УДК 004.9

Ярмолюк Р.С.

Хмельницький національний університет, Україна

ВЕРИФІКАЦІЇ ДАНИХ ЕЛЕКТРОННОГО КАТАЛОГУ НАУКОВО-ТЕХНІЧНОЇ БІБЛІОТЕКИ

Основною метою даного дослідження є визначення місця і ролі процесів верифікації інформації для забезпечення повноти та достовірності даних у бібліографічних базах даних.

The main purpose this research is to determine the role and place of the verification process information to ensure the completeness and accuracy of data in bibliographic databases.

Постановка проблеми. Електронний каталог, як основний довідниково-інформаційний модуль бібліотеки, являє собою складну метаінформаційну систему. Від якості даних, що містяться в електронному каталозі напряму залежить якість інформаційно-пошукових послуг бібліотеки. З іншого боку, на сучасній стадії розвитку інформаційних технологій на перше місце виходять задачі об'єднання локальних інформаційних ресурсів в рамках потужних інтегрованих інформаційних систем. Відповідно до концепції «Державної цільової національно-культурної програми створення єдиної інформаційної бібліотечної системи "Бібліотека-XXI"» в Україні функціонує понад 40 тис. бібліотек різного підпорядкування і форми власності. У той же час доступність цієї інформації залишається на дуже низькому рівні. Єдиним шляхом вирішення даної проблеми є створення розподіленої системи зберігання електронних бібліотечних ресурсів на основі єдиного центру каталогізації. При створенні такої потужної інформаційної системи одною з основних проблем є проблема якості даних та відсутності єдиного підходу оцінки та моніторингу наповненості електронного каталогу бібліотеки.

Для перевірки запропонованих у роботах [1-4] методів та засобів верифікації інформації в електронному каталозі бібліотеки було обрано бібліографічну базу даних електронного каталогу.

Електронний каталог НТБ працює під управлінням АБІС «УФД/Бібліотека». Представленням бібліографічної бази даних у АБІС «УФД/Бібліотека» є таблиця *document* з ключовим атрибутом *doc_id*.

Для подальшого дослідження було проведено аналіз типів атрибутів таблиці *document.* Результати представлені на рисунку 1.

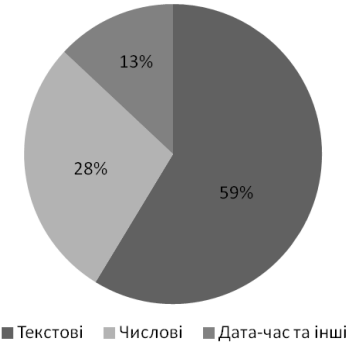

Рис. 1 **Діаграма співвідношення між типами атрибутів у таблиці document.**

З діаграми рис. 1 видно, що переважна більшість атрибутів мають текстовий тип даних, що дозволило при їх верифікації застосовувати запропоновані у [1-4] методи та засоби верифікації даних.

Попередній аналіз. Провівши аналіз атрибутів за призначеннями та врахувавши тип даних було зроблено наступне розбиття загальної множини атрибутів на підмножини:

1. Атрибути, що відображають час, дату та користувача, що змінив стан бібліографічного запису.
2. Атрибути, що слугують для забезпечення взаємозв'язку між службовими таблицями в структурі АБІС та атрибути допоміжного типу.
3. Основні атрибути бібліографічного запису.
4. Атрибути додаткової інформації.

Для подальшого дослідження була обрана підмножина основних атрибутів бібліографічного запису, які в свою чергу були розбиті на три групи у відповідності до застосовних до них методів верифікації даних:

- Атрибути до яких застосовуємо верифікацію за загальними правилами орфографічного правопису: анотація (annot1, annot2, annot3, annot4), назва документа (name, name_prefix).
- Атрибути які мають чітко визначену структуру запису, та до яких застосовуємо методи уніфікації та стандартизації за

допомогою регулярних виразів: ISBN(isbn), ISSN(issn), ББК(bbk), авторський знак(author_mark), УДК(udk), Шифр документа(cipher), рік видання(publ_year).

- Атрибути які мають власні назви і дозволяють вільну структуру запису, та до яких застосовуємо верифікацію, як за загальними правилами, так і на основі словників власних назв: автор(author), місце видання(publ_place), видавництво(publisher).

Визначення напрямків верифікації за аналізом значень атрибутів бібліографічної бази даних електронного каталогу НТБ. Для дослідження будо взято бібліографічну базу даних електронного каталогу НТБ. Потужність бібліографічної бази даних складає 284997 записів. Розподіл за мовою документа представлений на рисунку 2.

