УДК 004.052.42:01
DOI https://doi.org/10.32838/2663-5941/2021.4/19

Петренко М.В.
Відкритий міжнародний університет розвитку людини «Україна»

ДОСЛІДЖЕННЯ ОСОБЛИВОСТЕЙ ПОЛЯ «ISBN» ТА ЇХ ВПЛИВУ НА ПРОЦЕС ОБРОБКИ БІБЛІОГРАФІЧНИХ ДАНИХ

Стаття присвячена розгляду міжнародного книжкового номера ISBN як складової частини бібліографічного запису та бібліографічної бази даних у контексті процесів підвищення якості та інтеграції бібліографічних даних.

Нині актуальні процеси створення зведених електронних каталогів бібліотек та покращення якості даних, що в них зберігаються. Для забезпечення достатнього рівня якості зведених каталогів необхідно розробити, обґрунтувати та випробувати методи підготовки та обробки даних.

ISBN є важливою частиною бібліографічного запису та, окрім унікальності, несе у собі закодовану додаткову інформацію. Використання цієї інформації може принести багато користі. Але для цього необхідно розуміти всі нюанси ISBN як теоретичного плану, так і особливостей його застосування в реальних бібліографічних базах даних.

Стаття описує результати проведеного дослідження кількісних та якісних показників поля даних ISBN на основі реальних даних 141-ї публічної бібліотеки м. Києва, отриманих з двох різних джерел.

Відзначені особливості поля, як спільні, так і відмінні для обох бібліографічних баз. Зроблена оцінка процесів відновлення структури ISBN на основі правил. Оцінено потенціал використання поля ISBN у процесах інтеграції гетерогенних бібліографічних даних. Запропоновано стандартизацію ISBN та оцінено ефект від неї. Описано знайдену проблему з правилами ISBN, яка потребує уточнення на рівні агенції ISBN. Оцінено та пояснено динаміку змін поширеності ISBN. Багато уваги приділено дублям ISBN і зв'язку кількості дублів та унікальних записів ISBN з помилками та форматами представлення даних. Проаналізовано помилки ISBN, їх структуру та причини виникнення.

Отримані результати дозволять вибудувати стратегію обробки досліджуваного поля бібліографічних даних з метою досягнення кращих результатів у задачах очистки, стандартизації та інтеграції даних.

Ключові слова: *ISBN, структура ISBN, стандартизація ISBN, бібліографічні дані, дедублікація, помилки введення.*

Постановка проблеми. Натепер актуальним є питання створення зведених електронних каталогів бібліотек. Їх створення можливе шляхом інтеграції гетерогенних бібліографічних даних. Більшість систем електронних каталогів є пошуковиками, а це робить питання якості даних та її підвищення надзвичайно важливими.

Одним з питань, які виникають під час такої інтеграції, є питання об'єднання записів, що стосуються одного і того ж об'єкта. У дослідженні [1, с. 27] зазначено, що проблема видалення дублікатів є найбільш актуальною з невирішених проблем зведених електронних каталогів.

Одним з поширених полів бібліографічних даних є ISBN. Це поле не часто використовується у пошуку користувачами бібліотек, але може бути використане у задачах інтеграції даних, видалення дублів та підвищення якості даних. Але для ефективного використання необхідно дослідити та використовувати особливості ISBN, яким і присвячена стаття.

Аналіз останніх досліджень і публікацій. ISBN – міжнародний стандарт кодування книжок, який забезпечує унікальність коду книги та має контрольне число для контролю відсутності помилок. До 2007 року привласнювали коди з 10 десяткових цифр. З 1 січня 2007-го ISBN був уніфікований з EAN, міжнародною системою кодування товарів у роздрібній торгівлі з 13 десяткових цифр [2, с. 11]. Також з 2008 року уніфіковано було і міжнародний стандарт кодування нотних видань (ISMN) [3, с. 6].

Після уніфікації ISBN та ISMN є повністю сумісними [3, с. 5] і по кодах і по алгоритмах контролю відсутності помилок. ISBN починається з EAN 978 або 979. Коди ISMN починаються з EAN 979, за яким слідує 0.

