УДК 004.652
DOI https://doi.org/10.32782/2663-5941/2025.2.2/05

Василенко О. О.
Національна академія наук України

ОПТИМІЗАЦІЯ ОБРОБКИ БІБЛІОГРАФІЧНИХ ДАНИХ ЗА ДОПОМОГОЮ РЕГУЛЯРНИХ ВИРАЗІВ

Стаття присвячена дослідженню використання регулярних виразів (RegEx) у процесах обробки бібліографічних даних. У статті розкрито важливість забезпечення швидкості, точності та автоматизації процесів введення, редагування та структурування інформації. Розкрито механізм роботи регулярних виразів як ефективного інструменту для пошуку, заміни та видалення непотрібних елементів у текстових даних, що сприяє оптимізації процесів опрацювання бібліографічної інформації. З'ясовано, що використання регулярних виразів у бібліотечних інформаційних системах (БІС) дозволяє автоматизувати численні етапи обробки даних, включаючи очищення тексту від зайвих пробілів, символів і небажаних форматів, а також структурування змісту, що є критично важливим для формування та подальшого використання бібліографічних записів. Доведено, що цей підхід особливо актуальний для таких процесів, як обробка відсканованих текстів анотацій, змістів наукових видань, створення аналітичних записів і переробка списків статей. На прикладі бібліотечної системи Koha визначено, що застосування регулярних виразів у поєднанні з інструментами, такими як текстовий редактор Notepad++, значно спрощує та автоматизує процеси введення і редагування бібліографічних записів. Обґрунтовано, що використання регулярних виразів дозволяє не лише скоротити час, необхідний для обробки даних, але й підвищити їхню якість, мінімізуючи кількість помилок і забезпечуючи узгодженість форматування. Результати дослідження свідчать, що застосування регулярних виразів є особливо корисним для великих бібліотек, які працюють із значними обсягами даних, включаючи наукові видання, де необхідна висока точність і відповідність певним стандартам бібліографічного опису. З'ясовано, що автоматизація рутинних завдань завдяки регулярним виразам дозволяє бібліотекарям зосередитися на більш важливих аспектах роботи, таких як аналіз контенту та надання послуг користувачам.

Ключові слова: *регулярні вирази, оптимізація обробки даних, бібліографічні записи, Koha, Notepad++, обробка відсканованих даних.*

Постановка проблеми. В умовах стрімкого зростання обсягу цифрових бібліографічних записів бібліотеки стикаються з необхідністю впровадження нових технологій для підвищення ефективності обробки текстових даних. Регулярні вирази є одним із найпотужніших інструментів для виконання автоматизованих операцій пошуку, заміни й валідації даних, особливо в таких системах, як електронні каталоги бібліотек. Їх використання значно спрощує рутинні процеси, зокрема виправлення помилок форматування, очищення полів метаданих, стандартизацію назв і дат, а також аналіз масивів записів для виявлення шаблонів.

Зважаючи на обмежений бюджет і людські ресурси більшості бібліотек, впровадження регулярних виразів стає не лише бажаним, але й необхідним для оптимізації робочих процесів. Згідно з даними Міжнародної федерації бібліотечних асоціацій та установ (IFLA), кількість записів, які обробляються у великих бібліотеках, щороку зростає на 10–15 %, що робить питання автоматизації ключовим для забезпечення доступності та якості каталогів. Бібліотеки, які впроваджують технології регулярних виразів, знижують трудові витрати на 30–50 % при обробці тексту, що доводить їхню практичну значущість.

