



VOCABULARIES AND QUALITY IMPROVEMENT OF LIBRARY CATALOGUES

Péter Király | GWDG | 2023-03-27 | Centre Marc Bloch, Berlin

<https://bit.ly/qa-triple-2023>

A decorative graphic on the right side of the slide, featuring several overlapping diagonal bars in purple, red, and yellow, with circular accents at the intersections.

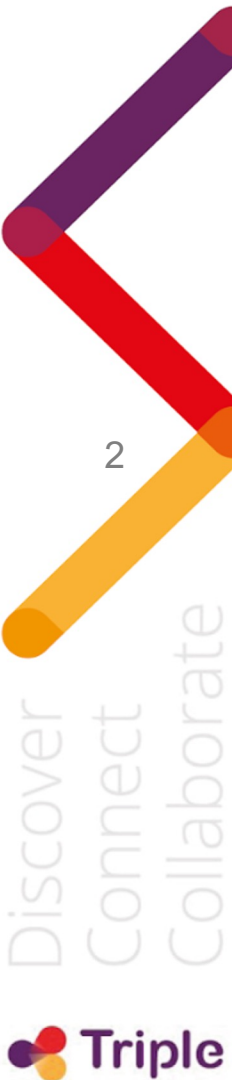
Discover
Connect
Collaborate





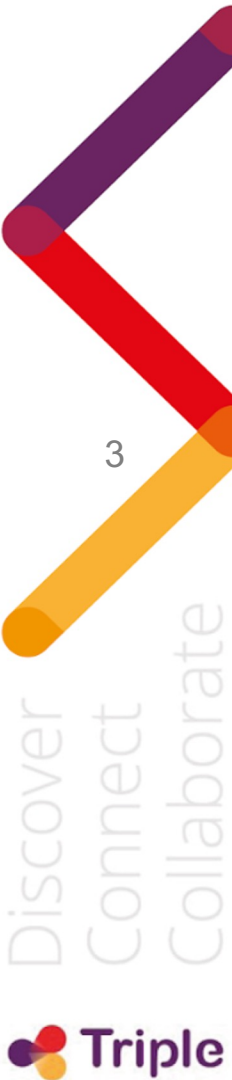
Transforming Research Through Innovative
Practices for Linked Interdisciplinary Exploration

The TRIPLE project was launched in October 2019. It develops the GoTriple discovery platform, which will be one of the dedicated services of OPERAS, the Research Infrastructure supporting open scholarly communication in the social sciences and humanities (SSH) in the European Research Area.



INTERNAL VOCABULARIES

IN BIBLIOGRAPHIC RECORDS (MARC, PICA)



Internal vocabularies

Leader/05 - Record status

a - Increase in encoding level

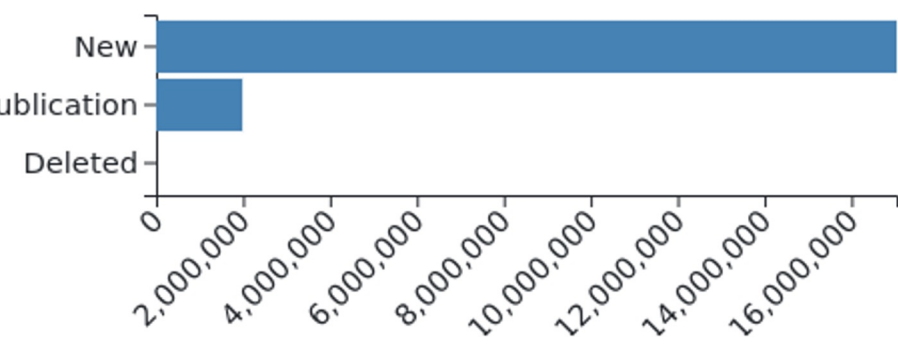
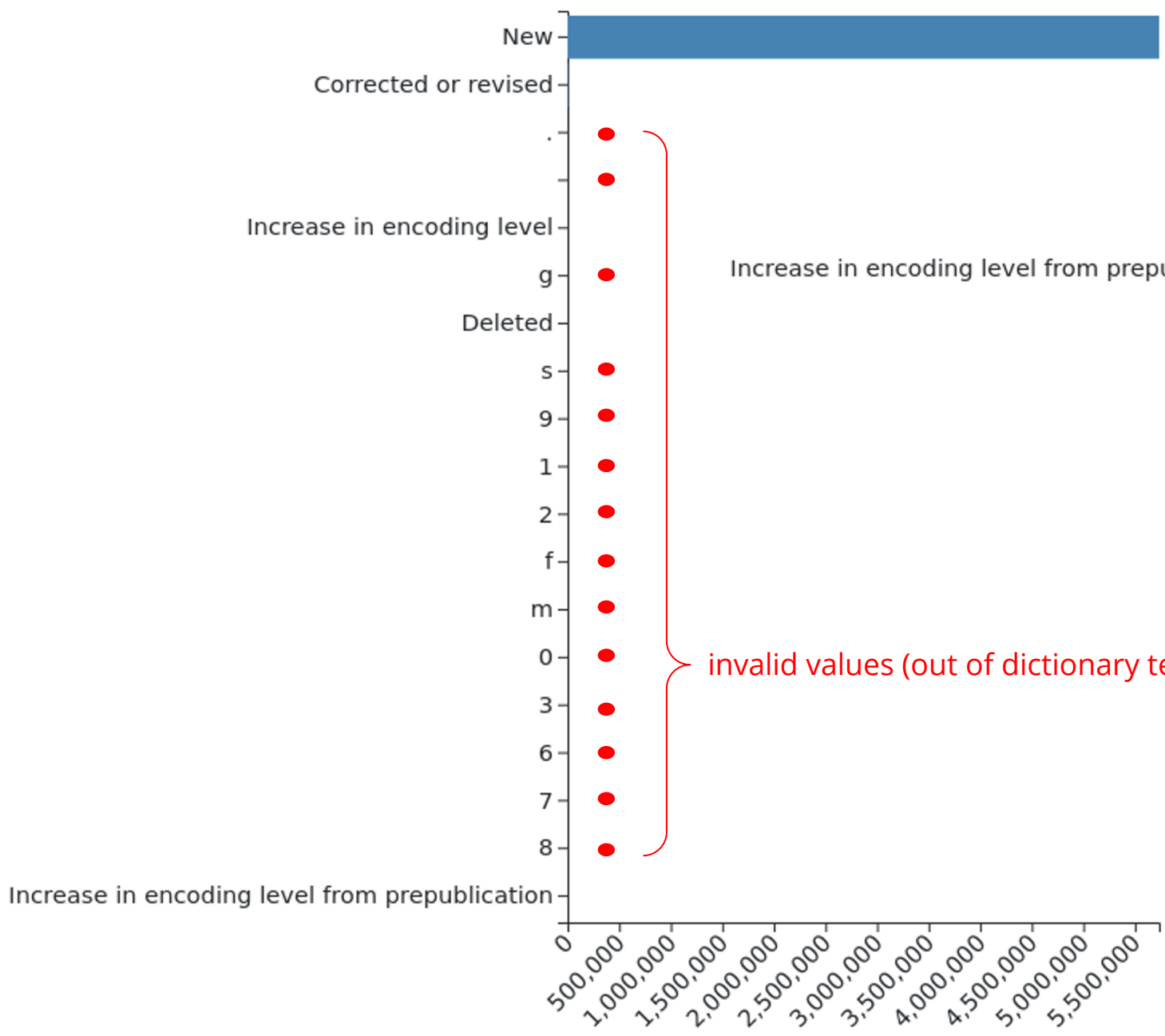
c - Corrected or revised

