КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (27), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.27.700
УДК 025.4.036

Василенко Олег Олегович
аспірант спеціальності 123 Комп'ютерна інженерія
Інститут кібернетики імені В.М.Глушкова НАН України, Київ, Україна
викладач кафедри Інформаційних систем, технологій, фінансів та менеджменту
Український гуманітарний інститут, Буча, Україна
ORCID ID: 0000-0001-8498-2950
international@ugi.edu.ua

АНАЛІЗ ОСНОВНИХ МЕТАДАНИХ ДЛЯ ПОШУКУ ДУБЛІКАТІВ БІБЛІОГРАФІЧНИХ ЗАПИСІВ

Анотація. У даному дослідженні розглянуто проблему дублювання бібліографічних записів у бібліотечних інформаційних системах, яка стає дедалі актуальнішою в умовах зростання обсягів цифрових каталогів. Зокрема, досліджено основні метадані, які використовуються для порівняння записів і виявлення дублікованих даних. Аналіз охоплює ключові поля метаданих, такі як назва, ISBN, видавництво, місце видання, дата публікації, пагінація, серії та додаткові атрибути, що застосовуються для ідентифікації видань. Особливу увагу приділено варіативності даних у цих полях, зокрема проблемам, які виникають через переплутані підполя (наприклад, місце видання замість року або навпаки) та використання різних форматів дат, включаючи авторські права, діапазони або приблизні дати. Розглянуто специфіку оформлення багаторівневих записів, особливо для журналів і багатотомних видань та помилки міграції даних між різними автоматизованими бібліотечними інформаційними системами (АБІС). Дослідження демонструє, що навіть за наявності стандартів ISBD, UNIMARC та інших, у бібліографічних записах зберігається значна частка невідповідностей, що ускладнюють автоматизовану обробку. Поля, які містять видавництва та місця видання, характеризуються високою варіативністю значень, де унікальні дані становлять від 9% до 38% від загальної кількості записів. Під час кластеризації даних методом найближчого сусіда за алгоритмом PPM (Prediction by Partial Matching) виявлено можливість зменшення кількості унікальних значень на 6–24%, що свідчить про потенціал автоматизації для підвищення ефективності ручного редагування записів. Результати дослідження є вагомим внеском у розробку ефективних підходів до вдосконалення автоматизованих систем управління бібліографічними даними, оптимізації пошуку дублікатів та підвищення загальної якості бібліотечних баз даних.

Ключові слова: бібліографічний запис; бібліографічні метадані; пошук дублікатів; автоматизовані бібліотечні інформаційні системи; Prediction by Partial Matching; багаторівневий бібліографічний опис.

ВСТУП

Актуальність теми дослідження зумовлена швидким розвитком цифрових бібліотек та автоматизованих бібліотечних систем (АБІС), а також необхідністю забезпечення точності та ефективності в обробці великих обсягів бібліографічних даних. У сучасних умовах існує постійне збільшення кількості бібліографічних записів через різноманітні джерела та платформи. Наприклад, за статистикою, лише у 2020 році в рамках національних бібліотечних систем світу було зареєстровано понад 2,8 мільярда нових бібліографічних записів, з яких значна частина стосується видань, що постійно оновлюються або мають додаткові випуски (наприклад, журнальні статті, монографії у серіях) [1, с. 378].

Пошук дублікатів записів є критично важливим завданням для підтримки якості баз даних. Дублікати створюють проблеми для користувачів, спотворюють результати

пошуку та ускладнюють підтримку актуальності інформації в бібліотечних системах. На практиці це може призводити до зайвих витрат часу для бібліотекарів, які повинні здійснювати ручне порівняння та видалення дублікованих записів, а також до зниження довіри до автоматизованих систем з боку користувачів. За даними досліджень, наявність дублікованих записів може становити до 15–25% від загальної кількості записів у великих бібліографічних базах, що робить важливою автоматизацію цього процесу [2, с. 2786].

З огляду на це, автоматизація пошуку дублікатів, зокрема за допомогою методів кластеризації, таких як PPM (Prediction by Partial Matching), може значно полегшити цей процес. Наприклад, в одній із національних бібліотек серед 1,5 мільйона записів кластеризація дозволила зменшити кількість унікальних значень у базі на 6–24%. Це дозволяє скоротити час, витрачений на перевірку та обробку даних вручну, а також покращити точність бібліографічного опису [3, с. 6].

Крім того, проблема ускладнюється великою варіативністю форматів бібліографічного опису. На практиці часто використовуються різні стандарти: ISBD, UNIMARC, MARC21, кожен з яких має свої особливості оформлення. Згідно з дослідженнями, в 25% бібліографічних записів наявні проблеми з переплутаними підполями, що значно ускладнює пошук дублікатів та обробку таких записів [4, с. 725]. Тому важливим є удосконалення процесів виявлення дублікованих записів через розробку інтелектуальних алгоритмів, які можуть працювати з такими неповними чи некоректно заповненими полями.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

У зв’язку з розширенням обсягів даних та постійними змінами в технологіях, використання новітніх методів кластеризації та алгоритмів для пошуку дублікатів є критично важливим для забезпечення якісної роботи бібліотечних систем. Тому вдосконалення цих процесів забезпечить оптимізацію не лише технічних, але й організаційних аспектів роботи бібліотек, а також підвищить їх ефективність у обслуговуванні користувачів [6].

