International Science Journal of Engineering & Agriculture
2025; 4(1): 111-121
https://isg-journal.com/isjea
doi: 10.46299/j.isjea.20250401.10
ISSN: 2720-6319

INTERNATIONAL SCIENCE GROUP

Розробка принципів суміщення бібліографічних записів для автоматизації зведеного каталогу

Олег Василенко
Інститут кібернетики імені В.М.Глушкова НАН України, м. Київ, Україна / кафедра Інформаційних систем, технологій, фінансів та менеджменту ПВНЗ Український гуманітарний інститут, м. Буча, Україна
ORCID 0000-0001-8498-2950

Для цитування цієї статті:
Василенко Олег. Розробка принципів суміщення бібліографічних записів для автоматизації зведеного каталогу. International Science Journal of Engineering & Agriculture. Vol. 4, No.1, 2025, pp. 111-121.doi: 10.46299/j.isjea.20250401.10.
Надійшла до редакції: 09 грудня 2024 р.; **Схвалено:** 11 січня 2025 р.;
Опубліковано: 01 лютого 2025 р.

Анотація: Розробка принципів суміщення бібліографічних записів для автоматизації зведеного каталогу є важливою частиною вдосконалення бібліотечних інформаційних систем, що сприяє полегшенню доступу до знань і інформації. З огляду на збільшення обсягів бібліографічних даних і необхідність їх швидкої обробки, важливою є розробка ефективних методів для інтеграції записів з різних джерел, що забезпечує точність і зручність їх використання. Одним із основних викликів є автоматизація процесів очищення і підготовки бібліографічних записів, що включає в себе перетворення даних із різних форматів, таких як MARC21 у UNIMARC, а також виявлення та корекцію помилок, включаючи друкарські. Для цього використовуються алгоритми, що дозволяють швидко обробляти великі масиви даних і підвищувати якість записів. Ключовим етапом є порівняння даних для виявлення дублюючих записів. Для цього застосовуються новітні методи штучного інтелекту, які можуть автоматично знаходити схожі або дубльовані записи, враховуючи контекст і специфіку кожного запису. Використання таких підходів значно покращує точність пошуку та зменшує ризик помилок при злитті даних. Об'єднання даних також автоматизується за допомогою методів штучного інтелекту, що дозволяє зменшити кількість ручних втручань і підвищити ефективність роботи бібліотек. В результаті цієї автоматизації бібліотечні працівники можуть зосередитися на більш важливих аспектах роботи, таких як аналітика даних, дослідження змісту та надання високоякісних послуг користувачам. Процес інтеграції даних з різних джерел і форматів, зокрема із застосуванням технологій штучного інтелекту та алгоритмів машинного навчання, дає можливість створювати зведені каталоги, що відповідають сучасним вимогам точності, швидкості і доступності інформації. Цей підхід дозволяє бібліотекам значно підвищити ефективність своєї діяльності, оптимізуючи рутинні процеси та покращуючи взаємодію з користувачами.

Ключові слова: суміщення бібліографічних записів, автоматизація зведеного каталогу, бібліографічні дані, MARC21, UNIMARC, Штучний інтелект.

1. Вступ

Автоматизація зведеного каталогу бібліографічних записів є важливою складовою сучасних бібліотечних інформаційних систем. Оскільки бібліотеки постійно збільшують свої ресурси, необхідність ефективного обміну бібліографічними даними стає все більш актуальною. Одним із ключових аспектів є інтеграція даних з різних бібліотек і баз даних для створення єдиного зведеного каталогу. Оскільки в традиційних бібліотечних системах

відсутня централізована платформа для обміну інформацією, використання спеціалізованих протоколів для автоматизованого збирання та обробки бібліографічних записів стало важливим напрямом розвитку. Зокрема, значним викликом є узгодження та суміщення бібліографічних записів з різних джерел, які можуть бути представлені в різних форматах та мати різні стандарти. Тому питання створення системи, що дозволяє автоматично з'єднувати дані з різних джерел, не порушуючи їхню цілісність і точність, є дуже важливим для бібліотекарів і розробників систем автоматизації бібліотечних процесів.

2. Об'єкт і предмет дослідження

Об'єктом дослідження є процес створення та функціонування зведених каталогів у бібліотечних і інформаційних системах. Предметом дослідження є принципи суміщення бібліографічних записів, які забезпечують автоматизацію інтеграції даних для зведеного каталогу.

3. Мета та задачі дослідження

Метою дослідження є розробка принципів і методик суміщення бібліографічних записів для ефективної автоматизації процесів створення та оновлення зведених каталогів. Для досягнення мети були поставлені такі задачі: визначити основні вимоги до суміщення бібліографічних записів у сучасних інформаційних системах; дослідити методи інтеграції бібліографічних даних, враховуючи міжнародні стандарти (MARC, RDA, FRBR); провести аналіз існуючих алгоритмів автоматизації та визначити їх переваги й обмеження; розробити рекомендації щодо застосування принципів суміщення записів у зведених каталогах.

