# Metadata Quality Benchmarks of ETDs in International Institutional Repositories: An Automated Appraisal

**ADITI ROY** (RESEARCH SCHOLAR)

&

**DR. SAPTARSHI GHOSH** (PROFESSOR)

DEPARTMENT OF LIBRARY & INFORMATION SCIENCE

UNIVERSITY OF NORTH BENGAL

RAJA RAMMOHUNPUR

WEST BENGAL, INDIA

# Introduction:

- Good metadata quality makes a record more discoverable, facilitating search and retrieval.
- Guy, Powell, and Day (2004) described metadata quality in terms of "**functional requirements**" or "**fitness for purpose**".
- According to Park's study, **consistency, correctness, and completeness** are the most critical factors in determining metadata quality (Park, 2009).
- This study provides a brief comparative account of Electronic Theses and Dissertation Metadata structure of Institutional Repositories and Libraries.

# Objectives:

The study aimed to focus on identifying the following objectives,

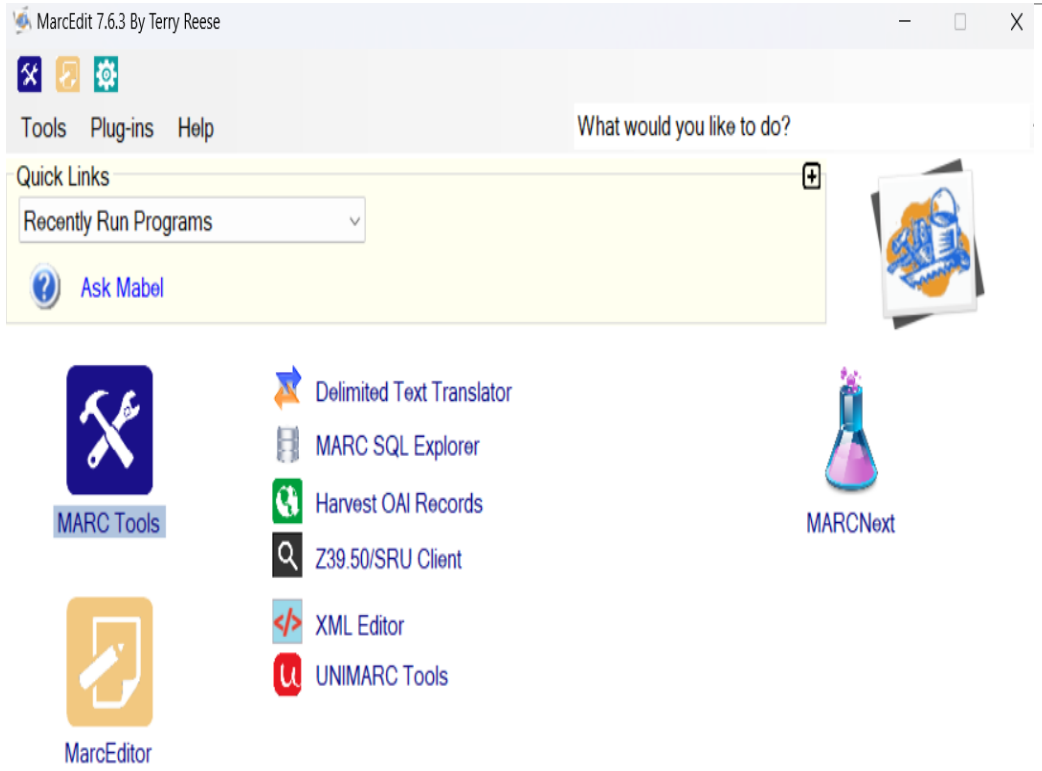
1. To identify the primary criteria that can be used to measure metadata quality.
2. To measure the metadata quality of ETDs in selected International Institutional Repositories.
3. To recognise the significant issues encountered in ensuring metadata quality.
4. To find out the primary mechanisms that can be used to improve metadata quality.