d - Deleted

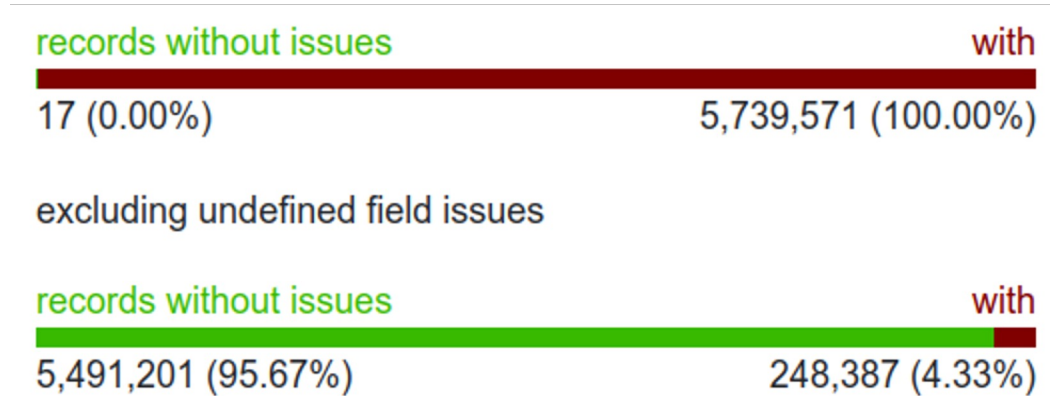
n - New

p - Increase in encoding level from prepublication





Internal vocabularies



path	message	url	instances	records	chart	%
record level issues			154	147	Q ↓	0.00
•	undetectable type (8 variants) [+]		33	33	Q ↓	0.00
	invalid linkage (17 variants) [+]		121	114	Q ↓	0.00
control field level issues			453,070	136,075	Q ↓	2.37
•	invalid code (421 variants) [+]		33,469	13,589	Q ↓	0.24
•	invalid value (716 variants) [+]		419,601	130,870	Q ↓	2.28
data field level issues			78,305,654	5,739,563	Q ↓	100.00
	missing reference subfield (880\$6) (1 variants) [+]		509	509	Q ↓	0.01
	repetition of non-repeatable field (29 variants) [+]		6,716	6,709	Q ↓	0.12
•	undefined field (226 variants) [+]		78,298,429	5,739,563	Q ↓	100.00
indicator level issues			62,932	52,251	Q ↓	0.91
•	obsolete value (11 variants) [+]		287	222	Q ↓	0.00
•	non-empty indicator (289 variants) [+]		5,241	4,497	Q ↓	0.08
•	invalid value (232 variants) [+]		57,404	48,355	Q ↓	0.84
subfield level issues			66,692	58,088	Q ↓	1.01
•	undefined subfield (412 variants) [+]		32,353	28,542	Q ↓	0.50
	invalid length (282 variants) [+]		515	484	Q ↓	0.01
•	invalid classification reference (10 variants) [+]		10,397	7,573	Q ↓	0.13
	content does not match any patterns (369 variants) [+]		1,054	519	Q ↓	0.01
	repetition of non-repeatable subfield (155 variants) [+]		7,393	7,146	Q ↓	0.12
	invalid ISBN (9 variants) [+]		9,970	9,832	Q ↓	0.17
	invalid ISSN (28 variants) [+]		4,300	3,844	Q ↓	0.07
	content is not well-formatted (22 variants) [+]		47	47	Q ↓	0.00
•	invalid value (284 variants) [+]		663	652	Q ↓	0.01