Аналіз останніх досліджень і публікацій. Українські та зарубіжні дослідники приділяли значну увагу можливостям застосування регулярних виразів у бібліотечній справі. У своїх дослідженнях акцентується увага на використанні регулярних виразів для стандартизації бібліографічних описів у системах на базі UNIMARC. Автор зазначає, що регулярні вирази дозволяють не лише автоматизувати виправлення помилок у форматуванні, але й виявляти дублікати записів [1]. Досліджується використання регулярних виразів для аналізу бібліографічних даних у процесах інтеграції каталогів, зокрема під час міграції даних між різними платформами автоматизованих бібліотечних систем (АБІС). Smyk також наголошує на можливостях регулярних виразів у вирішенні проблем, пов'язаних із різними форматами запи-

сів і мовними варіаціями [2]. Пропонується підхід до використання регулярних виразів для розпізнавання текстових шаблонів у бібліографічних записах. У своїх працях дослідники акцентують увагу на важливості правильного налаштування виразів для уникнення помилкових результатів під час очищення даних [3]. Інші джерела [4–6] обговорюють технічні аспекти роботи з регулярними виразами в бібліотечних інформаційних системах, їхню інтеграцію з мовами програмування (наприклад, Python або JavaScript), а також порівняння ефективності різних алгоритмів обробки тексту.

Разом із тим, аналіз наукових публікацій показує, що недостатньо вивченими залишаються аспекти практичного впровадження регулярних виразів у бібліотечні інформаційні системи, зокрема їх адаптація для обробки текстів різними мовами. Наявні дослідження здебільшого зосереджені на технічних аспектах використання регулярних виразів, проте менше уваги приділено оцінці їхньої ефективності у великих бібліотечних базах даних, що містять різноманітні типи бібліографічної інформації. Також залишається відкритим питання щодо оптимальних методик налаштування регулярних виразів для зменшення кількості помилково розпізнаних записів та підвищення точності автоматизованої обробки даних.

Постановка завдання. Метою статті є дослідження та обґрунтування ефективності використання регулярних виразів (RegEx) для оптимізації обробки бібліографічних даних. Для досягнення цієї мети передбачається вирішення таких завдань, а саме: проаналізувати існуючі підходи до автоматизації обробки бібліографічних записів у бібліотечних інформаційних системах (БІС); дослідити можливості використання регулярних виразів для очищення, стандартизації та структурування бібліографічних даних; оцінити ефективність регулярних виразів у вирішенні типових проблем, таких як виправлення форматування, виявлення дубльованих записів і уніфікація метаданих. Результати дослідження дозволять оцінити доцільність застосування регулярних виразів у бібліотечній практиці та розробити рекомендації щодо їх ефективного використання.

Виклад основного матеріалу. Регулярні вирази є важливим інструментом у сучасній бібліотечній практиці, який дозволяє значно покращити якість бібліографічних даних та знизити витрати на їх обробку. Огляд літератури свідчить про значний потенціал цієї технології для подальшого вдосконалення автоматизованих бібліотечних систем. У бібліотеці бібліографічні записи, як правило, створюються вручну каталогізаторами, що забезпечує детальне внесення інформації відповідно до стандартів. Кожен елемент бібліографічного опису – назва, автор, рік видання, місце видання, видавництво, ISBN тощо – розташовується в окремих полях або підполях електронного каталогу. Такий підхід дозволяє структурувати інформацію, забезпечуючи легкість доступу до потрібних даних під час пошуку або сортування.

Однак подібне структурування не виключає проблем, які виникають при обробці записів. У багатьох випадках з'являється необхідність виправлення помилок або внесення змін у текстову інформацію, наприклад, через використання різних форматів або помилкове введення даних. Зокрема, переплутані значення у полях (наприклад, рік замість місця видання), дублікати записів чи нестандартизовані скорочення створюють складнощі для автоматизованих бібліотечних систем.

