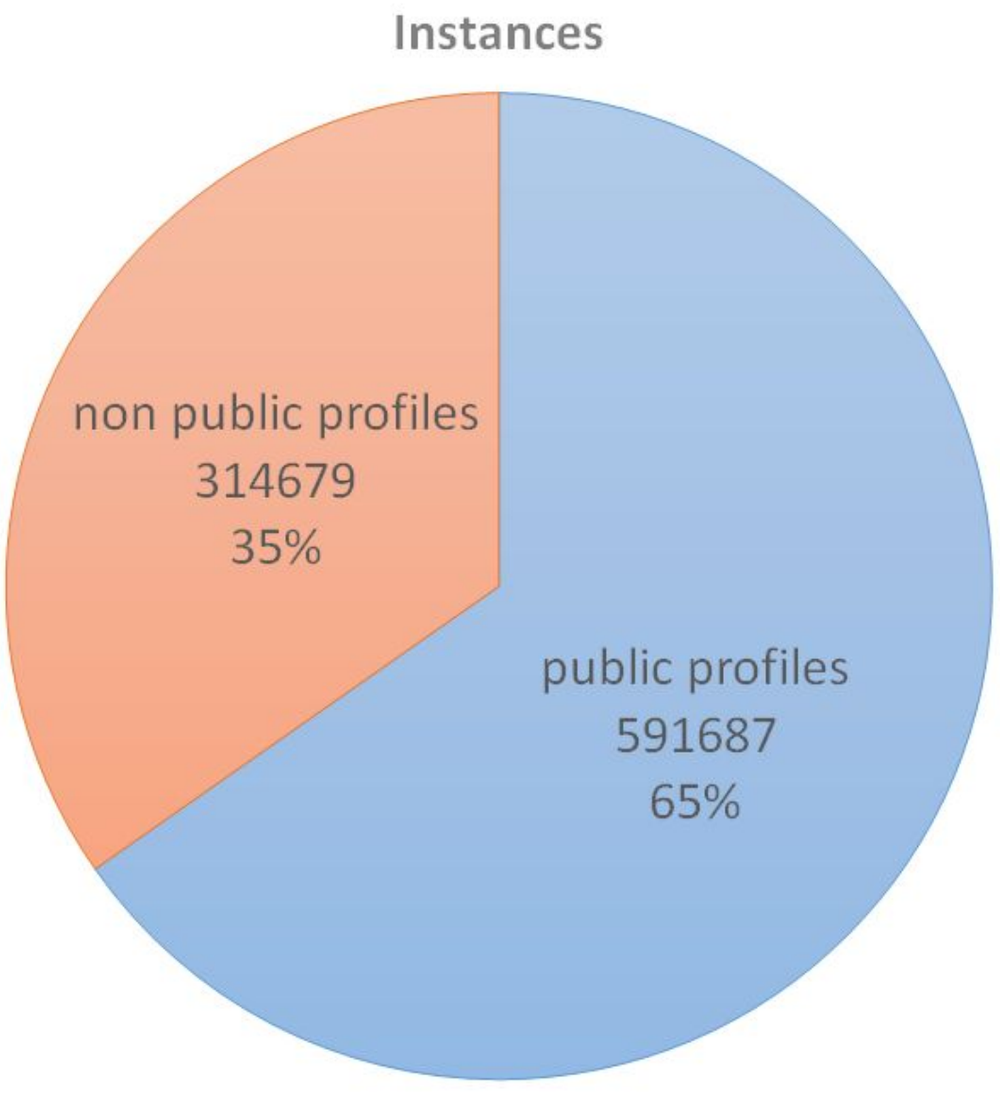

Profiles Score Distribution
2.9778
3.00
2.50
2.00
1.50
1.9944
1.00
0.50
0.7146
0.00
Avg Score
Score