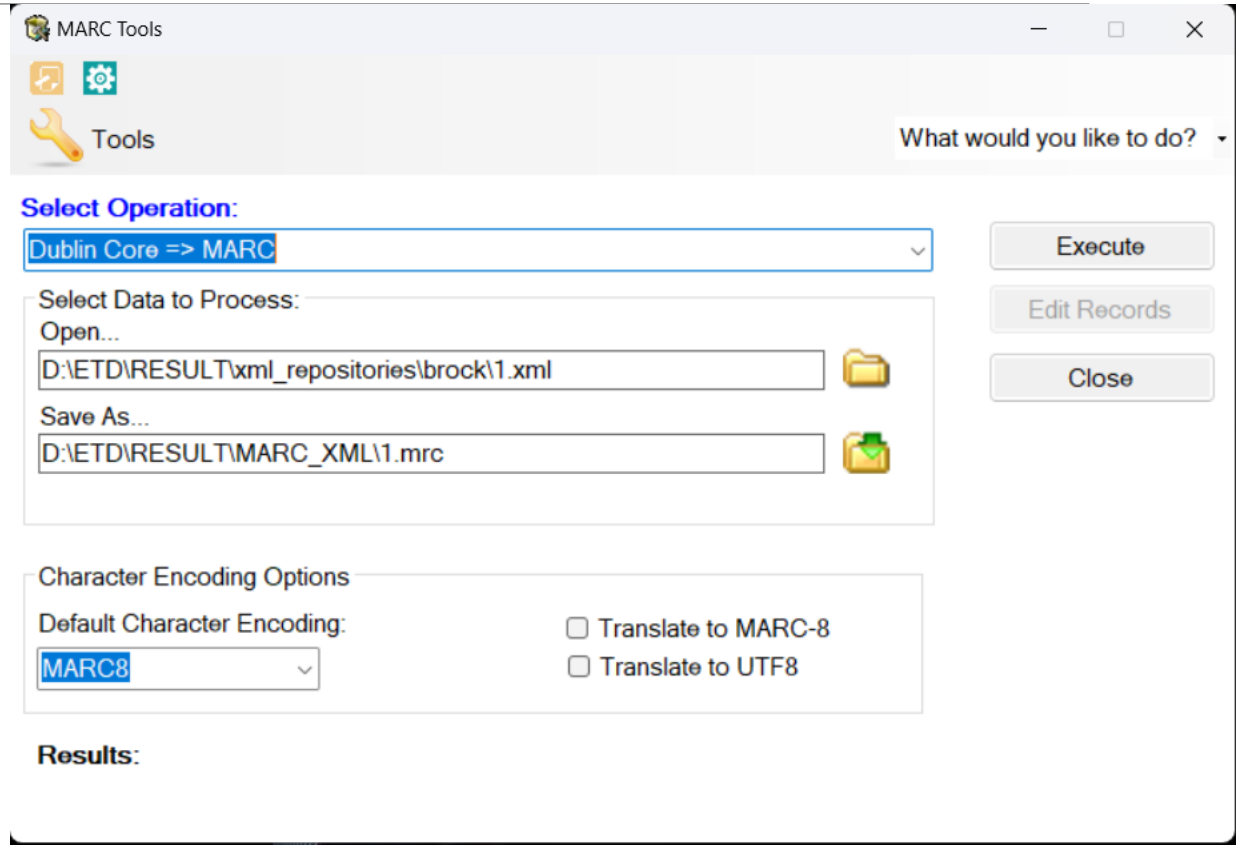
# Methodology:

# Data Analysis:

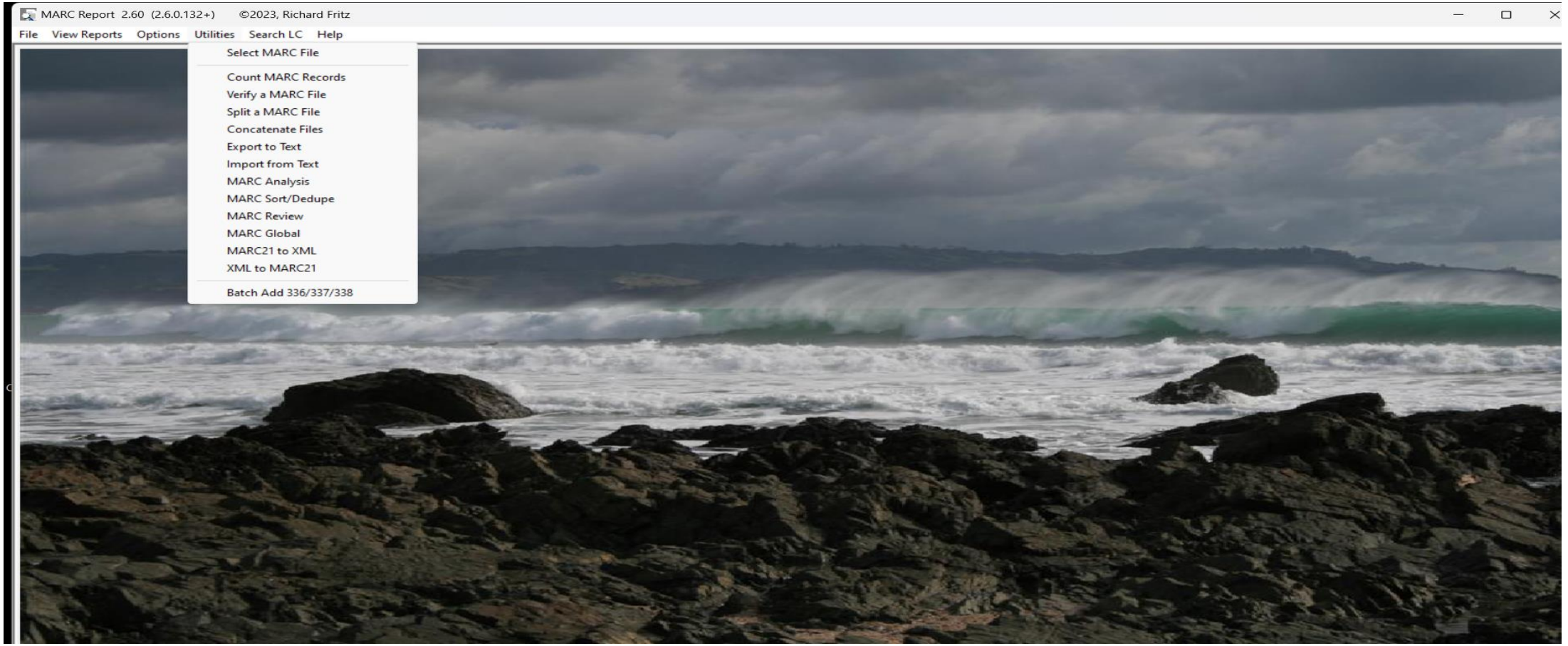
- **MarcEdit-** Using MarcEdit, we first converted the DCXML records of the Institutional Repositories to MARC21 records.
- **Marc Report -** We analysed each MARC21 record of IRs and Libraries using Marc Report utility plug-ins - Verify a MARC file and MARC Analysis.
- **Metadata-Analyzer-** A metadata-analyser is developed for the study to calculate the automated score of each IR and Library depending on the weightage of score distributions as prescribed on Metadata Quality Assessment provided by Consortium of data.europa.eu in July 2023.
- Lastly, a java based pre-compiled program by Peter Király has been used in this study. The program **"Metadata-Quality analysis-Marc"** is available on GitHub (https://github.com/pkiraly/metadata-qa-marc#configuration-1).

# MarcEdit:


MarcEdit 7.6.3 By Terry Reese
Tools
Plug-ins
Help
What would you like to do?
Quick Links
Recently Run Programs
Ask Mabel
MARC Tools
Delimited Text Translator
MARC SQL Explorer
Harvest OAI Records
Z39.50/SRU Client
XML Editor
UNIMARC Tools
MARCNext
MarcEditor


MARC Tools
Tools
What would you like to do?
Select Operation:
Dublin Core => MARC
Execute
Edit Records
Close
Select Data to Process:
Open...
D:\ETD\RESULT\xml_repositories\brock\1.xml
Save As...
D:\ETD\RESULT\MARC_XML\1.mrc
Character Encoding Options
Default Character Encoding:
MARC8
Translate to MARC-8
Translate to UTF8
Results:

# Marc Report


MARC Report 2.60 (2.6.0.132+) ©2023, Richard Fritz
File
View Reports
Options
Utilities
Search LC
Help
Select MARC File
Count MARC Records
Verify a MARC File
Split a MARC File
Concatenate Files
Export to Text
Import from Text
MARC Analysis
MARC Sort/Dedupe
MARC Review
MARC Global
MARC21 to XML
XML to MARC21
Batch Add 336/337/338

# Marc- Verify:

File Edit View

```
MARC VERIFY RESULTS     --  13-10-2023 11:21:59

Input file: C:\Users\aditi\OneDrive\Documents\MarcReport\Marc\marc_test_file.mrc
Output file: C:\Users\aditi\OneDrive\Documents\MarcReport\Marc\marc_test_file-verified.mrc
Config file: c:\users\aditi\onedrive\documents\marcreport\options\myVerifyOptions.cfg

100 records read from: C:\Users\aditi\OneDrive\Documents\MarcReport\Marc\marc_test_file.mrc
The size of the input file is 110113 bytes.

100 records written to: C:\Users\aditi\OneDrive\Documents\MarcReport\Marc\marc_test_file-verified.mrc
The size of the output file is 110113 bytes.


0 records marked as having UTF8 encoding (000/09).
0 records indicate the use of 'rda' description rules (040$e)

Format counts
  Bibliographic format: 100

Records processed per second = 2173
```

- 0 records marked as having UTF8 encoding (000/09).
- 0 records indicate the use of 'rda' description rules (040$e)

# Marc Analysis:

File Edit View

```
MARC Analysis, ©1996-2021, TMQ, Inc.

Run date: 28-07-2023 16:30

Report name: D:\ETD\RESULT\xml_repositories\all_agder-verified.mrc-analysis.txt

MARC Filename: D:\ETD\RESULT\xml_repositories\all_agder-verified.mrc

Config file: c:\users\aditi\onedrive\documents\marcreport\options\myAnalysisOptions.cfg

File size: 331972 bytes

MARC record count: 112

Total processing time: 61 milliseconds (00:00:00.061)

Records processed per second = 1836

--

Average record length: 2964

Mean Average record length: 7945 (1 records with this length)

Shortest record length: 613  (record number 35)

Longest record length: 7945  (record number 1)

Number of records with 040 $e = 'rda': 0

--

Total number of tags in the file: 2700

Average tags per record: 24

Mean Average tags per record: 19 (15 records with this tag count)

Most repeated tag in the file: 787  (12 times in record number 80)
```

File Edit View

```
Tag using the most bytes in the file: 520  (occupies 160140 bytes in 200 occurrences)

Longest single tag in the file: 6922 bytes  (tag 520 in record number 1)

--------------------------------------------------------------------------------

Number of unique tags in the file: 13

Tags present in the file:

024 042
245 246 260
520 540 546
653 655
720 786 787
856

Tags present in every record in the file:
024 042 245 260 546 655 720 856

--------------------------------------------------------------------------------

Tags not present in the file:
001 002 003 004 005 006 007 008 009
010 011 012 013 014 015 016 017 018 019 020 021 022 023 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 043 044 045 046 047 048
059 060 061 062 063 064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 080 081 082 083 084 085 086 087 088 089 090 091 092 093 094 095
100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136
147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183
194 195 196 197 198 199
200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236
249 250 251 252 253 254 255 256 257 258 259 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286
297 298 299
300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336
347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383
394 395 396 397 398 399
400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436
447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483
494 495 496 497 498 499
500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537
550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586
597 598 599
```