Така потреба може стати особливо актуальною при об'єднанні даних із кількох каталогів, підготовці записів до міграції між системами або інтеграції їх у централізовані бази даних. У цих випадках регулярні вирази стають ефективним

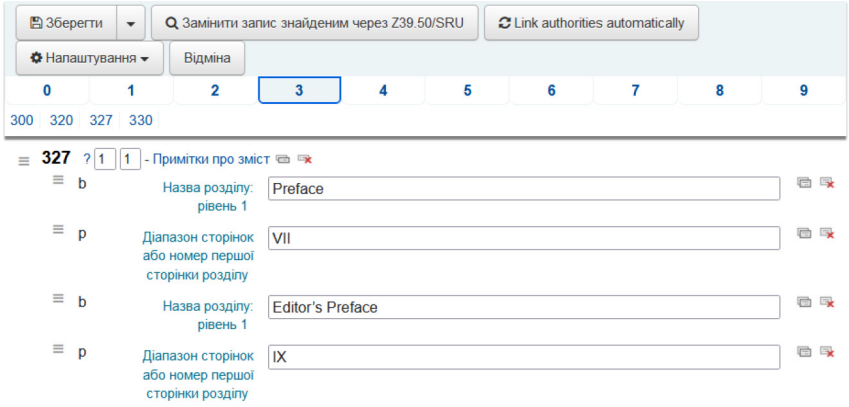

Рис. 1. Редагування бібліографічного запису

інструментом для автоматизації пошуку і заміни тексту, а також стандартизації записів у масштабі великих бібліотечних систем.

Однак потреба в інструментах пошуку та заміни стає критично важливою при роботі з відсканованими текстами, зокрема анотаціями чи змістом. Після розпізнавання такі тексти часто містять велику кількість зайвих символів і помилок форматування, які необхідно очистити перед подальшою обробкою або включенням у бібліографічні записи.

Очищення подібних текстів є більш ефективним, якщо виконувати його не у текстовому процесорі Word, адже цей інструмент зазвичай зберігає зайве форматування, яке ускладнює стандартизацію даних. Більше того, в Word стандартно відсутня повноцінна підтримка пошуку за регулярними виразами, що значно обмежує можливості автоматизації.

Натомість після розпізнавання текстів доцільно використовувати текстовий редактор, наприклад, Notepad++ [4], який підтримує регулярні вирази. У цьому редакторі можна виконати комплексне очищення тексту, включаючи:

– видалення спеціальних символів, що залишаються після сканування;

– заміну табуляцій на пробіли для забезпечення рівномірного форматування;

– видалення зайвих пробілів, що можуть створювати візуальний або структурний безлад.

Використання Notepad++ дозволяє автоматизувати ці завдання, що значно зменшує час, витрачений на підготовку текстів, та підвищує їх якість для подальшої обробки у бібліотечних системах.

Поява розширеного редактора бібліографічних записів в автоматизованій бібліотечній інформаційній системі Koha [5] значно розширила можливості обробки та редагування бібліографічних даних. Цей інструмент (див. рис. 2) забезпечує користувачів зручним інтерфейсом для роботи з метаданими, що дозволяє спрощувати створення структурованого змісту записів.

Розширений редактор надає низку переваг:

– можливість легкої навігації між полями та підполями запису;

– автоматизацію введення стандартних елементів опису;

– підтримку валідаторів для перевірки правильності введених даних відповідно до бібліографічних стандартів, таких як MARC;

– інтеграцію з інструментами для обробки текстів, що дозволяє уникати дублювання та помилок у бібліографічних записах.

Завдяки цим функціям система Koha не лише оптимізує процес створення нових записів, але й спрощує редагування наявних, забезпечуючи кращу організацію бібліографічної інформації. Це особливо важливо у великих бібліотеках, де ефективність роботи залежить від швидкого доступу до якісно оформлених даних.

Для якісного структурування бібліографічних записів, кожна сторінка та кожен заголовок будь-якого рівня мають бути записані в окреме підполе. Це особливо актуально для змісту книг і наукових видань, де обсяги даних надто великі для ручного введення. Навіть після сканування та розпізнавання тексту, без автоматизації процесу структурування працювати із змістом могли б лише ентузіасти, оскільки це вимагало б значних витрат часу та зусиль.

