

УДК 004.052.42:01

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ПОЛЯ «КІЛЬКІСТЬ СТОРІНОК» В ПРОЦЕСІ ІНТЕГРАЦІЇ ГЕТЕРОГЕННИХ БІБЛІОГРАФІЧНИХ ДАНИХ

DOI 10.36994/2788-5518-2021-01-01-14¹⁴

Петренко М. В., Відкритий міжнародний університет розвитку людини «Україна», mxms@ukr.net

Анотація. Стаття присвячена дослідженню особливостей поля даних «кількість сторінок» в контексті використання його в процесі інтеграції бібліографічних даних з різних джерел в єдиний зведений каталог. Увага приділяється можливості використання даних з цього поля в задачі пошуку дублікатів записів з метою їх об'єднання та підвищення якості бази даних.

Сьогодні в Україні актуальні процеси цифрової трансформації бібліотек та створення зведених електронних каталогів. Найважливішою проблемою зведених каталогів є наявність дублікатів. Дослідження реальних даних наших бібліотек та пошук нових методів їх інтеграції та покращення якості створених на їх основі зведених каталогів в цьому контексті є важливим.

Основою для дослідження стали бібліографічні дані 141 бібліотеки м. Києва. Дані були спеціально зібрані для аналізу. Наявність 378742 бібліографічних записів з різних джерел та представлені в різних форматах (в спеціалізованому машиночитаному MARC-форматі та у форматі web-сторінок каталогу) дозволила моделювати умови створення зведеної бази.

Використання друкарських стандартів дозволило знайти обґрунтування знайденим під час аналізу особливостям.

Описані результати дослідження особливостей інформації, що зберігається у полі «кількість сторінок». Висвітлені та пояснені знайдені закономірності та якість заповнення цього поля. Запропоновано практичне використання результатів проведеного дослідження у вигляді методу виявлення нечітких співпадінь у гетерогенних бібліографічних даних.

Використання результатів дослідження в цілому та даного методу зокрема дозволить більш широко використовувати поле даних, якому присвячене дослідження, не зважаючи на відсутність єдиного стандарту його заповнення. В результаті такого використання з'явиться можливість підвищити якість інтеграції гетерогенних бібліографічних даних, а отже зросте якість зведених електронних каталогів.

Ключові слова: нечіткі співпадіння, бібліографічні дані, дедублікація, кількість сторінок.

¹⁴ Петренко М. В.

PECULIARITIES OF USING THE "NUMBER OF PAGES" FIELD IN THE PROCESS OF INTEGRATION OF HETEROGENEOUS BIBLIOGRAPHY DATA

Petrenko M.V., Open International University of Human Development "Ukraine",
mxms@ukr.net

Abstract. The article is devoted to the research of the features of the data field "number of pages" in the context of its use in the process of integrating bibliographic data from different sources into a single consolidated catalog. Attention is paid to the possibility of using data from this field in the task of finding duplicate records in order to combine them and improve the quality of the database.

Today in Ukraine the processes of digital transformation of libraries and creation of consolidated electronic catalogs are relevant. The most important problem of consolidated directories is the presence of duplicates. Researching the real data of our libraries and finding new methods of their integration and improving the quality of the consolidated catalogs created on their basis is important in this context.

The basis for the study were bibliographic data of 141 libraries in Kyiv. Data were specially collected for analysis. The presence of 378742 bibliographic records from different sources and presented in different formats (in a specialized machine-readable MARC-format and in the format of web-pages of the catalog) allowed to model the conditions for creating a consolidated database.

The use of printing standards allowed to find a justification for the features found during the analysis.

The results of the study of the features of the information stored in the field "number of pages" are described. The found regularities and the quality of filling in this field are highlighted and explained. The practical use of the results of the study in the form of a method for detecting fuzzy matches in heterogeneous bibliographic data is proposed.

