УДК 004.052.42:01

М. В. Петренко
Відкритий міжнародний університет розвитку людини «Україна»
вул. Львівська, 23, 03115 Київ, Україна

Дослідження візуально непомітних помилок введення та їхнього впливу на якість і пошукову доступність бібліографічних даних

Статтю присвячено дослідженню особливого класу помилок при введенні бібліографічних даних до автоматизованої бібліотечної інформаційної системи, який є непомітним для користувача, але впливає на функціонування системи електронних каталогів. Причиною проблеми є помилкове введення візуально подібних символів латиниці замість символів кирилиці та навпаки. Дослідження відбувалося на основі бібліографічних даних зі 141 публічної бібліотеки м. Києва за період з 1993 року до початку 2021 року (отримано з двох джерел). Описано особливості помилок, метод автоматичної ідентифікації помилок, розповсюдженість помилок, вплив на пошукову доступність і пошук дублів, розподіл помилок за символами. Надано рекомендації щодо профілактики та усунення проблеми.

Ключові слова: *помилки введення, виправлення помилок, вплив помилок, кирилиця, латиниця, бібліографічні дані, дедублікація, пошукова доступність, автоматизована бібліотечна інформаційна система.*

Вступ

Частина бібліотек України мають свої автоматизовані бібліотечні інформаційні системи (АБІС) з модулями електронного каталогу, які надають користувачам інформацію про асортимент бібліотек через мережу Інтернет. Перегляд користувачем інформації про наявність книжок є можливим виключно через інтерфейс пошуку. У випадку, якщо бібліографічні дані про книгу введені з помилками, користувачі повністю чи частково втрачають можливість її знайти. Бібліотеки та користувачі під час користування системою помічають помилки (у випадку, якщо вдалося знайти книгу), і тоді з'являється можливість їх виправити.

Існує відомий тип помилок, які неможливо виявити візуально. Вони є наслідком того, що деякі символи латиниці виглядають абсолютно ідентичними символам кирилиці, але при цьому мають різні коди, й усіма програмними засобами

вважаються різними. Вводячи поміж символами однієї знакової системи символи іншої, оператор не помічає різниці.

Не зважаючи на те, що такі помилки є відомими, АБІС, які використовуються в Україні, допускають введення даних з помилками та не мають ефективних засобів пошуку та виправлення помилок у бібліографічних даних [1].

Відповідно, такі помилки накопичуються. А через модулі електронних каталогів стає складно, або і неможливо знайти сотні записів книг лише через цей клас помилок. Виявити такі помилки лише силами операторів, без спеціального ПЗ, неможливо.

У табл. 1 можна побачити наслідок наявності таких помилок на прикладі значень поля «Видавництво». Значення поля пройшли об'єднання дублів, але є дві групи візуально тотожних записів (виділено напівжирним накресленням).

Таблиця 1. Візуально однакові записи

Значення	Кількість дублів
КЛУБ СЕМЕЙНОГО ДОСУГА	30
КЛУБ СЕМЕЙНОГО ДОСУГА	2
КЛУБ СЕМЕЙНОГО ДОСУГА	3570
КЛУБ СІМЕЙНОГО ДОЗВІЛЛЯ	1
КЛУБ СЕМЕЙГОГО ДОСУГА	0
…	…
КНИЖКОВИЙ КЛУБ СІМЕЙНОГО ДОЗВІЛЛЯ	156
КНИЖНИЙ КЛУБ СЕМЕЙНОГО ДОСУГА	0
КНИЖНЫЙ КЛУБ СЕМЕЙНОГО ДОСУГА	0
КНИЖНЫЙ КЛУБ СЕМЕЙНОГО ДОСУГА	949

Окрім уже звичних помилок друку маємо дві групи з трьох значень, що є повністю ідентичними, але не були розпізнані як дублі.