Internal vocabularies

control field level issues			453,070	136,075	Q ↓	2.37
invalid code (421 variants) [+]			33,469	13,589	Q ↓	0.24
invalid value (716 variants) [+]			419,601	130,870	Q ↓	2.28
↓ data element ↑	↓ message ↑		↓ ↑	↓ ↑		
Leader/05 (leader05)	.	i	24	24	Q ↓	0.00
Leader/05 (leader05)	" "	i	5	5	Q ↓	0.00
Leader/05 (leader05)	g	i	5	5	Q ↓	0.00
Leader/05 (leader05)	s	i	4	4	Q ↓	0.00
Leader/05 (leader05)	9	i	3	3	Q ↓	0.00
Leader/05 (leader05)	1	i	2	2	Q ↓	0.00
Leader/05 (leader05)	2	i	2	2	Q ↓	0.00
Leader/05 (leader05)	f	i	2	2	Q ↓	0.00
Leader/05 (leader05)	m	i	2	2	Q ↓	0.00
Leader/05 (leader05)	0	i	1	1	Q ↓	0.00
Leader/05 (leader05)	3	i	1	1	Q ↓	0.00
Leader/05 (leader05)	6	i	1	1	Q ↓	0.00
Leader/05 (leader05)	8	i	1	1	Q ↓	0.00
Leader/05 (leader05)	7	i	1	1	Q ↓	0.00

count: 14 | filter: *Leader/05 (leader05)* | [list all](#) | [grouped by tag](#)

<https://bit.ly/qa-triple-2023>



EXTERNAL VOCABULARIES

IN BIBLIOGRAPHIC RECORDS (MARC, PICA)

9

Discover
Connect
Collaborate

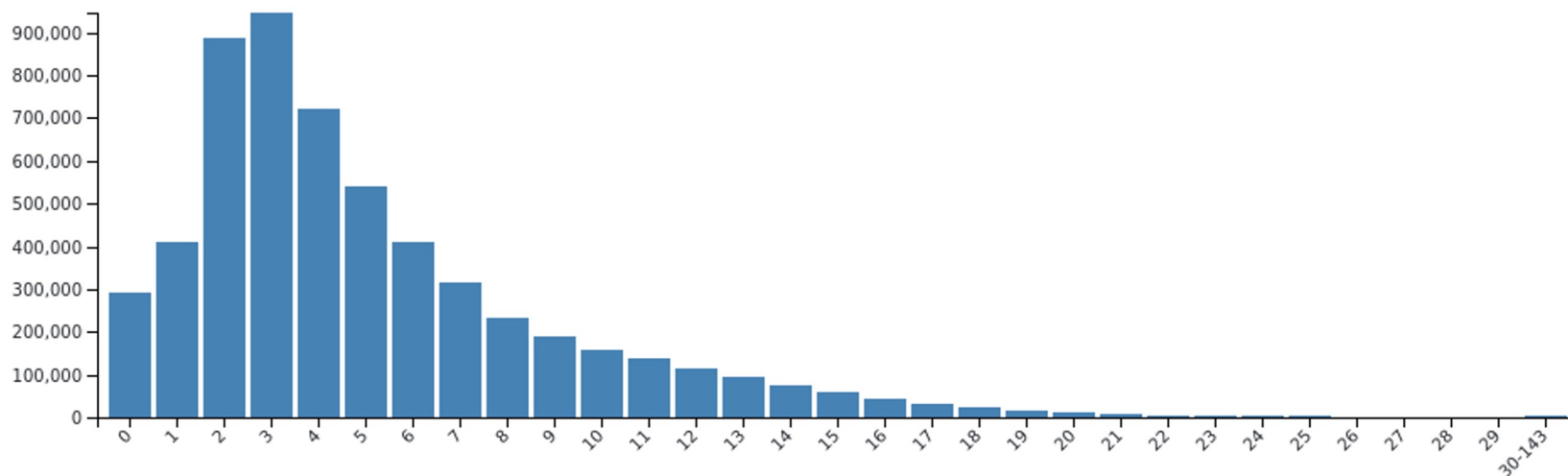


External vocabularies - subjects

records with classifications/subject headings



histogram



y: number of records

x: number of subjects in one record

example records (one record for each subject count): 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 68, 69, 73, 76, 77, 78, 89, 97, 98, 100, 102, 103, 116, 117, 118, 120, 121, 132, 143