The use of the results of the study in general and this method in particular will allow a wider use of the data field to which the study is devoted, despite the lack of a single standard for its completion. As a result of this use, it will be possible to improve the quality of integration of heterogeneous bibliographic data, and thus increase the quality of consolidated electronic catalogs.

Keywords: fuzzy matches, bibliographic data, deduplication, number of pages.

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ ПОЛЯ «КОЛИЧЕСТВО СТРАНИЦ» В ПРОЦЕССЕ ИНТЕГРАЦИИ ГЕТЕРОГЕННЫХ БИБЛИОГРАФИЧЕСКИХ ДАННЫХ

Петренко М. В., Открытый международный университет развития человека
«Украина», mxms@ukr.net

Аннотация. Статья посвящена исследованию особенностей поля данных «количество страниц» в контексте использования его в процессе интеграции библиографических данных из различных источников в единый сводный каталог. Внимание уделяется возможности использования данных с этого поля в задаче поиска дубликатов записей с целью их объединения и повышения качества базы данных.

Сегодня в Украине актуальные процессы цифровой трансформации библиотек и создание сводных электронных каталогов. Важнейшей проблемой сводных каталогов является наличие дубликатов. Исследование реальных данных

наших библиотек и поиск новых методов их интеграции и улучшения качества созданных на их основе сводных каталогов в этом контексте важно.

Основой для исследования стали библиографические данные 141 библиотеки г. Киев. Данные были специально собраны для анализа. Наличие 378742 библиографических записей из различных источников и представлены в различных форматах (в специализированном машиночитаемом MARC-формате и в формате web-страниц каталога) позволила смоделировать условия создания сводной базы.

Использование печатных стандартов позволило найти обоснование найденным при анализе особенностям.

Описанные результаты исследования особенностей информации, хранящейся в поле «количество страниц». Освещены и объяснены найдены закономерности и качество заполнения этого поля. Предложено практическое использование результатов проведенного исследования в виде метода выявления нечетких совпадений в гетерогенных библиографических данных.

Использование результатов исследования в целом и данного метода в том числе позволит более широко использовать поле данных, которому посвящено исследование, несмотря на отсутствие единого стандарта его заполнения. В результате такого использования появится возможность повысить качество интеграции гетерогенных библиографических данных, следовательно, возрастет качество возведенных электронных каталогов.

Ключевые слова: нечеткое совпадение, библиографические данные, дедубликация, количество страниц.

Вступ

В Україні та світі актуальним є питання створення зведених електронних каталогів бібліотек[1]. Їх створення можливе шляхом інтеграції гетерогенних бібліографічних даних.

Одним з питань, які виникають під час такої інтеграції, є питання об'єднання записів, що стосуються одного і того ж об'єкта (книги). У дослідженні[2] проблема дублікатів виявилася найбільш актуальною для користувачів невирішеною проблемою зведених електронних каталогів.

Одним з параметрів, які можуть допомогти у вирішенні проблеми дублікатів, є кількість сторінок. Кількість сторінок не є основним параметром для ідентифікації, але може застосовуватися для додаткового підтвердження ймовірності того, що два бібліографічних записи можуть стосуватися одного і того ж твору. Для такого застосування достатньо наближено знати кількість сторінок. В цьому випадку кількість сторінок виступає опосередкованим фізичний виявом об'єму інформації.

Зовсім інша справа, якщо нам необхідно використати кількість сторінок для автоматичного виявлення повних дублів книг (книг одного видання). В ідеальному випадку необхідно мати точне співпадіння по кількості сторінок.

Аналіз проблеми

Використання кількості сторінок для ідентифікації видання може бути або додатковим елементом, або і основним, за умови відсутності інформації про

ISBN, сумнівів у її достовірності, відсутності інформації про видавництво та рік видання.