Дослідивши наведені у табл. 1 значення, було виявлено, що частина літер низькочастотних значень (кількість дублів від 0 до 30) має в своєму складі символи латинського алфавіту, які виглядають тотожно кириличним, але мають інші коди символів (помилкові символи виділені та узяті в лапки):

— КЛУБ «**С**»ЕМЕЙНОГО ДОСУГА

— КЛУБ СЕМЕЙНОГ«**О**» ДОСУГА

— КНИЖНЫЙ КЛУБ «**С**»ЕМЕЙНОГО ДОСУГА

Операції пошуку, порівняння та сортування розрізняють такі символи, а їхнє помилкове введення ускладнює окрім іншого знаходження дублів у межах однієї системи та створення зведених електронних каталогів на основі даних з кількох систем. З огляду на це, важливо дослідити такі помилки та їхній вплив у контексті бібліографічних даних.

Огляд літератури

Заміна символів одного набору на візуально ідентичні символи іншого набору є питанням, яке добре відоме та застосовується в таких випадках:

— візуальне розпізнавання друкованих текстів (OCR) [2, 3];

— боротьба з плагіатом [4];

— кібербезпека (стеганографія, боротьба з фішингом) [5, 6].

Різні застосування мають свою специфіку та ряд синонімічних назв для проблеми.

Для питань кібербезпеки та боротьби з плагіатом характерна робота переважно з випадками штучної експлуатації відомих описаних особливостей. Боротьба з плагіатом також має задачу нечіткого пошуку подібності, вільного від впливу, в тому числі й описаної проблеми.

Проблема розпізнавання текстів може мати подібні та застосовні алгоритми виправлення помилок розпізнавання. Але природа помилок різна — в одному випадку вони введені з клавіатури оператором, а в іншому — згенеровані автоматично з використанням алгоритмів розпізнавання образів. У випадку розпізнавання образів важливими є шрифти та розширені набори символів, яких є дуже багато. Особливості шрифтів сприяють появі додаткової множини подібних символів, яка відсутня в часто використовуваних шрифтах, у той час, як у базах даних дані зберігаються у вигляді тексту, а відображаються одним зі стандартних шрифтів.

На відміну від описаних споріднених проблем, проблема якості введених користувачем бібліографічних даних в АБІС і їхньої пошукової доступності має наступні особливості:

1) природний характер:

— джерело — людина, а не алгоритми;

— відсутність умислу;

2) введення з клавіатури (наявні лише символи з розкладки клавіатури);

3) наявність помилок важлива навіть, якщо відсутні дублі, з якими можна порівнювати (на відміну від задачі пошуку плагіату);

4) дані вже пройшли валідацію користувачами та АБІС;

5) пошук відбувається лише вбудованими в АБІС засобами (інтелектуального пошуку, як в системах антиплагіату, немає).

Специфіка термінології:

— у фішингу візуально подібні назви доменів із використанням символів різних наборів називаються «IDN homograph» [5];

— у стеганографії це називається використанням «Multilingual characters» [6];

— у розпізнаванні образів щодо описаної проблеми вживаються формулювання: «Mixed-Alphabet recognition» [3], «Script Identification» [2];

— у розпізнаванні плагіату використання візуально подібних символів має назву «Fake characters» [4];

— спільними для усіх випадків є ключові слова «cyrillic», «latin», «character».

Проведене дослідження

Джерела даних

З метою проведення якісного аналізу, було вирішено розглянути проблему на даних отриманих за допомогою пошукового робота з електронних каталогів:

— бібліотеки ім. Т.Г. Шевченка для дітей (публічні бібліотеки м. Києва для дітей, далі База 1) [7];

— бібліотеки ім. Лесі Українки (публічні бібліотеки м. Києва для дорослих, далі База 2) [8].

Скановані дані були отримані як у вигляді звичайних бібліографічних записів, так і у машиночитному форматі MARC. Після цього дані були розділені, стандартизовані власним програмним забезпеченням, щоб їх можна було розглядати разом. Для подальшої обробки дані були розмішені у спеціально створеній базі даних. У дослідженні використовувалися лише записи книг (не періодики чи статей, які містить періодика).