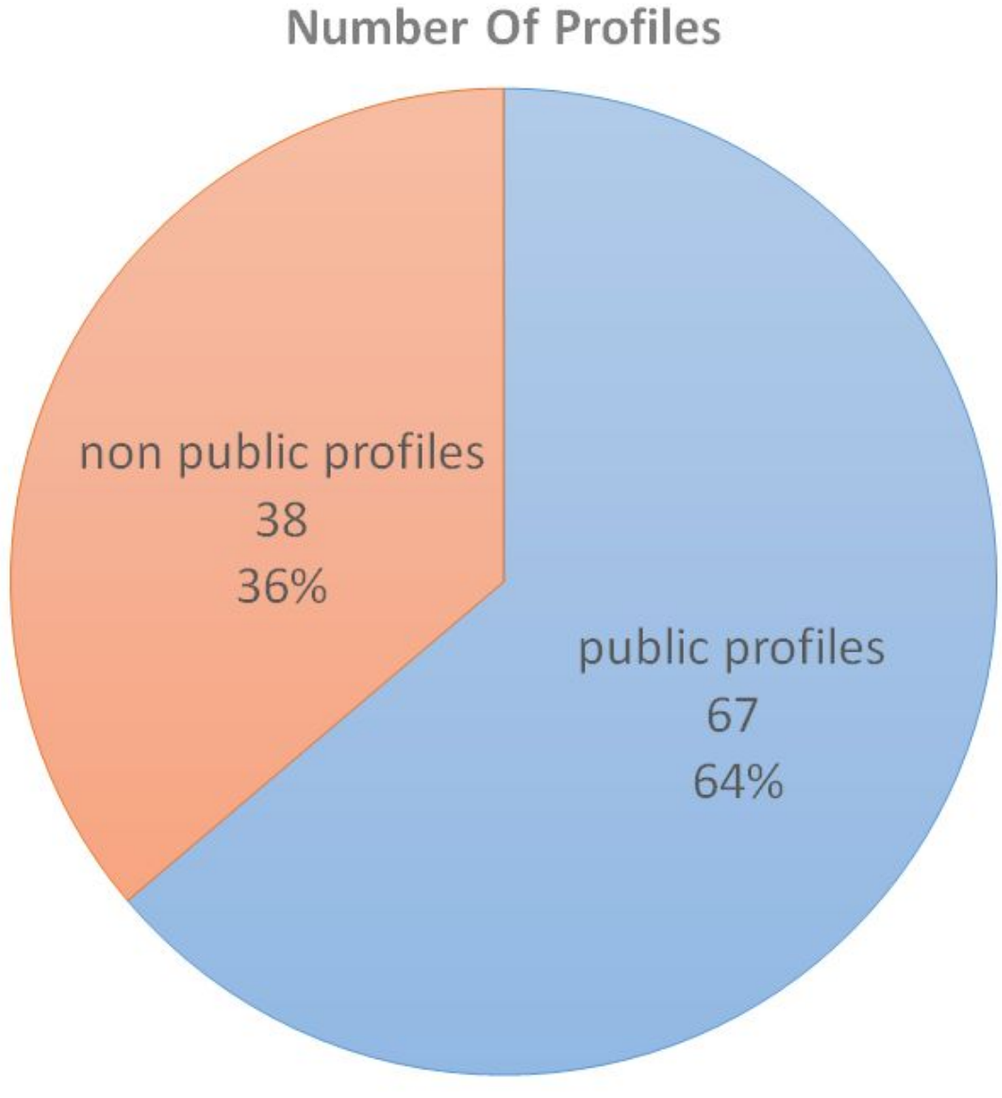

Top 10 Public Profiles regarding score
DGDCorpus
DGDEvent
TextCorpusProfile
LexicalResourceProfile
SpokenCorpusProfile
ToolProfile
media-corpus-profile
ExperimentProfile
media-corpus-profile
ResourceBundle
1 0.9778 1.0000 2.9778
1 0.9722 0.9615 2.9338
1 0.9955 0.9231 2.9185
1 0.9952 0.9231 2.9183
1 0.9549 0.9615 2.9164
1 0.9857 0.9231 2.9088
1 0.9649 0.9231 2.8880
1 0.9965 0.8846 2.8811
1 0.9574 0.9231 2.8805
1 0.9885 0.8846 2.8731
0
0.5
1
1.5
2
2.5
3
isPublic
perc of elemenets with concept
facet coverage