4. Аналіз літератури

Аналіз наукових праць та практичних розробок показав, що проблема суміщення бібліографічних записів активно досліджується в контексті стандартизації та автоматизації бібліотечних систем. Значний внесок зроблено у вивченні формату MARC 21 та принципів опису ресурсів за RDA. Роботи з розробки алгоритмів інтеграції записів переважно зосереджені на проблемах дублювання, відмінностей у метаданих і мовних варіантах описів. Особливу увагу привертають дослідження щодо реалізації концепцій FRBR для моделювання бібліографічних даних і створення їх узгоджених структур у зведених каталогах.

5. Методи досліджень

У процесі роботи були використані такі методи: аналіз і синтез для визначення вимог до суміщення бібліографічних записів. Порівняльний аналіз для вивчення існуючих алгоритмів і стандартів інтеграції записів. Моделювання для розробки принципів автоматизації суміщення записів. Емпіричний метод для тестування ефективності запропонованих підходів у реальних інформаційних системах. Програмна реалізація для розробки та перевірки алгоритмів автоматизації. Обрані методи дослідження створюють основу для побудови дослідження, спрямованого на вдосконалення процесів інтеграції бібліографічних даних.

6. Результати досліджень

Протокол Z39.50 є одним із стандартів, який використовується для обміну бібліографічними даними між бібліотечними інформаційними системами. Він був розроблений для того, щоб дозволити користувачам отримувати інформацію з різних бібліотек у форматі, сумісному з їхньою системою. Це дозволяє бібліотекам здійснювати пошук, запити

і обробку даних з віддалених джерел без необхідності фізичного перенесення даних чи використання складних інтеграційних механізмів. Протокол Z39.50 був широко впроваджений у 1990-х роках і досі залишається важливим інструментом для організації доступу до бібліографічних даних. Він дозволяє бібліотечним системам здійснювати пошук по індексах та каталогах різних бібліотек, а також обробляти запити на основі метаданих і передавати отриману інформацію у вигляді, зручному для користувача. Протокол включає в себе механізми для пошуку записів, відбору потрібної інформації та передачі результатів пошуку. Z39.50 здатен працювати з різними типами баз даних, що робить його універсальним інструментом для забезпечення доступу до бібліографічних ресурсів.

Незважаючи на те, що протокол Z39.50 був дуже корисним для інтеграції бібліотечних систем, він має деякі обмеження. Наприклад, Z39.50 часто використовує застарілу технологію, що ускладнює його інтеграцію з новими технологічними стандартами, такими як веб-сервіси або API. Крім того, для використання протоколу потрібно мати спеціалізоване програмне забезпечення та знання для налаштування й обробки запитів. Нове дослідження в галузі автоматизації обробки бібліографічних записів полягає у розробці альтернативних методів отримання та обробки даних без використання Z39.50. Зокрема, в останні роки активно використовуються REST API, які дозволяють здійснювати запити та отримувати дані у форматах JSON або XML без необхідності інтеграції старих бібліотечних стандартів. API забезпечують більшу гнучкість у порівнянні з протоколом Z39.50 і дають змогу обробляти дані з різних джерел значно швидше і ефективніше. Ці технології дозволяють більш безпосередньо інтегрувати дані з відкритих джерел, таких як бібліографічні бази даних, що доступні через інтернет. Така система дає змогу більш ефективно створювати зведені каталоги, об'єднуючи дані з різних бібліотечних джерел без використання застарілих протоколів. Окрім того, використання API може допомогти вирішити питання сумісності даних різних форматів і дозволить бібліотекам значно швидше оновлювати свої записи, що в свою чергу покращує процес надання доступу до актуальної інформації користувачам.

Очищення і підготовка даних — це критичний етап у процесі обробки бібліографічних записів, особливо коли йдеться про конвертацію даних з одного формату в інший. У контексті бібліотечних систем, цей процес включає в себе вирішення проблем, пов’язаних з неповними, некоректними або дубльованими записами. Успішне очищення та підготовка даних дозволяє забезпечити високий рівень точності та сумісності з іншими бібліографічними записами.

Одним із завдань, яке виникає в процесі обробки бібліографічних записів, є конвертація даних з формату MARC21 в UNIMARC. Це важливий крок для забезпечення сумісності між різними бібліотечними системами та форматами. MARC21 і UNIMARC — це два основні стандарти для представлення бібліографічної інформації, проте між ними існують значні відмінності в структурі записів, форматах і специфікаціях.

- MARC21 (Machine-Readable Cataloging) — це широко використовується в США та інших країнах формат, який підтримує складні метадані, що включають дані про авторів, видання, каталоги та інші елементи бібліографії. Він має велику кількість полів і підполів для детального опису документів.

