

ВЕРИФІКАЦІЯ ІНФОРМАЦІЇ В ЕЛЕКТРОННИХ КАТАЛОГАХ БІБЛІОТЕК

© Руслан Ярмолюк, 2010

In this paper the problem of verification, data cleaning in the electronic library catalogs. The classification of major types of errors in the electronic catalog records. The sources of their origin. The basic methods and means of verification data in the electronic library catalogs.

Keywords - electronic catalog, input errors, retrospective conversion, and import data.

У роботі піднімається проблема верифікації, очищення даних в електронних каталогах бібліотек. Проведена класифікація основних типів помилок у записах електронного каталогу. Проаналізовані джерела їх виникнення. Розглянуті основні методи та засоби верифікації даних в електронних каталогах бібліотек.

Ключові слова – електронний каталог, помилки вводу, ретроспективна конверсія, імпорт даних.

Вступ

Кількість інформації в світі невпинно росте, не показуючи тенденції до уповільнення темпів росту. Об'єми даних і складність взаємозв'язків давно перевищили можливості людського організму по їх обробці, і стрімко наближаються до границь можливостей ЕОМ. Використання ЕОМ для обробки інформації, як не парадоксально, привело до різкого збільшення «інформаційного сміття» в наслідок спрощення операцій тиражування. Наявність великої кількості помилок в даних, дублювання інформації, територіальна розподіленість баз даних, різні формати опису даних і способів їх представлення для кінцевого користувача, легкість доступу до великих масивів інформації, швидкий ріст кількості її користувачів (як правило, погано підготовлених) – ось далеко не повний перелік факторів, які гостро ставлять проблему верифікації інформації та автоматичної обробки інформації. Сучасна бібліотека являється складною інформаційною системою, що складається як із традиційних, так і нових нетрадиційних інформаційних підсистем. Важливу роль серед нових інформаційних підсистем бібліотеки грає електронний каталог і його створення являється пріоритетною ціллю автоматизації бібліотек. Саме електронний каталог відкриває швидкий і

якісний доступ до інформаційних ресурсів бібліотеки. Якість і ефективність електронного каталогу обумовлюються комплексом методів і засобів, що визначають технологію його створення і використання. [1]

Теоретичні і практичні проблеми створення електронного каталогу розробляли такі західні вчені, як Henriette D. Avram, Hugh C. Atkinson, Cyril Cleverdon, Donald S. Culbertson, Richard de Gennaro, Franc W. Lancaster, Gerard Salton, L. Syre та інші.[2]

На пострадянському просторі основи рішення проблеми створення електронного каталогу закладені в працях Р.С. Гиляревского, Д.Г. Лахути, В.П. Леонова, А.В.Соколова, А.І. Черного, Ю.І. Шемакина. Теоретичні і практичні питання створення електронного каталогу розглядаються в роботах А.Б. Антопольского, Ф.С. Воройского, Б.С. Елепова, Н.Е. Каленова, О.А. Лавреновой, Я.Л. Шрайберга.[2]

На даний час в Україні та закордоном розроблено чимало автоматизованих бібліотечних систем(АБІС) різного рівня складності та масштабу. Серед таких систем зокрема 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. Однак аналіз описових можливостей переважної більшості перелічених АБІС показав, що в ній, як правило, відсутні ефективні засоби верифікації інформації в електронних каталогах. Тому проблема пошуку та виправлення помилок у бібліографічних записах електронного каталогу є досить актуальною.

Виклад основного матеріалу

Під час всього життєвого циклу електронного каталогу, як інформаційної системи відбувається процеси створення, редагування та видалення бібліографічних записів. Отже, електронний каталог можна розглядати як множину бібліографічних записів відповідного формату або структури. 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).

електронного каталогу бібліографічними записами відбувається різними способами (рисунок 1). Потрібно відмітити, що основна частина нової бібліографічної інформації вводиться за допомогою клавіатури ЕОМ і таким чином дані, отримані з зовнішніх джерел, також введені з клавіатури. Цей факт необхідно враховувати при оцінці якості отриманих із зовнішніх джерел даних. Досвід показує, що навіть незначні помилки, які ніяк не впливають на

результати пошуку в традиційних каталогах, мають велике значення при пошуку в електронному каталозі.[1]

Оскільки записи електронного каталогу являють собою інформацію про певне джерело інформації (книгу, статтю, дисертацію, монографію, тощо), то і розглядати ми їх будемо як набір метаданих. Помилки які виникають у записах електронного каталогу можна розділити на три основних типи:

- типографічні помилки у самих значеннях полів запису електронного каталогу (орфографічні, синтаксичні, семантичні);
- логічні помилки, узгодження між значеннями окремих полів бібліографічного опису, протиріччя даних;
- помилки дублювання записів в електронному каталозі.

Відповідно до способу створення або редагування запису (рисунок 1) в електронному каталозі можна визначити основні типи та джерела помилок, що притаманні кожному із способів.

