УДК 004.9

Р.С. Ярмолюк
Хмельницький національний університет

ОСНОВНІ ТИПИ ТА ДЖЕРЕЛА ПОМИЛОК У ЗАПИСАХ ЕЛЕКТРОННОГО КАТАЛОГУ

Електронний каталог розглядається як метаінформаційна система. Представлені основні формати опису електронного каталогу. Вказано основні типи помилок у записах електронного каталогу та джерела їх виникнення.

Ключові слова: електронний каталог, теорія множин, дублінське ядро, помилки вводу, ретроспективна конверсія, метадані.

The article OPAC regarded as objective information system. The main formats of the electronic catalog descriptions. The basic types of errors in the electronic catalog records, and indicated their source of origin.

Keywords: OPAC, set theory, Dublin Core, input errors, retrospective conversion, metadata.

Вступ

Глобальний процес, що отримав назву «інформатизація суспільства», впливає на всі аспекти суспільного життя. Головне, що відрізняє цей процес, полягає в отриманні інформацією статусу фундаментального фактора існування людства. Якщо раніше життя і прогрес людства залежали переважно від матеріального виробництва, то тепер вони неможливі без максимального використання інформації у всіх її проявах. Інформаційний фактор швидко набуває такої самої ваги, як і матеріальний. Змінились і суспільні потреби в інформації, одним із наслідків чого стала докорінна трансформація бібліотечної справи. Бібліотека все більше перетворюється на одну із потужних та важливих галузей індустрії інформації, що оснащується найновішою комунікаційною технікою, нетрадиційними носіями інформації, високоефективними автоматизованими технологіями її обробки і використання [1].

Постановка проблеми

Сучасна бібліотека є 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 є 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].

Початок процесу створення електронного каталогу належить до 1960-1962 р.р., коли декілька середніх і малих бібліотек США, переважно в навчальних закладах, незалежно одна від однієї почали розробкляти систем машиночитних каталогів і засобів доступу до них. У 1963р. з прийняттям програми MARC (Machine-Readable Cataloging, «машиночитна каталогізація») в Бібліотеці Конгресу США цей процес набув впорядкованого і направленого характеру.

Аналіз останніх досліджень та публікацій

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. Гиляревського, 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. Воройського, Б.С. Елепова, Н.Е. Каленова, О.А. Лавренової, Я.Л. Шрайберга [2].

Формулювання цілей статті та актуальність досліджень

Сьогодні 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. Однак аналіз описових можливостей переважної більшості перелічених АБІС показав, що в них відсутні ефективні засоби верифікації інформації в електронних каталогах. Тому проблема пошуку та виправлення помилок у бібліографічних записах електронного каталогу є доволі актуальною.

Виклад основного матеріалу

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

Використовуючи поняття метаінформації як інформації про інформацію, введене Ю.А. Шрейдером [3], можемо трактувати бібліографічний запис як метаінформацію. Отже, з повною підставою можемо визначити електронний каталог як метаінформаційну систему. З іншого боку, не слід змішувати поняття "метаінформація" і "метадані". Дані, за допомогою яких можуть бути описані об'єкти даних, називають метадані, що означає "дані про дані". Метадані слугують для ідентифікації, визначення та опису характеристик даних. Як приклад описових метаданих можна навести бібліографічні формати, що визначають правила створення машиночитного бібліографічного запису [1].

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).

Запис у форматі MARC (рис. 1) починається з маркера (24 символи). Потім йде довідник, який складається з декількох статей, довжиною 12 символів (по кількості полів в запису). Перші три символи в статті – номер поля. В кінці довідника ставиться символ – роздільник полів. Далі записується зміст полів, розділених на підполя. Після кожного поля ставиться роздільник полів. В кінці запису ставиться символ – роздільник записів.

Дамо логічну інтерпретацію схеми MARC формату. Для створення логічної моделі будемо використовувати математичний апарат теорії множин [4,5].

Електронний каталог C визначається як множина бібліографічних записів:

$$C = \left\{ B_k \,\middle|\, k \in N \right\}, \tag{1}$$

де N – множина натуральних чисел. Кількість бібліографічних записів в електронному каталозі можна визначити, як потужність множини C.