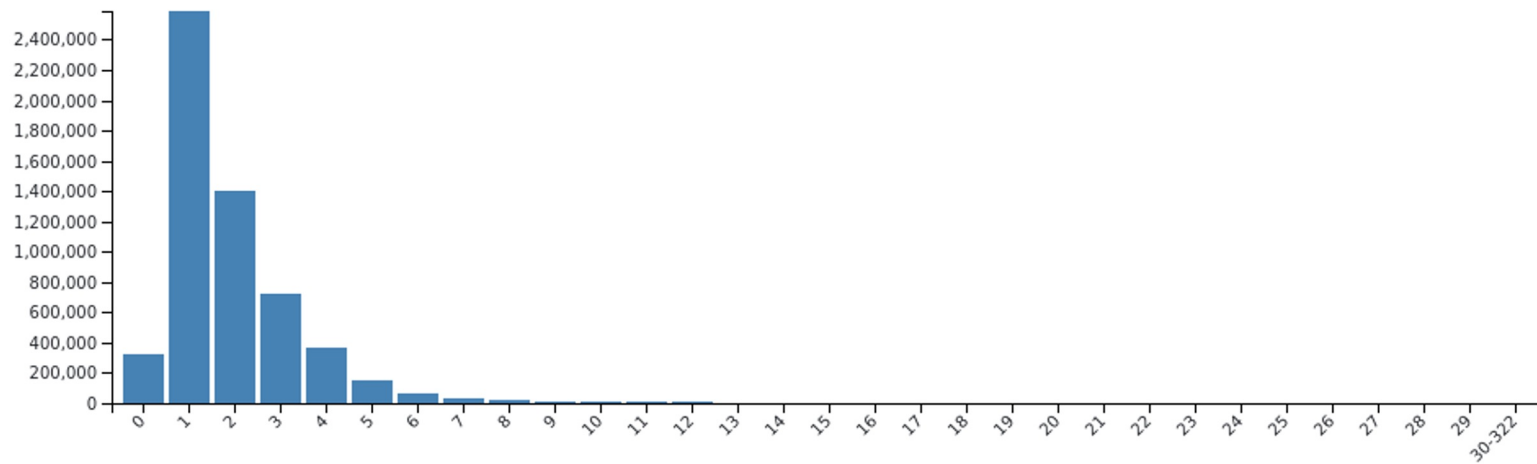


External vocabularies - names

records with authority names



histogram



y: number of records

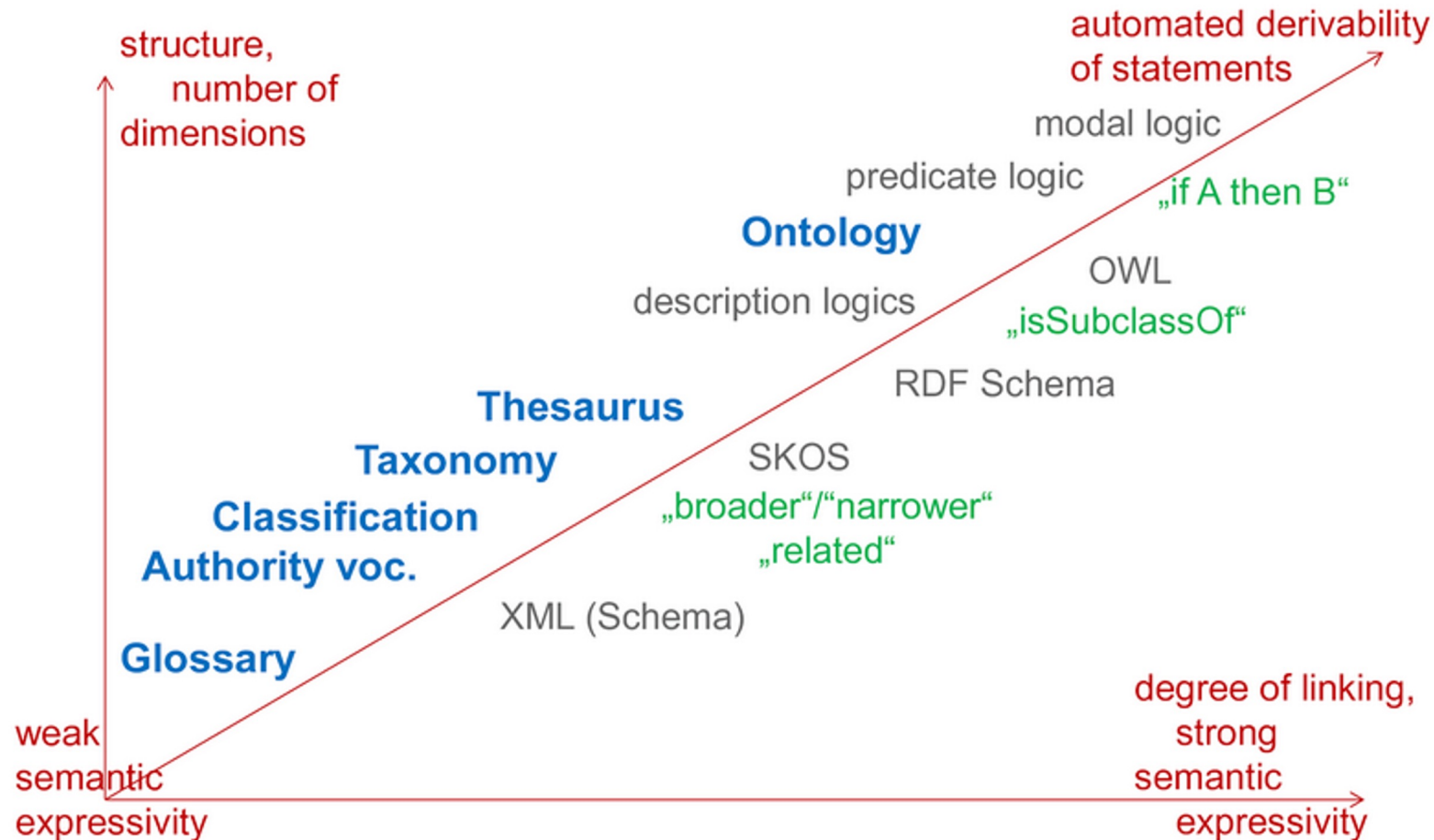
x: number of authority names in one record