У дослідженні використовувалися 378742 бібліографічні записи двох різних каталогів публічних бібліотек м. Києва[3,4]. Разом вони містять дані 141-ї публічної бібліотеки м. Києва. Дані отримані з різних джерел, які відрізняються за підходом до введення даних та їх форматом.

Дослідження параметру ISBN показало:

- Параметр відсутній у 34% бібліографічних записів
- Існують книги з однаковим ISBN, які мають:
 - різні назви;
 - різних авторів;
 - різні роки випуску;
 - суттєво різну кількість сторінок.

Неможливість покладатися на ISBN додатково підкреслює важливість використання такого параметра, як кількість сторінок.

У ході дослідження бібліографічних записів книг, що за усіма ознаками належать до одного видання, було помічено, що у деяких книг дані про кількість сторінок можуть відрізнятися на кілька одиниць. Відмінності трапляються навіть в межах однієї бази даних, хоча більш характерно це все ж для порівняння даних з різних баз.

Консультації з фахівцями підтвердили гіпотезу про можливість застосування бібліографами різних підходів до визначення кількості сторінок в книзі. Бібліограф може ввести:

- кількість сторінок, що вказана в описі, що знаходиться в книзі (за наявності);
- номер останньої пронумерованої сторінки;
- номер останньої пронумерованої сторінки + кількість сторінок після неї.

Після приведення отриманих даних до числових значень маємо ситуацію, описану в таблиці 1.

Таблиця 1.
Оцінка структури даних

	База 1	База 2
Не відповідає формату	515 (0,5%)	2333 (0,8%)
Унікальних значень	1055 (1%)	1373 (0,5%)
Дублів	100537	272929

Поглянемо на абсолютні значення цього параметру:

- 1055 унікальних значень (База 1);
- 1373 унікальних значень (База 2).

Незважаючи на різницю в кількості записів в 2,7 рази, кількість унікальних числових значень поля «Сторінки» відрізняється лише в 1,3 рази.

Окрім реально більшої варіативності через більший асортимент книг, тут може працювати фактор різних підходів до реєстрації даних, коли одна і та сама книга реєструється з дуже близькими, але не однаковими значеннями цього поля.

Для розуміння закономірностей даних поглянемо на розподіл кількості записів в залежності від вказаної кількості сторінок на рисунку 1.

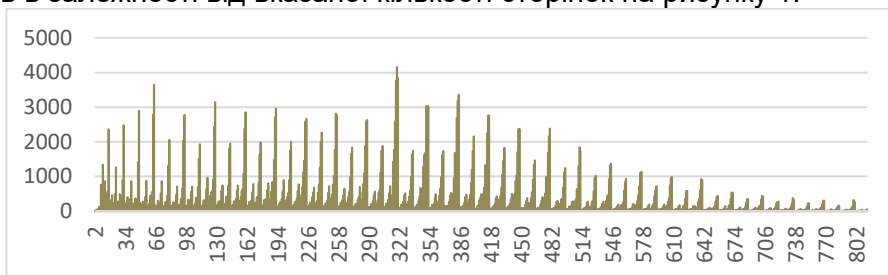


Рис. 1: Розподіл кількості записів в залежності від кількості сторінок

Розподіл дублів демонструє існування закономірностей у вигляді повторюваних шаблонів з екстремумами. Це пов'язано з видавничими стандартами, де комплектація книг іде «зошитами», кожен з яких може мати 8, 16 або 32 аркуша залежно від формату (16, 32 та 64 сторінки відповідно).[5]

За потреби, можливе комплектування блоків неповними зошитами. Неповні зошити можуть мати кількість аркушів кратну 2-ом через особливість збирання зошита (відповідає 4-м сторінкам).

Якщо узяти значення поля «Сторінки» по модулю 16 та просумувати кількість книг, що відповідають відповідним значенням, то отримаємо розподіл на рисунку 2.