Було проаналізовано 23 962 бібліографічних записи з трьох бібліотек вищих навчальних закладів. У цих бібліотеках використовується машинозчитуваний формат MARC — UKRMARC, який є адаптованим різновидом UNIMARC. У цьому форматі бібліографічний запис зберігається у спеціалізованих полях, що можуть містити дані в окремих підполях. Перша цифра тризначного номера кожного поля визначає його належність до одного з десяти блоків полів [7]. Основна частина бібліографічного опису зберігається у полях другого блоку, які містять інформацію про авторів, назви, дати видання тощо.

Четвертий блок використовується для збереження інформації про зв’язки між записами, зокрема бібліографічного опису оригіналу твору, з якого зроблено переклад, або аналітичного опису статей, включених до збірників. Цей блок є критично важливим для інтеграції даних і забезпечення взаємозв’язків між різними елементами бібліографічних ресурсів.

Для оцінки якості бібліографічних записів важливим є аналіз повноти заповнення полів. У контексті машинного навчання і тренування нейронних мереж найбільш ефективним є набір даних, у якому всі поля заповнені повністю [8]. Незаповнені або частково заповнені стовпці суттєво погіршують якість навчання нейронної мережі, знижують її здатність робити точні прогнози та приймати обґрунтовані рішення.

Таким чином, особливу увагу необхідно приділити аналізу ступеня заповненості даних у базах бібліографічних записів. Це дозволить виявити слабкі місця в системах каталогізації, підвищити якість даних і забезпечити їх придатність для використання у сучасних інформаційних технологіях, включаючи алгоритми машинного навчання.

Дослідження А. Сітаса та С. Капідакіса, присвячене аналізу ефективності десяти алгоритмів знаходження дублікатів серед бібліографічних записів (CITAC, с. 293), показало, що найбільш часто для ідентифікації дублікатів використовуються такі поля даних: назва, автор, дата видання, пагінація, ISBN, видавець та контрольний номер Бібліотеки Конгресу (LCCN) [9]. Ці поля застосовувалися щонайменше у 50% випадків, оскільки вони містять ключову ідентифікаційну інформацію, яка дозволяє з високою ймовірністю встановити унікальність чи повторюваність записів.

Менш поширеним, але все ж значущим, було використання таких полів, як видання, місце публікації та серія. Ці поля застосовувалися у щонайменше 30% випадків. Їх менша популярність зумовлена тим, що інформація в цих полях може варіюватися навіть у рамках одного і того ж видання (наприклад, при перевиданні книги у різних містах чи створенні нової серії) [10]. Крім того, у деяких випадках ці поля можуть бути відсутніми у записах або їх заповнення може бути неповним чи неоднозначним.

Також важливо зазначити, що успішність алгоритмів значною мірою залежить від рівня стандартизації даних у полях та відсутності помилок у записах. Поля, такі як ISBN та контрольний номер Бібліотеки Конгресу, забезпечують найбільш надійну ідентифікацію, оскільки ці дані унікальні та стандартизовані на міжнародному рівні. У той же час, поля, пов’язані з текстовим описом (наприклад, назва чи автор), можуть створювати виклики через різні варіанти написання, використання символів, мовні відмінності тощо.

Загалом, результати дослідження підкреслюють необхідність об’єднання декількох полів для забезпечення максимальної точності алгоритмів знаходження дублікатів. Крім того, рекомендується враховувати специфіку бібліографічних записів у конкретних базах даних для налаштування алгоритмів під їхні особливості.

Для аналізу повноти заповнення даних у бібліографічних записах були обрані наступні ключові поля: назва, ISBN, місце видання, видавець, дата видання, сторінки (область кількісної характеристики) [10]. Ці поля були визначені на основі їхньої важливості для ідентифікації запису, а також їхньої поширеності в бібліографічних базах даних.

Оскільки більшість бібліотек України зберігають книги, написані не англійською мовою, контрольний номер Бібліотеки Конгресу (LCCN) не є характерним для більшості бібліографічних записів. Тому це поле було виключене з аналізу, оскільки воно не використовується для українських видань. Для спрощення процесу обробки даних також було вирішено не проводити аналіз полів, що містять інформацію про авторів, оскільки це поле може містити варіативність у написанні імен та прізвищ, що значно ускладнює автоматичну обробку та аналіз. У результаті, були зібрані та проаналізовані лише ті поля, що мають чіткі та стандартизовані характеристики, які не залежать від суб’єктивних інтерпретацій або мовних особливостей.

Кількісні характеристики проведеного аналізу представлено в таблиці нижче, що дозволяє оцінити ступінь заповненості кожного з вибраних полів у бібліографічних записах за кількістю наявних та відсутніх значень. Ця таблиця є важливою для подальшого визначення проблемних зон і розробки методів покращення заповнення даних у базах бібліографічних записів.

ISSN 2663 - 4023

Таблиця 1

Кількісні характеристики аналізу бібліографічних записів

	Абсолютні показники			Відносні показники		
	Lib1	Lib2	Lib3	Lib1	Lib2	Lib3
Всього бібліографічних записів	10371	10029	3562	100%	100%	100%
Записи, що містять *n* полів порівняння (Назва, ISBN, місце видання, видавець, дата, сторінки)						
— 6 полів/запис	4730	6284	2829	46%	63%	79%
— 5 полів/запис	3629	808	608	35%	8%	17%
— 4 поля/запис	1439	146	70	14%	1%	2%
— 3 поля/запис	513	99	35	5%	1%	1%
— 2 поля/запис	26	28	19	0%	0%	1%
— 1 поле/запис	34	2664	1	0%	27%	0%
Всього записів з виданням (205)	358	713	594	3%	7%	17%
Всього записів з ISSN (011)	1	10	2	0%	0%	0%

(Таблиця може містити, наприклад, такі стовпці: поле, кількість записів, що містять дані, кількість записів, де дані відсутні, процент заповненості для кожного поля тощо).