- UNIMARC (Universal MARC) — це міжнародний стандарт, який забезпечує обмін бібліографічною інформацією між бібліотеками у різних країнах. Він більш орієнтований на міжнародний контекст і є стандартом, прийнятим в багатьох європейських країнах.

Процес перетворення між цими двома форматами вимагає точності і уважності, оскільки деякі дані можуть не бути прямо переведені або потребують коригування під іншу структуру. Очищення даних, що містять помилки у форматуванні, дублювання чи неактуальні записи, допомагає полегшити процес перетворення.

Одним із аспектів очищення даних є виявлення і виправлення друкарських помилок, які можуть виникати під час введення даних або при конвертації з одного формату в інший. Друкарські помилки можуть серйозно вплинути на точність пошукових запитів і доступність бібліографічних записів, тому їх виявлення є важливим кроком у підготовці даних.

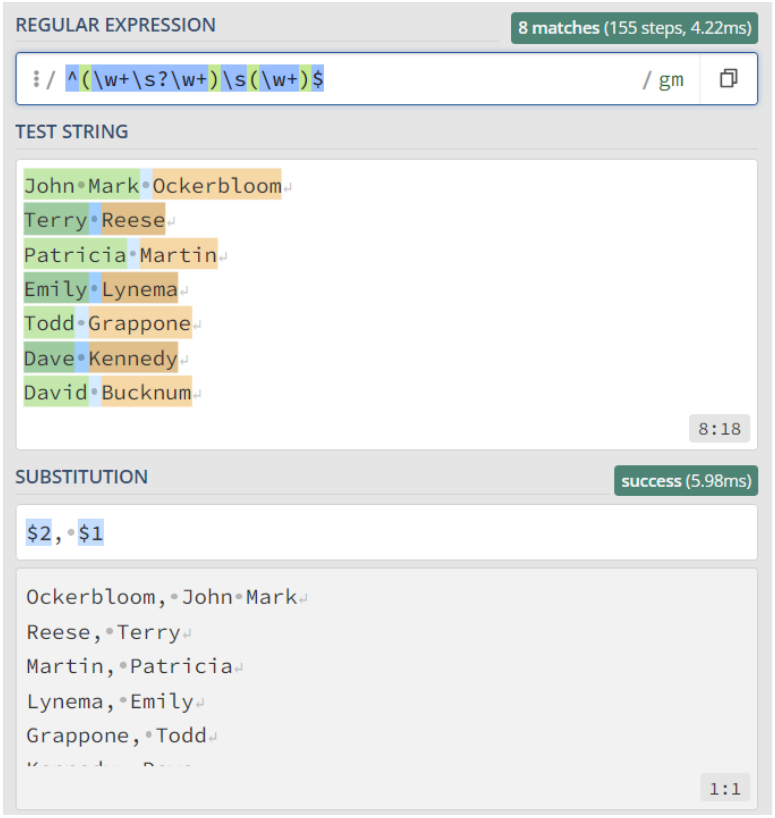

Рис 1. Зразок як працює регулярний вираз на невідібраних даних (регулярний вираз, який замінює імена авторів з прямого порядку на зворотній).

Алгоритми для пошуку друкарських помилок зазвичай включають такі методи як алгоритм Левенштейна (Levenshtein distance) – це метод, який визначає кількість операцій (вставка, видалення, заміна), необхідних для перетворення одного рядка в інший. Цей алгоритм може бути використаний для виявлення і виправлення дрібних помилок у написанні слів. Регулярні вирази (Regular Expressions, RegEx) – регулярні вирази дозволяють знаходити певні шаблони в тексті, що допомагає виявляти можливі помилки, такі як відсутні пробіли, додаткові символи або непотрібні знаки пунктуації. Словники і бази даних синонімів – для більш точного виявлення помилок можна використовувати спеціалізовані словники, які допомагають знайти найбільш поширені варіанти написання слів та їх правильні форми.