10-значні коди ISBN можна перетворити на новий формат, додавши перед кодом цифри «978», також необхідно за новим алгоритмом розрахувати контрольну цифру та замінити нею стару.

До 2008 року ISMN мав інший формат (коди починалися з «M-060», але екземплярів кодів, які могли б належати до цього формату у досліджуваних базах знайдено не було).

Коди ISBN та ISMN мають структуру. Старий формат має 4 блоки, а новий – 5 (додався EAN). Блоки цифр розділяються дефісами [2, с. 11].

Структура:

- Код EAN (для нового формату) – 3 цифри (978 або 979);
- Код країни/об'єднання;
- Код видавця;
- Код видання;
- Контрольна цифра.

Незмінну довжину мають лише перший та останній блоки. Розміри інших блоків визначаються згідно з правилами. Правила можуть змінюватися з часом.

Збереження структури коду може допомогти встановити країну видання та видавця.

Один видавець може мати декілька кодів блоку видавця, кожен з відповідним діапазоном кодів, що можуть використовуватися для видань.

Для задач машинної обробки часто прибирають дефіси і зберігають дані не структуровано, що відповідає стандарту MARC 21 [4]. При цьому унікальність номерів зберігається.

В Україні за видачу кодів видавцям відповідає Книжкова палата України. Згідно з її вимогами код без поділу на блоки не є дійсним кодом ISBN [5].

Коди, як і будь-яка бібліографічна інформація, можуть містити помилки введення. Такі помилки можуть з'являтися як на етапі введення даних бібліографом, так і на етапі підготовки до друку.

Стандарт бібліографічних даних MARC визначає такі варіанти недійсності ISBN кодів, як [4]:

1) структурна недійсність:

а) неправильна довжина;

б) неправильна структура;

в) контрольна цифра не відповідає розрахованій за алгоритмом;

2) недійсність застосування:

а) однакові коди, що привласнені різним ресурсам.

У задачах інтеграції даних з різних джерел ISBN може застосовуватись для пошуку дублів через коди книг; виправлення значення поля «Видавець».

Основне та очевидне завдання використання ISBN – можливість ідентифікувати унікальні книги. В ідеальних умовах, коли всі унікальні видання мають унікальні коди, які введені без помилок, питання ідентифікації повних дублів було б вирішеним. Досить було б порівняти коди і згрупувати записи, коди яких однакові. Так, ISBN використовується у багатьох системах обробки бібліографічних даних, хоча це і пов'язано з проблемами [6].

Але пошук дублів книг через співпадіння ISBN має такі проблеми, як:

1) відсутність кодів;

2) структурні помилки;

3) код може бути неунікальним для видання, оскільки:

а) може стосуватися серії книг;

б) може стосуватися набору книг;

в) може бути помилково привласненим видавцем різним виданням;

г) може містити помилки введення.

Також код може бути представлений у 10 цифрах або у 13 цифрах, порівняння яких не дасть точного співпадіння.

Також завдяки властивостям ISBN, а саме наявності в його складі групи, шо має прив'язку до видавця, з'являється можливість використовувати код у завданнях покращення поля «Видавець».

Постановка завдання. Мета статті – розглянути такі актуальні для описаних вище завдань пошуку дублів та виправлення значення поля «Видавець» питання, як:

- наявність ISBN;
- структурна недійсність ISBN;
- збіги ISBN;
- вплив стандартизації ISBN на пошук збігів;
- структура ISBN.

Надалі все, що стосується ISBN-13, стосуватиметься і ISMN-13. Виокремлювати його немає потреби через ідентичність алгоритмів та невелику кількість екземплярів ISMN-13.

У дослідженні використовуються 378742 бібліографічні записи двох різних каталогів публічних бібліотек м. Києва – Каталог бібліотеки ім. Тараса Шевченка для дітей [7] (далі – База 1) та Каталог публічної бібліотеки ім. Лесі Українки [8] (далі – База 2). Разом вони містять дані 141-ї публічної бібліотеки м. Києва.

Виклад основного матеріалу дослідження.