Більшість записів у представлених вище базах даних відносяться до монографічних видань, що є типовими для бібліографічних баз, які зберігають основний обсяг інформації. Серіальні видання зустрічаються рідше — у середньому в кожній бібліотеці налічується від 1 до 10 записів на серіальне видання, що свідчить про меншу поширеність таких матеріалів у вищезгаданих базах даних. За результатами попередніх досліджень, записів, що містять інформацію про видання, може бути від 3 до 17% від загальної кількості, що вказує на те, що це поле не є обов'язковим або найпоширенішим у кожному бібліографічному записі.

Залежно від визначення поняття «дублікат» і вимог до каталогізації, дані про видання можуть не використовуватися під час пошукових запитів для об'єднання різних версій одного твору. Оскільки інформація про видання може бути не настільки критичною для ідентифікації запису, її відсутність не впливає на загальну можливість об'єднання записів. У цьому контексті важливість поля «Видання» стає менш очевидною, особливо коли відсутність цього поля не перешкоджає зв'язку записів між собою.

Поле «Видання» має особливості формулювання при введенні даних згідно з міжнародним стандартом бібліографічного опису ISBD (International Standard Bibliographic Description) та національними стандартами, такими як ГОСТ 7.1, що вимагають чітких правил запису даних. Наприклад, номер видання записується арабськими цифрами (ISBD, с. 120), що є стандартом для багатьох міжнародних систем каталогізації. Цей підхід забезпечує єдність і стандартизованість бібліографічних записів, незважаючи на можливу варіативність формулювань, з якими можуть бути представлені різні видання в залежності від мови, країни або видавництва.

Незважаючи на таку варіативність, система автоматизованої обробки даних дозволяє обробляти записану інформацію навіть у випадку різних формулювань, що значно полегшує процес конвертування бібліографічних записів між різними форматами. Використання алгоритмів нечіткого пошуку і перевірки даних також дозволяє враховувати можливі помилки введення або варіації у формулюваннях, що знижує ймовірність виникнення помилок і дублювання записів при їх автоматичному обробленні.

За результатами кількісного аналізу всіх шести обраних полів порівняння, дані виявилися наступними: п’ять полів містять від 46% до 79% записів, тоді як шість полів/записів забезпечують від 71% до 96% заповненості даними. Ці результати дають змогу зробити певні припущення щодо заповненості окремих полів, зокрема, щодо можливості відсутності ISBN. Однак, варто зазначити, що використання ISBN як ідентифікуючої ознаки не є таким однозначним, оскільки кожен запис може мати свій власний набір ідентифікуючих ознак, що залежать від типу видання, формату та специфіки каталога.

Серед усіх перелічених полів, тільки поле «Назва» є обов’язковим для заповнення у бібліографічних записах. Без цього поля неможливо зберегти запис у базі даних. Це підкреслює важливість забезпечення наявності цього поля у кожному записі, адже воно є ключовим для ідентифікації документа. Відсутність інших полів, таких як ISBN або пагінація, не заважає збереженню запису, хоча це може вплинути на його повноту та точність для подальшої обробки та використання.

Особливо цікавою є ситуація з пагінацією документа. В першій базі даних спостерігається найменша кількість інформації щодо пагінації, що може бути пов’язано з різними практиками введення даних в окремих бібліотеках. Такі відмінності підкреслюють необхідність врахування специфіки кожної бази даних і її особливостей при проведенні аналізу. Кожна з баз даних має свої характеристики щодо структури записів, що визначає рівень їх повноти та точності при конвертуванні між різними форматами.

Таблиця 2

Ключові метадані бібліографічного запису

Полів/запис	6	5	4	3	2	1	Всього
Видавець	4730	3610	1397	494	0	0	10231
Місце видання	4730	3575	1378	489	1	0	10173
Дата	4730	3480	1257	13	4	0	9484
ISBN	4730	2922	170	9	8	0	7839
Сторінки	4730	929	115	21	13	0	5808
Всього	4730	3629	1439	513	26	35	10372

Цікавим є те, що в другій базі даних 27% записів містять лише одне поле для порівняння. Аналіз цих записів показує, що вони відповідають окремим журналам і містять посилання на аналітичні дані. Такі записи складені згідно з правилами багаторівневого бібліографічного опису, при якому другий рівень бібліографічного опису зберігається в окремому записі і додається до основного запису журналу через поле зв’язку. Така структура дозволяє більш детально відображати складові частини видання, зокрема окремі статті або розділи, що містяться в збірнику, зберігаючи їх зв’язок з основним записом.

Друга база даних, що спочатку велася в автоматизованій бібліотечній інформаційній системі (АБІС) Ірбіс, після чого мігрувала в АБІС Koha, має складнішу організацію, що зумовлено специфікою обробки даних в обох системах. Міграція між різними АБІС може спричиняти певні ускладнення, зокрема, через відмінності в методах організації даних, що, своєю чергою, впливає на структуру та доступність бібліографічних записів. Це пояснює більш складну структуру записів у другій базі даних порівняно з іншими, де таке багаторівневе посилання на додаткові записи не використовується.