Своєю чергою, бібліографічний запис також визначається як множина полів, що входять в конкретний бібліографічний запис:

$$B = \left\{ H_j \,\middle|\, j \in N \right\}, \tag{2}$$

причому кожне поле представляє собою також множину H для повторів цього самого поля:

$$H = \left\{ A_i \,\middle|\, i \in N \right\}, \tag{3}$$

де A – поле бібліографічного запису, яке представляє собою також множину атомарних текстових стрічок, що називаються підполями, в які заноситься безпосередньо інформація про бібліографічний опис документа:

$$A = \{ a_i \mid a_i \in S \}, \tag{4}$$

де S – множина цифрових та буквених символів відповідної таблиці UNICODE [5].

```
Маркер *****nam#a22*****1i#4500
001      15275
003      RuMoRGB
005      20001225101010.0
008      001220s1999####ru#####gr#####000#0#rus#d
017  ###≠a99-4336≠bРКП
020  ###≠a585251053X
040  ###≠aRuMoRGB
041  1##≠arus≠ager≠hger
084  ###≠aЧ33(4Ш)6-8Штейнер Р.я77-2≠2rubbk
084  ###≠aЧ33(0)6-я77-2≠2rubbk
084  ###≠aЮ3(4Ш)6-6768Штейнер Р.я77-2≠2rubbk
100  1#≠aШтайнер, Рудольф.
245  00≠aОбщее учение о человеке как основа педагогики =≠bAllgemine menschenkude als grundlage der
paradodic : Учеб. курс лекций для преподавателей Свобод. вальдорф. шк., прочит. 21.VIII-5.IX 1919 г. в
Штудгарте /≠cРудольф Штайнер ; Пер. с нем. Д. М. Виноградова.
246  31≠aAllgemine menschenkude als grundlage der paradodic
250  ###≠a2-е изд., доп. нем. текстом.
260  ###≠aМ. :≠bПарсифаль,≠c1999.
300  ###≠a399 с. ;≠c20 см.
500  ###≠aДругая форма имени автора: Штейнер, Рудольф.
546  ###≠aТекст парал. нем., рус.≠bкирилл., латин.
546  ###≠aДанные тит. л. частично парал. нем.≠bлатин.
700  1##≠aВиноградов, Д. М.≠4trl
600  17≠aШтейнер, Рудольф ≠c(философ ;≠d1861-1925)≠2<код системы предметизации>
650  #7≠aНародное образование≠zШвейцария≠xИстория≠y20 в.≠xПедагогические взгляды≠vЛекции для
повышения квалификации≠2<код системы предметизации>
650  #7≠aНародное образование≠zСтраны мира≠xИстория≠y20 в.≠xСистемы, школы,
направления≠xВальдорфская педагогика≠vЛекции для повышения квалификации≠2<код системы
предметизации>
650  #7≠aИстория философии≠zШвейцария укон. 19-20 вв.≠xАнтропософия≠vЛекции для повышения
квалификации≠2<код системы предметизации>
852  4##≠aРГБ bФБ≠j2:99-2/214-5
852  4##≠aРГБ bФБ≠j2:99-2/215-3
```

Рис. 1. Приклад бібліографічного запису у форматі MARC21

```
Штайнер, Рудольф.
   Общее учение о человеке как основа педагогики = Allgemine menschenkude als grundlage der paradodik :
Учеб. курс лекций для преподавателей Свобод. вальдорф. шк., прочит. 21.VIII-5.IX 1919 г. в Штудгарте /
Рудольф Штайнер ; Пер. с нем. Д. М. Виноградова. - 2-е изд., доп. нем. текстом. - М. : Парсифаль, 1999. - 399
с. ; 20 см. - Другая форма имени авт.: Штейнер, Рудольф. - Текст парал. нем., рус. - Данные тит. л. частично
парал. нем. - ISBN 5-85251-053-X.
   99-4336
2:99-2/214-5        2:99-2/215-3        ББК      Ч33(4Ш)6-8Штейнер Р.я77-2
                    Ч33(0)6-я77-2
                    Ю3(4Ш)6-6768Штейнер Р.я77-2
```

Рис. 2. Приклад виводу даних у вигляді традиційної каталожної форми

Система Dublin Core розроблялась з 1995 року співробітниками Online Computer Library Center (OCLC). Сьогодні розроблено набір метаданих Dublin Core, що складається з п'ятнадцяти елементів (табл.1).