1. Наявність ISBN

Цікаво розглянути динаміку змін наявності номера ISBN у книгах за роком видання (рис. 1) на основі даних з обох баз, які розглядаються в цій роботі.

У колишньому СРСР стандарти використання ISBN регламентувалися і були введені в дію 1 січня 1988 р. [9]. Підтвердження цьому можна спостерігати у вигляді різкого стрибка (рис. 1, точка 1) в наявності номерів у 1988 році.

У 1996 році Україна підписала Угоду про співпрацю з Міжнародним агентством ISBN. У 1997 році вимоги щодо обов'язкового використання ISBN були закріплені у Законі України «Про

видавничу справу» від 05.06.97 р. та було створено Національне агентство ISBN (при Книжковій палаті України) [9]. Відповідне зростання відсотка книг з ISBN теж можна помітити на графіку (рис. 1, точка 2).

Останній стрибок (рис. 1, точка 3) можна пояснити запровадженням 1 січня 2007 року нового формату ISBN, що одночасно і розширив кількість доступних номерів, і став сумісним зі штрихкодами, наявність яких зробило ISBN машинозчитуваним форматом.

Також видно, що з часом актуальність використання цього поля зростає і графік наявності ISBN прямує до 100%.

У розрізі ISBN маємо:

1) 313574 – загальна кількість записів ISBN (унікальних у межах записів книг, до яких належать);

а) База 1 – 69736;

б) База 2 – 243838;

2) 233015 унікальних ISBN у межах двох баз;

3) 51218 записів ISBN мають два чи більше дублі в межах двох баз.

У розрізі книг маємо:

4) 272831 записів книг мають хоча б один ISBN (72% від кількості загальної кількості записів книг);

а) База 1 – 69383 (68% від загальної кількості);

б) База 2 – 203488 (73% від загальної кількості);

5) 35044 записів книг мають більше одного відмінного ISBN (13% від загальної кількості книг з ISBN);

а) База 1 – 345 (0,3% від загальної кількості);

б) База 2 – 34699 (13% від загальної кількості);

6) Найбільша кількість різних ISBN у одного запису книги – 6.

Велика кількість записів книг з більш ніж одним ISBN більш характерна для Бази 2, ніж для Бази 1 (різниця у 43 рази).

Загалом у разі об'єднання записів обох баз виявлено 51218 ISBN, що мають не менше 2 співпадінь записів, що стосуються різних записів книг (19% від книг, що мають ISBN). Максимальна кількість співпадінь – 183.

2. Наявність у ISBN структурних помилок

Наступним важливим параметром є структурна дійсність [4]. Введемо такі класи помилок, як:

- кількість символів ISBN не відповідає формату;
- контрольна цифра вказує на помилку в коді;
- код не відповідає правилам ISBN.

Правила ISBN можна отримати у міжнародній агенції [10], а правила ISMN у інструкції міжнародної агенції ISMN [3].

Кількість записів ISBN з помилками цих видів – 9018 (2,9% від загальної кількості записів ISBN);

База 1 – 2475 (3,5% від загальної кількості);

База 2 – 6543 (2,7% від загальної кількості).

Розподіл за типом помилки:

1) кількість символів ISBN не відповідає формату – 2663 (29,5%):

а) База 1 – 1009 (41% від загальної кількості по базі);

б) База 2 – 1654 (25% від загальної кількості по базі);

2) контрольна цифра вказує на помилку в коді – 5044 (56%):

а) База 1 – 1465 (59% від загальної кількості по базі);