Top 10 profiles regarding number of instances in VLO

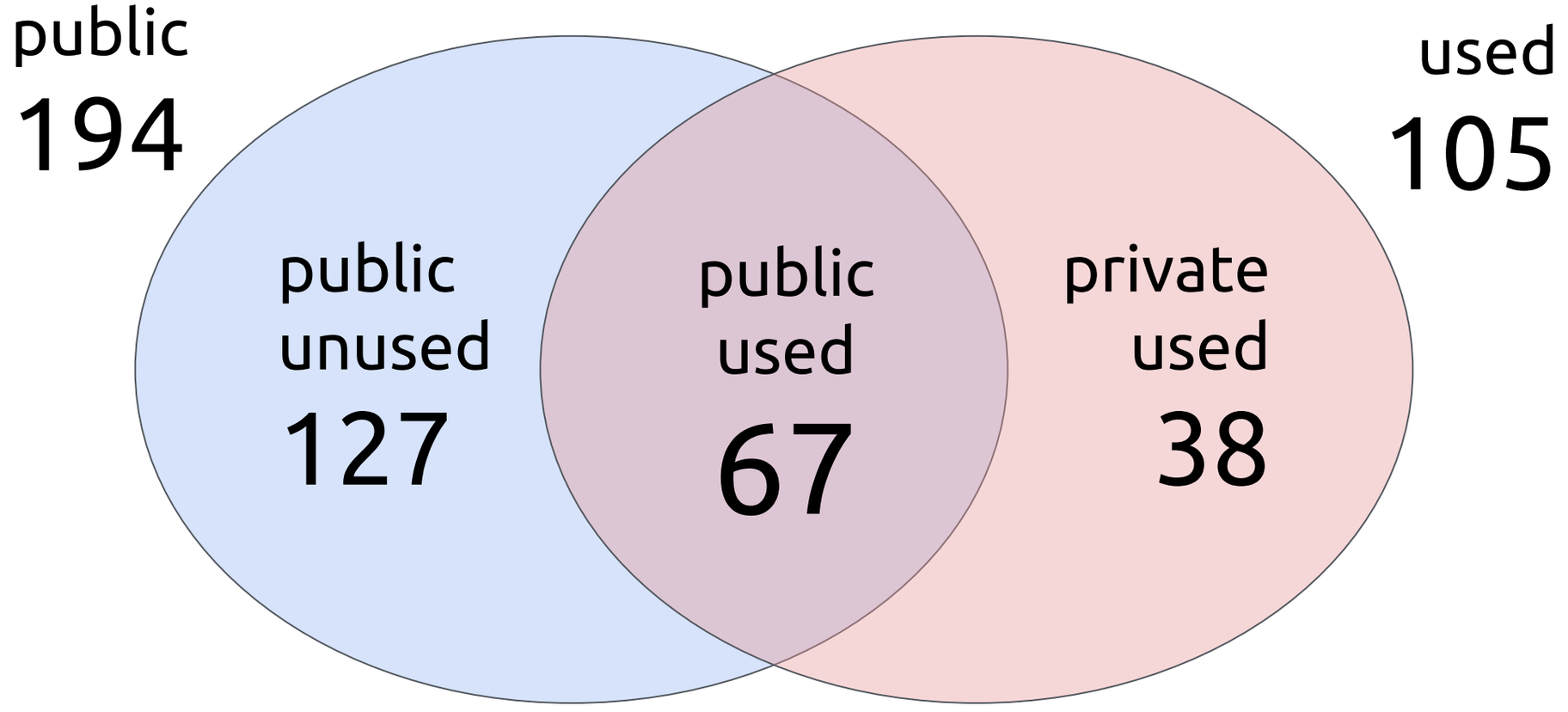


Public vs. private profiles

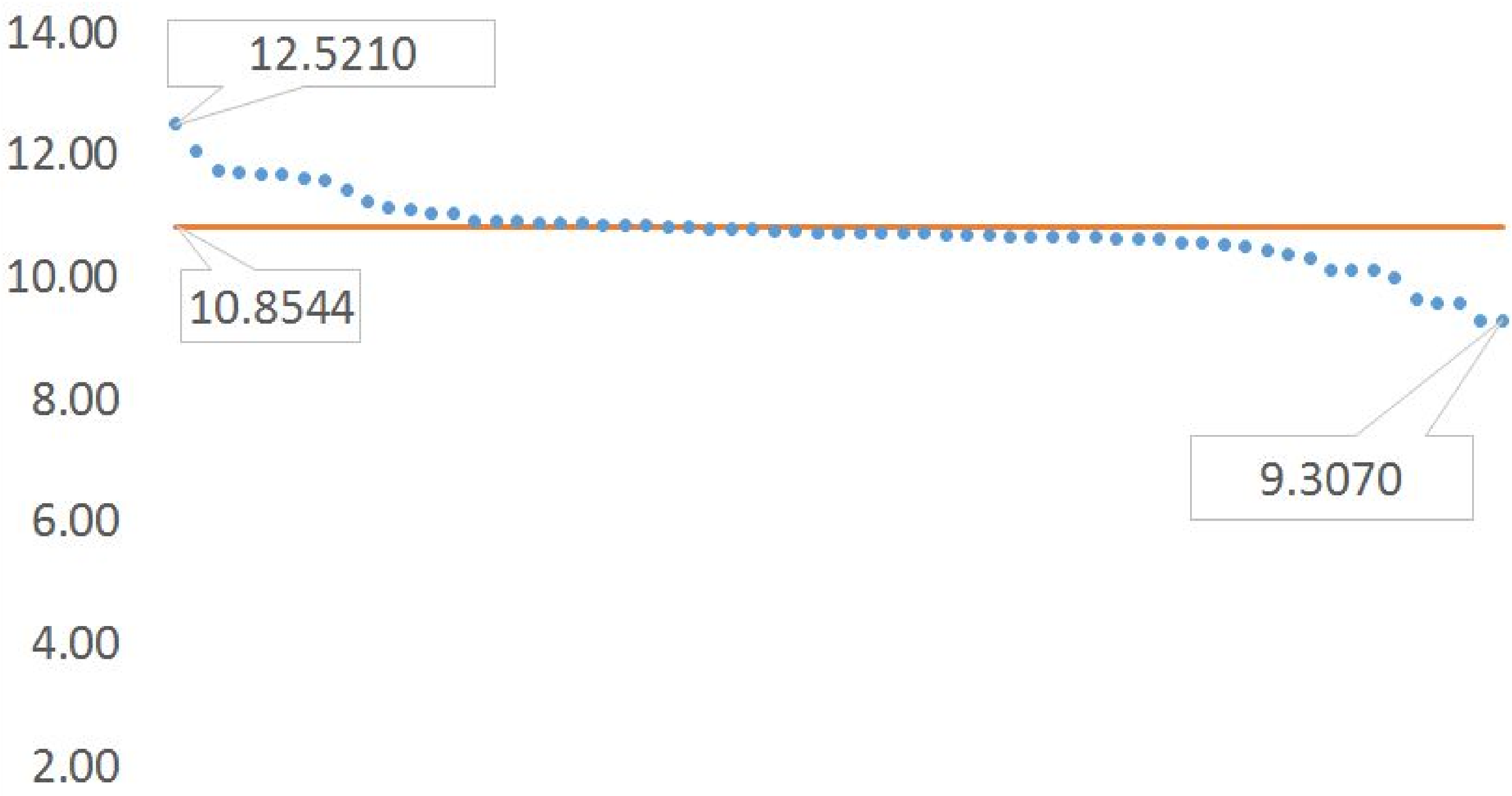


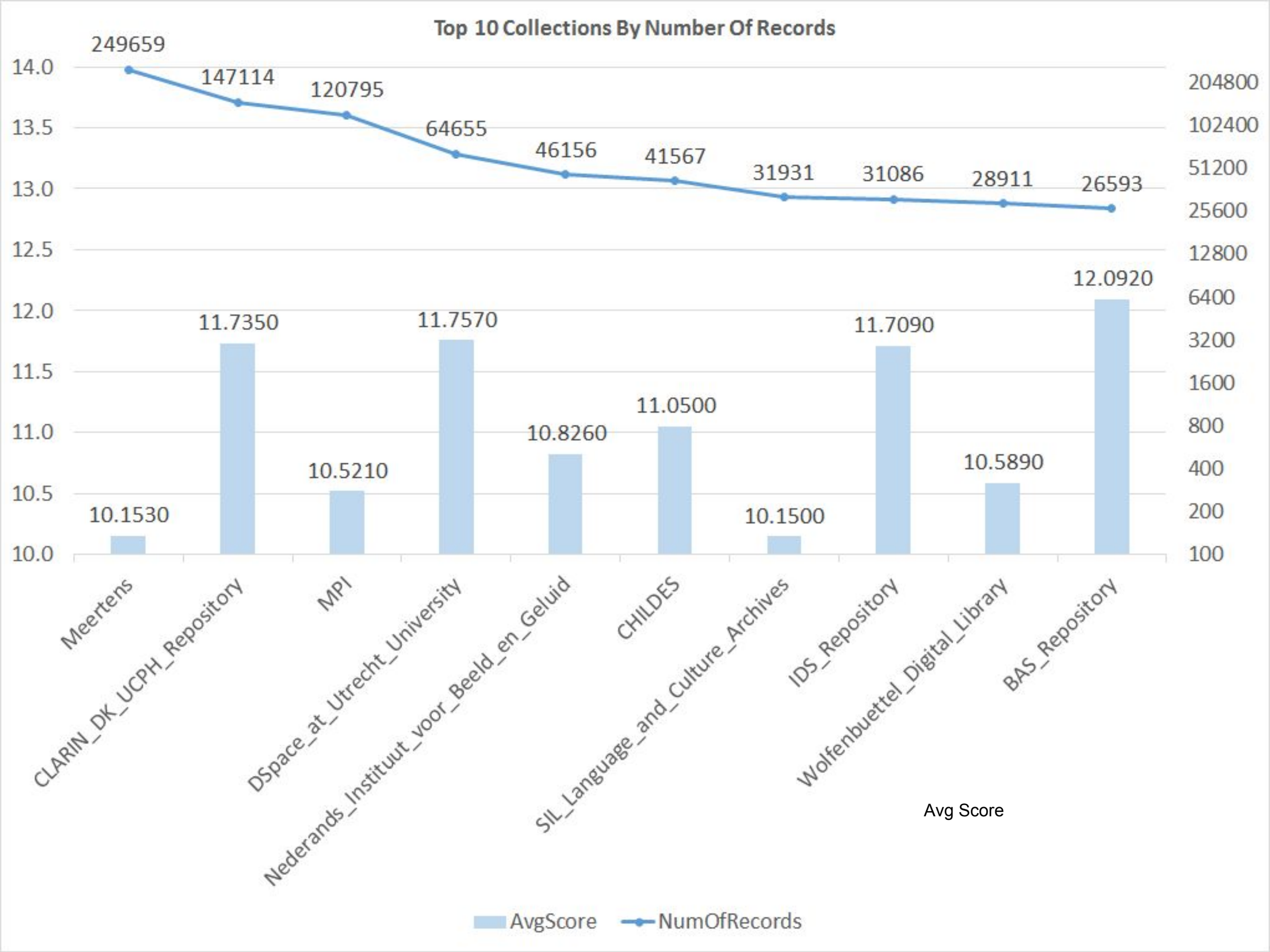