ISSN 2663 - 4023

Нижче наведено скорочений приклад двох MARC записів, які разом складають багаторівневий бібліографічний опис випуску журналу. Цей приклад демонструє, як різні рівні інформації взаємодіють у межах одного бібліографічного запису, що дозволяє створювати більш точні та детальні описи серіальних видань.

```
000 00851nam2a2200157 i 4500
001 7857
090 _ _ ‡a 7857
100 _ _ ‡a 20070510d2004    |||y0rusy50
101 0 _ ‡a rus
102 _ _ ‡a RU
200 0 _ ‡a 2004. N 1
461 _ 0 ‡0 RU/TCI1/001064 ‡t Рідна школа ‡v 1
464 _ 0 ‡0 RU/TCI1/005028 ‡t Демократична культура особистості: сутність та шляхи формування ‡f Є. Хриков ‡a Хриков, Євген Луганск
464 _ 0 ‡0 RU/TCI1/005029 ‡t Особливості сучасної освітньої системи в Великій Британії ‡f С. Старовойт ‡a Старовойт, Світлана Украина
464 _ 0 ‡0 RU/TCI1/005030 ‡t Микола Амосов: філософія здоров'я ‡f Т. Мостова ‡a Мостова, Таїсія
801 _ 0 ‡b UA-KsTCI ‡a UA ‡c 20070510
000 00522nas1a2200145 i 4500
001 7856
011 _ _ ‡a 0131-6788
090 _ _ ‡a 7856
100 _ _ ‡a 20070510a19229999|||y0rusy50
101 0 _ ‡a ukr
102 _ _ ‡a UA
110 _ _ ‡a afu||||0|||
200 1 _ ‡a Рідна школа ‡e Щомісячний науково-педагогічний журнал ‡f Міністерство освіти і науки України
210 _ _ ‡a Київ ‡c Видавниче підприємство "Деміур" ‡d 1922-
801 _ 0 ‡b UA-KsTCI ‡a UA ‡c 20070510
```

Рис. 1. Зразок багаторівневого бібліографічного опису випуску журналу

Однією з проблем міграції між автоматизованими бібліотечними інформаційними системами (АБІС) є те, що ідентифікатор загального запису, який зберігається в підполі $0 поля зв’язку 461, не відповідає ідентифікатору загального запису, що зберігається в полі 001. Це ускладнює процес пошуку дублікатів, оскільки виникає потреба у додатковому пошуку зв’язків між записами всередині однієї бази даних. Без належного зв’язку між ідентифікаторами записів у різних полях неможливо безпечно визначити наявність дубліката або точного зв’язку між версіями одного і того ж твору, що призводить до втрат інформації і порушення процесу злиття записів. Для отримання повної інформації про кожен твір необхідно враховувати ці зв’язки, що потребує більш складних алгоритмів і додаткових перевірок у межах однієї системи.

Подальший аналіз показує, що кожне з полів має свої особливості використання, які також впливають на якість та повноту бібліографічних записів. Зокрема, деякі поля, такі як «ISBN» або «видавець», можуть бути неповними або відсутніми в записах про книги, видані приватними особами чи малими видавництвами. Це може статися через різні причини, зокрема через неформальний характер видавничого процесу або відсутність фінансових можливостей для повного опису видання за стандартами, передбаченими для великих видавців. Відсутність певних елементів бібліографічного опису у таких записах ускладнює процес автоматизованої обробки даних, зокрема, виявлення і злиття дублікованих записів, що може негативно вплинути на якість бібліографічного обліку і зручність користувачів під час пошуку потрібної інформації.

Дата публікації є важливим елементом бібліографічного опису, і в результатах аналізу показано, що від 72% до 98% записів у розглянутих бібліотеках містять таку інформацію. Більшість записів (від 47% до 91%) мають лише одну дату публікації, що значно спрощує процес порівняння записів, оскільки дозволяє однозначно ідентифікувати видання за датою. Однак варто відзначити, що до 25% записів містять більше однієї дати. Це може бути викликано кількома факторами, зокрема:

1. ***Додаткові дати в полях зв’язку:*** Одним із основних джерел додаткових дат є поля зв’язку, де може бути вказана дата публікації твору мовою оригіналу або дата першого видання. Це особливо характерно для перевидань або для

книг, що були перекладені. У більшості випадків додається лише одна додаткова дата, однак є випадки, коли кількість таких дат може бути більшою.

2. ***Множинні твори в одній книзі:*** Поточні видання можуть містити кілька творів, які раніше публікувалися окремо. Це також може призвести до кількох дат у одному записі, оскільки кожен твір може мати свою дату публікації.
3. ***Багаторівневий бібліографічний опис:*** У деяких випадках застосовується багаторівневий бібліографічний опис, який передбачає, що один запис може містити зв'язок не лише з набором книг, але й з наступним томом багаточастинного твору. У такому випадку кожен том або частина твору можуть мати свою окрему дату публікації, що також призводить до наявності кількох дат у бібліографічному описі.

Ці особливості роблять процес аналізу і порівняння бібліографічних записів більш складним, оскільки для точного визначення дубліката або відповідності між різними версіями твору необхідно враховувати всі дати, які можуть бути присутні в записі.