example records (one record for each authority count): 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 108, 109, 110, 111, 112, 113, 114, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 128, 130, 132, 133, 134, 135, 136, 137, 138, 142, 143, 144, 147, 154, 158, 168, 169, 175, 176, 179, 185, 195, 210, 211, 231, 322
























Knowledge Organisation Systems

Knowledge organization systems – the spectrum



Typology of Knowledge Organisation System (KOS) visualisation by Anna Kasprzik based on (Souza et al., 2012)


Knowledge Organisation Systems

Location	Classification/subject headings scheme	Instances	Records chart	%
082 — Dewey Decimal Classification Number				
\$a	Dewey Decimal Classification (ddc) ▼	7,365,118	7,365,118 	38.71%
083 — Additional Dewey Decimal Classification Number				
\$a	Dewey Decimal Classification (ddc) ▼	2,328,241	2,134,189 	11.22%
084 — Other Classification Number				
\$2	Systematik der Deutschen Nationalbibliographie (sdnb) ▼	4,506,851	4,506,851 	23.69%
\$2	Subject Classes of the Schlagwortnormdatei (sswd) ▼	1,151,297	1,151,297 	6.05%
\$2	Sondersammelgebiets-Nummer (ssgn) ▼	44,339	44,268 	0.23%
\$2	Regensburger Verbundklassifikation (RVK) (rvk) ▼	21,427	21,255 	0.11%
\$2	Rheinland-Pfälzische Bibliographie (rpb) ▼	15,668	15,661 	0.08%
\$2	Fachinformationsdienste für die Wissenschaft (FID) (fid) ▼	13,871	12,664 	0.07%
\$2	Nordrhein-Westfälische Bibliographie (Köln: hbz - Hochschulbibliothekszentrum NRW) (nwbib) ▼	7,815	7,707 	0.04%
\$2	Systematik der Bayerischen Bibliographie (sbb) ▼	7,170	7,140 	0.04%
\$2	Systematik der TUB München (stub) ▼	5,974	5,970 	0.03%
\$2	Systematik der IfZ-Bibliothek (ifzs) ▼	3,063	3,048 	0.02%
\$2	Basisklassifikation (bkl) ▼	60	27 	0.00%
\$2	Dewey decimal classification and relative index (Dublin, Ohio: OCLC Online Computer Center) (ddc) ▼	38	30 	0.00%
\$2	DOPAED der UB Erlangen (dopaed) ▼	15	15 	0.00%
\$2	Library of Congress classification (Washington , D.C.: Library of Congress, CDS) (lcc) ▼	6	6 	0.00%
\$2	DDC22ger (DDC22ger) ▼	5	5 	0.00%
\$2	ddc22ger (ddc22ger) ▼	4	4 	0.00%
\$2	rbp (rbp) ▼	3	3 	0.00%
\$2	ifz (ifz) ▼	2	2 	0.00%
\$2	Universal decimal classification (London: British Standards Institute) (udc) ▼	2	2 	0.00%

-
-
-
-



Properties of term entries

Location	Classification/subject headings scheme	Instances	Records chart	%
082	Dewey Decimal Classification Number			
\$a	Dewey Decimal Classification (ddc) ^	7,365,118	7,365,118 	38.71%

Which subfields are available in the individual instances of this field?

\$a	Classification number	7,365,118
\$q	Assigning agency	7,365,118
\$2	Edition number	7,365,077
\$8	Field link and sequence number	3,560,734

subfields	count
\$a, \$q, \$2, \$8	3,236,870
\$a, \$q, \$2	2,704,629
\$a+, \$q, \$2	1,099,755
\$a+, \$q, \$2, \$8	323,823
\$a, \$q, \$8	41

notes:

- + sign denotes multiple instances

14

Discover
Connect
Collaborate



Vocabulary terms: encoded

vocabulary: "**Dewey Decimal Classification**"

082\$a: Dewey Decimal Classification Number / Classification number

B (745,890)

K (247,702)

330 (228,764)

610 (176,348)

S (157,873)

370 (148,048)

650 (142,257)

620 (141,677)

050 (130,219)

914.3 (126,181)

340 (122,330)

781.64 (111,031)

910 (102,304)

570 (101,717)

530 (95,001)