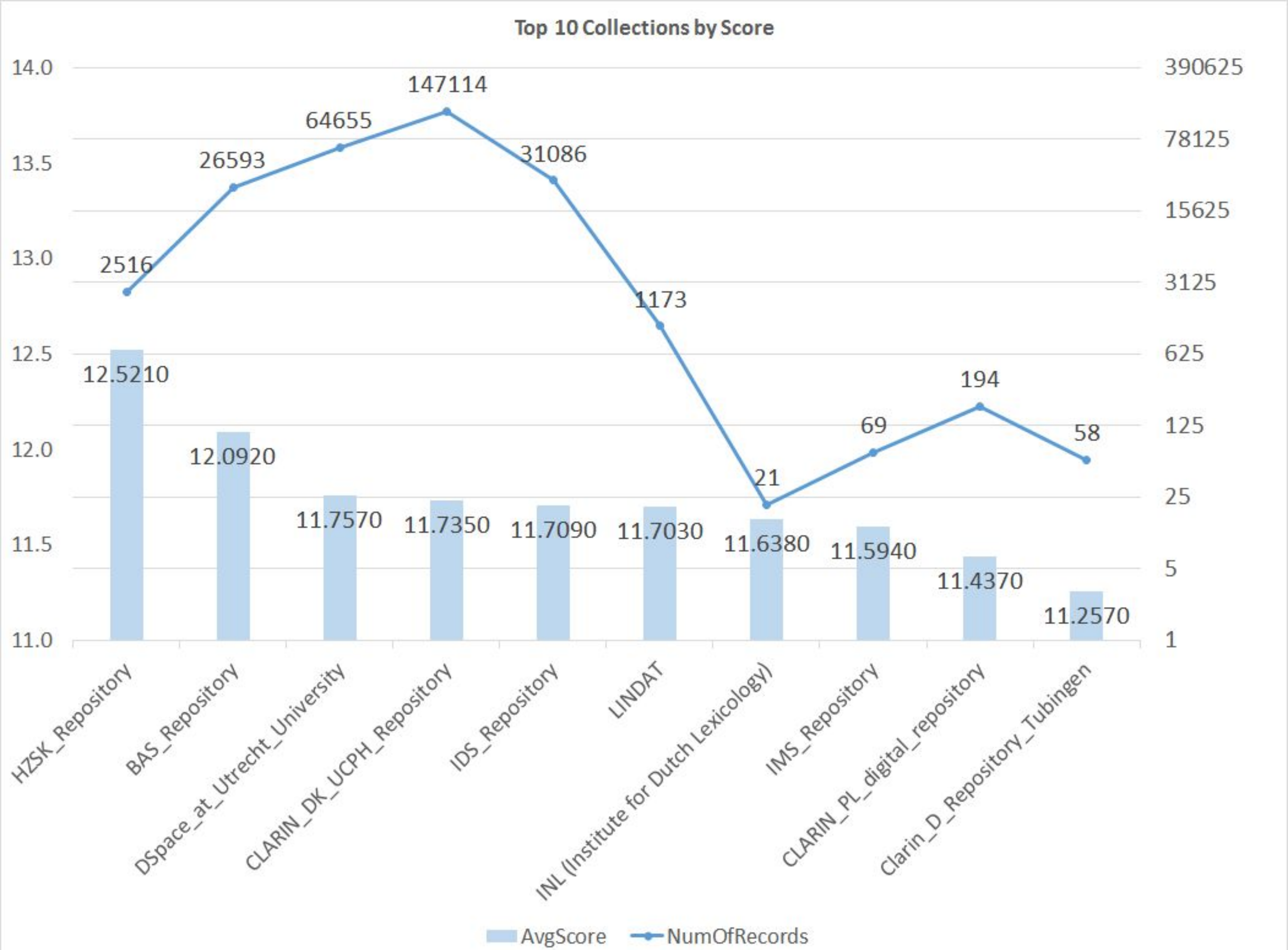
Public vs. used profiles



Collections Score Distribution




Top 10 Collections By Number Of Records
249659
147114
120795
64655
46156
41567
31931
31086
28911
26593
11.7350
10.5210
11.7570
10.8260
11.0500
10.1500
11.7090
10.5890
12.0920
10.1530
14.0
13.5
13.0
12.5
12.0
11.5
11.0
10.5
10.0
204800
102400
51200
25600
12800
6400
3200
1600
800
400
200
100
Meertens
CLARIN_DK_UCPH_Repository
MPI
DSpace_at_Utrecht_University
Nederands_Instituut_voor_Beeld_en_Geluid
CHILDES
SIL_Language_and_Culture_Archives
IDS_Repository
Wolfenbuettel_Digital_Library
BAS_Repository
Avg Score
AvgScore
NumOfRecords