База 1 містить 102107 записів книг, а База 2 — 276635. Разом вони містять інформацію по книжкам 141-ї публічної бібліотеки м. Києва з 142. Найстарші записи датовані 1993 роком, а останні, що використані в дослідженні, — початком 2021 року. Це дозволяє докладно дослідити особливості цього класу помилок і наслідки їхньої наявності.

З кириличних символів у бібліографічних базах найбільше представлені ті, шо використовуються українською, російською та білоруською мовами. Українська підмножина кириличних символів включає в себе усі символи російського та білоруського набору символів, які можуть виглядати тотожно символам латиниці, а також має літеру «І», яка в російському наборі відсутня. Тому за основу взято саме українську підмножину символів: АаВСсЕеНІіМТКХхОоРр, що виглядає тотожно латинським: AaBCcEeHIiMTKXxOoPp. Окремо розглянемо римські цифри, які теж можуть бути написані з використанням символів з різних наборів.

Метод ідентифікації помилок

Найближча зі споріднених до теми статті задач — виправлення помилок візуального розпізнавання символів (OCR). Але для OCR це питання не є окремим, а вирішується в комплексі з пошуком інших помилок із використанням статистичних даних за мовами та методів штучного інтелекту. Наприклад, у роботі [3] пропонується використання адаптованого методу знаходження відстані Левенштейна та частот біграм.

Аналіз, що проводиться в даній роботі, є значно більш вузьким завданням, ніж ті, які вирішуються в OCR. Тому було прийнято рішення шукати помилки спираючись на специфіку, яка випливає з причин їхньої появи, а саме — з властивостей візуальної подібності чи неподібності. Такий підхід не потребує перебору комбінацій біграм і підрахунку відстаней редагування. Перевіряється лише належність чи неналежність символів до 5 множин, а над результатами виконуються логічні операції. Це не потребує додаткових статистичних даних, навчання нейронних мереж та є більш обчислювально простим завданням.

Для ідентифікації помилкового використання символів, значення полів розбиваються на слова (за розділовими знаками та зміною регістра символів). Після цього до кожного слова застосовуються наступні правила.

1. *R* є правдою, якщо у слові наявні виключно символи, візуально подібні до валідних римських чисел, частина з яких — кирилиця (крім чисел «Х» та «І», які не можна однозначно відрізнити від слів кирилиці).

2. *C* є правдою, якщо у слові наявні символи кирилиці, які не схожі на символи латиниці.

3. C_L є правдою, якщо у слові наявні символи кирилиці, які схожі на символи латиниці.

4. L є правдою, якщо у слові наявні символи латиниці, які не схожі на символи кирилиці — позначається.

5. L_C є правдою, якщо у слові наявні символи латиниці, які схожі на символи кирилиці.

6. Висновки по слову:

1) слово є кириличним з латинськими символами, якщо: $\neg(R \vee L) \wedge C \wedge L_C$;

2) слово є кириличними символами в латиниці, якщо: $\neg(R \vee C) \wedge L \wedge C_L$;

3) слово не потребує корекції, якщо воно є:

— словом на кирилиці: $\neg(R \vee L_C \vee L) \wedge (C \vee C_L)$;

— словом на латиниці: $\neg(R \vee C_L \vee C) \wedge (L \vee L_C)$.

Описані вище правила були використані для ідентифікації помилок у програмі аналізу даних, написаній мовою Delphi у середовищі RadStudio 10.4

Слабким місцем методу є існування невизначеностей, коли у слові є як символи, які точно не є символами кирилиці, так і символи, що точно не є символами латиниці. В таких випадках існує вірогідність помилок, але неможливо точно їх ідентифікувати автоматично. При аналізі реальних даних таких випадків були одиниці і вони всі швидко пройшли ручну перевірку та були визнані вірними.