Таблиця 3

Аналіз і порівняння бібліографічних записів

	Абсолютні показники			**Відносні показники**		
	Lib1	Lib2	Lib3	Lib1	Lib2	Lib3
Всього бібліографічних записів	10371	10029	3562	100%	100%	100%
Всього записів з датами	9484	7240	3486	91%	72%	98%
— 3 дати/запис	1	142	6	0%	1%	0%
— 2 дати/запис	42	2400	476	0%	24%	13%
— 1 дата/запис	9441	4698	3004	91%	47%	84%
містять нецифрові символи	3850	166	239	37%	2%	7%

Окрім різної кількості дат у записі, кожне поле може мати власні ускладнення, зокрема пов'язані з наявністю недрукованих символів або інших нестандартних елементів, які ускладнюють їх подальшу обробку. Аналіз показав, що від 2% до 37% записів містять такі недоліки, що може серйозно вплинути на якість і точність обробки даних. Ось основні причини наявності недрукованих символів у полях, що містять дату:

У деяких випадках запис містить дату авторських прав, яка не завжди має стандартний вигляд. Наприклад, символи «©2015», «c1999«, або «cop. 1990» можуть бути використані для позначення року публікації, але такі варіанти складно обробляти за допомогою автоматизованих систем. Однак ці дати можна привести до більш стандартного формату, що дозволить їх використовувати для подальшої обробки.

Інколи у записах міститься кілька дат, що можуть бути написані через кому або тире, наприклад: «1991, 1996» або «2008–2014». Такий формат ускладнює обробку, оскільки системи, що використовуються для пошуку дублікатів, часто не можуть правильно інтерпретувати діапазони дат або множинні значення без додаткових налаштувань або алгоритмів для порівняння.

За правилами каталогізації, якщо дата публікації є приблизною, то вона записується в квадратних дужках, наприклад: «[?]» або «[2007]». Такі записи можуть бути проблемними, оскільки ці дати є менш точними та їх важко використовувати при пошуку дублікатів або при автоматичному порівнянні записів. Крім того, їх наявність може призвести до помилок у процесі каталогізації або створення невідповідностей у базах даних.

У деяких випадках, підполя в бібліографічному записі можуть бути переплутані між собою. Наприклад, підполя, що містять інформацію про назву видавництва та дату публікації, можуть бути записані в неправильному порядку, що ускладнює автоматизовану обробку та може призвести до непередбачуваних результатів. Іноді переплутані можуть бути навіть не лише поточні підполя, а й сусідні, як, наприклад, інформація про сторінки або назву в полях зв'язку.

У деяких записах може бути зазначено, що дата публікації невідома, наприклад, позначка «[б.г.]» (без дати). Це є стандартною практикою в каталогізації, коли дата не визначена, однак для подальшої обробки таких записів це є проблемою, оскільки відсутність точної дати ускладнює процес порівняння записів і пошук дублікатів. За правилами каталогізації такі записи зазвичай не включають додаткової інформації про дату.

Ці проблеми підкреслюють важливість ретельного налаштування процесів автоматизованої обробки бібліографічних записів, зокрема для правильного інтерпретування та обробки дат, що можуть бути представлені в нестандартних формах або з неточними даними.

Назва видавництва зустрічається в 72–99% бібліографічних записів розглянутих бібліотек. Зокрема, від 47 до 99% записів містять назву видавництва лише в одному блоці, що майже завжди відповідає полю другого блоку опису. Однак у деяких випадках (до 25% записів) назва видавництва може бути вказана в двох різних блоках. Це може бути наслідком різних підходів до структурування бібліографічного запису або необхідності вказати додаткову інформацію в окремих полях для більш детального опису видання. Важливо зазначити, що це може ускладнити процес автоматизованої обробки записів, оскільки системи можуть мати труднощі з правильною інтерпретацією або з'єднанням даних, що містяться в різних блоках.

Таблиця 4

Процес автоматизованої обробки бібліографічних записів

	Абсолютні показники			**Відносні показники**		
	Lib1	Lib2	Lib3	Lib1	Lib2	Lib3
Всього бібліографічних записів	10371	10029	3562	100%	100%	100%
Всього записів з видавництвом	10231	7246	3469	99%	72%	97%
— з них заповнено 2 блоки/запис	0	2536	96	0%	25%	3%
— з них заповнено 1 блок/запис	10231	4710	3373	99%	47%	95%
Унікальних значень	3333	2742	1292	32%	27%	36%
Унікальних значень після кластеризації	2543	2516	1156	76%	92%	89%
Кількість кластерів:						
— Методом колізії ключів Fingerprint	383	77	96			
— Методом найближчого сусіда за відстанню Левенштейна	749	164	120			
— Методом найближчого сусіда за відстанню PPM	449	269	149			

Дещо дивним є те, що в першій бібліотеці не було жодного запису, що містить назву видавництва в 4 блоці. Причиною цього може бути як відсутність відповідної інформації в самих книгах, так і рішення бібліотеки не заповнювати відповідне поле. Однак детальніший аналіз показав, що хоча частина записів і містить потрібну інформацію, вона була занесена не в правильне підполе. Замість підполя $n, яке має

містити назву видавництва в 4 блоці, більшість із 41 знайденого запису містять назву видавництва в підполі $e, що за стандартами повинно містити відомості про видання (наприклад, місце видання, рік видання). Це свідчить про помилки при заповненні полів або недотримання стандартів при створенні бібліографічних записів, що може ускладнити подальшу обробку даних та пошук дублікатів.

Частина записів містить більше ніж одне видавництво в кожному полі. У випадку з полем 210 ситуація дещо простіша, оскільки кожне видавництво зберігається в окремому підполі. Однак у 4 блоці використовується UNIMARC техніка стандартних підполів, згідно з якою кілька видавництв можуть бути збережені в одному полі. Це створює додаткові складнощі при обробці даних, оскільки для правильного зчитування і аналізу таких записів необхідно чітко розрізняти різні видавництва в межах одного запису.