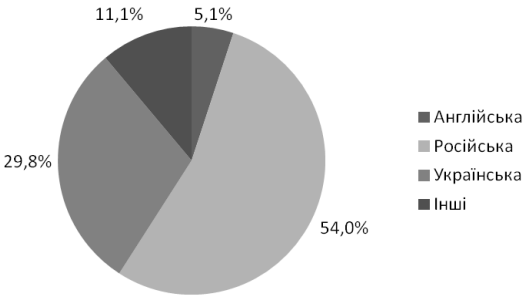

Рис. 2 **Діаграма розподілу записів ЕК НТБ за мовою документа.**

Отже, для верифікації 90% текстових даних ЕК НТБ необхідно застосовувати лише 3 набори загальнолексичних словників для української, російської та англійської мов.

За допомогою СВДЕК [5] проведемо верифікацію ЕК НТБ, щодо наявності NULL-значень у атрибутах, також оцінимо оригінальність значень та можливість запозичення. Дослідження проводились, як для множини всіх записів, так і для частин, що представляють відповідну мову документа. Результати дослідження наведені у таблиці 1.

Таблиця 1

Відсоткове відношення NULL-значень атрибутів від загальної кількості для ЕК НТБ

Для усіх записів	Записів англійською мовою	Записів російською мовою	Записів Українською мовою
41,66%	44,84%	39,73%	44,07%

З результатів дослідження слідує, що приблизно 40% усіх значень атрибутів не мають інформаційного наповнення і потребують їх редагування коректорами електронного каталогу. Також, очевидно, що дане відсоткове відношення не залежить від мови документу і має постійне середнє значення з проміжку [35%, 45%].

Однак відсутність даних у атрибутах, що не використовуються для забезпеченні інформаційно-пошукових функцій електронного каталогу не є критичною. Тому, для визначення впливу даної помилки на результат інформаційно-пошукових можливостей було визначено відсоткове відношення NULL-значень атрибутів, що входять до однорівневого бібліографічного опису документа (назва, автор, місце видання, рік видання, видавець). Результати представлені у таблиці 2.

Таблиця 2

Відсоткове відношення NULL-значень атрибутів, що входять до однорівневого бібліографічного опису, від загальної кількості для ЕК НТБ

Для усіх записів	Записів англійською мовою	Записів російською мовою	Записів українською мовою
19,85%	28,82%	11,73%	30,95%

З результатів дослідження слідує, що приблизно 20% усіх значень атрибутів (що входять до однорівневого бібліографічного опису документа) не мають інформаційного наповнення. Також, очевидно, що дане відсоткове відношення залежить від мови документу і є найбільшим для україномовних та англомовних записів і найменшим для російськомовних.

Дослідження текстових атрибутів, що входять до однорівневого бібліографічного опису на наявність орфографічних помилок представлено у таблицях 3,4,5.

Таблиця 3

Перевірка орфографії атрибутів, що входять до однорівневого бібліографічного опису ЕК НТБ для англомовних записів

Назва атрибуту	Загальна кількість орфографічних помилок	Загальна кількість помилкових записів	Відсоткове відношення помилкових записів	Середня кількість орфографічних помилок у одному записі
name	14554	6081	42,21%	2,39
publisher	9198	5659	39,28%	1,63
publ_place	4991	4076	28,29%	1,22
author	16665	4011	27,84%	4,15
publ_year	73	37	0,26%	1,97

Таблиця 4

Перевірка орфографії атрибутів, що входять до однорівневого бібліографічного опису ЕК НТБ для російськомовних записів

Назва атрибуту	Загальна кількість орфографічних помилок	Загальна кількість помилкових записів	Відсоткове відношення помилкових записів	Середня кількість орфографічних помилок у одному записі
name	47069	33853	21,98%	1,39
publisher	80271	63537	41,25%	1,26
publ_place	106156	103201	67,00%	1,03
author	168519	77513	50,32%	2,17
publ_year	2241	2095	1,36%	1,07

Таблиця 5

Перевірка орфографії атрибутів, що входять до однорівневого бібліографічного опису ЕК НТБ для україномовних записів

Назва атрибуту	Загальна кількість орфографічних помилок	Загальна кількість помилкових записів	Відсоткове відношення помилкових записів	Середня кількість орфографічних помилок у одному записі
name	32053	19977	23,56%	1,60
publisher	47110	26618	31,39%	1,77

publ_place	50427	49829	58,76%	1,01
author	103031	42851	50,53%	2,40
publ_year	2337	2257	2,66%	1,04

Отриманий у результаті дослідження великий відсоток орфографічних помилок для англомовних записів пояснюється тим, що сам бібліографічний запис у базі даних зроблений не англійською а російською або українською мовою. Наприклад для усіх англо-російських або англо-українських словників бібліографічний запис зроблено або на українській або на російській мові.

Також можливо пояснити великий відсоток помилок для атрибутів, що мають у своєму записі власні назви (Автор, Назва видавництва, Місце видавництва) відсутністю у складі підсистеми перевірки орфографії СВДЕК спеціалізованих словників прізвищ, імен та географічних назв міст.