Оскільки метод розпізнає як помилки всі слова, які містять одночасно і символи кирилиці, і символи латиниці, то пропуск помилок неможливий. Можлива лише невірна їхня інтерпретація. Оскільки алгоритм інтерпретації завжди однаковий, то на успішність порівняння при пошуку дублів він не впливає.

З метою оцінки кількості хибних спрацювань пропонованого методу, було переглянуто вручну слова, ідентифіковані як помилки. Серед них було кілька сумнівних слів, які навіть людина не може однозначно ідентифікувати, де які символи мали би бути. Наприклад:

— «**Та**kki, H. (Настя)»

— «**КО**M KON».

Також було виявлене з імовірністю майже 100 % хибне розпізнавання у слові «PRO**МОВА**».

Отже, пропонований метод є достатньо точним для дослідження помилок. Невелика кількість хибних (0,1 % від загальної кількості знайдених помилок) і ймовірно хибних спрацювань (< 1 % від загальної кількості знайдених помилок) не може суттєво вплинути на результати навіть без етапу ручної перевірки, яка мала місце в цьому дослідженні. Водночас метод є простим для реалізації і для обчислень.

Аналіз кількості записів з помилками

Серед текстових полів, що описують об'єктивні характеристики книги (на які не впливає бібліограф) є поля: «Назва книги», «Автор», «Видавництво». Такі текстові поля як «Анотація», «Зміст», «Ключові слова» і т.д. наявні не завжди та не завжди вводяться безпосередньо з книги. Тому у статті аналіз відбувається саме за полями «Назва книги», «Автор», «Видавництво».

Із 704 записів, що містять помилки, у 88 ~~записах~~ знайдено більше однієї помилки (12,5 %).

Написання римських чисел з використанням символів кирилиці чи кирилиці та латиниці разом є найпоширенішою помилкою. Помилок з римськими числами трапляється в 2 рази більше, ніж зі звичайними словами. Особливо багато проблем з римськими числами у полі «Назва книги», а от у полі «Автор» вони не трапляються через специфіку поля.

Таблиця 2. Кількість записів з помилками за полями, базами та типом помилки

	Помилки у Базі 1				Помилки у Базі 2				Усього Cyr+Lat
	Cyr римські числа	Cyr	Lat	Cyr+Lat	Cyr римські числа	Cyr	Lat	Cyr+Lat	
«Назва книги»	216	10	61	**71** (0,070 %)	1336	62	158	**220** (0,080%)	**291** (0,077 %)
«Автор»		1	29	**30** (0,029 %)		35	115	**150** (0,054 %)	**180** (0,048 %)
«Видавництво»	73	7	5	**12** (0,012 %)	168	81	140	**221** (0,080 %)	**233** (0,062 %)
Усього по базам	289	**18**	**95**	**113**	1504	**178**	**413**	**590**	**704**

У табл. 2 видно, що по усім полям відсоток помилок вище у Базі 2. Це корелює з тим, що База 2 має більший відсоток унікальних значень цих полів, порівняно з Базою 1 (відносна різниця унікальності від 5 до 40 %). Це логічно, оскільки частину унікальних записів унікальними робить саме наявність помилок.

Вплив на пошук

У задачах пошуку бібліографічних записів найчастіше використовуються поля «Назва книги» та «Автор». Пошук за полем «Видавництво» відноситься до рідкісних специфічних випадків. За наявності помилок у цих полях, знайти книгу стає значно важче чи навіть неможливо.

Пошук відбувається словами. Часто значення характеристики книги складається з кількох слів, і пошук може відбуватись або за всіма словами, або за обраними користувачем. Поля можуть містити слова різної довжини. Коротші частини мови зазвичай є високочастотними, а довші — більш низькочастотними, за якими легше знайти саме те, що потрібно.

Для оцінки проблем зі знаходженням записів книг, що містять помилки, введемо дві метрики:

— відношення слів з помилками до загальної кількості слів поля;

— відношення суми довжин слів з помилками до суми довжин усіх слів поля.