Слід зазначити, що поле, яке містить назву видавництва, характеризується високою варіативністю даних. Аналіз показує, що з усієї кількості записів, що містять інформацію про видавництво, лише 33–38% значень є унікальними (що складає від 27 до 36% від загальної кількості записів). Це свідчить про те, що на кожне видавництво припадає близько трьох найменувань книг, що дає можливість групувати записи за видавництвами.

Після кластеризації за методом найближчого сусіда (KNN — K-Nearest Neighbors) за відстанню PPM (Prediction by Partial Matching — передбачення на основі часткового збігу) кількість унікальних значень зменшилася на 8–24%, що в абсолютних цифрах відповідає 136–790 унікальним назвам. Це свідчить про можливість значного зменшення варіативності при обробці даних. Автоматизація цього процесу, навіть при такому невеликому зменшенні відсотку унікальності, може суттєво заощадити час, який зазвичай витрачається на ручне редагування записів. Це, у свою чергу, підвищує ефективність роботи бібліотек і зменшує ймовірність помилок при введенні і обробці бібліографічної інформації. На рисунку нижче наведено приклад одного зі знайдених кластерів, що ілюструє подібні групи даних.

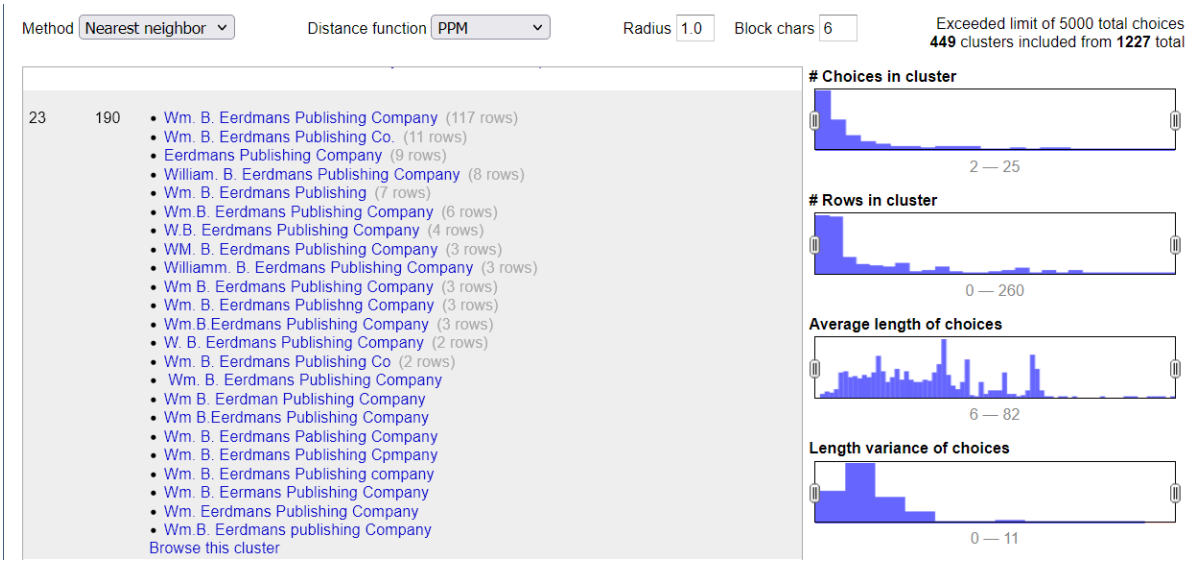

Рис. 2. Зразок знайдених кластерів для ілюстрації подібних груп даних

Місце видання містять від 69 до 98% бібліографічних записів розглянутих бібліотек. Однак важливо зазначити, що до 8% від загальної кількості записів містять інформацію про невідоме місце видання. У таких випадках використовуються або стандартні скорочення, як-от [б.м.] (без місця видання) чи [s.l.] (лат. sine loco — без

місця), або певні значення, що зазначаються в квадратних дужках, наприклад, [невідомо] або [не вказано]. Це свідчить про необхідність додаткової уваги до таких записів при їх подальшій обробці, оскільки наявність невизначених значень у полі місця видання може ускладнити пошук дублікатів або автоматизовану обробку бібліографічних даних. Для підвищення точності аналізу таких записів доцільно здійснювати їх подальше уточнення або коригування на етапі редагування даних.

Таблиця 5

Уточнення або коригування на етапі редагування даних

	Абсолютні показники			**Відносні показники**		
	Lib1	Lib2	Lib3	Lib1	Lib2	Lib3
Всього бібліографічних записів	10371	10029	3562	100%	100%	100%
Всього записів з місцем видання	10173	6965	3495	98%	69%	98%
— з них місце не відомо	877	25	16	8%	0%	0%
Унікальних значень	1146	622	304	11%	9%	9%
Унікальних значень після кластеризації	909	570	287	79%	92%	94%
Кількість кластерів						
— Методом колізії ключів Fingerprint	120	13	19			
— Методом найближчого сусіда за відстанню Левенштейна	318	18	20			
— Методом найближчого сусіда за відстанню PPM	350	62	22			