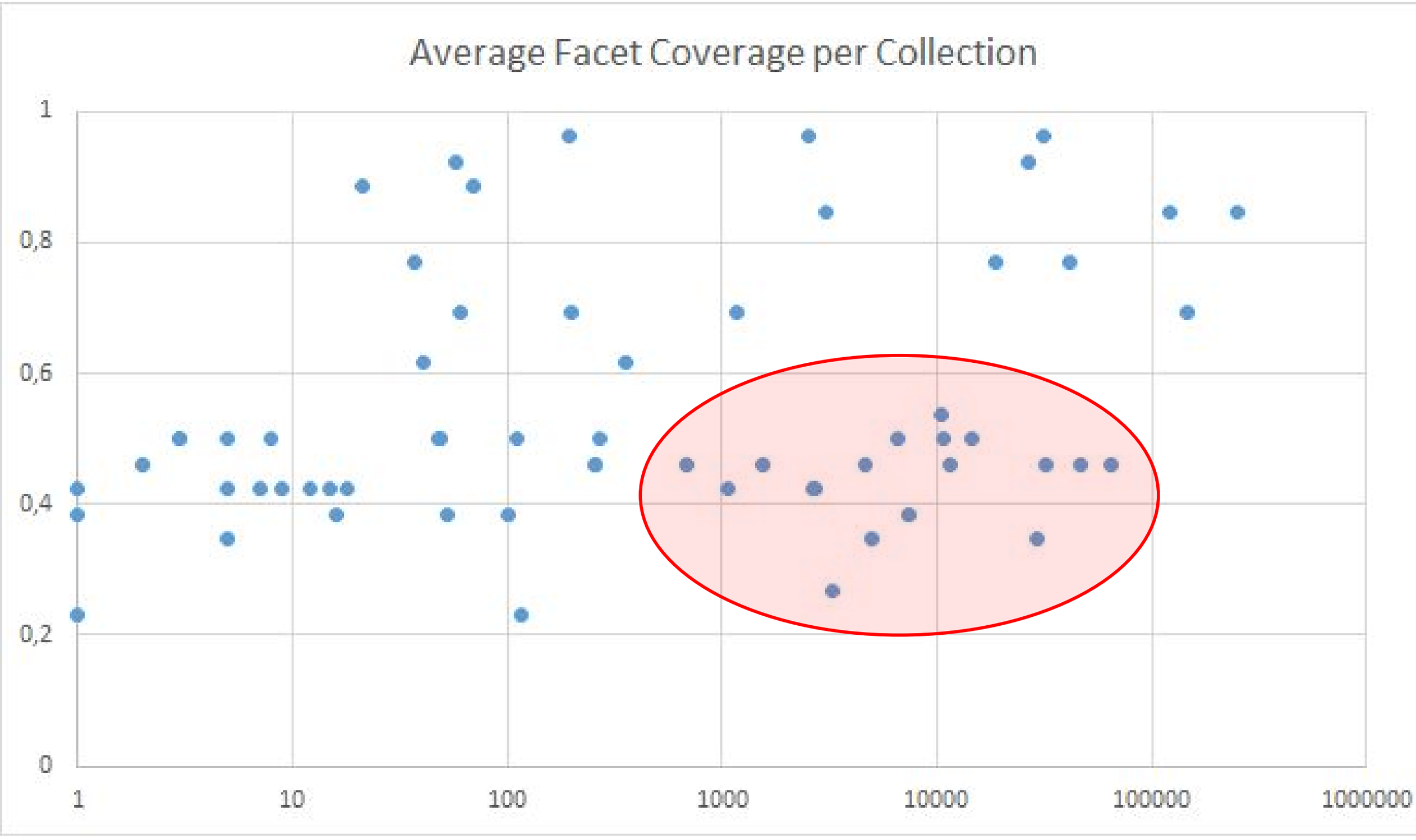

Top 10 Collections by Score
14.0
13.5
13.0
12.5
12.0
11.5
11.0
390625
78125
15625
3125
625
125
25
5
1
2516
26593
64655
147114
31086
1173
21
69
194
58
12.5210
12.0920
11.7570
11.7350
11.7090
11.7030
11.6380
11.5940
11.4370
11.2570
HZSK_Repository
BAS_Repository
DSpace_at_Utrecht_University
CLARIN_DK_UCPH_Repository
IDS_Repository
LINDAT
INL (Institute for Dutch Lexicology)
IMS_Repository
CLARIN_PL_digital_repository
Clarin_D_Repository_Tubingen
AvgScore
NumOfRecords

