УДК 004.912

Ярмолюк Р.С.

Хмельницький національний університет, Україна

РОЛЬ ВЕРИФІКАЦІЇ ДЛЯ ЗАБЕЧПЕЧЕННЯ ЯКОСТІ ДАНИХ В БІБЛІОГРАФІЧНИХ БАЗАХ ДАНИХ ЕЛЕКТРОННОГО КАТАЛОГУ 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. Відповідно до [1, 2] під електронним каталогом будемо розуміти бібліотечний каталог в машиночитаній формі, що працює в реальному режимі часу і надається в розпорядження читачів бібліотеки.

Поняття електронного каталогу, як певної підсистеми сучасної бібліотеки пов'язане з впровадженням у бібліотечну практику комп'ютерно-інформаційних технологій. Перші спроби автоматизації бібліотечно-бібліографічних процесів розпочались у 1961 році на базі Бібліотеки Конгресу США. З того часу еволюційний процес розвитку електронних каталогів йшов у ногу із розвитком, як апаратного так і програмного забезпечення ЕОМ.

На даний час в Україні та за кордоном розроблено чимало автоматизованих бібліотечних систем (АБІС) різного рівня складності та масштабу. Серед таких систем можна виділити УФД/Бібліотека, ІРБІС, МАРК-SQL, КАБІС, UNILIB, LIBER, ALEPH, Руслан. Чимало АБІС є open-source продуктами, зокрема, Koha, ISIS, CDS Invenio, OpenBiblio, Evergreen. Усі перераховані АБІС реалізують функціональні можливості електронного каталогу на своїх програмних платформах.

На рисунку 1 запропонована узагальнена організаційно-функціональна схема АБІС.

Відповідно до схеми представленої на рисунку 1 електронний каталог, як підсистема АБІС, бере участь у всіх технологічних процес, що проходять у сучасній бібліотеці і є основним засобом для

забезпечення інформаційно-пошукових потреб користувачів бібліотеки.

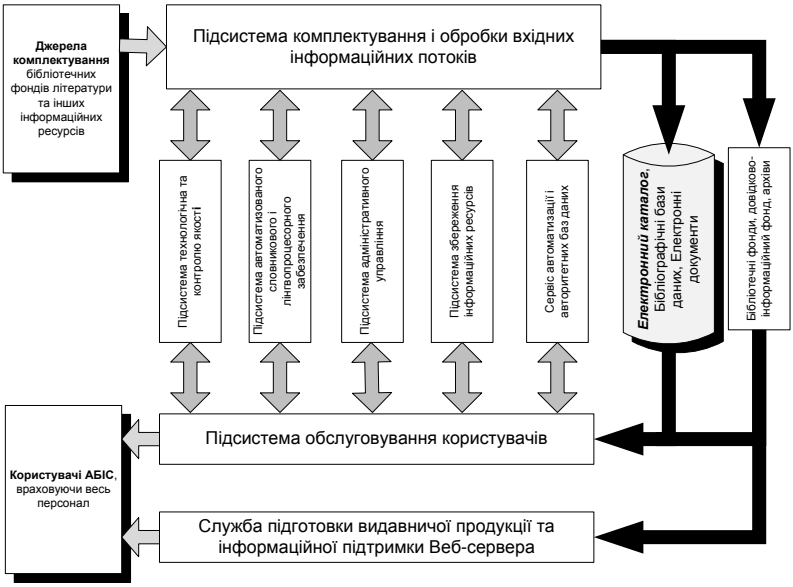

Рис. 1.**Узагальнена організаційно-функціональна схема АБІС.**

Отже від функціонального стану підсистеми електронного каталогу АБІС залежить якість надання інформаційно-пошукових та довідникових послуг сучасною бібліотекою.