Notepad++ пропонує інструменти для автоматизації рутинних завдань, таких як структурування даних, за допомогою функціоналу макросів. Макроси дозволяють записувати послідовності команд для виконання повторюваних операцій. У поєднанні з використанням регулярних виразів

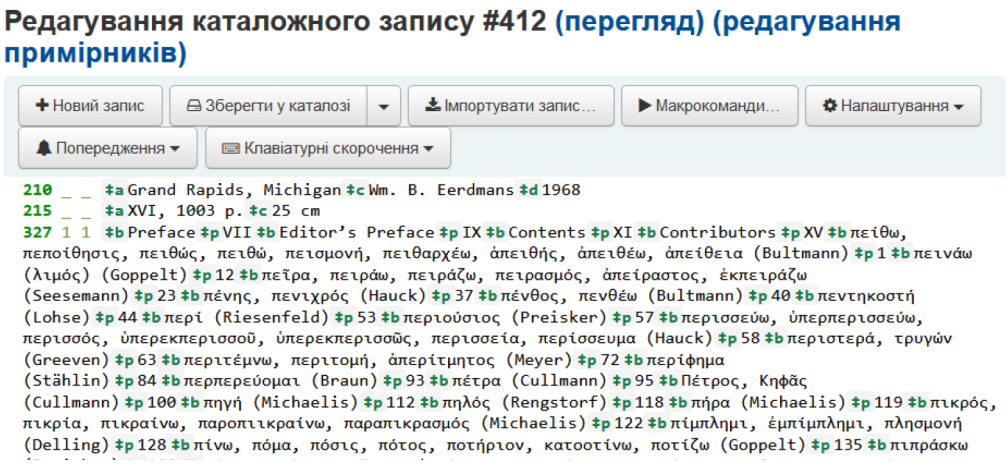

Рис. 2. Розширений редактор

це відкриває можливість ефективної обробки тексту навіть для користувачів, які не є технічними фахівцями.

Наприклад, регулярні вирази можуть бути застосовані для:

– автоматичного розділення тексту на підполя залежно від заголовків або маркерів;

– видалення зайвих символів чи форматування;

– упорядкування та форматування змісту для подальшого використання в бібліотечній системі.

Завдяки таким інструментам забезпечується висока швидкість і точність обробки текстів, що значно скорочує час, необхідний для підготовки структурованих бібліографічних записів, і підвищує їхню якість.

Після очищення тексту зміст книги потребує додаткової підготовки для правильного структурування та подальшої обробки. Основні вимоги до підготовленого тексту включають:

1. Окремий рядок для кожного елемента змісту, де в кінці рядка має бути зазначений номер сторінки.

2. Заголовки 1 рівня розташовуються без додаткових налаштувань, тобто без відступів.

3. Заголовки 2 рівня повинні мати одну табуляцію на початку рядка.

4. Заголовки 3 рівня потребують двох табуляцій на початку рядка.

Після ручного налаштування або за допомогою автоматизації цих правил можна запускати підготовлений макрос. Макрос виконує послідовність команд із пошуку та заміни, що автоматизують подальшу обробку тексту. Основні етапи роботи макросу включають:

– Вирівнювання форматування за визначеними правилами (наприклад, видалення зайвих пробілів чи заміна дефісів на тире).

– Автоматичне розставлення табуляцій для відповідних рівнів заголовків.

– Упорядкування номерів сторінок та перевірка їх правильності.

Цей підхід значно зменшує ручну роботу, підвищує точність даних і забезпечує їх сумісність із системами автоматизованої обробки, такими як Koha або інші АБІС.

Використання регулярних виразів дозволяє реалізувати складні трансформації тексту, які включають збереження та використання знайдених даних у процесі заміни. Зокрема, перший регулярний вираз виконує такі дії:

1. Знаходить в кінці рядка послідовність арабських або римських чисел, що представляють номери сторінок.

2. Замінює пробіли перед номером сторінки на мітку підполя, яка використовується для позначення сторінок у бібліографічному записі.