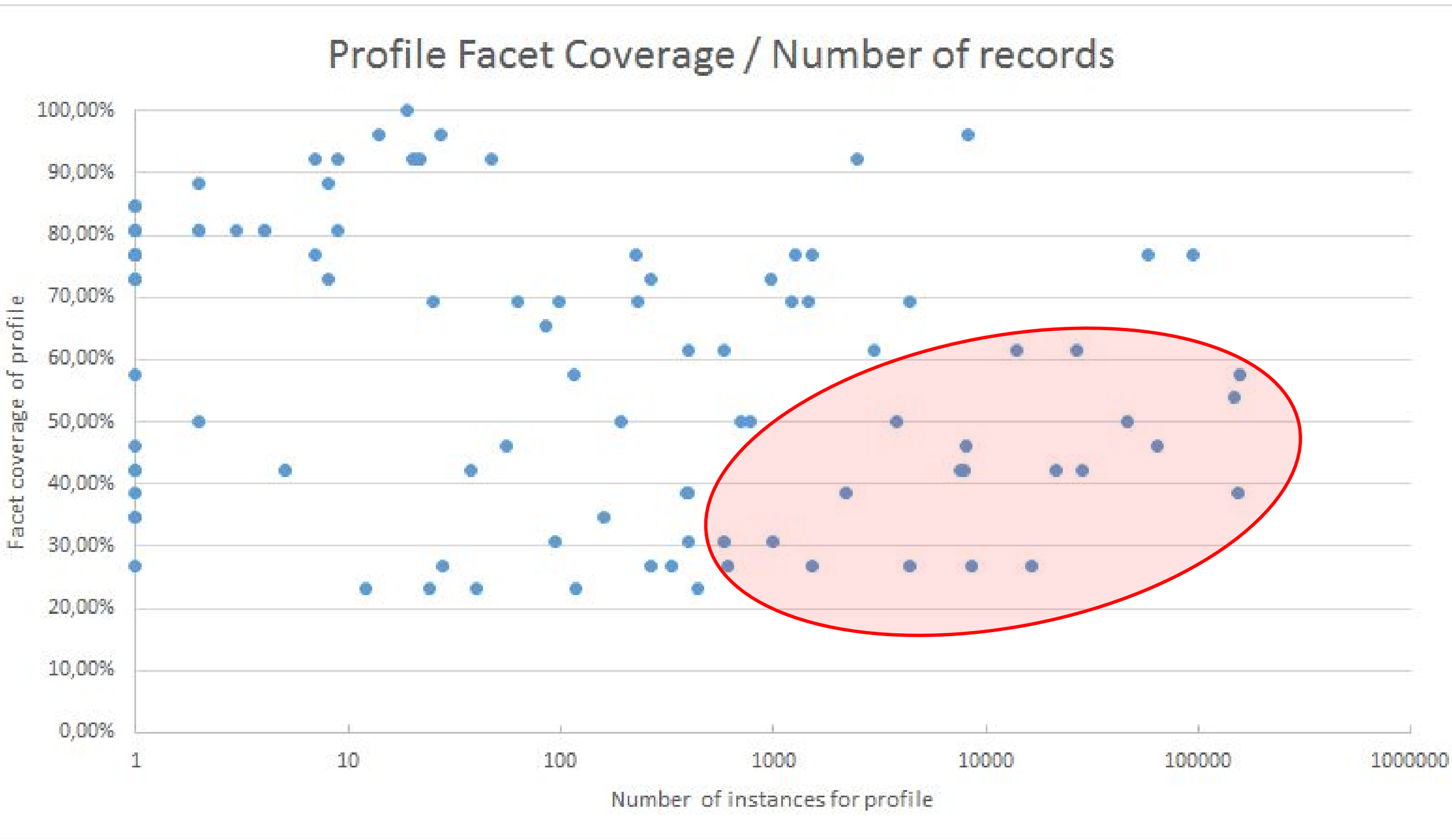
ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

Comparison of facet **un**coverage

	2015		2016		
Nr. Records	**631000**	ratio	**880973**	ratio	change
Language Code	240183	38%	224423	25%	-13%
Collection	0	0%	0	0%	0%
Resource Type	482935	77%	520290	59%	-17%
			174170	20%	-57%
Continent	472048	75%	-	-	-
Country	474637	75%	669885	76%	1%
Modality	490195	78%	706971	80%	3%
Genre	329114	52%	478582	54%	2%
Subject	503233	80%	611406	69%	-10%
Format	62381	10%	37714	4%	-6%
Organisation	520560	82%	687504	78%	-4%
Availability	580907	92%	704595	80%	-12%
National Project	104316	17%	312475	35%	19%
Keywords	567347	90%	816140	93%	3%


Average Facet Coverage per Collection
1
0,8
0,6
0,4
0,2
0
1
10
100
1000
10000
100000
1000000


Profile Facet Coverage / Number of records
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Facet coverage of profile
1
10
100
1000
10000
100000
1000000
Number of instances for profile

5. Into the future

- Upload file/copy&paste function for the assessment
- API support and/or CSV export (on its way)
- Automatic email notification to data providers (**curation reports**)
- Visualise results in a user friendly way
- **Calibration** of scoring
- Concentrate on the facet-value **variability / normalisation**
- Add facet-oriented view
- **VLO Dashboard** as integrated environment for VLO curators and administrators for work with CM and other components (harvester, validator, mapping and normalisation, VLO importer etc.)
- *Any other suggestions?*

ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

Demo is ready for you to see more stats of the metadata (after the break)

Austrian Centre for Digital Humanities (ACDH-OEAW)

www.oeaw.ac.at/acdh

Davor Ostojic davor.ostojic@oeaw.ac.at

Go Sugimoto go.sugimoto@oeaw.ac.at

Matej Ďurčo matej.durco@oeaw.ac.at