Помилки у написанні римських чисел поширені більше, ніж всі інші, але використання в пошуку римських чисел є малоймовірним. Тому при розрахунку наступних характеристик помилки у римських числах не враховуються.

З метою простішого оцінювання верхньої і нижньої меж недоступності, обрано наступні похідні від метрик характеристики:

— мінімальне значення з двох метрик (оптимістична межа);

— максимальне значення з двох метрик (песимістична межа).

Для спрощення оцінок, приймемо, що:

— наявність помилок у 25–49 % інформації, створює користувачу **проблеми з** пошуковою **доступністю**;

— наявність помилок у 50–99 % інформації, робить запис **важкодоступним** для пошуку;

— наявність помилок у 100 % інформації, робить відповідний запис **недоступним** для пошуку.

Розрахувавши значення та відсортувавши за песимістичним значенням, отримуємо графіки (рис. 1–3). На графіках проблем доступності інформації крива оптимістичних очікувань позначена числом 1, а песимістичних очікувань — числом 2.

Проблеми доступності записів книг за полем «Назва книги» (рис. 1):

— загалом є проблеми: 291 запис;

— проблеми з доступністю: 76 записів;

— важкодоступні: 57 записів;

— недоступні: 14 записів.

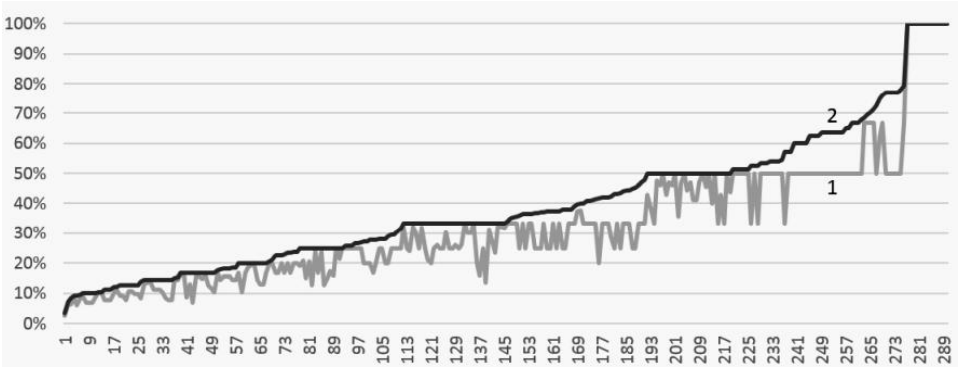

Рис. 1. Відсоток зіпсованих даних за полем «Назва книги»

Проблеми доступності записів книг за полем «Автор» (рис. 2):

— загалом з проблемами: 180 записів;

— проблеми з доступністю: 72 записи;

— важкодоступні: 78 записів;

— недоступні: 16 записів.

Проблеми доступності за полем «Видавництво» (рис. 3):

— загалом з проблемами: 231 запис;

— проблеми з доступністю: 84 записи;

— важкодоступні: 38 записів;

— недоступні: 85 записів.

При розгляді разом полів «Назва книги» та «Автор» виявилося, що у 7-ми записах є одночасно помилки в обох полях.

Загальні характеристики проблем пошукової доступності записів книг за полями «Назва книги» та «Автор» такі:

— загалом з проблемами: 463 записи;

— проблеми з доступністю: 146 записів;

— важкодоступні: 132 записи;
— недоступні: 30 записів.