3. Вставляє знайдені номери сторінок у відповідне місце структури запису.

Інші пошукові запити, які виконуються, належать до типу Розширений пошук і включають:

– Використання розширених спеціальних символів для виявлення попереднього нового рядка.

– Аналіз кількості табуляцій на початку рядка, що відповідають рівню заголовку (одна табуляція для заголовків другого рівня, дві – для третього і так далі).

– Видалення зайвих пробілів або інших небажаних символів, які можуть порушувати структуру.

– Автоматичну заміну цих елементів мітками відповідних підполів, таких як мітка для заголовків чи підзаголовків.

Ця технологія дозволяє ефективно структурувати зміст, значно скорочуючи час на ручну обробку. Крім того, автоматизовані дії забезпечують високу точність і відповідність стандартам опису, що важливо для інтеграції даних у бібліотечні системи або інші бази даних.

Навіщо	Режим пошуку	Шукати	Замінити
Сторінки в кінці рядку	Регулярні вирази	\s+([\divxcdlVXCD]+)\s*$	‡p\1
Заголовок 5 рівня	Розширений	\r\n\t\t\t\t	‡f
Заголовок 4 рівня	Розширений	\r\n\t\t\t	‡e
Заголовок 3 рівня	Розширений	\r\n\t\t	‡d
Заголовок 2 рівня	Розширений	\r\n\t	‡c
Заголовок 1 рівня	Регулярні вирази	^	327 _ 1 ‡b

Рис. 3. Зміст в UNIMARC 327.
Цей макрос є послідовністю пошуку та замін

Навіщо	Режим пошуку	Шукати	Замінити
Поле і індикатори	Регулярні вирази	327\s[_012]\s[_01]\s	
Заголовок 1 рівня	Розширений	‡b	\r\n
Заголовок 2 рівня	Розширений	‡c	\r\n\t
Заголовок 3 рівня	Розширений	‡d	\r\n\t\t
Заголовок 4 рівня	Розширений	‡e	\r\n\t\t\t
Заголовок 5 рівня	Розширений	‡f	\r\n\t\t\t\t
Сторінки в кінці рядку	Звичайний	‡p	(пробіл)

Рис. 4. UNIMARC 327 в зміст.
Цей макрос є послідовністю пошуку та замін

Останній етап обробки тексту за допомогою регулярних виразів забезпечує додавання правильно оформлених міток поля до кожного заголовка першого рівня. Це особливо актуально для рядків, які починаються з нового рядка. Використання регулярних виразів дозволяє створити більш елегантний і гнучкий спосіб обробки, ніж простий пошук спеціальних символів, які вказують на новий рядок. Регулярні вирази також дозволяють додавати текст на початку першого рядка, що є важливою перевагою.

Аналогічно, регулярні вирази застосовуються для обробки відсканованих списків статей з метою формування посилань на аналітичні записи [6]. Перед застосуванням регулярних виразів список потребує попередньої підготовки. Однак, цей процес є дещо складнішим через специфіку списків статей. Зокрема, потребується:

1. Зміна порядку розташування прізвища та власного імені для приведення записів до стандартного формату.

2. Оформлення підполів для авторів додавання прізвища та ініціалів автора як заголовка.

3. Внесення перших відомостей про відповідальність, наприклад, зазначення співавторів, редакторів або перекладачів у відповідних підполях.

Ці дії спрямовані на стандартизацію записів, яка забезпечує відповідність аналітичних записів встановленим бібліографічним нормам. Регулярні вирази в цьому контексті стають незамінним інструментом, адже дозволяють швидко і автоматично виконувати масові трансформації тексту, які потребували б значних витрат часу в разі ручної обробки. Таким чином, застосування регулярних виразів є ключовим кроком для автоматизації та покращення обробки даних у бібліотечних системах, забезпечуючи ефективність і точність навіть у найскладніших сценаріях.