Відповідно до рекомендацій, всі елементи Dublin Core, описані в таблиці, розбиваються на три групи (табл.2):

• елементи, що відносяться до змісту ресурсу (Content);

• елементи, що описують цифровий ресурс з точки зору інтелектуальної власності (Intellectual Propery);

• елементи, що відносяться до конкретного екземпляру ресурсу (Instanitation).

Очевидно, що на основі DublinCore не можна створити повноцінного бібліографічного опису. З іншого боку, перевагою є простота опису. Бібліографічний формат сімейства MARC не замінити DublinCore, але каталог електронної бібліотеки за його допомогою може створити і звичайний користувач, а не тільки кваліфікований бібліотекар [1].

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

Особливістю електронного каталогу як метаінформаційної системи є надмірність метаінформації, що містить опис документа, яка ддає змогу однозначно ідентифікувати об'єкт опису за підмножиною ознак (полів). Крім того, перестановка символів в полі запису (наприклад, в заголовних даних та інших текстових полях) не змінює її інформативності. Тобто, якщо запис знайдено, то розпізнати помилки, прочитати запис, зрозуміти сенс і знайти відповідний документ можливо. Цього не можна зробити в базах даних, що містять числові поля. У них перестановка цифр змінює цілком зміст даних, і помилку не можна виявити візуально. Тобто, визначення електронного каталогу як метаінформаційної системи дає змогу зробити такі висновки [1]:

• загальні принципи організації електронного каталогу не повинні залежати від його програмно-апаратної реалізації;

• модель даних електронного каталогу повинна бути настільки абстрактною, щоб дозволити не змінювати її при різних способах доступу до електронних каталогу;

• засоби введення інформації в електронний каталог повинні мінімізувати кількість помилок введення і мати можливість для їх розпізнання та корекції;

• електронний каталог повинен містити, крім засобів введення інформації, засоби оперативного контролю та обробки даних;

• електронний каталог повинен містити засоби автоматизованого та автоматичного індексування;

• електронний каталог повинен містити пошукові засоби, що враховують наявність помилок різного походження та виду в записах.

Ігнорування цих вимог призводить до створення електронного каталогу, що не задовольняє повною мірою користувача. Як правило, це виражається у звуженні пошукових можливостей, призводить до великої кількості помилок і, як наслідок, до втрати інформації.

При інтенсивній роботі з інформаційною системою виникає безліч помилкових даних різного рівня, які можуть вплинути не тільки на конкретну транзакцію, але і на всю інформаційну систему. У загальному випадку помилки роботи автоматизованої системи можна класифікувати як [6]:

• Модульні помилки. Зустрічаються при неправильному погоджені модулів автоматизованих систем. Впливають на взаємодію окремих блоків, а також на взаємодію з іншими аналогічними системами при обміні даними. Наприклад, неправильне написання програми перенесення записів з однієї підсистеми в іншу або помилки конвертації. У цьому випадку виникають помилки систематичного характеру, які проявляються за низкою ознак. Такі помилки присутні у всіх або в багатьох записах системи.

• Функціональні помилки. З'являються при неправильному описі або розумінні технологічних етапів, ланцюжків. Впливають на працездатність системи, особливо в логічному плані при аналізі даних. Наприклад, при неправильному розумінні технологічного процесу виникають помилки заповнення полів записів і настроювання алгоритмів роботи. Такі помилки також мають систематичний характер.

• Додаткові помилки. Являють собою помилки, що виникають при автоматичному генеруванні даних інформаційної системи з помилковими початковими даними. Впливають на всі документи, джерелами яких були помилкові документи. Наприклад, неправильна генерація аналітичних описів при помилковому описі номера періодичного видання.