Аналіз останніх досліджень і публікацій. В У роботі [3] представлено аналіз якісного наповнення електронного каталогу Харківської державної наукової бібліотеки ім. В. Г. Короленка. Відповідно до аналізу основну частину помилок складають [3]:

32% – технологічні, пов'язані в основному з неправильним визначенням виду документа (наприклад видання, що продовжується, монографія);

27% – пропущені елементи опису (видавництво, вихідні дані, ISBN і т. ін.);

22% – бібліографічні помилки, пов'язані з порушенням правил БО (БЗ зроблений під заголовком, а слід – під колективом, БЗ зроблений під індивідуальним автором, який є укладачем і т. д.);

9% – граматичні помилки;

5% – змістові (смислові, коли зі словників беруться неправильні відомості, помилки в термінології);

3% – складають прогалини, що засмічують пошукові словники; 2% – неправильна пунктуація (наприклад плутання дефісу і тире, крапки та крапки з комою).

Проведений Вершиніним М.Й. у [4] аналіз статистики помилок дозволяє зробити наступні висновки:

- в середньому в записах ЕКБ частота помилок складає 0,1% в тому числі одно літерні помилки складають 85-95%;
- найбільш ймовірне спотворення початку слова; для слів довжини 3-8 символів найбільш ймовірні помилки в трьох-чотирьох позиціях;
- приблизний розподіл помилок: пропуск літер – 30-40%, вставка – 25-35%, заміна – 15-20%, перестановка – 10-15%;
- помилки в голосних (вставка і пропуск) зустрічаються частіше ніж в приголосних;
- найбільш ймовірні помилки в початкових лексичних одиницях полів бібліографічного запису.

З точки зору місця появи помилки у бібліографічній базі даних їх можна класифікувати наступним чином:

- Помилки в значення окремих атрибутів;
- Помилки у кортежах, що представляють бібліографічний запис;

Отже представлений огляд підходів щодо класифікації помилок дозволяє зробити висновки про можливість появи помилок на всіх рівнях функціонування електронного каталогу. У наукових дослідженнях різних авторів основну увагу приділяється знаходженню та виправленню лише символьних спотворень у текстових полях бібліографічного запису, а проблема пошуку помилок на рівні самого бібліографічного запису залишається невирішеною.

Формулювання цілей статті та актуальність досліджень. Відповідно до [1,2,4,5] електронний каталог бібліотеки складається з двох основних частин:

- системи лінгвістичного забезпечення;
- баз даних.

Схема структури електронного каталогу представлена на рисунку 2.

У роботі [4] відмічається, що головною частиною електронного каталогу є бібліографічна база даних. Також у [6] дається визначення електронному каталогу, як бібліографічній базі даних , що має ознаки каталогу. Звичайно, повністю ототожнювати

електронний каталог, як бібліографічну базу даних не можна. Але у контексті задачі верифікації інформації в електронному каталозі, дане припущення є природнім. Тому надалі електронний каталог розглядається, як структурована певним чином множина бібліографічних записів.

Рис. 2 **Структура електронного каталогу бібліотеки.**

Виклад основних матеріалів дослідження.

На даний час основними стандартами представлення бібліографічних записів в електронному каталозі є формати сімейства MARC: MARC21 (www.loc.gov/marc), UNIMARC (www.ifla.org/unimarc) та стандарт представлення описових метаданих для цифрових об’єктів Dublin Core (www.dublincore.org).

Як метаінформація, повний бібліографічний запис, що містить багатоаспектний опис документа (видання), характеризується значною інформаційною надмірністю. Це означає, що для однозначної ідентифікації об'єкта опису досить частини полів (підполів), а не всього запису.

Також доцільно відмітити, що бази даних не є єдиною технологією зберігання і забезпечення доступу до інформації. Вершинін М.Й. у своїх роботах відмічає і широке поширення технології XML та перспективність даного напрямку у майбутньому. Однак, оскільки засоби верифікації інформації для електронних

каталогів бібліотек повинні будуватись для АБІС, що функціонують в даний час. А переважна більшість АБІС, як зазначалось вище, реалізовані на реляційних СКБД. Тому у подальшому зупинимось лише на реляційній моделі даних.

Отже у контексті верифікації інформації електронний каталог бібліотеки можливо розглядати, як бібліографічну базу даних, що реалізована на реляційній моделі даних. А сам процес верифікації інформації, як верифікацію даних у бібліографічних базах даних.