Практичний кейс: *Оптимізація обробки бібліографічних даних за допомогою регулярних виразів*

У бібліотечних інформаційних системах (БІС) важливим аспектом є обробка та структуризація бібліографічних записів. Багато з них формуються вручну, що може призводити до помилок і зайвої витрати часу. Для прискорення процесу та зменшення помилок важливо використовувати інструменти автоматизації. Один із таких інструментів – регулярні вирази (RegEx). Вони дозволяють швидко здійснювати пошук, заміну та видалення елементів даних, оптимізуючи таким чином процес обробки великих масивів інформації.

У бібліографічних записах часто виникають такі проблеми, як:

– Невідповідність форматів запису (наприклад, різні формати дати).

– Дублікатні записи, створені через помилки при введенні або скануванні.

Навіщо	Режим пошуку	Шукати	Замінити
Сторінки в кінці рядку	Регулярні вирази	\s([cCpP])?.?\s*([\divxcdlVXCD]+)\s*$	‡p\u\1. \2
Назва	Розширений	\r\n\t	‡t
Автор	Регулярні вирази	^	464 _ 1 ‡a

Рис. 5. Список статей в UNIMARC 464.
Цей макрос є послідовністю пошуку та замін

– Наявність непотрібних символів, пробілів або табуляцій, що ускладнюють подальшу обробку даних.

Крім того, багато бібліографічних записів містять складні структури, наприклад, зміст, який складається з багатьох рівнів заголовків. Для ручної обробки таких записів потрібен значний час і зусилля.

Метою кейсу є продемонструвати, як можна за допомогою регулярних виразів оптимізувати обробку бібліографічних даних, автоматизуючи такі процеси, як:

– Очищення даних від непотрібних символів.

– Переведення записів у єдиний формат.

– Створення аналітичних записів на основі відсканованого тексту.

Початкові бібліографічні записи часто містять зайві символи, пробіли або неправильне форматування після сканування або ручного введення. Для прикладу, припустимо, що є відсканований зміст книги, який містить багаторівневі заголовки та численні пробіли, розміщені не відповідно до структури.

1. Видалення зайвих пробілів та символів. Регулярні вирази дозволяють швидко знайти непотрібні символи, які залишаються після сканування або неправильного форматування. Наприклад:

– Видалення зайвих пробілів між словами.

– Замінити декілька пробілів на один.

– Видалення спецсимволів, які не є частинами бібліографічного опису.

Приклад регулярного виразу:

```
arduino
Копировать код
\s+ → « »
```

Цей вираз знаходить всі зайві пробіли (окрім одного) і замінює їх на єдиний пробіл.

2. Оформлення заголовків. Якщо необхідно автоматично перетворити зміст на структуровану форму, можна використовувати регулярні вирази для додавання табуляцій або інших символів для визначення рівнів заголовків.

Приклад регулярного виразу для заміни заголовків:

```
scss
Копировать код
^([0-9]+)\. (.+) → \t$2
```

Цей вираз дозволяє замінити заголовки другого рівня, додаючи табуляцію на початку кожного заголовка.

Автоматичне створення аналітичних записів Регулярні вирази можуть також бути використані для автоматизації створення аналітичних записів. Наприклад, при обробці сканованого списку статей можна за допомогою регулярних виразів автоматично змінювати порядок прізвища та імені авторів, або додавати необхідні дані для створення підполів «Автор» та «Відповідальність».

Приклад регулярного виразу для зміни порядку імені та прізвища:

```
bash
Копировать код
([A-Za-z]+), ([A-Za-z]+) → $2, $1
```

Цей вираз змінює місцями прізвище та ім’я автора.

3. Переваги використання регулярних виразів

1) *Швидкість обробки*. Регулярні вирази дозволяють значно зменшити час обробки великих обсягів даних, оскільки операції заміни та пошуку виконуються дуже швидко.