Таблиця 19

Елементи Дублінського ядра

Title (Заголовок)	назва, присвоєна ресурсу автором або видавцем
Creator (Автор)	юридичні чи фізичні суб'єкти, відповідальні за зміст ресурсу (первинна інтелектуальна відповідальність)
Subject (Предмет)	предметна область ресурсу
Description (Опис)	текстовий опис змісту ресурсу (реферат, анотація і т.д.)
Publisher (Видавець)	юридичний чи фізичний суб'єкт, відповідальний за створення ресурсу
Contributor(Учасники створення)	юридичні чи фізичні суб'єкти, що не являються авторами але зробили значний інтелектуальний внесок в створення ресурсу (вторинна інтелектуальна відповідальність)
Date (Дата)	дата створення чи публікації ресурсу (у доступному вигляді)
Type (Тип)	категорія ресурсу, наприклад, жанр
Format (Формат)	формат представлення даних ресурсу (тип програмного забезпечення, тип комп'ютера, інших приладів, які можуть бути потрібні для відображення і роботи ресурсу)
Identifier (Ідентифікатор)	набір букв або цифр, який зазвичай використовують для унікальної ідентифікації ресурсу, наприклад ISBN, URL і т.д.
Source (Джерело)	інформація про вторинне джерело із якого було отримано ресурс
Language (Мова)	мова, на якій викладено зміст ресурсу
Relation (Зв'язок)	ідентифікатор вторинного ресурсу і його зв'язок з даним ресурсом, наприклад видавництво книги, глава книги
Coverage (Покриття)	характеристика місця розташування і часу доступності ресурсу
Rights (Права)	затвердження авторських прав та управління ними

Таблиця 2

Групування елементів Дублінського ядра

Content	Intellectual Propery	Instanitation
Title (Заголовок)	Creator (Автор)	Date (Дата)
Subject (Предмет)	Publisher (Видавець)	Format (Формат)
Description (Опис)	Contributor(Учасники створення)	Identifier (Ідентифікатор)
Type (Тип)	Rights (Права)	Language (Мова)
Source (Джерело)		
Relation (Зв'язок)		
Coverage (Покриття)		

• Помилки, що призводять до роботи системи в екстремальних ситуаціях. Цей тип помилок виникає при появі в системі записів і даних, які близько або перевершують фізичні обмеження чи

цілісні обмеження даних. Наприклад, фізичні обмеження на обсяг запису 99 кБайт і 999 (кількість) міток полів при передачі бібліографічних описів документів у форматі ISO-2709, на якому засновані всі формати сімейства MARC.

• Помилки, що виникають при додатковому навантаженні на систему. Автоматизована система може змінювати дані відповідно до свого стану. Можливі випадки появи помилкових даних при фізичних порушеннях апаратних і програмних засобів, а також при роботі всього програмно-апаратного комплексу на граничній потужності.

• Помилки, що виникають при зміні продуктивності. Такі помилки з'являються при нерівномірному або імпульсному навантаженні і спричинені «людським чинником», а саме тим, що люди втрачають знання і навички, якщо деякий час знаходяться «без діла».

Поповнення електронного каталогу бібліографічними записами відбувається різними способами (рис. 3). Потрібно зазначити, що основна частина нової бібліографічної інформації вводиться за допомогою клавіатури ЕОМ і дані, отримані з зовнішніх джерел, також введені з клавіатури. Цей факт необхідно враховувати, оцінюючи 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].