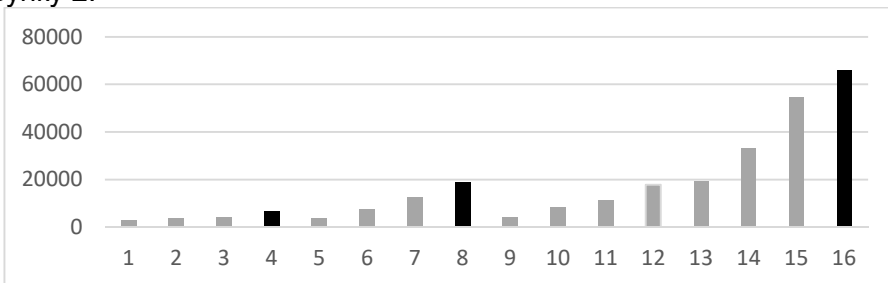


Рис. 2: Розподіл кількості записів в залежності від кількості сторінок по мод. 16

Якщо перейти до більшого масштабу, то на рисунку 3 помітно, що сторінок кратних 64 найбільше, далі йдуть 32, а там вже кратні 16.

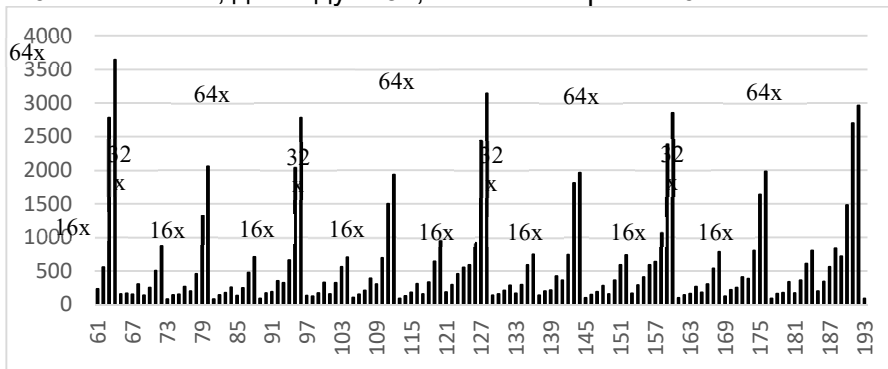


Рис. 3: Розподіл кількості записів в залежності від кількості сторінок (уринок)

Віднайдену особливість, яка має підтвердження і теоретичне і практичне, можна використати з метою очищення поля від помилок інтерпретації бібліографом кількості сторінок. Таким чином ми отримуємо точний інструмент для задач ідентифікації дублів записів книг певного видання.

Пропоноване рішення

Проаналізувавши реальні випадки введення різних даних щодо кількості сторінок книги одного видання, маємо:

- максимальна різниця між введеним числом та найближчим достовірним – 13 сторінок;
- максимальна різниця між двома введеними бібліографами значеннями – 10 сторінок;
- медіана різниці – 5 сторінок.

Розподіл достовірної кількості сторінок по кратності на основі проаналізованих дублів видно на рисунку 4.

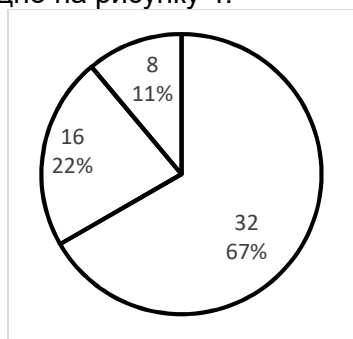


Рис. 4: Розподіл достовірної кількості сторінок по кратності

З урахуванням того, що максимальна різниця складає 13 сторінок, можна прийняти за максимальну кратність - 16. Під такий опис потрапить 89% випадків.

Таким чином реальна кількість сторінок книги може дорівнювати одному з трьох значень:

$$N_{\min}(N) = 4 * (\text{ceil}(N \bmod 4));$$

$$N_{\text{mid}}(N) = 8 * (\text{ceil}(N \bmod 8));$$

$$N_{\max}(N) = 16 * (\text{ceil}(N \bmod 16));$$

де:

N – значення поля «кількість сторінок»;

$\bmod$ – оператор взяття числа по модулю;

ceil – функція, що повертає округлене вгору число.