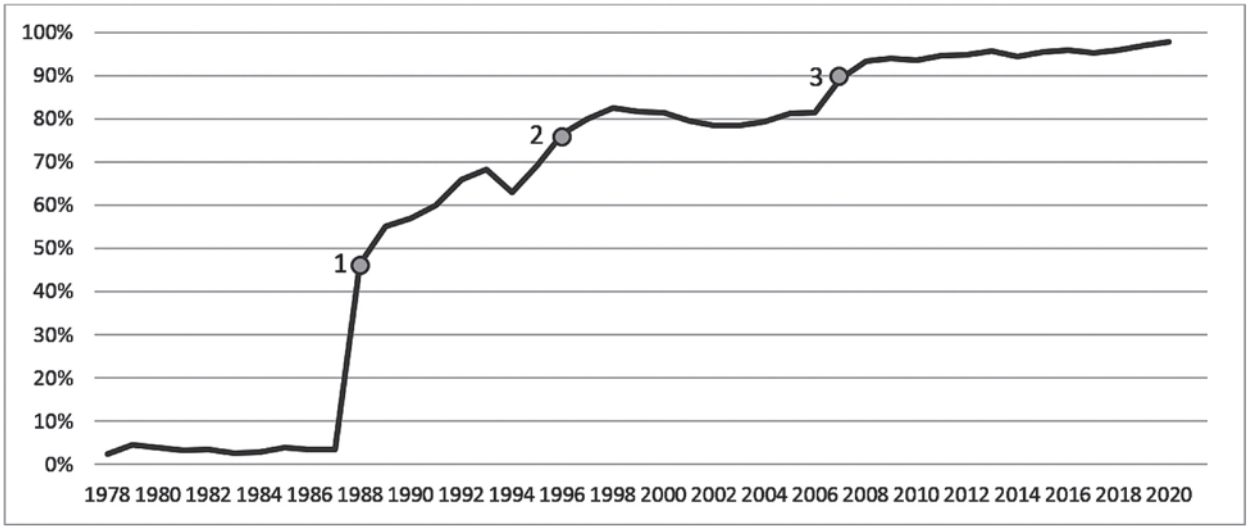

Рис. 1. Відсоток наявності в записах книг ISBN за роками видання

б) База 2 – 3579 (55% від загальної кількості по базі);

3) код не відповідає правилам ISBN – 1311 (14,5%):

а) База 1 – 1 (0,04% від загальної кількості по базі);

б) База 2 – 1310 (20% від загальної кількості по базі).

Абсолютна більшість помилок в обох базах пов'язана з тим, що контрольна цифра не відповідає коду, а отже, код містить помилку або в значимій частині, або в контрольній цифрі. Варто зазначити, що частина кодів стали помилковими ще на етапі підготовки книги до друку.

Серед записів без структурної недійсності дублі має 50107 записів ISBN (16% від загальної кількості ISBN без помилок).

Серед записів зі структурною недійсністю дублі має 1170 записів (12,3% від загальної кількості зі структурними помилками).

Помилкові ISBN не можуть мати співпадінь з правильними за визначенням (і це підтвердилося на практиці). Випадкові співпадіння кількох помилкових записів між собою малоймовірні. У разі помилок введення кількість дублів неправильних ISBN мала б прямувати до нуля. Отримані результати свідчать про велику кількість помилок, створених під час підготовки до друку.

Записи ISBN, що містять помилки, можна розділити на такі множини:

1) помилки введення даних з книги:

а) записи, які повинні були мати дублі

− мають дублі через співпадіння кількох помилкових ISBN (малоймовірно);

− не мають дублів через помилку;

б) Записи, які не повинні були мати дублів

− не мають дублів;

2) помилки підготовки книги (той самий відсоток дублів ISBN, що і в іншій базі).

Оскільки дублі серед записів ISBN з помилками можливі лише у разі, якщо джерелом помилок є неправильно надрукований номер на виданні, то мінімальна кількість ISBN з помилками підготовки книги дорівнює кількості дублів – 1170 (12,3% від кількості ISBN з помилками).

Якщо ж врахувати те, що з об'єктивних причин не всі ISBN можуть мати дублі і серед ISBN без помилок дублі мало лише 16% кодів, то ймовірний відсоток помилок видавця серед загальної кількості помилок можна розрахувати так:

$$N*0+K*16 = 12,3$$

$$K*16 = 12,3$$

$$K = 12,3/16 = 0,77 = 77\%,$$

де N – кількість записів ISBN з помилками введення, а K – кількість записів ISBN з помилками підготовки книги.

Отже, маємо орієнтовну частину помилок, пов'язаних з додрукарською підготовкою, – 77% за мінімальної 12,3%.