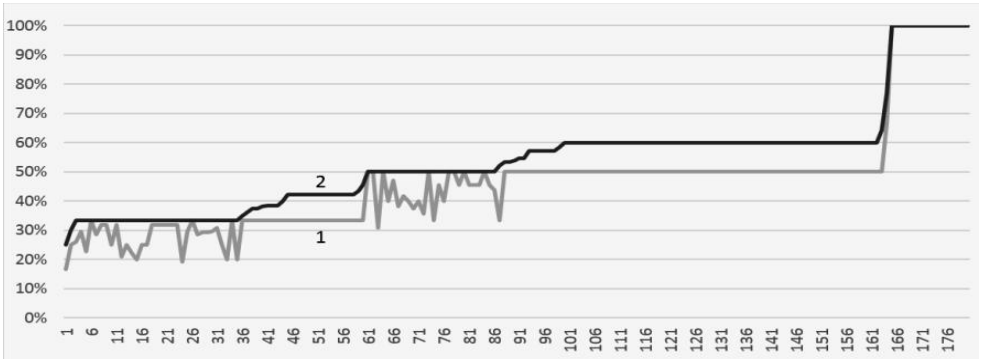

Рис. 2. Відсоток зіпсованих даних за полем «Автор»

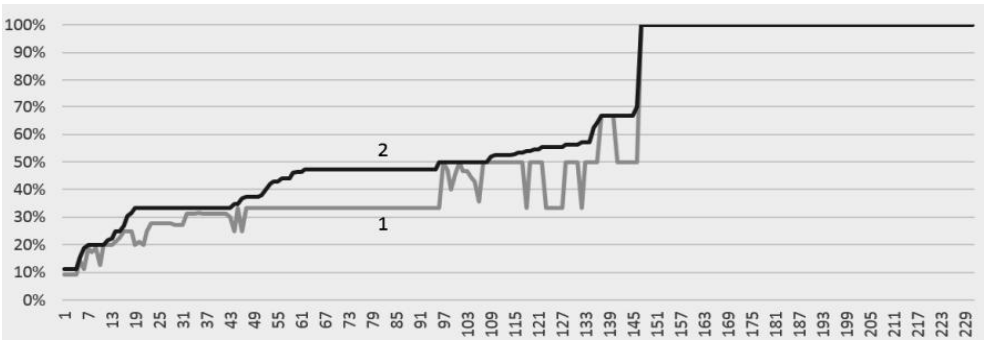

Рис. 3. Відсоток зіпсованих даних за полем «Видавництво»

У Базі 1 є 51 бібліотека (та відділення), у Базі 2 — 90 бібліотек (та відділення). У кожній бібліотеці/відділенні може бути від одного екземпляра книги до кількох. Тому одному запису може відповідати від 1 книги до кількох сотень реальних книг. У даному випадку 30 недоступних хоча б за одним показником записів відповідають не менш як 173 екземплярам книг (173 — якщо в усіх бібліотеках, де є книга, є лише один її екземпляр).

Вплив на пошук дублів записів

Будь-які помилки збільшують кількість унікальних низькочастотних значень певного поля та заважають виявленню дублів записів. У табл. 3 показане зменшення кількості псевдоунікальних низькочастотних значень після автоматичного виправлення помилок описаного у статті типу. Необхідно зазначити, що на відміну від пошуку, у задачах ідентифікації дублів поле «Видавництво» є достатньо застосовним.

Таблиця 3. Зменшення кількості унікальних значень полів після виправлення помилок

	Зменшення абсолютне	Зменшення відносне
Поле «Назва книги»	234	0,11 %
Поле «Автор»	110	0,1 %
Поле «Видавництво»	82	0,35 %

Розподіл помилок за символами

На початку статті висувалася гіпотеза про те, що усі візуально однакові символи кирилиці та латиниці можуть використовуватися помилково у словах іншої множини символів і впливати на якість даних. Ця гіпотеза підтверджується у табл. 4. Як ми бачимо, ймовірність використання символів суттєво різниться, але помилкове використання усіх їх має місце.

Для спрощення символи, що відповідають великим і малим літерам, об'єднані.

Таблиця 4. Кількість помилок за символами

Кирилиця		Латиниця	
Літера	Помилкових використань	Літера	Помилкових використань
С	46	C	244
А	46	I	143
М	29	A	50
О	28	O	44
Х	24	P	32
Т	22	M	19
Е	20	E	17
І	17	B	10
К	10	X	7
Р	9	T	4
В	7	K	2
Н	2	H	1

У Базі 1 надзвичайно поширеною помилкою є використання латинських літер «I» замість кириличних. Так буває за відсутності української розкладки чи підтримки мови в програмному забезпеченні.