300 (94,493)



15

Discover
Connect
Collaborate

 Triple

Vocabulary terms: human readable

vocabulary: **Library of Congress subject headings (Washington, DC: LC, Cataloging Distribution Service)**

600\$a Subject Added Entry - Personal Name / Personal name

Willems, Jan Frans, (1,438)

Shakespeare, William, (1,237)

Aristotle. (601)

Dante Alighieri, (578)

Plato. (536)

Jesus Christ (523)

Napoleon (500)

Louis (474)

Charles (454)

Mary, (451)

Homer (430)

Goethe, Johann Wolfgang von, (420)

Nietzsche, Friedrich Wilhelm, (412)

Homer. (343)



Vocabulary terms: mixed

Vokabel: **"Allgemeine Systematik für Öffentliche Bibliotheken"**

045B: Allgemeine Systematik für Bibliotheken (ASB)

R 11 (39,240)

S (37,704)

Zba (35,230)

ZAA (34,705)

Erzählende Literatur: Gegenwartsliteratur ab 1945 (22,884)

Erzählende Literatur (20,294)

Krimis, Thriller, Spionage (15,676)

Kinderbücher bis 11 Jahre (15,561)

I J 0 (13,124)

4.1 (12,225)

1 (11,294)

Erzählerische Bilderbücher (9,870)

5.1 (9,777)

II J 0 (8,701)





17


Discover
Connect
Collaborate

 Triple

Vocabulary terms in the record

001220824  

The pre-school years.


 Published in Harmondsworth : by Penguin, in 1967.

144 p. : tables. ; 18 1/2 cm.

Series:

[Penguin educational special ES1](#)

Authority names

main personal names:  [Van der Eyken, Willem,](#) dates: [1933-](#), authority ID: [\(viaf\)102142410](#)

identifier / link

Subjects

Dewey Decimal [# 372.21/0942](#)

Classification:

topics:

[# Child development.,](#) vocabulary: [lcsh](#)

[# Education, Preschool.,](#) vocabulary: [lcsh](#)

different
vocabularies



Cataloging Source: original cataloging agency: [DLC](#), transcribing agency: [DLC](#), modifying agency: [DLC](#)

18


Discover
Connect
Collaborate



Vocabulary terms in the record

001220824  

The pre-school years.


 Published in Harmondsworth : by Penguin, in 1967.

144 p. : tables. ; 18 1/2 cm.

Series:

[Penguin educational special ES1](#)

Authority names

main personal names:  [Van der Eyken, Willem,](#) dates: [1933-](#), authority ID: [\(viaf\)102142410](#)

identifier / link

Subjects

Dewey Decimal Classification: [# 372.21/0942](#)

different vocabularies

topics:

[# Child development.,](#) vocabulary: [lcsh](#)

[# Education, Preschool.,](#) vocabulary: [lcsh](#)

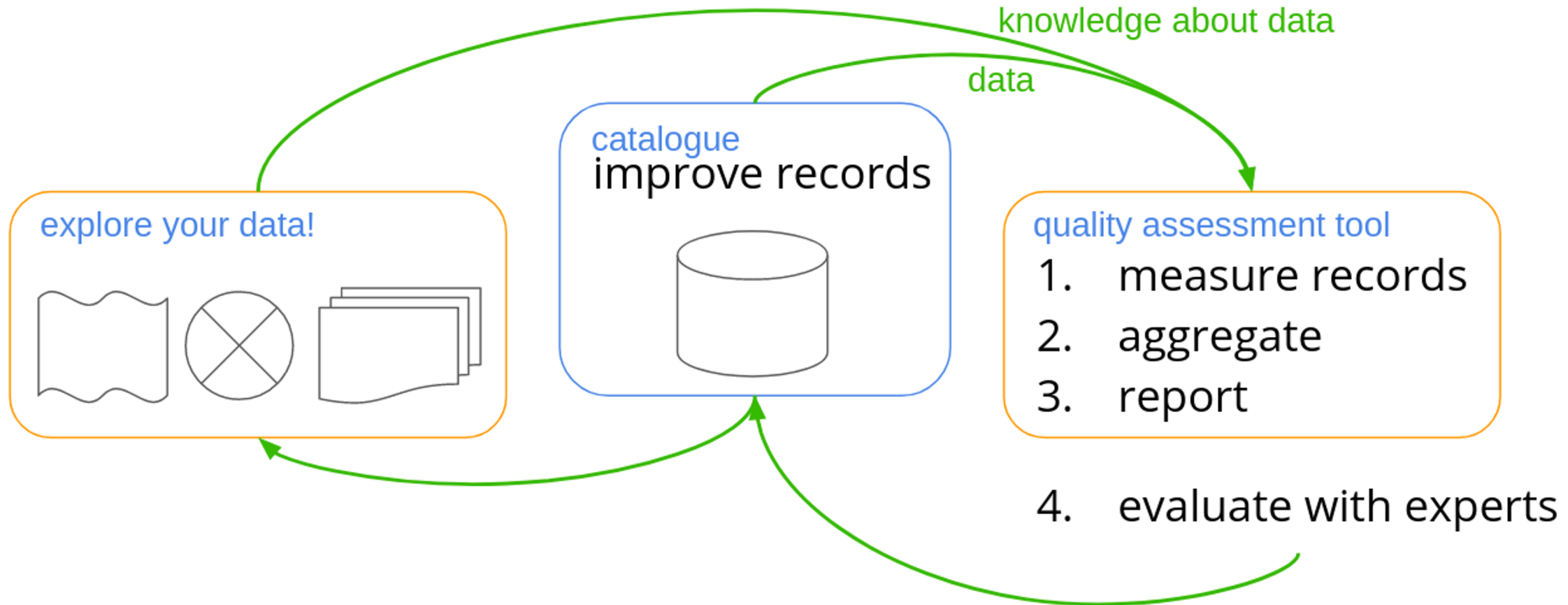
Cataloging Source: original cataloging agency: [DLC](#), transcribing agency: [DLC](#), modifying agency: [DLC](#)