Клавіатурний ввід даних. З моменту появи прототипів сучасних ЕК і по нинішній час клавіатурний ввід інформації являється основним способом створення ЕК. В ролі операторів виступає комплектатор, каталогізатор, систематизатор або звичайний робітник бібліотеки. Вдосконалення цього процесу зводиться до наступного:[1]

- візуальний контроль вводу;
- використання засобів прискорення вводу, наприклад, словників;
- застосування шаблонів для вводу типової інформації.

Не зважаючи на це, помилки при ручному вводі даних неминучі.[3]

В роботах різних авторів приводяться класифікації можливих помилок при вводі даних. Об'єднання та узагальнення результатів цих робіт дозволяють запропонувати наступну класифікацію помилок: [1]

1. Помилки каталогізації: невірне виділення бібліографічного опису із джерела інформації.

2. Помилки структури запису: не вірні теги полів або невірні мітки підполів як результат технічної помилки.

3. Помилки структури підполів: пропуск одного чи більше слів, абзаців, перестановка слів, помилки пунктуації.

4. Символьні помилки: вставка, заміна, видалення одного чи більше символів та перестановці сусідніх символів.

5. Орфографічні помилки: неправильне написання слів іноземного походження, винятків правопису, підсвідома заміна невідомого слова на подібне по напису.

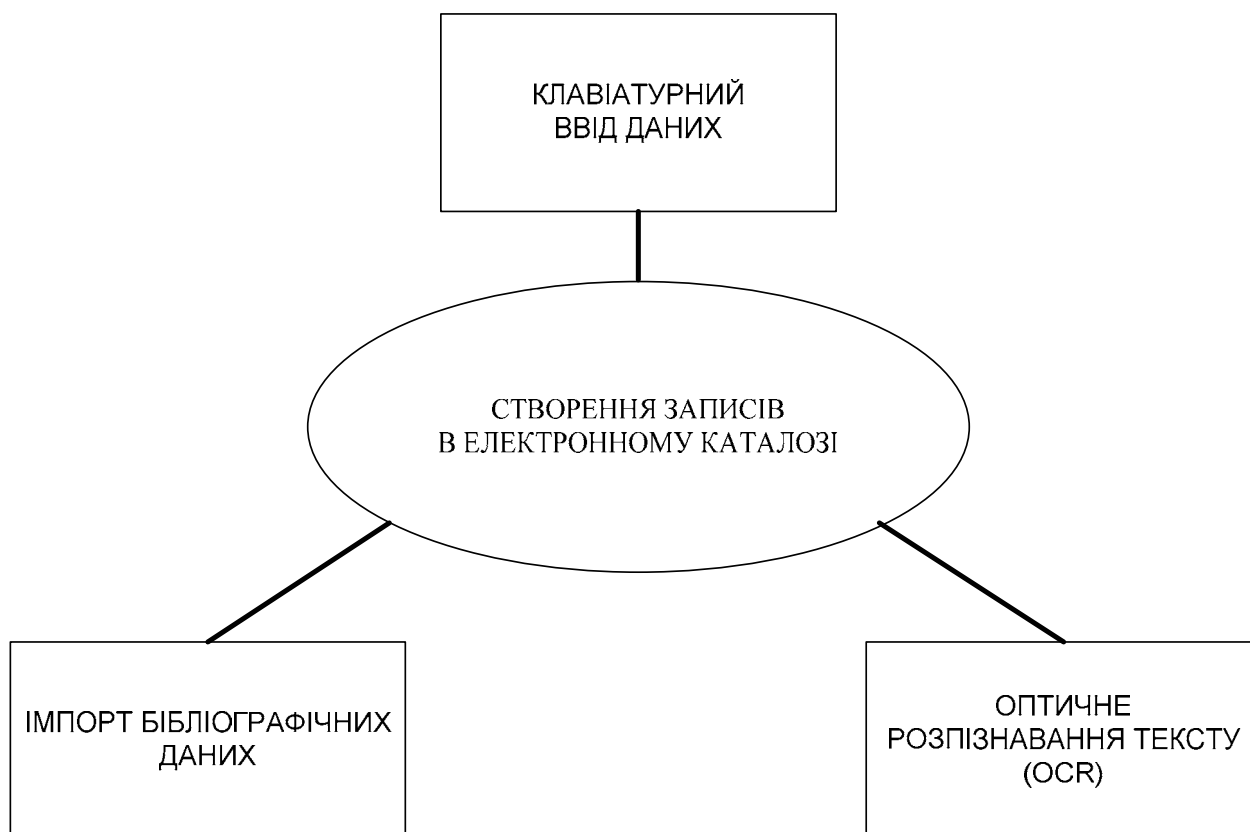


Рис. 1. Засоби створення записів в ЕК

Оптичне розпізнавання тексту (OCR). Оптичне розпізнавання тексту в основному використовується при ретроспективній конверсії. Ретроспективна конверсія називається технологічний процес переводу традиційних каталогів (карточних, книжкових и т.д.) в цифрову форму.[1]

Окрім помилок характерних для клавіатурного вводу, появляються специфічні, які залежать від наступних факторів:

- якість носія інформації;
- ступінь збереження тексту на носієві;
- шифру що використовувався.