2) *Автоматизація*. Використання регулярних виразів дозволяє автоматизувати процеси, які раніше виконувались вручну, що мінімізує людську помилку.

3) *Гнучкість*. Регулярні вирази можуть бути налаштовані для виконання різних типів операцій на текстових даних, що дозволяє застосовувати їх до різних типів записів та форматів.

4) *Зручність*. Застосування регулярних виразів у таких текстових редакторах, як Notepad++

Навіщо	Режим пошуку	Шукати	Замінити
Сторінки в кінці рядку	Звичайний	‡p	(пробіл)
Назва	Розширений	‡t	\r\n\t
Автор	Звичайний	‡a	
Поле і індикатори	Регулярні вирази	464 [_01] [_01]	

Рис. 6. UNIMARC 464 в список статей.
Цей макрос є послідовністю пошуку та замін

або в БІС Koha, дозволяє швидко адаптувати інструмент до специфічних потреб бібліотеки.

Висновки. Використання регулярних виразів для обробки бібліографічних записів є важливим кроком до автоматизації процесів у бібліотечних системах. Аналіз показав, що, хоча введення бібліографічного опису вручну в окремі підполя менш залежить від функцій пошуку та заміни, наступна обробка записів, зокрема для підготовки структурованого змісту або створення аналітичних записів, значно виграє від застосування регулярних виразів. Розширений редактор, доступний в автоматизованій бібліотечній інформаційній системі (АБІС) Koha, у поєднанні з макросами текстового редактора Notepad++, відкриває широкі можливості для ефективної автоматизації таких завдань, як: вставка структурованого змісту в окремі підполя; створення аналітичних записів після обробки сканованого тексту; очищення тексту від зайвих символів, видалення помилок форматування, а також впорядкування заголовків різних рівнів.

Автоматизація створення аналітичних записів вимагає ширшого використання регулярних виразів для швидкої обробки масивів даних. Ці інструменти дозволяють: здійснювати складні пошуково-замінні операції, що включають трансформацію даних у структурований формат. Підготовлювати списки статей, змінюючи порядок прізвищ та ініціалів авторів, розподіляючи їх по відповідних підполях та додаючи інші елементи бібліографічного опису. Обробляти зміст багатотомних видань і наукових книг, які традиційно є трудомісткими для каталогізаторів.

Регулярні вирази також дозволяють бібліотечним фахівцям з мінімальними технічними знаннями виконувати завдання, які раніше вимагали б участі програмістів. Завдяки можливості автоматизації рутинних процесів вони сприяють економії часу, покращенню якості каталогів і зниженню ймовірності помилок, спричинених людським фактором. Узагальнюючи, можна стверджувати, що використання регулярних виразів у бібліотечній практиці є перспективним напрямом для вдосконалення автоматизованих бібліотечних інформаційних систем, що підвищує ефективність обробки бібліографічних даних і забезпечує високу якість управління інформацією.

Перспективи подальших досліджень у цьому напрямі можуть включати удосконалення алгоритмів застосування регулярних виразів для складніших завдань обробки бібліографічних записів, таких як семантичний аналіз даних та визначення зв'язків між записами. Інтеграцію регулярних виразів із технологіями машинного навчання для автоматичного коригування бібліографічних описів. Розробку спеціалізованих модулів для автоматизованих бібліотечних інформаційних систем, які дозволять користувачам з мінімальними знаннями програмування ефективно застосовувати регулярні вирази. Порівняльний аналіз ефективності регулярних виразів у поєднанні з іншими методами обробки текстових даних, зокрема нейронними мережами та алгоритмами природної мови. Подальше дослідження цих аспектів сприятиме ще більшій автоматизації бібліотечних процесів і підвищенню якості управління бібліографічною інформацією.