19

Discover
Connect
Collaborate



QA lifecycle



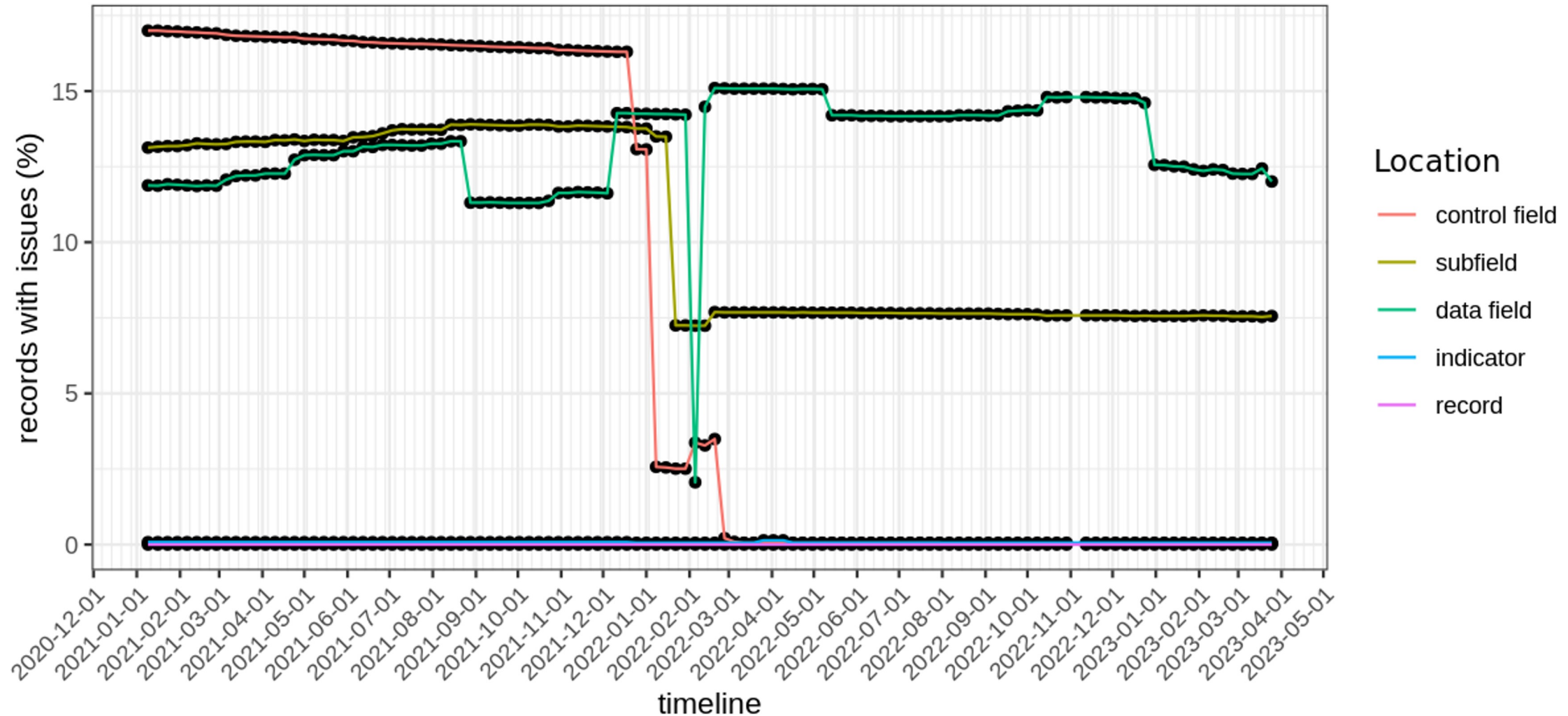
20

Discover
Connect
Collaborate



QA lifecycle

How different MARC issues changed over time



PLANS

REGARDING QA CATALOGUE

22

Discover
Connect
Collaborate



Plans

The quality of the subject terms in the metadata records

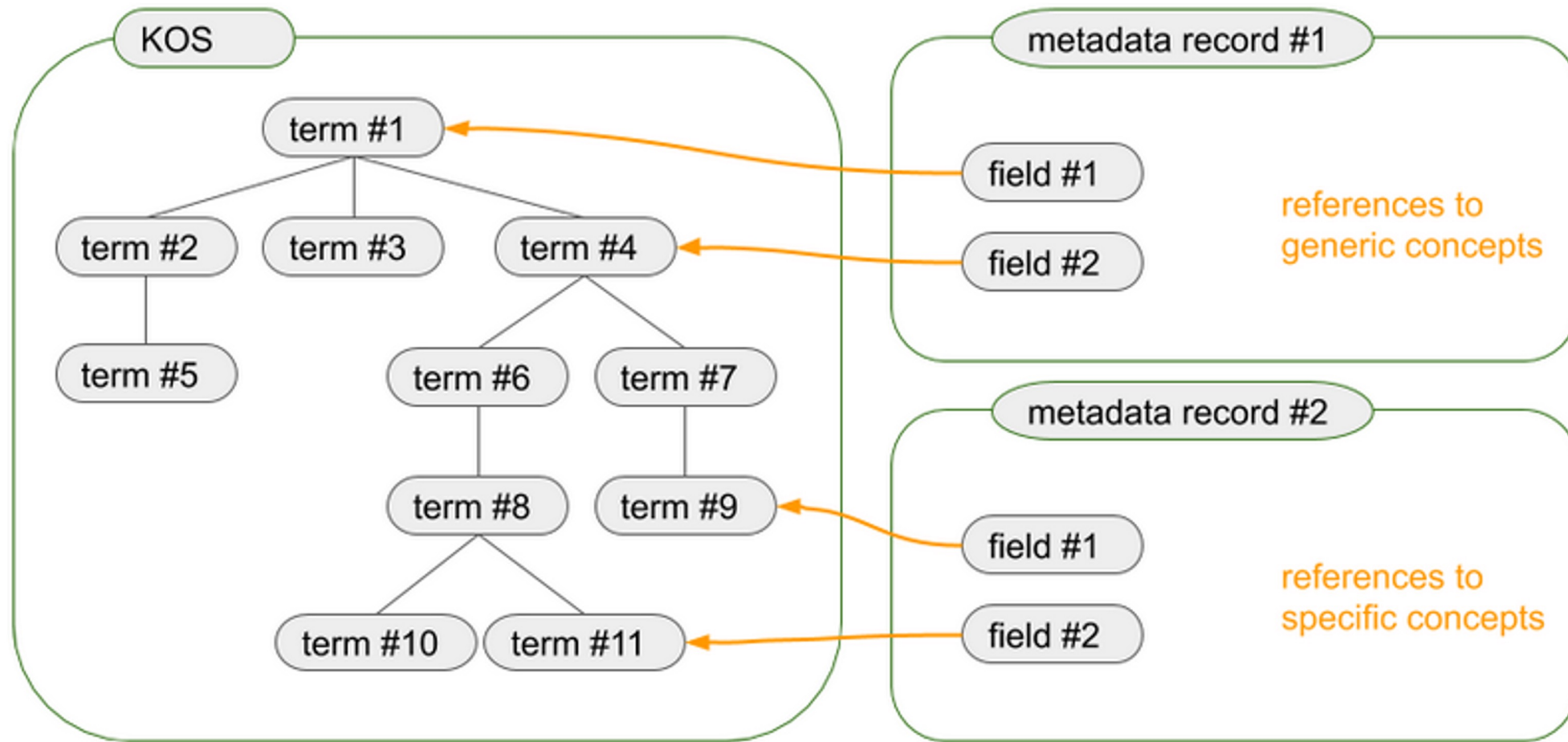
- Are the subject terms in the metadata record properly formatted?
- Do they represent a clear reference to a KOS?
- Are they specific or rather generic?
- Do multiple generic terms improve the specificity, and if so, how?

The quality of the connected KOS

- What is the expressive power of a KOS and what are its functionalities?
- What are the intrinsic qualities of a KOS?

<https://pkiraly.github.io/topics>

Plans



Plans

vocabulary: "**Dewey Decimal Classification**"

082\$a: Dewey Decimal Classification Number / Classification number

B (745,890)

K (247,702)

330 (228,764)

610 (176,348)

S (157,873)

370 (148,048)

650 (142,257)

620 (141,677)

050 (130,219)

914.3 (126,181)

340 (122,330)

781.64 (111,031)

910 (102,304)

570 (101,717)

530 (95,001)

300 (94,493)

validate and resolve with coli-conc/bartoc/UDC validator

25

Discover
Connect
Collaborate



About

QA catalogue

- backend: <https://github.com/pkiraly/metadata-qa-marc>
- UI: <https://github.com/pkiraly/metadata-qa-marc-web>
- web service: <https://github.com/pkiraly/metadata-qa-marc-ws>
- running instance: <http://gent.qa-catalogue.eu/metadata-qa/>

Partners

British Library, Royal Library of Belgium, Europeana, Gemeinsamer Bibliotheksverbund, Deutsche Digitale Bibliothek, Flemish Institute for Archives), Gent University Library, Victoria and Albert museum, German, Swedish, Israeli, Czech, Polish, Dutch, Hungarian, Scottish national libraries



26



**pkiraly.
github.io/
about/**



Follow us on:





Triple



*The **GoTriple** platform will be the
Discovery Service of the OPERAS
Research Infrastructure.*



The TRIPLE project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement number 863420.

PÉTER KIRÁLY

PKIRALY@GWDG.DE

CC BY 4.0 International Licence 