Унікальних значень місця видання становить від 9 до 11% від загальної кількості записів, що мають заповнене місце видання. Після кластеризації методом найближчого сусіда за відстанню PPM (Prediction by Partial Matching — передбачення щодо часткового збігу) кількість унікальних значень зменшилась на 6–21%, що в абсолютних показниках відповідає 17-237 унікальним місцям видання. Важливо зазначити, що при використанні цього методу кластеризації за замовчуванням застосовувалося 6 символів на блок, тому не був утворений кластер для варіацій скорочень відсутності місця видання, таких як (б. м.; б./м.; б.м.; [б. м.]), а також для коротших назв місць видання, як-от U. S. A.; U.S.A.; USA, або різні варіанти написання назв міст, наприклад, Київ; Київ.; Київ; Київ, Л.; Л. Додатково було виявлено кілька помилок у даних, зокрема переплутані поля (наприклад, рік замість місця видання) або поєднання двох різних місць видання в одному полі (наприклад, Київ-Одеса). Ці проблеми можуть ускладнити точність кластеризації та потребують подальшого очищення і коригування даних перед використанням в автоматизованих системах для пошуку дублікатів чи інших цілей.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В результаті проведеного аналізу основних метаданих, що використовуються для пошуку дублікатів бібліографічних записів, було виявлено, що найбільш ефективними для порівняння є поля, такі як назва, ISBN, видавництво, місце видання та дата публікації. Проте, значні труднощі виникають через варіативність записів, зокрема у полях, що стосуються даних видавництва та дати публікації, а також через переплутані підполя в бібліографічних записах, що ускладнює автоматизовану обробку даних. Метод кластеризації PPM виявив свою ефективність у зменшенні кількості унікальних значень

і значно покращив процес автоматизованої обробки бібліографічних даних, допомагаючи усунути дублікати та знизити витрати часу на ручне редагування. Однак, варто зазначити, що для організації багаторівневих записів, зокрема для журналів, необхідний особливий підхід, оскільки такі записи мають складні зв'язки між собою, що потребує додаткової уваги при їх обробці. Однією з важливих проблем є процес міграції між різними автоматизованими бібліотечними інформаційними системами (АБІС), що створює додаткові складнощі через непослідовність ідентифікаторів записів. Це ускладнює пошук дублікатів і потребує застосування більш точних методів для синхронізації даних між різними системами. Тому важливо впроваджувати новітні алгоритми та технології для ефективного вирішення цих проблем, що забезпечить більшу точність і надійність бібліографічних записів у сучасних бібліотечних системах.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Baas, J., Schotten, M., Plume, A., & Côté, G. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative science studies*, *1(1)*, 377–386. https://doi.org/10.1162/qss_a_00019
2. Beesley, L., Bondarenko, I., Elliot, M., & Kurian, A. (2021). Multiple imputation with missing data indicators. *Stat Methods Med Res.*, *30(12)*, 2685–2700. https://doi.org/10.1177/09622802211047346
3. Burnham, J. F. (2006). Scopus database: a review. *Biomedical digital libraries*, *3(1)*, 1–8. https://doi.org/10.1186/1742-5581-3-1
4. Ceasar, S. A., & Ignacimuthu, S. (2023). CRISPR/Cas genome editing in plants: Dawn ofAgrobacterium transformation for recalcitrant and transgene-free plants for future cropbreeding. *Plant Physiology and Biochemistry*, *196*, 724–730. https://doi.org/10.1016/j.plaphy.2023.02.030
5. Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5(1)*, 31–49. https://doi.org/10.1162/qss_a_00286
6. Elango, B. (2024). Duplication issues with the new interface of Scopus. *INFONOMY*, *2*. https://doi.org/10.3145/infonomy.24.015
7. Elango, B., & Matilda, S. (2023). Mapping thecybersecurity research: A scientometric analysis of Indian publications. *Journal of ComputerInformation Systems*, *63(2)*, 293–309. https://doi.org/10.1080/08874417.2022.2058644
8. Elango, B., Kozak, M., & Rajendran, P. (2019). Analysis of retractions in Indian science. *Scientometrics*, *119(2)*, 1081–1094. https://doi.org/10.1007/s11192-019-03079-y
9. Hammer, B., Virgili, E., & Bilotta, F. (2023). Evidence-based literature review: De-duplication a cornerstone for quality. *World J Methodol*, *13(5)*, 390–398. https://doi.org/10.5662/wjm.v13.i5.390
10. Krauskopf, E. (2018). An analysis of discontinued journals by Scopus. *Scientometrics*, *116(3)*, 1805–1815. https://doi.org/10.1007/s11192-021-03948-5
11. Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: acomparative analysis. *Scientometrics*, *106*, 213–228. https://doi.org/10.1007/s11192-015-1765-5
12. Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographicinformation in today's academic world. *Publications*, *9(1)*. https://doi.org/10.3390/publications9010012
13. Tennant, J. P. (2020). Web of Science and Scopus are not global databases of knowledge. *European Science Editing*, *46*. https://doi.org/10.3897/ese.2020.e51987
14. Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of informetrics*, *12(2)*, 430–435. https://doi.org/10.1016/j.joi.2018.03.006
15. Thelwall, M., & Sud, P. (2022). Scopus 1900–2020: Growth in articles, abstracts, countries,fields, and journals. *Quantitative Science Studies*, *3(1)*, 37–50. https://doi.org/10.1162/qss_a_00177
16. АПН України & Держ. наук.-пед. б-ка України ім. В. О. Сухомлинського. (2010). *Упровадження в практику роботи бібліотек освітянської галузі ДСТУ ГОСТ 7.1:2006 «Бібліографічний запис. Бібліографічний опис. Загальні вимоги та правила складання» та ДСТУ ГОСТ 7.80:2007 «ССІБВС. Бібліографічний запис. Заголовок. Загальні вимоги та правила складання»*.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (27), 2025
ISSN 2663 - 4023