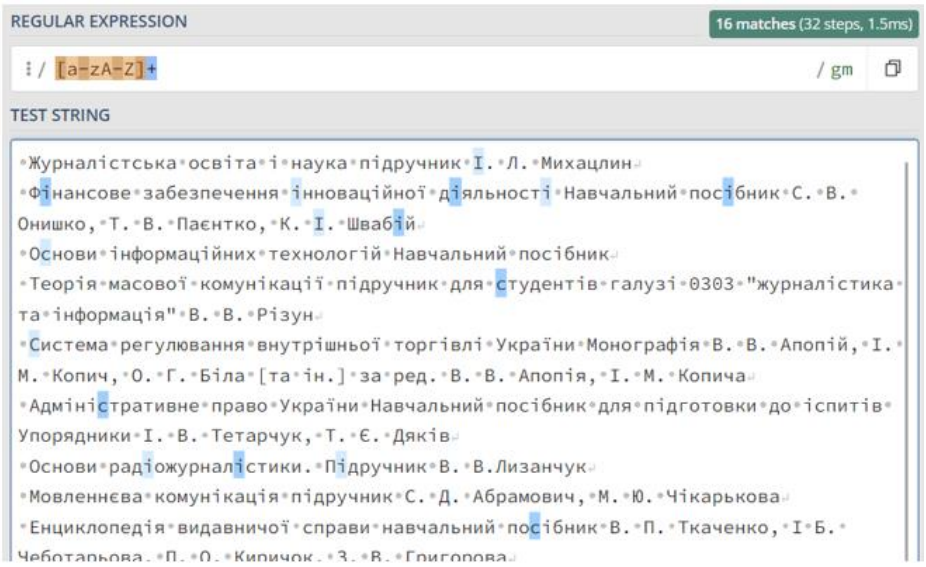

Рис. 2. Візуалізація пошуку латинських символів в кириличних текстах.

Один із практичних кейсів застосування очищення даних можна розглянути на прикладі використання системи **Koha** для бібліографічних записів. У цій системі можна застосувати алгоритми для автоматичного виявлення та виправлення друкарських помилок.

Бібліотека отримує дані з декількох джерел, у тому числі з MARC21. Для перетворення їх у формат UNIMARC використовуються спеціальні інструменти, які автоматично змінюють поля, коригують форматування та вирівнюють записи під міжнародні стандарти. Після перетворення даних проводиться перевірка на наявність поширених друкарських помилок, таких як неправильне написання авторів, назв книг чи журналів. Алгоритм Левенштейна використовується для порівняння назв і виявлення варіантів, що можуть бути помилково записані. За допомогою регулярних виразів можна швидко знаходити й виправляти непотрібні пробіли або символи в бібліографічних записах, що можуть бути результатом некоректного введення даних. Бібліотека також використовує словник для виявлення та виправлення найпоширеніших друкарських помилок у назвах авторів або видавців.

Рис. 3. XSL файли перетворення з формату MARC21 в UNIMARC.

Використання алгоритмів Левенштейна і регулярних виразів дозволило автоматично виправити понад 90% дрібних помилок у бібліографічних записах, що значно підвищило точність бази даних. Час на обробку одного запису скоротився в 2-3 рази порівняно з ручною перевіркою і коригуванням даних. Конвертація даних з MARC21 в UNIMARC сприяла створенню зведеного каталогу, який можна безперешкодно обмінювати з іншими бібліотеками, що використовують цей стандарт.

Таким чином, очищення і підготовка бібліографічних даних, особливо в контексті перетворення між різними форматами, є важливим етапом для створення ефективних бібліотечних систем. Використання алгоритмів для виявлення та виправлення друкарських помилок забезпечує високу якість даних, що є критичним для подальшого використання і зберігання бібліографічних записів.

Пошук і виявлення дублікатів є важливою частиною обробки бібліографічних даних, оскільки дублікати можуть впливати на ефективність пошуку та знижувати точність даних у бібліотечних системах. Сучасні технології для пошуку дублікатів часто використовують різні алгоритми порівняння даних, що дозволяють автоматично знаходити та об'єднувати схожі записи.

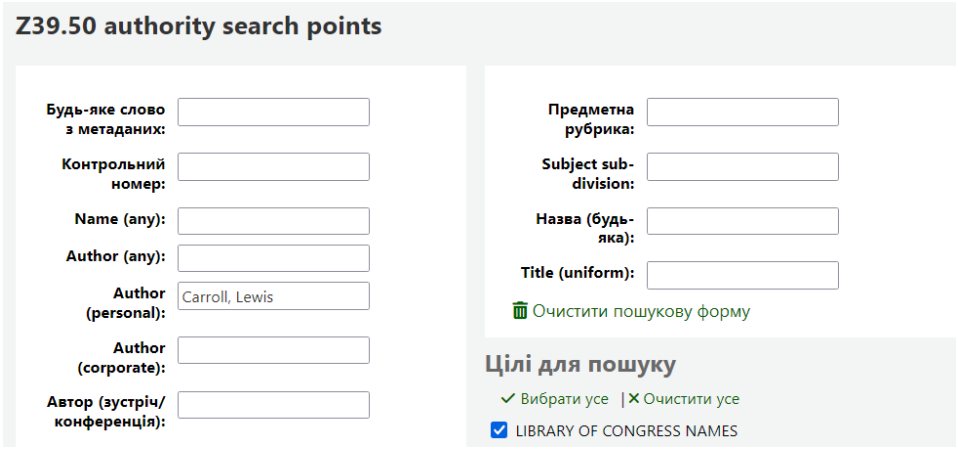

Рис. 4. Пошук автора в авторитетному файлі імен Бібліотеки конгресу (На скріншоті можна побачити, що основний заголовок (ім'я автора) розташоване в полі 200, що відповідає формату UNIMARC, на відміну від MARC21, в якому дана інформація зберігається в полі 100).