| Libraries | Total number of tags | Most repeated tag | Tags present in every record | Total number of subfield codes | Most repeated subfield code |
|---|---|---|---|---|---|
| British Library | 121582 | 650 (41 times in record number 664) | 001 005 008 245 | 214894 | 505 $t (96 times in record number 564) |
| Library of Congress | 5516 | 991 (20 times in record number 61) | 001 005 008 010 040 050 245 650 | 13105 | 850 $a (49 times in record number 64) |
| University of Colorado | 34827 | 020 (14 times in record number 131) | 008 040 245 300 | 67182 | 505 $t (58 times in record number 550) |
| University of Exeter | 39860 | 653 (100 times in record number 570) | 008 245 907 | 68354 | 505 $t (55 times in record number 774) |
| University of Maryland | 116547 | 852 (76 times in record number 2021) | 001 005 008 245 | 275134 | 505 $t (154 times in record number 2263) |
| Trent University | 41256 | 992 (32 times in record number 1404) | 001 008 035 245 300 852 992 993 | 78913 | 040 $d (25 times in record number 1492) |
| University of Virginia | 134484 | 700 (90 times in record number 62) | 001 008 245 926 | 222624 | 040 $d (149 times in record number 4697) |
| Institutional Repository | | | | | |
| Agder | 2700 | 787 (12 times in record number 80) | 024 042 245 260 546 655 720 856 | 2970 | 024 $a (1 time in record number 1) |
| Brock | 49887 | 653 (15 times in record number 2153) | 024 042 245 260 | 59800 | 024 $a (1 time in record number 1) |
| Brunel | 7627 | 856 (52 times in record number 293) | 024 042 245 260 655 | 9057 | 024 $a (1 time in record number 1) |
| Cranfield | 14653 | 856 (30 times in record number 792) | 024 042 245 260 655 720 | 19182 | 024 $a (1 time in record number 1) |
| Darius | 10 | 260 (3 times in record number 1) | 024 042 245 260 520 546 720 856 | 11 | 024 $a (1 time in record number 1) |
| IPB | 14443 | 856 (19 times in record number 146) | 024 042 245 260 720 856 | 16368 | 024 $a (1 time in record number 1) |

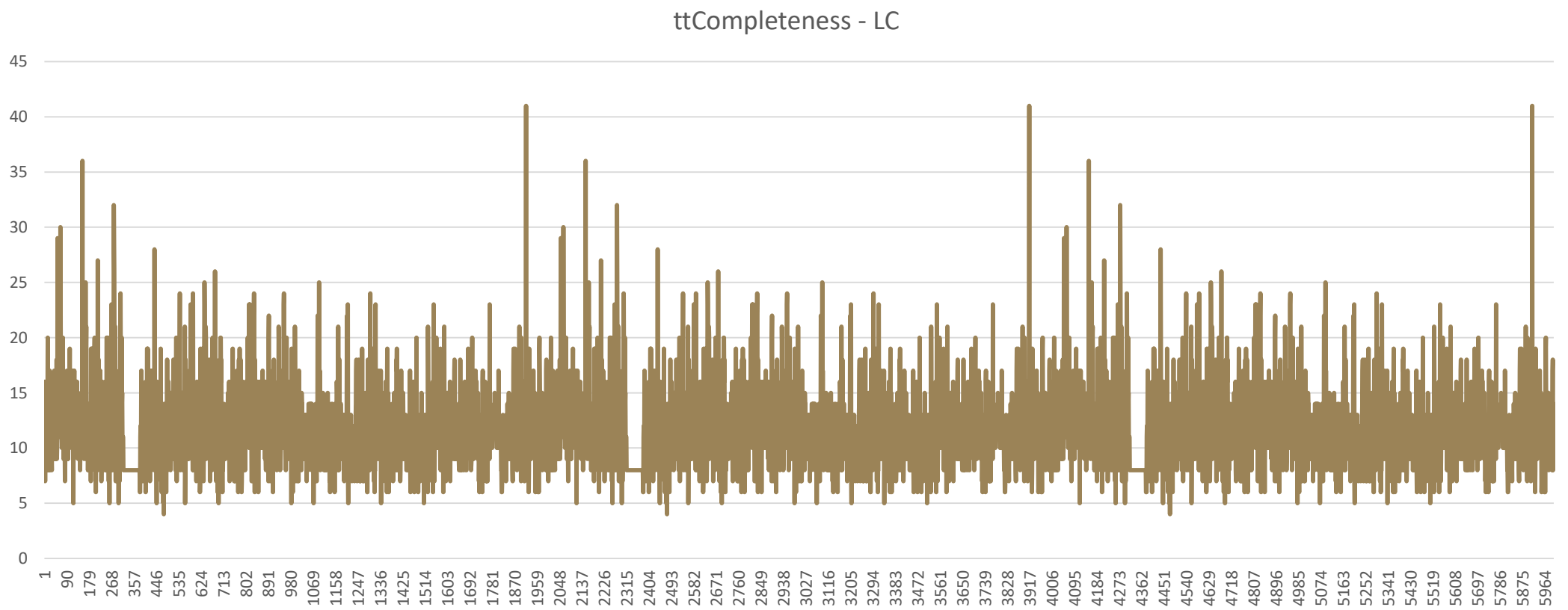
# Pitfalls Identified:

- The 040 $e=rda is absent in all records of the institutional repository metadata.
- Only 008 (Fixed Length Data Elements) and 245 (Title Statement) are present in each library record we studied.
- Though in the case of theses, *metadata publication statement is a mandatory field to identify the publisher, which is not given in each record of the libraries*.
- The record structure varies in the case of each record of the libraries, but the record structure is nearly identical in all records in the Institutional repositories.

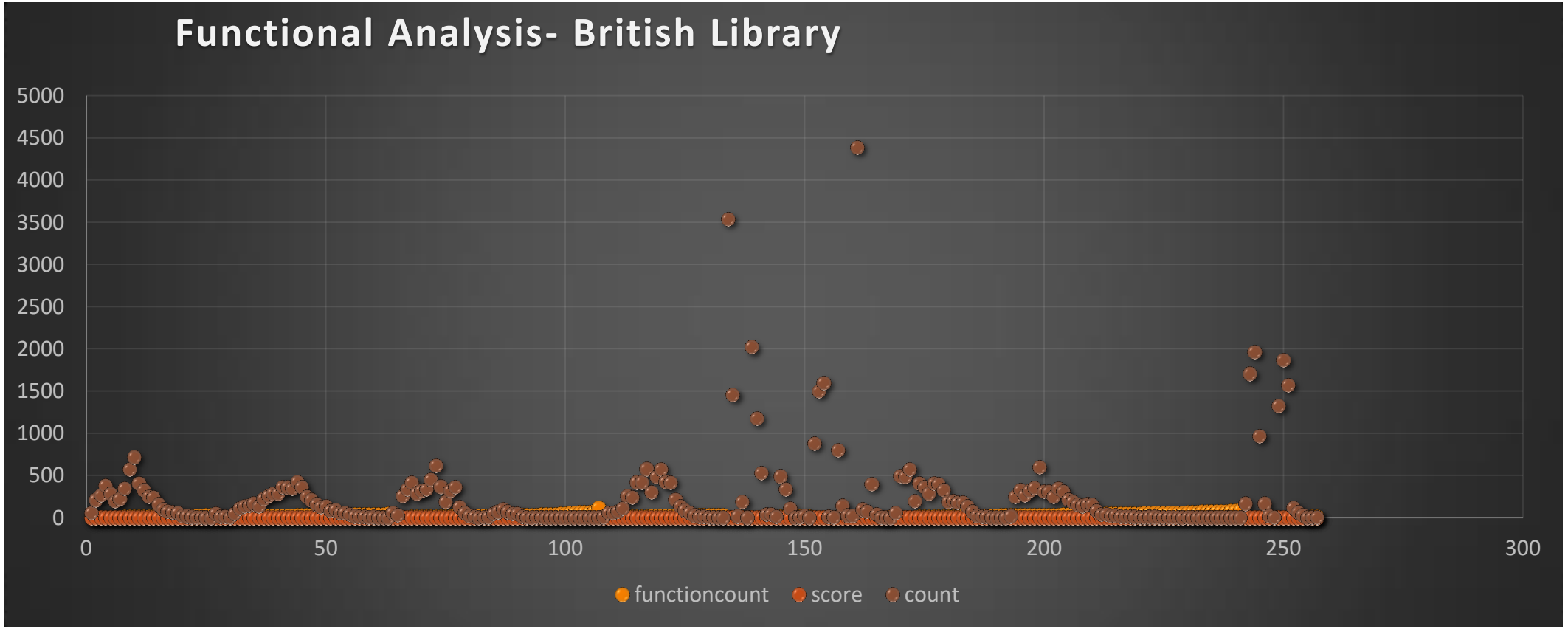
# Match Key Analysis:

| Libraries | Records without any Match Keys | Records with an LCCN | Records with an ISBN | Records with an ISSN | Records with an OCLC |
|---|---|---|---|---|---|
| British Library | 1463 | 1145 | 3215 | 87 | 375 |
| Library of Congress | 0 | 170 | 30 | 33 | 90 |
| University of Colorado | 0 | 144 | 613 | 0 | 503 |
| University of Exeter | 2 | 33 | 998 | 0 | 169 |
| University of Maryland | 1271 | 1731 | 1441 | 112 | 515 |
| Trent University | 9 | 15 | 437 | 0 | 1470 |
| University of Virginia | 15 | 340 | 739 | 4 | 4753 |

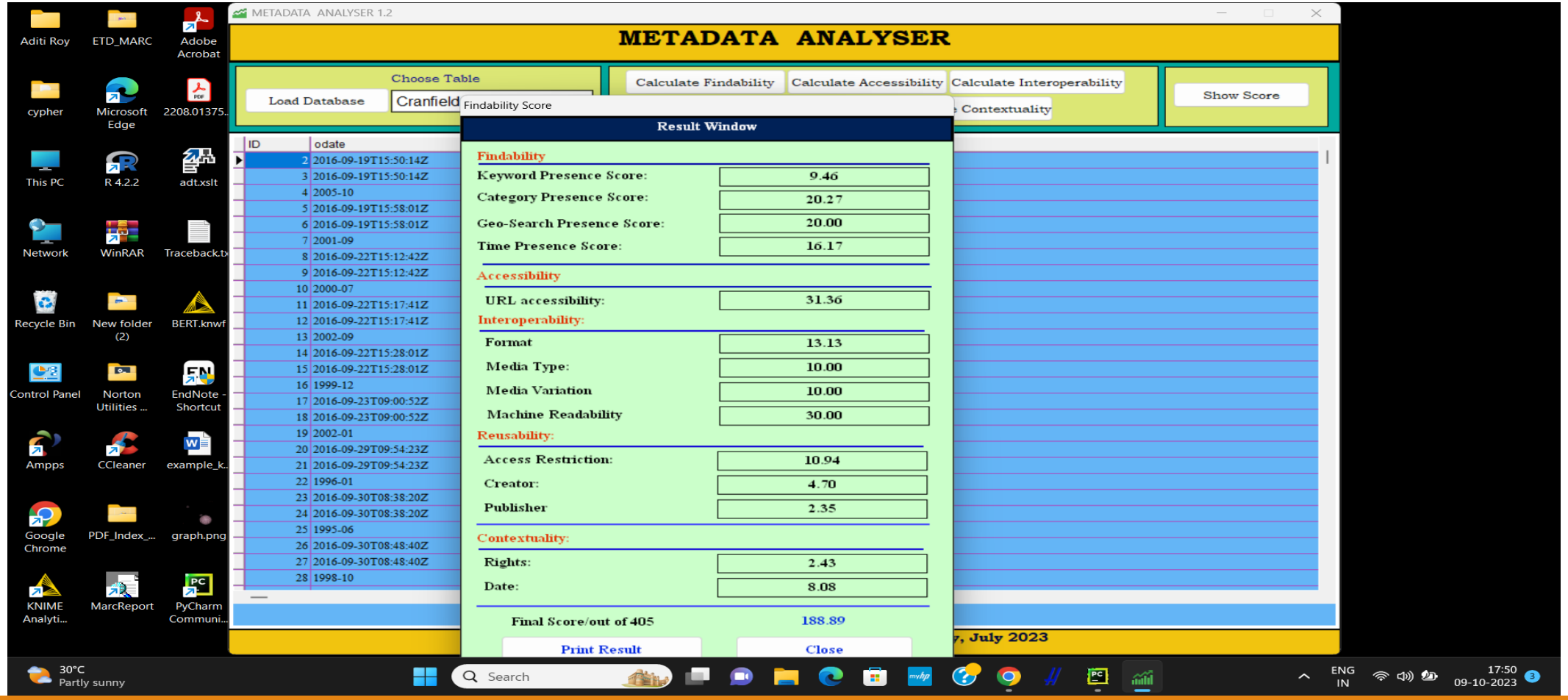
# Thompson Trail Completeness (tt-Completeness):