Список літератури:

1. Bazoge A. Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review. *JMIR Med Inform*. 2023. No. 11. URL: https://doi.org/10.2196/42477

2. CICERO: A Domain-Specific Architecture for Efficient Regular Expression Matching / D. Parravicini et al. *ACM Trans. Embed. Comput. Syst*. 2021. P. 1–24. URL: https://surl.lu/asusnz (date of access: 21.11.2024).

3. Comparison of bibliographic data sources: Implications for the robustness of university rankings / C.-K. Huang et al. *Quantitative Science Studies*. 2020. No. 1(2). P. 445–478. URL: https://doi.org/10.1162/qss_a_00031

4. Grapefruit: An Open-Source, Full-Stack, and Customizable Automata Processing on FPGAs / R. Rahimi et al. *2020 IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. New York, NY, 2020. P. 138–147. URL: https://doi.org/10.1109/FCCM48280.2020.00027

5. Heng G., Han M.-J. Revamping Metadata Maker for 'Linked Data Editor': Thinking Out Loud. *The Code4Lib Journal*. 2023. No. 55. URL: https://journal.code4lib.org/articles/16925 (date of access: 02.12.2024).

6. Muddana A., Vinayakam S. Regular Expressions. *Python for Data Science*. 2024. URL: https://doi.org/10.1007/978-3-031-52473-8_11

7. Nagaraju S., Shanmugham B., Baskaran K. High throughput token driven FSM based regex pattern matching for network intrusion detection system. *Materials Today: Proceedings*. 2021. Vol. 47, no. 1. P. 139–143. URL: https://doi.org/10.1016/j.matpr.2021.04.028

8. New trends in bibliometric APIs: A comparative analysis / A. Velez-Estevez et al. *Information Processing & Management*. 2023. Vol. 60, no. 4. URL: https://doi.org/10.1016/j.ipm.2023.103385

9. Optimizing Regular Expressions via Rewrite-Guided Synthesis / J. McClurg et al. *PACT '22: Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*. 2022. P. 426–438. URL: https://doi.org/10.1145/3559009.3569664

10. Regex Matching with CountingSet Automata / L. Turoňová et al. *Proceedings of the ACM on Programming Languages*. 2020. Vol. 4. 218. URL: https://doi.org/10.1145/3428286

11. Software-Hardware Codesign for Efficient In-Memory Regular Pattern Matching / L. Kong et al. *PLDI*. 2022. URL: https://arxiv.org/pdf/2209.05686

Vasylenko O. O. **OPTIMIZATION OF BIBLIOGRAPHIC DATA PROCESSING USING REGULAR EXPRESSIONS**

The article is dedicated to the study of the use of regular expressions (RegEx) in the processing of bibliographic data. The article reveals the importance of ensuring speed, accuracy, and automation in the processes of data entry, editing, and structuring. The mechanism of regular expressions as an effective tool for searching, replacing, and removing unnecessary elements in textual data is explored, contributing to the optimization of bibliographic information processing. It has been established that the use of regular expressions in library information systems (LIS) enables the automation of numerous stages of data processing, including text cleaning from unnecessary spaces, symbols, and unwanted formatting, as well as content structuring, which is critically important for the formation and further use of bibliographic records. It has been proven that this approach is particularly relevant for processes such as the processing of scanned annotation texts, tables of contents of scientific publications, the creation of analytical records, and the reformatting of article lists. Based on the example of the Koha library system, it has been determined that the application of regular expressions in combination with tools such as the Notepad++ text editor significantly simplifies and automates the processes of entering and editing bibliographic records. It has been substantiated that the use of regular expressions not only reduces the time required for data processing but also improves its quality by minimizing errors and ensuring consistency in formatting. The research results indicate that the application of regular expressions is particularly useful for large libraries that handle significant volumes of data, including scientific publications, where high accuracy and compliance with specific bibliographic description standards are required. It has been established that the automation of routine tasks through regular expressions allows librarians to focus on more important aspects of their work, such as content analysis and providing services to users.

Key words: *regular expressions, data processing optimization, bibliographic records, Koha, Notepad++, scanned data processing.*