До традиційних способів порівняння даних відносяться такі алгоритми як один з найбільш простих методів, де два записи порівнюються за наявністю точних збігів у полях (наприклад, ім'я автора, назва книги). Однак цей метод не враховує можливих варіантів написання та граматичних помилок. Визначає мінімальну кількість операцій (вставка, видалення, заміна), які необхідні для перетворення одного рядка в інший. Це дозволяє знаходити схожі записи, навіть якщо вони містять незначні помилки. Регулярні вирази (RegEx) використовуються для пошуку певних шаблонів у тексті, що допомагає автоматично виявляти і обробляти повторювані елементи, наприклад, однакові чи схожі частини бібліографічних записів.

Results for authority records

Сервер	Heading	Тип авторитетного джерела	Дії
LIBRARY OF CONGRESS NAMES	Carroll Lewis 1832-1898	NP	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Alice's adventures in Wonderland	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Alice's adventures in Wonderland. Czech	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Alice's adventures in Wonderland. French	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Alice's adventures in Wonderland. Swedish	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Bambine de Carroll. English	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Bedside book	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Hunting of the snark	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Hunting of the snark. French	SAUTTIT	MARC
LIBRARY OF CONGRESS NAMES	Carroll, Lewis (1832-1898) Jabberwocky	SAUTTIT	MARC

Рис. 5. Налаштування сервера авторитетних записів.

Сучасний підхід до пошуку дублікатів в бібліографічних даних включає застосування методів штучного інтелекту (ШІ), зокрема машинного навчання **та** нейронних мереж**,** які здатні покращити точність і ефективність процесу. Використання моделей машинного навчання для виявлення схожих записів є більш ефективним, оскільки вони можуть автоматично навчатися на основі великих обсягів даних, адаптуючись до різних варіантів написання та структур бібліографічних записів. Наприклад, застосування класифікації текстів на основі характеристик слів і фраз дозволяє ефективно визначати схожі записи.

Глибокі нейронні мережі (наприклад, нейронні мережі на основі моделей трансформерів**, як** BERT або GPT) можуть аналізувати бібліографічні записи не тільки на рівні символів або слів, але й на більш глибокому рівні, враховуючи контекст і зміст кожного запису. Це дозволяє точніше виявляти дублікати навіть за наявності варіативних форм записів або помилок. Один із методів, що застосовуються разом з машинним навчанням – це векторизація тексту**,** де кожен запис перетворюється на набір числових значень. Потім за допомогою алгоритмів, таких як K-means або кластеризація, можна визначити схожість записів і виявити дублікати на основі їх векторних представлень.

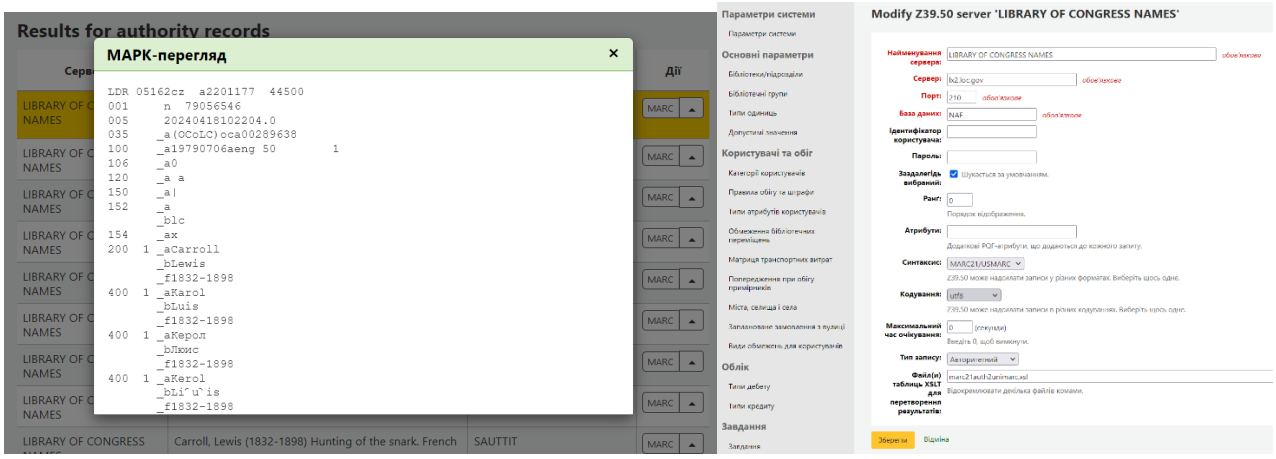

Рис. 6. Конвертація відбувається завдяки підготовленому файлу XSL «на льоту» (при цьому слід зазначити, що на даний момент, Koha без модифікацій підтримує XSL перетворення лише для бібліографічних записів. Зроблені модифікації в подальшому планується додати в спільну кодову базу Koha).