Помилки класифікують за типами наступним чином:

1. Заміна символу на подібний по напису (I=1, O=0).
2. Заміна сусідніх символів на один (I I=П, M И=Ш).
3. Заміна одного символу на два (Ю=I O, Ж=} K).

Помилки оптичного розпізнавання тексту мають системний характер тому для їх усунення достатньо використовувати системі утиліти для усунення типових помилок за шаблонами.

Імпорт бібліографічних даних. Даний засіб створення бібліографічних записів використовується при отриманні бібліографічних описів із зовнішніх джерел (бази даних видавництв, журналів, тощо) та при обміні даними між каталогами. Звичайно при обміні даними між каталогами, помилки, що були у записах одного каталогу перейдуть у інший. Тому розробка інструментарію очистки даних при імпорті бібліографічних даних є дуже важливою проблемою. Також до помилок, що вже присутні у даних, що імпортуються, додаються помилки конвертації із одного формату представлення даних в інший.

Отже основними проблемами та причинами помилок у ЕК є: відсутність належної спеціалізації у операторів, незнання іноземної мови, неправильне розуміння термінології, недостатнє володіння комп'ютерною технікою, ввід даних без оригіналу екземпляру літературного джерела або по його електронній копії, в якій можуть бути відсутні титульні сторінки, недосконалість програмного інтерфейсу вводу, помилки розпізнавання рукописних знаків, неправильні бібліографічні описи, бібліографічні описи з різними стандартами, проблеми конвертації бібліографічної інформації. Пошук помилок в будь-якій системі слід починати з можливого аналізу їх появи. Основні варіанти пошуку і виявлення помилок можна визначити наступним чином: Пошук помилок за запитом бібліотекарів або «автоматизаторів»: автоматичний запит - використання типових запитів на часті помилки; випадкові запити - використання змінюваних запитів для пошуку нестандартних помилок. Використання зворотного зв'язку для знаходження помилок. Найбільш частий випадок - виправлення помилок згідно з відгуками користувачів системи. Після знаходження та локалізації помилки її слід виправити, а також знайти і «покарати» винних. Сам процес виправлення помилок може вестися різними методами: [4]

1. Автоматичне виправлення часто зустрічаються помилок. Помилки в даному випадку виправляє програмне забезпечення для автоматизації або зовнішнє програмне забезпечення.
2. Ручне виправлення помилок. У міру знаходження помилок про них повідомляється оператора або бібліотекаря, який їх виправляє.
3. Напівавтоматичне виправлення помилок. У даному випадку при знаходженні деякої кількості однакових помилок у системі є можливість виправити їх пакетним способом.
4. Виправлення помилок зовнішніми користувачами системи. У цьому випадку може бути організований доступ користувачів для організації виправлення помилкових даних.

Висновки

У роботі піднімається проблема верифікації, очищення даних в електронних каталогах бібліотек. Вказано основні засоби та методи створення записів в електронному каталозі. Розглянуто та проведена класифікація помилок, що виникають при ручному (клавіатурному) вводі інформації та ретроспективній конверсії та імпорту даних в електронному каталозі. Представлено основні причини виникнення даних помилок. Розглянуті деякі методи та засоби верифікації даних в електронному каталозі бібліотек.

1. Вершинин М.И. *Электронный каталог проблемы и решения*. Изд. «ПРОФЕССИЯ», СПб, 2007. 2. Вершинин М.И. *Проблемы создания электронного каталога в научных библиотеках*. Дис. канд. пед. наук, СПб, 2002 3. Nielsen R. *Lost articles: Filing problems with initial articles in data bases*/ Nielsen R. Pyle J.M.// *Libr. Resources a. techn. Services* – 1995 – Vol 39 №3 P.291-293 4. Крауш А.С., *Утилиты для проверки и коррекции электронных каталогов*/ Крауш А.С., Копытков Д.Ю., Макаревич А.С.// *Библиотечное дело* – 2005 – №6(30) С. 21-24.

УДК 004.9

Андрій Андрухів, Дмитро Тарасов

Науково-технічна бібліотека Національного університету «Львівська політехніка»

ІНТЕГРАЦІЯ БІБЛІОТЕЧНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ У ІНФОРМАЦІЙНЕ СЕРЕДОВИЩЕ УНІВЕРСИТЕТУ

© Д.О. Тарасов, А.І. Андрухів, 2010

Запропоновано схему інтеграції автоматизованої бібліотечної системи (АБІС) у інформаційне середовище університету з метою підвищення якості інформаційного забезпечення навчального та управлінського процесів, автоматизації формування та поширення списків рекомендованої літератури.

Ключові слова: АБІС, інформаційне середовище університету, інформаційні потоки навчального процесу.

The chart of integration library information system and university information environment is offered with the purpose of upgrading of the informative providing of educational and administrative processes, automation of forming and distribution of lists of the recommended literature.