Серед записів ISBN з дублями, що містять помилки, є такий розподіл помилок за типами:

− кількість символів ISBN не відповідає формату – 223 (20%);

− контрольна цифра вказує на помилку в коді – 642 (58%);

− код не відповідає правилам ISBN – 2 46 (22%).

Отже, видавець на 32% рідше використовує неправильну кількість символів, ніж у середньому в базах. Помилки з неправильною кількістю символів простіше виявити та виправити на етапі підготовки.

До помилок, спричинених неправильним введенням даних бібліографами, ймовірно належать 2074 записів ISBN з помилками. Якщо врахувати, що загальна кількість записів ISBN 313574, то їх вага незначна – 0,7%.

Існують методи усунення помилок введення із застосуванням контрольної цифри. Вони не гарантують відновлення даних, хоча здебільшого можуть допомогти. Деякі види помилок введення потребують ручного втручання або розробки інтелектуального алгоритму на базі правил ISBN.

3. Пошук потенційних дублів записів книжок по ISBN

Оскільки з одним записом книг може бути пов'язано декілька записів ISBN, статистика по записах книг, що мають співпадіння з іншими книгами по ISBN, суттєво відрізняється від того, що ми мали у разі порівняння ISBN:

− 121275 – усього записів книг, що мають співпадіння з іншими книгами по ISBN у межах двох баз (32% усіх записів книг, або 44% від книг, що мають ISBN);

− 118705 – записів книг, що мають співпадіння з іншими книгами по ISBN без помилок у межах двох баз (118 587 виключно на основі записів без помилок);

− 2688 – записів книг мають співпадіння з іншими книгами виключно по ISBN зі структурними помилками.

Кількість потенційних дублів книг на основі порівняння ISBN є значною. Не всі записи книг з однаковими ISBN є дублями. Але групування на основі співпадіння ISBN може допомогти перевірити відповідність інших параметрів та знайти частину справжніх дублів.

4. Стандартизація ISBN

Після переходу на новий стандарт ISBN продовжили перевидаватися книги, які отримали ще

10-значний код. Такі номери мали змінити формат згідно з новими вимогами. Також переведення номерів на новий формат могло відбуватися з інших міркувань. Крім іншого, як зазначає Бібліотека Конгресу [11], з 2005 по 2007 рік існував перехідний період, коли видавці могли друкувати номери різного формату, а бібліографи в різний час могли вибирати для введення в базу той чи інший формат.

Якщо порівнювати ISBN у тому вигляді, в якому він введений у БД, то з поля зору може зникнути певна кількість книжок, ISBN яких введено в іншому стандарті.

Приведемо всі 10-значні ISBN до 13-значного формату (додамо префікс «978» та замінимо значення контрольної цифри на розраховане за відповідним алгоритмом). Порівнювати будемо лише ISBN без помилок, яких є 303770 записів.

Порівняння у вихідному форматі:

− 225228 унікальних значень;

− 49766 унікальних ISBN мають дублі та об'єднують 128308 записів.

Порівняння після приведення до 13-значного формату:

− 223227 унікальних значень (-2001 запис, або -1%);

− 50266 унікальних ISBN мають дублі та об'єднують 130759 записів.

Відбулося скорочення кількості унікальних значень (-2001, або -1%) та зростання кількості кодів, що мають дублі (+500, або +1%) та кількості об'єднаних записів (+2451, або +2%).

Без приведення ми мали 118705 записів книг, що мали однакові ISBN. Після приведення стало 120573. Приріст на 1,5%, або на 1868 записів книг.

Отже, зміна формату є виправданим засобом стандартизації значень поля та дозволяє підвищити якість бази за рахунок об'єднання записів, ISBN яких знаходилися в різних форматах.

5. Структура ISBN

Наявність структури ISBN важлива для отримання доступу до окремих груп цього номера. Наприклад, для ідентифікації видавця.

Як зазначалося у вступі, ISBN складається з кількох блоків різної довжини, розділених дефісами. Довжина блоків варіюється і має відповідати зведенню правил, розроблених міжнародною агенцією [10].