Для прикладу, у бібліотеці з великою кількістю бібліографічних записів можна застосувати методи машинного навчання та нейронних мереж для пошуку дублікатів у колекціях, які містять схожі записи, але мають варіативні формати. Спочатку створюється база даних із бібліографічними записами, що містить різні варіанти записів для одних і тих самих джерел. Дані можуть бути отримані із різних джерел, зокрема з інших бібліотек, або через автоматичні системи імпорту. Кожен бібліографічний запис перетворюється на вектор за допомогою алгоритмів векторизації тексту (наприклад, TF-IDF або Word2Vec). Це дозволяє представити кожен запис у вигляді числового вектора, що відображає його контекст і зміст. Для виявлення схожих записів можна використати алгоритми класифікації або регресії, що дозволяють «навчити» модель знаходити дублікати. Наприклад, модель може бути тренована на виявлення схожих записів через певний поріг схожості між векторами. Після того як модель навчається на векторизованих даних, вона застосовується для перевірки нових бібліографічних записів на наявність дублікатів. Модель на основі аналізу схожості між записами може автоматично об'єднувати дублікати, знижуючи кількість ручної роботи.

Використання штучного інтелекту дозволяє знизити кількість помилкових збігів, що може статися при використанні лише традиційних методів порівняння, таких як алгоритм Левенштейна або регулярні вирази. Методи машинного навчання здатні обробляти великі обсяги даних за короткий час, що робить їх особливо корисними для великих бібліотечних баз

даних. ШІ методи здатні адаптуватися до різних варіантів записів, навіть у випадку, коли стандартні алгоритми не можуть точно виявити схожість. В результаті застосування штучного інтелекту в процесі пошуку та виявлення дублікатів бібліографічних записів, бібліотеки можуть значно покращити якість своїх даних, автоматизувати процеси та знизити витрати часу на ручну перевірку і коригування.

Об'єднання даних – важливий етап у процесі обробки бібліографічних записів, оскільки багато джерел можуть містити подібну або дублюючу інформацію, що потребує злиття в єдиний структурований формат. Це особливо актуально при зведенні різних бібліографічних систем у зведений каталог або при інтеграції даних з різних джерел. Зазвичай, при об'єднанні даних застосовують такі підходи як:

- Порівняння за полями використання стандартних полів, таких як ім'я автора, назва твору, рік публікації. Це дозволяє порівнювати записи та виявляти їх дублікати.
- Механізми злиття після виявлення схожих записів створюється механізм для злиття їх в один запис, що включає об’єднання даних з різних джерел.
- Статистичні методи використовуються для обчислення ймовірності того, що два записи є дублікатами на основі певних характеристик, таких як схожість назв або авторів.
- Застосування штучного інтелекту для об'єднання бібліографічних даних відкриває нові можливості для автоматизації та підвищення точності обробки. ШІ допомагає не тільки виявляти дублікати, але й злиття даних з різних джерел з урахуванням контексту, значення та взаємозв'язків між елементами.

Основні методи ШІ для об'єднання даних методи на основі навчання з учителем. Вони включають тренування моделей, які можуть автоматично розпізнавати схожі записи та комбінувати їх у один. Ці моделі можуть використовувати класифікацію для оцінки, чи є два записи однаковими, чи потрібно їх об'єднувати. Глибокі нейронні мережі дозволяють аналізувати складні шаблони у великих наборах даних. Вони здатні визначати схожість навіть у випадку, коли записи мають різні формати чи варіанти написання. Методи природної мови (NLP) – це використання методів обробки природної мови допомагає моделі розуміти контекст і значення бібліографічних даних, що дозволяє більш точно виконувати об'єднання записів. Наприклад, може бути застосовано семантичне порівняння текстів, де навіть різні формулювання одних і тих самих ідей можуть бути визначені як схожі.

Уявімо ситуацію, коли два бібліографічні записи з різних бібліотечних систем містять інформацію про одну і ту ж книгу, але з різними варіантами написання авторів та назв. Наприклад, один запис може містити ім'я автора як «J.K. Rowling», а інший – «Rowling, J.K.» або навіть «Joanne Rowling».

Крок 1 – векторизація даних. Для початку кожен бібліографічний запис перетворюється на вектор, використовуючи методи векторизації тексту, такі як Word2Vec або BERT. Це дозволяє подати записи у вигляді числових векторів, де схожість між записами може бути визначена на основі векторних відстаней.

Крок 2 – навчання моделі. Модель машинного навчання, наприклад, класифікаційна модель, тренується на історичних даних, де вже вказано, чи потрібно зливати записи. Після навчання модель може автоматично оцінювати ймовірність того, що два записи належать до одного і того ж джерела.