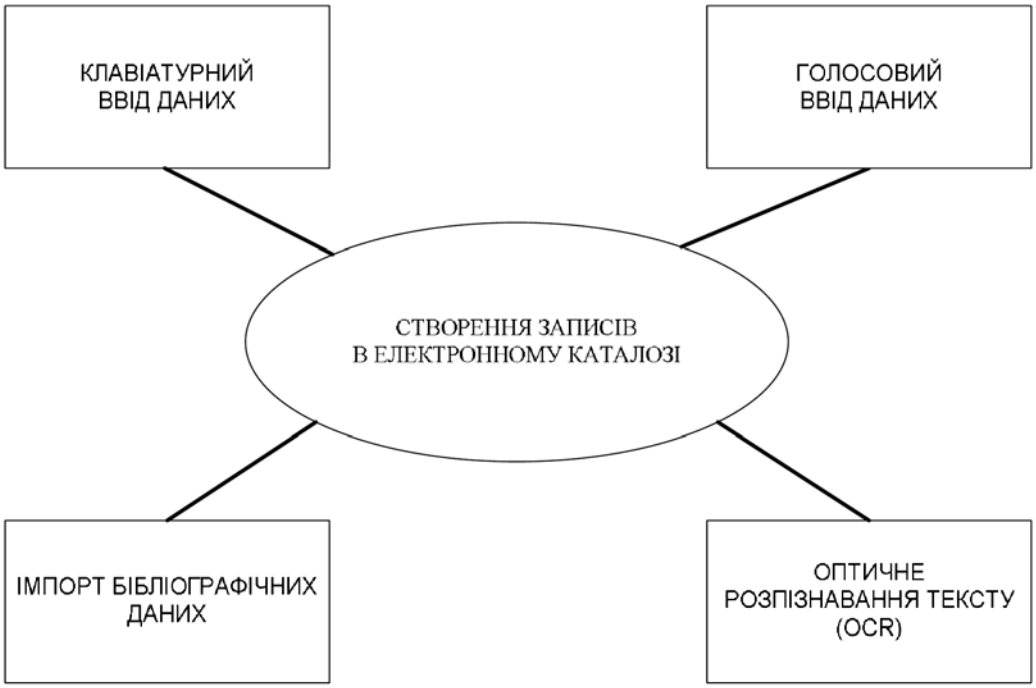

Рис. 3. Засоби створення записів в електронному каталозі

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 основні 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 (рис. 3) в електронному каталозі можна визначити основні типи та джерела помилок, що притаманні кожному із способів.

Клавіатурний ввід даних. З моменту появи прототипів сучасних електронних каталогів і по нинішній час клавіатурний ввід інформації являється основним способом створення електронних

каталогів. В ролі операторів виступає комплектатор, каталогізатор, систематизатор або звичайний робітник бібліотеки. Вдосконалення цього процесу зводиться до наступного [1]:

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

Не зважаючи на це, помилки при ручному введенні даних неминучі [7].

У роботах різних авторів наводяться класифікації можливих помилок під час введення даних. Об'єднання та узагальнення результатів цих робіт дають змогу запропонувати таку класифікацію помилок [1]:

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

2. Помилки структури запису: неправильні теги полів або 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 перестановка сусідніх символів.

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

Основною причиною помилок каталогізації є неправильне трактування правил каталогізації. Зниження числа помилок каталогізації можна добитись навчанням персоналу, розміщенням на екрані додаткової інформації за правилами каталогізації з прикладами. Використання шаблонів типових записів також дають змогу знизити рівень таких помилок [1].

Причиною помилок структури запису переважно є погано прорблений інтерфейс введення даних, що дає змогу в явному вигляді змінювати теги та мітки полів і підполів. Використовуючи ретельно розроблені формати введення (робочі листи), що давали б змогу вводити дані без зазначення у явному вигляді структурної інформації, також дали змогу знизити рівень даного типу помилок [1].

Очевидно, що помилки структури підполя та пунктуаційні помилки не впливають на пошук записів. Щодо символьних помилок, то їх частково можна усунути використовуючи словники типових фрагментів та авторитетні файли. У цьому випадку оператор вибирає дані із запропонованого списку варіантів [1].

Орфографічні помилки найважче усунути, і задачу їх виправлення остаточно не розв'язано. Методом, заснованим на використанні списку типових помилок (список Балларда [8]), також не можна вирішити цю проблему [1].

У статті Р. Нільсон виділив основні причини помилок в електронному каталозі [7]:

- різниця в трактовці стандартів як операторами, так і бібліотекарями;
- збереження доступу до всіх полів електронного каталогу, що призводить до ненавмисної зміни даних;
- неналежне володіння персоналом іноземними мовами (мається на увазі процес ручного індексування електронного каталогу за заголовком).

Ці причини також актуальні і для вітчизняних бібліотек. Можна також ще додати і причину слабкого володіння термінологією предметної галузі видання, що підлягає каталогізації [1].