# Functional Analysis:



# Metadata-Analyzer-



# Metadata Quality Score:

| Repository | Total Score (405) | Rating |
|---|---|---|
| **Agder** | 197.48 | Sufficient |
| **Brock** | 224.12 | Good |
| **Brunel** | 203.97 | Sufficient |
| **Cranfield** | 188.89 | Sufficient |
| **Darius** | 282.15 | Good |
| **IPB** | 200.19 | Sufficient |
| **Library of Congress** | 395 | Excellent |
| **British Library** | 197.38 | Sufficient |
| **University of Exeter** | 138.91 | Sufficient |
| **University of Trent** | 150.59 | Sufficient |
| **University of Virginia** | 193.36 | Sufficient |
| **University of Maryland** | 118.01 | Bad |
| **University of Colorado** | 171.54 | Sufficient |

# Outcome-based Implementation:

- Precisely, this metadata quality analysis underscores the importance of robust metadata management for effective data utilisation.
- Repository managers can use this evaluation as a roadmap to enhance metadata quality, promote data discoverability, ensure accurate interpretation, and foster more informed decision-making processes.
- It also emphasises the need to continuously monitor and refine metadata practices to adapt to evolving data needs and technological advancements.

# Conclusion:

- As all of the required tags are not present in the case of Institutional Repository theses metadata, the java-based pre-compiled program by Peter Király cannot be used in those records to determine the tt-completeness test and functional analysis of the records.
- By maintaining robust metadata quality, organisations can mitigate the risk of erroneous interpretations and foster greater trust in data-driven initiatives.
- This, in turn, enhances collaboration across departments, aids compliance with regulations, and contributes to developing more accurate and valuable data assets.
- The assessment can serve as a foundation for targeted improvements in metadata management practices, ultimately contributing to more effective data-driven operations and decision-making.

# References:


- Choudhury, M. H., Salsabil, L., Jayanetti, H. R., Wu, J., Ingram, W. A., & Fox, E. A. (2023). MetaEnhance: Metadata Quality Improvement for Electronic Theses and Dissertations of University Libraries. http://arxiv.org/abs/2303.17661
- Day, M., Guy, M., & Powell, A. (2004). Improving the quality of metadata in eprint archives. Ariadne, 38.
- Elouataoui, W., El Alaoui, I., Gahi, Y. (2022). Metadata Quality in the Era of Big Data and Unstructured Content. In: Maleh, Y., Alazab, M., Gherabi, N., Tawalbeh, L., Abd El-Latif, A.A. (eds) Advances in Information, Communication and Cybersecurity. ICI2C 2021. Lecture Notes in Networks and Systems,vol 357. Springer, Cham https://doi.org/10.1007/978-3-030-91738-8_11
- Király, P. (2019). Measuring metadata quality. http://hdl.handle.net/21.11130/00-1735-0000-0003-C17C-8
- Metadata Quality Assurance. https://doi.org/10.08.2023
- Park, J.-R. (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Cataloging & Classification Quarterly, 47(3-4), 213–228. doi:10.1080/01639370902737240
- Romero-Pelaez, A., Segarra-Faggioni, V., & Alarcon, P. P. (2018). Exploring the provenance and accuracy as metadata quality metrics in assessment resources of OCW repositories. ACM International Conference Proceeding Series,
- Spencer, S., & White, H. (2019). Automated Techniques for Measuring Metadata Quality. https://zenodo.org/record/3612497