Крок 3 – пошук схожих записів. Коли нові записи додаються в систему, вони автоматично порівнюються з існуючими. Моделі на основі глибоких нейронних мереж (наприклад, Siamese Networks) можуть бути використані для пошуку схожих записів. Ці моделі добре працюють з випадками, коли записи не збігаються за точними характеристиками, але мають схожу семантику.

Крок 4 – об'єднання даних. Якщо система визначає, що два записи є схожими, то вони автоматично зливаються в один. У цей процес можуть бути інтегровані додаткові алгоритми для виправлення помилок (наприклад, заміна різних варіантів написання імені автора або уточнення рік публікації).

Результататом є точність – штучний інтелект дозволяє досягти високої точності при об'єднанні даних, навіть коли записи містять дрібні помилки або варіації. Швидкість – алгоритми ШІ значно пришвидшують процес об'єднання даних, особливо при роботі з великими масивами інформації. Адаптивність – моделі можуть навчатися на нових даних і постійно вдосконалювати свої результати.

Переваги використання ШІ для об'єднання даних – це зниження людських помилок. Автоматизоване об'єднання даних мінімізує ймовірність помилок, що можуть виникати при ручному злитті. Економія часу, коли алгоритми можуть працювати з великими обсягами даних, швидко зливаючи записи без необхідності ручного втручання. Підвищення якості баз даних, коли об'єднання записів забезпечує узгодженість і точність у зведених каталогах.

Таким чином, використання методів штучного інтелекту для об'єднання бібліографічних записів дозволяє значно підвищити ефективність та точність процесу обробки даних у бібліотечних системах, а також знизити трудозатрати на ручну обробку записів.

7. Перспективи подальшого розвитку досліджень

Подальше вдосконалення алгоритмів суміщення бібліографічних записів із врахуванням мовних, регіональних і технічних особливостей різних інформаційних систем. Це сприятиме створенню універсальних рішень для використання у глобальних зведених каталогах. Застосування методів машинного навчання та штучного інтелекту для автоматичного розпізнавання та обробки записів, усунення дублювання та підвищення точності суміщення. Дослідження можливостей інтеграції нових форматів метаданих, таких як BIBFRAME, і їх сумісності з традиційними стандартами MARC та RDA.

8. Висновки

Розробка принципів суміщення бібліографічних записів для автоматизації зведеного каталогу є важливою складовою вдосконалення бібліотечних систем та оптимізації роботи з бібліографічними даними. Враховуючи постійне зростання обсягу та різноманіття інформації, яку обробляють бібліотеки, ефективне злиття записів з різних джерел стає критично важливим для забезпечення точності і зручності доступу до знань. Протокол Z39.50 залишається основним інструментом для обміну бібліографічними даними між різними бібліотеками. Однак новизна цього дослідження полягає в оптимізації отримання даних без його використання, що дає можливість знижувати залежність від традиційних протоколів та відкриває нові шляхи для інтеграції сучасних бібліографічних баз.

Очищення і підготовка даних за допомогою алгоритмів для виявлення та виправлення помилок, а також перетворення форматів даних (з MARC21 у UNIMARC), є важливим кроком для забезпечення сумісності та якості даних. Використання алгоритмів для автоматичного очищення та трансформації даних допомагає підвищити точність і значно зменшити час, витрачений на ручну обробку.

Порівняння даних і пошук дублікатів за допомогою методів штучного інтелекту, таких як глибокі нейронні мережі та обробка природної мови, забезпечують ефективне виявлення дублюючих записів і покращення результатів пошуку. Завдяки таким технологіям, система може навчатися на великій кількості даних, постійно підвищуючи свою точність та адаптивність. Об'єднання даних також стає більш точним та автоматизованим завдяки штучному інтелекту. Використання методів глибокого навчання та інших підходів, орієнтованих на контекстуальне розпізнавання схожих записів, дозволяє знижувати кількість помилок, покращувати узгодженість інформації та значно економити час на злиття даних.

У результаті впровадження таких технологій та методів, бібліотеки можуть значно покращити якість своїх каталогів, підвищити ефективність обробки даних і забезпечити кращий доступ до інформації для користувачів. Використання ШІ та алгоритмів машинного

навчання робить ці процеси швидкими, точними та адаптивними, що дозволяє забезпечити більш високий рівень автоматизації та зручності для бібліотечних працівників та користувачів.