Оптичне розпізнавання тексту (OCR). Оптичне розпізнавання тексту використовується при ретроспективній конверсії. Ретроспективною конверсією називається технологічний процес переведення традиційних каталогів (карткових, книжкових тощо) в цифрову форму. Основні задачі, що виникають при розпізнаванні інформації, що зберігається на каталожних карточках і в книжних каталогах, такі [1]:

- власне розпізнавання символів і лексичних одиниць (ЛО);

- структуризація інформації, тобто виділення даних, що відповідають полям бібліографічного формату.

Перша задача розв'язується використанням словників для вибору ЛО-кандидатів при візуальному контролі з боку оператора.

Розв'язання другої задачі потребує створення складних процедур семантичного аналізу інформації, враховуючи особливості правил каталогізації у різні періоди часу і активного втручання оператора.

Окрім помилок, характерних для клавіатурного введення, з'являються специфічні, які залежать від таких факторів [1]:

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

Помилки класифікують за типами так:

1. Заміна символу на подібний за написанням (І=1,О=0).
2. Заміна сусідніх символів на один (І І=П, М И=Ш).
3. Заміна одного символу на два (Ю=І О,Ж=} К).

Якщо з'являється неправильне слово (слово, яке відсутнє у словнику), система розпізнавання реагує адекватно, позначаючи його. При заміні правильного слова на правильне помилку може виправити лише оператор при візуальному контролі [1].

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

Імпорт бібліографічних даних. Цей 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 розроблення інструментарію очистки даних при імпорті бібліографічних даних є дуже важливою проблемою. Також до помилок, які вже присутні у даних, що імпортуються, додаються помилки конвертації із одного формату подання даних в інший.

Ще одним проблемним питанням є імпорт непідготовлених (неформатних) бібліографічних описів. Наприклад, імпорт даних з бази книгодрукарень або приватних колекцій. При цьому формат подання даних може бути абсолютно довільним. Завдання полягає у знаходженні і відокремленні з бази даних довільної форми представлення саме бібліографічних описів та форматування їх у відповідні формати.

Також зрозуміло, що при імпорті великої кількості бібліографічних записів виникає проблема дублювання інформації.

Оскільки помилки при імпорті даних мають технічний та системний характер, то можливо розробити інструментарії знаходження та автоматичної корекції помилок імпорту даних.

Голосове введення даних. Зростання потужності комп'ютерів сприяло активному розвитку і такого порівняно нового способу введення даних, як голосове введення. Побудова голосового інтерфейсу поділяється на три завдання [1].

Перше завдання полягає в тому, щоб комп'ютер міг "зрозуміти" те, що йому каже людина, тобто він повинен уміти витягати з промови людини корисну інформацію. Поки що, на нинішньому етапі, ця задача зводиться до того, щоб витягти з промови смислову частину, текст (розуміння таких складових, як, скажімо, інтонація, поки взагалі не розглядається). Отже, ця задача зводиться до заміни клавіатури на мікрофон [1].

Друге завдання полягає в тому, щоб комп'ютер сприйняв сенс сказаного. Якщо мовне повідомлення складається зі стандартного набору зрозумілих комп'ютеру команд (скажімо, дублюючих пункти меню), нічого складного в її реалізації немає. Однак навряд такий підхід буде зручніший, ніж введення цих самих команд з клавіатури або за допомогою "миші". В ідеалі комп'ютер повинен чітко "осмислювати" природну мову людини [1].

Третє завдання полягає в тому, щоб комп'ютер міг перетворити інформацію, з якою він оперує, на мовне повідомлення, зрозуміле людині. Поки остаточне рішення існує тільки для третьої задачі. ПО суті, синтез мови – це суто математична задача, яку сьогодні на доволі високому рівні [1].

Перешкодою для остаточного вирішення першого завдання є те, що досі точно невідомо, як можна розчленувати нашу мову, щоб витягти з неї складові, у яких міститься сенс. У звуковому потоці мовлення не можна розрізнити ні окремих букв, ні складів: навіть, здавалося б, однакові букви і склади в різних словах на спектрограмах виглядають по-різному [1].