$N_{\max}$ одночасно є і верхньою межею і найбільш імовірною реальною кількістю сторінок даного видання книги.

Якщо $N_{\max}$ для двох бібліографічних записів співпадає, ймовірність того, що вони належать до одного видання однієї і тієї ж книги є високою. В інакшому випадку вони не належать до одного видання. Таким чином, ми отримуємо

новий метод виявлення нечітких співпадінь, альтернативний простому порівнянню чисел. Наприклад:

183 і 191 – різниця 8, але для обох чисел $N_{\max} = 192$, а отже ймовірність, що такі значення можуть належати одній книзі дуже висока.

191 і 193 – різниця 2, але $N_{\max}(191) = 192$, а $N_{\max}(193) = 208$ вони належать до різних проміжків. Отже ці значення не можуть належати до одного видання книги.

Висновки

В ході дослідження були виявлені та пояснені закономірності інформації в полі «кількість сторінок». Результатом дослідження є метод виявлення нечітких співпадінь, альтернативний простому порівнянню числових значень. Використання запропонованого методу дозволить успішно використовувати дане поле для задач пошуку дублікатів навіть за умов роботи з гетерогенними бібліографічними даними.

Література

1. Бруй О. Не база даних, а пошуковик: як виглядає сучасний Зведений каталог для бібліотек [Електронний ресурс] – Режим доступу до ресурсу: <https://chytomo.com/ne-baza-danykh-a-poshukovyk-iak-vyhladaie-suchasnyj-zvedenyj-kataloh-dlia-bibliotek/>;
2. Online catalogs: what users and librarians want: an OCLC report.: OCLC, 2009. Web. [Електронний ресурс] – Режим доступу до ресурсу: <https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf>;
3. Електронний каталог бібліотеки ім. Лесі Українки (публічні бібліотеки для дорослих м. Києва). [Електронний ресурс] – Режим доступу до ресурсу: <http://ecatalog.kiev.ua>;
4. Електронний каталог бібліотеки ім. Тараса Шевченка для дітей (публічні бібліотеки для дітей м. Києва). [Електронний ресурс] – Режим доступу до ресурсу: <http://zra.kiev.ua:8081/MarcWeb>;
5. Дурняк Б.В. Стандарти в поліграфії та видавничій справі : довідник / Б. В. Дурняк, В. П. Ткаченко, І. Б. Чеботарьова. — Львів : Вид-во Укр. акад. друкарства, 2011. —320 с.

References

1. Bruy, O. (2021). Ne baza danykh, a poshukovyk: jak vyghljadaie suchasnyj zvedenyj katalogh dlja bibliotek. Retrieved from <https://chytomo.com/ne-baza-danykh-a-poshukovyk-iak-vyhladaie-suchasnyj-zvedenyj-kataloh-dlia-bibliotek/>;
2. OCLC. (2009). Online catalogs: What users and librarians want: an OCLC report. Retrieved from <https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf>;
3. Biblioteka im. Lesi Ukrayinky. (n.d.). Elektronnyy katalog biblioteky im. Lesi Ukrayinky. Retrieved from <http://ecatalog.kiev.ua>;
4. Biblioteky im. Tarasa Shevchenka. (n.d.). Elektronnyy katalog biblioteky im. Tarasa Shevchenka dlya ditey. Retrieved from <http://zra.kiev.ua:8081/MarcWeb>;
5. Durnyak, B. V., Tkachenko, V. P., & Chebotar'ova, I. B. (2011). Standarty v polihrafiyi ta vydavnychiy spravi : dovidnyk. L'viv, Ukraine: Vyd-vo Ukr. akad. drukarstva.