Список літератури:

1) Daquino, M., Peroni, S., Shotton, D., & Colavizza, G. (2020). The OpenCitations Data Model. *ArXiv*. https://arxiv.org/abs/2005.11981

2) Bologna, F., Iorio, A., Peroni, S., & Poggi, F. (2021). Do open citations inform the qualitative peer-review evaluation in research assessments? An analysis of the Italian National Scientific Qualification. *ArXiv*. https://arxiv.org/abs/2103.07942

3) Bologna, F., Peroni, S., & Poggi, F. (2021). Can we assess research using open scientific knowledge graphs? A case study within the Italian National Scientific Qualification. *ArXiv*. https://arxiv.org/abs/2105.08599

4) Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021). Data citation and the citation graph. *Quantitative Science Studies*, *2*, 1399–1422. https://doi.org/10.1162/qss_a_00166

5) Färber, M., & Lamprecht, D. (2021). The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies*, (2), 1324–1355. https://doi.org/10.19181/smtp.2023.5.2.4

6) Giambattista, C., Heibi, I., Peroni, S., & Shotton, D. (2022). OpenCitations: an Open e-Infrastructure to Foster Maximum Reuse of Citation Data. *Int. J. Digit. Curation*, *17*(5).

7) Jiang, Y., & Liu, X. (2024). A new method of calculating the disruption index based on open citation data. *Journal of Information Science*. https://doi.org/10.1177/01655515241263545

8) Lim, W. M., Kumar, S., & Donthu, N. (2024). How to combine and clean bibliometric data and use bibliometric tools synergistically: Guidelines using metaverse research. *Journal of Business Research*, *182*. https://doi.org/10.1016/j.jbusres.2024.114760

9) Malínek, V., Umerle, T., Gray, E., Heibi, I., & Király, P. (2024). Open Bibliographical Data Workflows and the Multilinguality Challenge. *Journal of Open Humanities Data*, *10*(27), 1–14. https://doi.org/10.5334/johd.190

10) Massari, A., & Heibi, I. (2022). How to structure citations data and bibliographic metadata in the OpenCitations accepted format. In *ULITE@JCDL*.

11) Massari, A., Mariani, F., Heibi, I., Peroni, S., & Shotton, D. (2024). OpenCitations Meta. *Quantitative Science Studies*, *5*(1), 50–75. https://doi.org/10.1162/qss_a_00292

12) Mazov, N., & Gureyev, V. (2023). Open Access Bibliographic Resources for Maintaining a Bibliographic Database of Research Organization. *Scientific and Technical Information Processing*, *50*, 211–223.

13) Petrovich, E., Verhaegh, S., Bös, G., & Cristalli, C. (2024). Bibliometrics beyond citations: introducing mention extraction and analysis. *Scientometrics*, *129*, 5731–5768. https://doi.org/10.1007/s11192-024-05116-x

14) Rizzetto, E., & Peroni, S. (2023). Mapping bibliographic metadata collections: the case of OpenCitations Meta and OpenAlex. In *Italian Research Conference on Digital Library Management Systems*.

15) Rodrigues, N., Mariano, A., & Ralha, C. (2024). Author name disambiguation literature review with consolidated meta-analytic approach. *Int. J. Digit. Libr.*, *25*, 765–785. https://doi.org/10.1007/s00799-024-00398-

16) Teixeira da Silva, J., Huang, C., & Ortega, J. (2023). Open Citations as a Tool for Bibliometric Verification and Transparency and for Correcting Erroneous References. *Journal of Scholarly Publishing*, *54*, 60–79.

Development of principles for merging bibliographic records for the automation of union catalogs

Oleh Vasylenko
V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences (NAS) of Ukraine, Kyiv, Ukraine / Department of Information Systems, Technologies, Finance and Management Ukrainian Institute of Arts and Sciences Bucha, Ukraine

Abstract: The development of principles for merging bibliographic records for the automation of union catalogs is an important part of improving library information systems, which facilitates easier access to knowledge and information. Given the increasing volume of bibliographic data and the need for its rapid processing, it is crucial to develop effective methods for integrating records from various sources, ensuring accuracy and ease of use. One of the main challenges is the automation of the cleaning and preparation of bibliographic records, which involves converting data from different formats, such as MARC21 to UNIMARC, as well as identifying and correcting errors, including typographical ones. Algorithms are employed to quickly process large datasets and improve the quality of records. A key stage is data comparison to identify duplicate records. The latest artificial intelligence (AI) methods are applied to automatically find similar or duplicate records, considering the context and specifics of each record. The use of such approaches significantly improves search accuracy and reduces the risk of errors when merging data. Data merging is also automated through AI techniques, which helps reduce manual interventions and increase the efficiency of library operations. As a result of this automation, library staff can focus on more critical aspects of their work, such as data analytics, content research, and providing high-quality services to users. The process of integrating data from various sources and formats, especially using AI technologies and machine learning algorithms, enables the creation of union catalogs that meet modern requirements for accuracy, speed, and accessibility of information. This approach allows libraries to significantly enhance the efficiency of their operations by optimizing routine processes and improving user interaction.

Keywords: Merging bibliographic records**,** Union catalog automation**,** Bibliographic data**,** MARC21**,** UNIMARC**,** Artificial Intelligence.