Під верифікацією даних в електронному каталозі бібліотеки будемо розуміти процес пошуку різного типу помилок у даних бібліографічної бази даних. Щодо питання автоматичного виправлення знайдених у процесі верифікації помилкових або недостовірних даних, то даний аспект проблеми повинен залишатись на розсуд відповідальної особи.

Основне завдання процесу верифікації - забезпечення високого рівня якості даних у електронному каталозі бібліотеки. У свою чергу поняття «якість даних» різними науковцями трактується по-різному в залежності від сфери використання. У даній роботі поняття «якість даних», відповідно до стандартів ISO для інформаційних систем [7], використовується як деякий критерій відповідності даних або інформації потребам користувача. Тоді якість даних у контексті бібліографічної бази даних електронного каталогу – це критерій достовірності даних, тобто чи бібліографічний запис відповідає реальному документу з фондів бібліотеки.

Відкритою залишається проблема визначення місця верифікації даних у життєвому циклі електронного каталогу. Відповідно до запропонованого у [4] життєвого циклу електронного каталогу бібліотеки даний процес незалежно від моделі розробки складається п’яти етапів, представлених на рисунку 3.

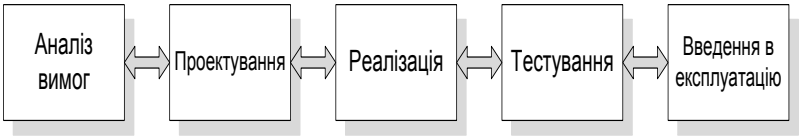

Рис. 3 **Етапи життєвого циклу електронного каталогу.**

На кожному з етапів життєвого циклу можлива поява певного роду технічних помилок проектувальника, що не були виявлені у процесі міжетапної перевірки і які мають суттєвий вплив на функціонування електронного каталогу. Очевидно, що після етапу «Введення в експлуатацію», дані помилки виправити практично

неможливо без змін у програмній частині електронного каталогу. До основник таких технічних помилок можна віднести:

- неповне або часткове врахування бізнес-правил, що притаманні даній предметній області і як наслідок недостатня реалізація обмежень цілісності у базах даних;
- помилки у проектуванні користувацького інтерфейсу, як читача так і службового;
- помилки в проектуванні структури бази даних, наприклад відсутність механізмів контролю цілісності, повноти та достовірності даних у базах даних.

Наявність помилок проектування у процесі роботи електронного каталогу є одним із основних чинників появи недостовірних, неповних та помилкових даних у записах бібліографічної бази даних електронного каталогу, що спричиняють появу потужного інформаційного шуму при роботі з електронним каталогом бібліотеки.

Отже верифікація даних повинна проводитись паралельно процесу функціонування електронного каталогу бібліотеки і бути непомітною для користувачів.

Оскільки, основним елементом бібліографічної бази даних є бібліографічний запис, то розглянемо докладніше життєвий цикл бібліографічного запису представлений на рисунку 4.

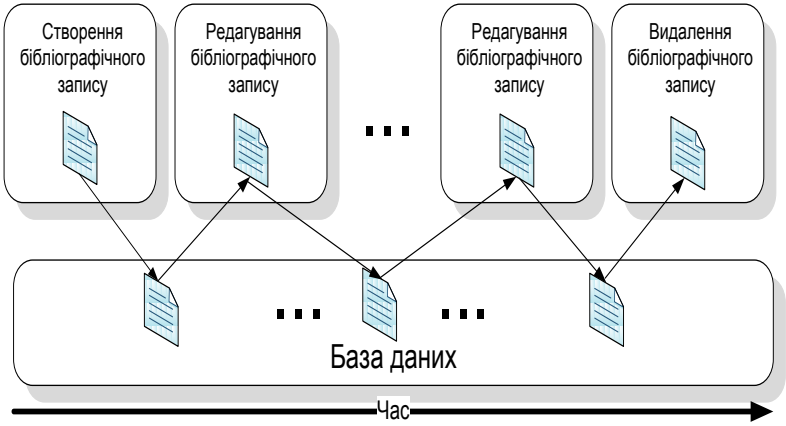

Рис. 4 **Життєвий цикл бібліографічного запису [8].**

Очевидно, що переважна більшість помилок виникають підчас етапу створення бібліографічного запису. Однак, на етапі редагування