Отже, пошук орфографічних помилок є коректним лише для атрибуту, що відповідає за назву документа (name). Для даного атрибуту відсоткове відношення орфографічних помилок становить приблизно 22%.

Розрахунок інтегральної оцінки якісного наповнення ЕК НТБ. За представленим у [6] методом обрахуємо інтегральну оцінку якісного наповнення ЕК НТБ. Розглядати будемо лише множину атрибутів, що входять до однорівневого бібліографічного опису документа, тобто

$$A = \{name,\ author,\ publisher,\ publ\_place,\ publ\_year\}$$

Розглянемо також 3 набори кортежів (англомовні, російськомовні, україномовні):

1.Для англомовних отримаємо:

Граничне значення	Кількість кортежів	Оцінка
Помилки відсутні	3356	23,29%
1 помилки з 5	4676	55,75%
2 помилки з 5	4549	87,32%
3 помилки з 5	1223	95,81%
4 помилки з 5	599	99,97%
5 помилки з 5	5	100,00%

2.Для російськомовних отримаємо:

Граничне значення	Кількість кортежів	Оцінка
Помилки відсутні	8132	5,28%
1 помилки з 5	49446	37,38%
2 помилки з 5	61659	77,41%
3 помилки з 5	31753	98,03%
4 помилки з 5	3034	99,99%
5 помилки з 5	8	100,00%

3.Для україномовних отримаємо:

Граничне значення	Кількість кортежів	Оцінка
Помилки відсутні	3606	4,25%
1 помилки з 5	35859	46,54%
2 помилки з 5	31847	84,09%
3 помилки з 5	12010	98,25%
4 помилки з 5	1451	99,97%
5 помилки з 5	29	100,00%

З аналізу отриманих результатів для трьох наборів кортежів слідує, що для забезпечення прийнятної (>95%) інтегральної оцінки якісного наповнення ЕК НТБ достатньо встановити граничне значення для кількості помилкових атрибутів рівним 3 помилки з 5, або >60%.

Отже проаналізувавши показники якісного наповнення ББД ЕК НТБ можливо зробити висновки про доцільність проведення перевірки електронного каталогу. Особливу увагу необхідно звернути на дотримання дисципліни заповнення співробітниками бібліотеки неосновних полів бібліографічного запису.

Також більшу увагу приділяти перевірці правопису при введені бібліографічних записів на іноземній мові.

Висновки.

1.При верифікації ЕК НТБ були відмічені великі показники відсоткового відношення на NULL-значення для основних атрибутів бібліографічного опису, зокрема кодів (Авторський знак-43%, УДК-47%, ISSN-85%, ISBN-86%, ББК-94%), що не дають можливості користувачам у повному обсязі використовувати розширені поля для пошуку літературних джерел. Також слід відмітити наявність малої кількості анотацій до літературних джерел, що пояснюється високими затратами часу на їх обробку та введення в ББД. Але основною проблемою є високе відсоткове відношення NULL-значень для атрибуту Автор, що складає 46%.

2.Коректору електронного каталогу необхідно почати з перевірки атрибутів основного пошуку: це - Рік видання (publ_year-12%) та Автор (author-46%), що дозволить підвищити інформаційно-пошукові можливості ЕК НТБ

Література

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

2.Ярмолюк Р.С. Задача аналізу текстових атрибутів в електронному каталозі / Р.С. Ярмолюк //Вісник Хмельницького Національного Університету, серія Технічні науки, - 2011. - № 3 – С.225-228.

3.Ярмолюк Р.С. Уніфікація атрибутів кортежу бази даних засобами регулярних виразів на прикладі електронного каталогу бібліотеки / Р.С. Ярмолюк //Вісник Хмельницького Національного Університету, серія Технічні науки, - 2012. - № 1 – С.186-190 .

4.Ярмолюк Р. С. Підсистема перевірки орфографії електронного каталогу бібліотеки на основі технології HUNSPELL / Р. С. Ярмолюк // Вісник Національного Університету «Львівська політехніка» Інформаційні системи та мережі,- 2012. - № 743. – С.219-230.

5.Ярмолюк Р.С. Структурно-функціональна модель верифікації даних електронного каталогу / Р.С. Ярмолюк // Сучасні проблеми діяльності бібліотеки в умовах інформаційного суспільства: матеріали третьої науково-практичної конференції, 29 вересня 2011р., Львів/ Національний університет "Львівська політехніка"; - Львів : Видавництво Національного університету "Львівська політехніка", 2011. – С. 217-224.

6.Ярмолюк Р. Підходи до розрахунку інтегральної оцінки якісного наповнення електронного каталогу бібліотеки / Р. Ярмолюк // Інноваційні комп'ютерні технології у вищій школі : матеріали 3-ї науково-практичної конференції, 18–20 жовтня 2011 р., Львів / Національний університет "Львівська політехніка"; - Львів : Видавництво Національного університету "Львівська політехніка", 2011. – С. 132-136.