Друге завдання, на думку більшості фахівців, не може бути вирішене без допомоги систем штучного інтелекту. Тому поки частка мовного інтерфейсу – всього лише дублювання голосом команд, які можуть бути введені з клавіатури або за допомогою миші, але тут його переваги сумнівні. Є ще одна область застосування, яка може бути дуже привабливою. Це голосове введення текстів у комп'ютер. У цьому випадку не потрібно, щоб комп'ютер осмислював почуте, а завдання перекладу мови в текст більш-менш вирішене [1].

При введенні бібліографічного запису на семантичному рівні виникає проблема, яка полягає в тому, що дані повинні бути розпізнані й структуровані відповідно до вимог бібліографічного формату, і помилки, які виникають при цьому, важко усунути. Отже, застосування мовного інтерфейсу для масового введення бібліографічного запису є завданням майбутнього [1].

Отже, основними проблемами та причинами помилок у електронному каталозі є:

- відсутність належної спеціалізації у операторів, незнання іноземної мови, неправильне розуміння термінології, недостатнє володіння комп'ютерною технікою;
- введення даних без оригіналу екземпляру літературного джерела або за його електронною копією, в якій можуть бути відсутні титульні сторінки;
- недосконалість програмного інтерфейсу введення;
- 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 так [6]:

- 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. Найчастіший 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 [6]:

1. Автоматичне виправлення помилок, які часто зустрічаються. Помилки в даному випадку виправляє програмне забезпечення для автоматизації або зовнішнє програмне забезпечення.

2. Ручне виправлення помилок. У міру знаходження помилок про них повідомляється оператор або бібліотекар, який їх виправляє.

3. Напівавтоматичне виправлення помилок. У цьому 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.

Отже, якісний процес верифікації інформації в електронному каталозі не можливий без аналізу типів, механіки, джерел та причин виникнення помилок в записах електронного каталогу.

Висновки

Електронний каталог представлено як метаінформаційну систему. Розглянуто системи представлення бібліографічних даних у машиночитній формі, такі як MARC та DublinCore. Для математичного опису логічного представлення формату MARC було використано апарат теорії

множин. Дано приклад представлення даних у форматі MARC. Представлено опис та групування полів формату метаданих DublinCore. Вказано основні засоби та методи створення записів в електронному каталозі. Розглянуто та здійснено класифікацію помилок, що виникають при ручному (клавіатурному) введенні інформації, голосовому введенні, імпорті даних та при оптичному розпізнаванні текстів. Докладно проаналізована можливість використання методів голосового введення даних. Представлено основні причини виникнення даних помилок. Задані узагальнені принципи пошуку помилок у записах електронного каталогу та методи їх усунення.

1. Вершинин М. И. Электронный каталог проблемы и решения / М. И. Вершинин. – СПб. : ПРОФЕССИЯ, 2007. – 233с. 2. Вершинин М.И. Проблемы создания электронного каталога в научных библиотеках: дис. канд. пед. наук: 05.25.03 / Вершинин Михаил Иосифович. – СПб., 2002. – 221с. 3. Шрейдер Ю.А. Информация и метаинформация / Ю.А. Шрейдер // НТИ. Информ. процессы и системы – 1982. – Сер. 2 №4. – С.3–10. 4.Крауш А.С. Модель корпоративного создания и тиражирования электронных каталогов библиотек: дис. канд. тех. наук: 05.25.05 / Крауш Александр Сергеевич. – Новосибирск, 2004. – 157с. 5. Шрайберг Я.Л. Руководство по UNIMARC / Я.Л. Шрайберг, А.И. Земсков – М.: ГПНТБ, 1992. – 319с. 6. Крауш А. С. Утилиты для проверки и коррекции электронных каталогов / А. С. Крауш, Д. Ю. Копытков, А. С. Макаревич // Библиотечное дело. – 2005. – № 6 (30). – С. 21 – 24. 7. Nielsen R. Lost articles: Filing problems with initial articles in data bases / R. Nielsen, J. M. Pyle // Libr. Resources a. techn. Services. – 1995. – Vol. 39, №3. – P. 291 – 293. 8. Ballard T. Spelling and typographical errors in library databases: one libr. System for noting out spelling error / T. Ballard // Computer in libr. – 1992 – Vol. 12, №6. – P. 14 – 19.