Цікава закономірність була віднайдена у 3-му класі структурних помилок – код не відповідає правилам ISBN. Перевірка на такий клас помилок відбувалась виключно у випадках, коли код був потрібної довжини та контрольна цифра підтверджувала відсутність помилок. Усього таких помилок в обох базах 992. Причому 972 (98%) з них суперечать лише одному з правил – у правилі сказано, що код з першими блоками «978-611» належить країні Таїланд та заборонений до використання. Решта 2% є проблемами введення даних, але з правильною контрольною цифрою.

На підтвердження суперечностей кодів України та Таїланду знайдено дві книги на один ISBN 978-611-01-0039-7:

− Банківські інновації С.Б. Єгоричева;

− การวิจัยดำเนินงาน (Operations Research).

Один і той самий код був знайдений у різних джерелах. Щодо українського коду можна навіть наочно пересвідчитись у тому, що це не помилка бібліографів.

Можна вважати такі ISBN умовно придатними, оскільки настільки багато однотипних помилок з правильною контрольною цифрою бути не може. Також ручною перевіркою не було знайдено випадків, коли б частина коду «978-611» була наслідком помилки введення (основний кандидат «978-617»). Усі книги, які були знайдені в Інтернеті, мали код саме «978-611».

Можливо, це є підтвердженням того, що зміни правил ISBN можуть мати наслідки. Лише за другий квартал 2021 року було впроваджено 73 зміни діапазонів, а відповідно і правил [10].

Наслідки внесення змін до правил:

а) відсутня гарантія відтворення структури ISBN (щодо «978-611» це вже неможливо);

б) немає певності, що відтворена структура є правильною;

в) дійсні ISBN можуть не пройти перевірку за правилами.

Таким чином, правило використання у машинозчитуваних форматах даних ISBN без поділу на блоки має ризики втрати даних без можливості автоматичного відновлення. В тому числі до цього можуть призвести інструкції стандарту MARC 21 [4]. А відсутність достовірної структури суттєво ускладнює використання поля ISBN у процесах інтеграції гетерогенних бібліографічних даних.

Деякі коди ISBN введені зі збереженням структури. Для таких кодів можна проаналізувати співпадіння відновленої за правилами структури та тієї, яка була введена в бібліографічну базу.

Записів ISBN зі збереженням структури:

− усього – 71847 (23% від записів ISBN двох баз);

− у Базі 1 – 45335 (65% від наявних у Базі 1);

− у Базі 2 – 26512 (11% від наявних у Базі 2).

У Базі 2 у 6 разів менший відсоток записів зі структурою. Така невелика кількість пов'язана

з тим, що вони вводилися так лише на початку ведення бази. Серед сучасних записів таких лише кілька десятків.

У Базі 1 ISBN зі збереженням структури вводяться рівномірно, хоча третина кодів вводиться без поділу на блоки.

Така різниця вкотре показує, що при роботі з обраними бібліографічними базами ми маємо справу з гетерогенними даними.

Для записів ISBN, в яких була збережена структура, можна провести перевірку, чи автоматична генерація зможе відтворити саме таку структуру, яка була введена.

Якщо взяти ISBN, які не мають структурних помилок, то співпадіння введеної та відновленої структури буде у 69057 випадках. Відсутня вона буде у 786 записах ISBN (1,1% від загальної кількості без помилок). 2004 записи мають помилки.

Отже, спостерігається невисока наявність структурованих записів ISBN. Трапляються випадки, коли за діючими правилами відновити структуру неможливо. А наслідки відновлення за правилами в 1,1% випадків не співпали з введеним у базу значенням. Також заважають відновленню структури і помилки, які роблять ISBN недійсним.

Висновки. Були зібрані великі обсяги бібліографічних даних, масив правил ISBN та алгоритми перевірки ISBN. Розроблено програмне забезпечення, яке реалізовувало правила та алгоритми ISBN для використання в обробці даних. За допомогою програмного забезпечення проаналізовано дані, знайдено особливості поля. Зроблена оцінка процесів відновлення структури ISBN на основі правил. Оцінено потенціал використання поля ISBN у процесах інтеграції гетерогенних бібліографічних даних. Оцінено та пояснено динаміку змін поширеності ISBN. Проаналізовано помилки ISBN, їх структуру та причини виникнення.