також можлива поява помилок у записах. Тому, процес верифікації кожного бібліографічного запису потрібно проводити, як на етапі створення так і при кожному редагуванні бібліографічного запису. Для розв'язання даної проблеми необхідно створення системи online-моніторингу якості даних бібліографічної бази даних електронного каталогу бібліотеки.

Висновки.

1.Електронний каталог, як підсистема АБІС, бере участь у всіх технологічних процес, що проходять у сучасній бібліотеці і є основним засобом для забезпечення інформаційно-пошукових потреб користувачів бібліотеки. Від функціонального стану бібліографічної бази даних електронного каталогу залежить якість надання інформаційно-пошукових та довідникових послуг сучасною бібліотекою.

2.Реляційна модель даних повністю сумісна із форматами представлення бібліографічної інформації Dublin Core та MARC-сімейства. У контексті верифікації інформації електронний каталог бібліотеки можливо розглядати, як бібліографічну базу даних, що реалізована на реляційній моделі даних. А сам процес верифікації інформації, як верифікацію даних у бібліографічних базах даних.

3.Якість даних у контексті бібліографічної бази даних електронного каталогу – це критерій достовірності даних, тобто чи бібліографічний запис відповідає реальному документу з фондів бібліотеки.

4.Верифікація даних проводиться паралельно процесу функціонування електронного каталогу бібліотеки і непомітна для користувачів. У процесі верифікації даних бібліографічної бази даних необхідно проведення попереднього аналізу джерел та технологій створення бібліографічних записів. Повний бібліографічний запис, що містить багатоаспектний опис документа (видання), характеризується значною інформаційною надмірністю.

5.Відомі методи та засоби верифікації, очистки та підвищення достовірності даних у бібліографічних базах даних не вирішують задачу комплексно, вимагають додаткових відомостей про предметну область, яка представлена в електронному каталозі, що не завжди є доступним та не враховують специфіку бібліографічних баз даних.

6.Розробка уніфікованого підходу до вирішення задачі верифікації бібліографічних баз даних та необхідність створення ефективних програмних засобів очистки та підвищення достовірності бібліографічних даних є нагальною потребою для бібліотеки 21-го століття.

Література

1.ДСТУ 2394-94 Інформація та документація. Комплектування фонду, бібліографічний опис, опис, аналіз документів. Терміни та визначення. – Чинний з 01.01.1995. – 89с.

2.ГОСТ 7.76-96 СИБИД Комплектование фонда документво. Библиографирование. Каталогизация. Термины и определения. - Чинний з 01.01.1998, Б.м., Б.г. – 56с.

3. Поліщук О. Редагування електронних каталогів – новий напрямок підвищення якості інформаційного обслуговування користувачів бібліотек /О.Поліщук //Бібл. форум України.–2006.–№ 4. – С.27 – 30.

4. Вершинин М. И. Электронный каталог проблемы и решения / М. И. Вершинин. – СПб. : ПРОФЕССИЯ, 2007. – 233с.

5.Воройский Ф.С. Основы проектирования автоматизированных бибилиотечно-информационных систем / Ф.С. Воройский – М.: ГПНТБ России, 2002. – 389с.

6.Воройский Ф.С. Информатика: новый систематизированный толковый словарь-справочник / Ф.С. Воройский. – М.: Физ. мат. лит., 2003. – 706 с.

7.ISO 8000-110:2009, Data quality — Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification.

8.Карауш А.С. Вопросы обеспечения обратной связи при работе с библиографическими записями // «Информационные технологии, компьютерные системы и издательская продукция для библиотек»: Доклады и тез. докладов. – М.: ГПНТБ России, 2003. – С. 113-116.