Oleh Vasylenko
Postgraduate student of specialty 123 Computer Engineering
V.M. Glushkov Institute of Cybernetics of the
National Academy of Sciences (NAS) of Ukraine, Kyiv, Ukraine
Lecturer at the Department of Information Systems, Technologies, Finance and Management
Ukrainian Institute of Arts and Sciences, Bucha, Ukraine
ORCID ID: 0000-0001-8498-2950
international@ugi.edu.ua

ANALYSIS OF KEY METADATA FOR IDENTIFYING DUPLICATES IN BIBLIOGRAPHIC RECORDS

Abstract. This study addresses the issue of duplicate bibliographic records in library information systems, a problem that is becoming increasingly relevant with the growth of digital catalogs. It specifically examines the key metadata fields used for comparing records and identifying duplicate entries. The analysis includes critical metadata fields such as title, ISBN, publisher, place of publication, publication date, pagination, series, and additional attributes used for identifying editions. Special attention is given to the variability of data within these fields, including issues arising from misplaced subfields (e.g., place of publication instead of year, or vice versa) and the use of various date formats, such as copyright dates, ranges, or approximate dates. The study explores the specifics of multi-level records, particularly for journals and multi-volume publications, as well as errors caused by data migration between different automated library information systems (ALIS). The research demonstrates that, despite the existence of ISBD, UNIMARC, and other standards, a significant proportion of inconsistencies persist in bibliographic records, complicating automated processing. Fields containing publishers and places of publication exhibit a high degree of variability, with unique values accounting for only 9% to 38% of total records. During data clustering using the nearest neighbor method with the Prediction by Partial Matching (PPM) algorithm, the number of unique values was reduced by 6–24%, highlighting the potential of automation to improve the efficiency of manual record editing. The findings make a substantial contribution to the development of effective approaches for enhancing automated bibliographic data management systems, optimizing duplicate detection, and improving the overall quality of library databases.

Keywords: bibliographic record; bibliographic metadata; duplicate detection; automated library information systems; Prediction by Partial Matching; multi-level bibliographic description.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Baas, J., Schotten, M., Plume, A., & Côté, G. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative science studies*, *1(1)*, 377–386. https://doi.org/10.1162/qss_a_00019
2. Beesley, L., Bondarenko, I., Elliot, M., & Kurian, A. (2021). Multiple imputation with missing data indicators. *Stat Methods Med Res.*, *30(12)*, 2685–2700. https://doi.org/10.1177/09622802211047346
3. Burnham, J. F. (2006). Scopus database: a review. *Biomedical digital libraries*, *3(1)*, 1–8. https://doi.org/10.1186/1742-5581-3-1
4. Ceasar, S. A., & Ignacimuthu, S. (2023). CRISPR/Cas genome editing in plants: Dawn ofAgrobacterium transformation for recalcitrant and transgene-free plants for future cropbreeding. *Plant Physiology and Biochemistry*, *196*, 724–730. https://doi.org/10.1016/j.plaphy.2023.02.030
5. Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5(1)*, 31–49. https://doi.org/10.1162/qss_a_00286
6. Elango, B. (2024). Duplication issues with the new interface of Scopus. *INFONOMY*, *2*. https://doi.org/10.3145/infonomy.24.015
7. Elango, B., & Matilda, S. (2023). Mapping thecybersecurity research: A scientometric analysis of Indian publications. *Journal of ComputerInformation Systems*, *63(2)*, 293–309. https://doi.org/10.1080/08874417.2022.2058644

8. Elango, B., Kozak, M., & Rajendran, P. (2019). Analysis of retractions in Indian science. *Scientometrics, 119(2),* 1081–1094. https://doi.org/10.1007/s11192-019-03079-y
9. Hammer, B., Virgili, E., & Bilotta, F. (2023). Evidence-based literature review: De-duplication a cornerstone for quality. *World J Methodol*, *13(5),* 390–398. https://doi.org/10.5662/wjm.v13.i5.390
10. Krauskopf, E. (2018). An analysis of discontinued journals by Scopus. *Scientometrics*, *116(3),* 1805–1815. https://doi.org/10.1007/s11192-021-03948-5
11. Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: acomparative analysis. *Scientometrics*, *106*, 213–228. https://doi.org/10.1007/s11192-015-1765-5
12. Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographicinformation in today's academic world. *Publications*, *9(1).* https://doi.org/10.3390/publications9010012
13. Tennant, J. P. (2020). Web of Science and Scopus are not global databases of knowledge. *European Science Editing*, *46*. https://doi.org/10.3897/ese.2020.e51987
14. Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of informetrics*, *12(2),* 430–435. https://doi.org/10.1016/j.joi.2018.03.006
15. Thelwall, M., & Sud, P. (2022). Scopus 1900–2020: Growth in articles, abstracts, countries,fields, and journals. *Quantitative Science Studies*, *3(1),* 37–50. https://doi.org/10.1162/qss_a_00177
16. APN Ukrainy & Derzh. nauk.-ped. b-ka Ukrainy im. V. O. Sukhomlynskoho. (2010). *Uprovadzhennia v praktyku roboty bibliotek osvitianskoi haluzi DSTU HOST 7.1:2006 «Bibliohrafichnyi zapys. Bibliohrafichnyi opys. Zahalni vymohy ta pravyla skladannia» ta DSTU HOST 7.80:2007 «SSIBVS. Bibliohrafichnyi zapys. Zaholovok. Zahalni vymohy ta pravyla skladannia» [Implementation in the practice of libraries in the educational sector of DSTU GOST 7.1:2006 "Bibliographic record. Bibliographic description. General requirements and rules of compilation" and DSTU GOST 7.80:2007 "SSIBVS. Bibliographic record. Title. General requirements and rules of compilation"]*.