Описано знайдену проблему з правилами ISBN, яка потребує уточнення на рівні агенції ISBN. Запропоновано стандартизацію ISBN та оцінено ефект від неї.

Незважаючи на те, що лише 72% книг з бібліографічних баз мають ISBN, а 2,9% наявних ISBN містять помилки, це поле несе у собі корисну інформацію та є перспективним для використання в процесах обробки бібліографічних даних. В тому числі в процесах інтеграції гетерогенних бібліографічних даних.

Список літератури:

1. Online catalogs: what users and librarians want: an OCLC report: OCLC, 2009. URL: https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf.
2. International ISBN Agency. ISBN Users' Manual. URL: https://www.isbn-international.org/sites/default/files/ISBN; International Users Manual-7th edition_absolutely_final.docx.
3. International ISMN Agency. ISMN Users' Manual URL: https://www.ismn-international.org/files/Web_ISMN_Users_Manual_2016.pdf.
4. Library of Congress. MARC 21. International Standard Book Number. URL: https://www.loc.gov/marc/bibliographic/bd020.html.
5. Книжкова палата. ISBN/ISMN URL: http://www.ukrbook.net/agentstvo.html.
6. Library of Congress. Use of ISBNs and LCCNs in MARC 21 Bibliographic Records. URL: https://www.loc.gov/marc/marbi/2004/2004-dp04.html.
7. Електронний каталог бібліотеки ім. Тараса Шевченка для дітей (публічні бібліотеки для дітей м. Києва). URL: http://zra.kiev.ua:8081/MarcWeb.
8. Електронний каталог бібліотеки ім. Лесі Українки (публічні бібліотеки для дорослих м. Києва). URL: http://ecatalog.kiev.ua.
9. Українська бібліотечна енциклопедія. Міжнародний стандартний номер книги URL: https://ube.nlu.org.ua/article/Міжнародний стандартний номер книги (ISBN = International Standard Book Number).
10. International ISBN Agency. ISBN Ranges. URL: https://www.isbn-international.org/range_file_generation.
11. Library of Congress. LC Plan to Accommodate 13-Digit ISBN. URL: https://www.loc.gov/cds/notices/notisbn13.html.

Petrenko M.V. RESEARCH OF FEATURES OF THE "ISBN" FIELD AND THEIR INFLUENCE ON USE IN THE PROCESSES OF BIBLIOGRAPHICAL DATA PROCESSING

The article is devoted to the consideration of the international book number ISBN as an integral part of the bibliographic record and bibliographic database in the context of the processes of quality improvement and integration of bibliographic data.

Today, the current processes of creating consolidated electronic catalogues of libraries and improving the quality of data stored in them. To ensure a sufficient level of quality of consolidated directories, it is necessary to develop, justify and test methods of data preparation and processing.

ISBN is an important part of the bibliographic record and in addition to uniqueness contains encoded additional information. Using this information can bring many benefits. But for this it is necessary to understand all the nuances of ISBN as a theoretical plan and the features of its application in real bibliographic databases.

The article describes the results of the study of quantitative and qualitative indicators of the ISBN data field on the basis of real data of the 141st public library of Kyiv, obtained from two different sources.

The features of the field, both common and different for both bibliographic databases, are noted. An assessment of the processes of restoring the structure of the ISBN on the basis of rules. The potential of using the ISBN field in the processes of integration of heterogeneous bibliographic data is estimated. ISBN standardization is proposed and its effect is evaluated. Describes a problem with ISBN rules that needs to be clarified at the ISBN level. The dynamics of changes in ISBN prevalence is estimated and explained. Much attention is paid to ISBN duplicates and the association of the number of duplicates and unique ISBN entries with errors and data presentation formats. ISBN errors, their structure and causes are analyzed.

The obtained results will allow to build a strategy for processing the studied field of bibliographic data in order to achieve better results in the tasks of data purification, standardization and integration.

Key words: *ISBN, ISBN structure, ISBN standardization, bibliographic data, deduplication, input errors.*