Помилки та використання довідкових таблиць

Помічено, що в Базі 1 має місце обмежена кількість варіантів назв для таких полів як «Автор» та «Видавництво». Зазвичай, наявні лише варіанти різними мовами. Для одного і того ж видавництва у Базі 1 може бути 2 варіанти (російський та український), а у Базі 2 більше 10. Це видає широке використання в Базі 1 довідкових таблиць.

Використання довідкових таблиць дозволяє уникати зайвих помилок. Або помічати та виправляти їх усі в одному місці — словнику. Проте це не працює з помилками, які не помітні для оператора. У випадку потрапляння такої помилки до довідкової таблиці, вона впливає одразу на велику кількість записів.

Висновки

Розглянуто тип візуально непомітних помилок на великому масиві реальних бібліографічних даних. Описано метод автоматичного виявлення. Визначено розподіл помилок за полями, їхні особливості та вплив.

За важливими для пошуку полями («Назва книги» та «Автор») вказані помилки впливають лише на 0,10–0,13 % записів книг. Якщо додати до них поле

«Видавництво», отримаємо 0,11–0,21 %. Але у кількісному вимірі це сотні записів книг. А кожному запису може відповідати від однієї реальної книги до сотні та більше книг.

Важливість книг не є рівномірною (як і кількість екземплярів на один запис), будь-яка з тисяч книг, які отримали проблеми з доступністю через даний клас помилок, може бути саме тією, яка потрібна читачам.

Вирішення проблем подібного роду можливе наступними методами:

1) проведенням інструктажу бібліографів, які вводять дані, про вплив помилок;

2) удосконаленням АБІС та унеможливленням введення нових помилок;

3) розробкою інструментів пошуку та виправлення помилок;

4) використання електронних каталогів, які:

— надають користувачу інтелектуальний пошук, який враховує такі помилки (ПЗ має орієнтуватися на український ринок);

— дозволяють користувачеві не лише здійснювати пошук, але і переглядати за певними критеріями з фасетною навігацією.

Ідеальним варіантом є використання всього перерахованого вище.

За відсутності описаних вище програмних засобів, особливо уважними потрібно бути при введенні даних до довідкових таблиць, які можуть впливати на велику кількість записів, а побачити та виправити їх не буде можливості. У випадку з пошуком дублів, варто враховувати ймовірність таких помилок і виправляти їх автоматично. Ефект зменшення кількості унікальних записів становить 0,1–0,35 % залежно від поля.

1. Ярмолюк Р.С. Основні типи та джерела помилок у записах електронного каталогу. *Вісник Національного Університету «Львівська політехніка» Інформаційні системи та мережі*, 2010. № 689. С. 348–357.

2. Genzel D. HMM-based Script Identification for OCR. Proceedings of the 4th International Workshop on Multilingual OCR. 2013. P. 1–5.

3. Ringlstetter C. The same is not the same-postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE, 2005. P. 406–410.

4. Kakkonen T., Mozgovoy M. Hermetic and web plagiarism detection systems for student essays — an evaluation of the state-of-the-art. *Journal of Educational Computing Research*. 2010. P. 135–159.

5. Hu H. Assessing Browser-level Defense against IDN-based Phishing. 30th Security Symposium Security 21, 2021.

6. Rahma A., Bhaya W., Al-Nasrawi D. Text steganography based on unicode of characters in multilingual. *International Journal of Engineering Research and Applications*. 2013. 3.4. P. 1153–1165.

7. Електронний каталог бібліотеки ім. Лесі Українки (публічні бібліотеки для дорослих м. Києва). URL: http://ecatalog.kiev.ua

8. Електронний каталог бібліотеки ім. Тараса Шевченка для дітей (публічні бібліотеки для дітей м. Києва). URL: http://zra.kiev.ua:8081/MarcWeb

Надійшла до редакції 25.05.2021