DOI https://doi.org/10.15589/znp2020.2(480).14
УДК 004.652.4:930.25

AUTOMATION OF DATA PREPARATION FOR THE GEOGRAPHIC BLOCK OF THE ARCHIVAL DATABASE

АВТОМАТИЗАЦІЯ ПІДГОТОВКИ ДАНИХ ДЛЯ ГЕОГРАФІЧНОГО БЛОКА АРХІВНОЇ БАЗИ ДАНИХ

Yuliia V. Prokop[1]
yulia13.prokop@gmail.com
ORCID: 0000-0002-6608-3668

Olena G. Trofymenko[2]
egt@ukr.net
ORCID: 0000-0001-7626-0886

Artem S. Prokop[3]
artem.gpay@gmail.com
ORCID: 0000-0003-4246-9635

Ю. В. Прокоп[1],
канд. іст. наук, ст. викладач

О. Г. Трофименко[2],
канд. техн. наук, доцент

А. С. Прокоп[3],
ліцеїст

[1]***O. S. Popov Odessa National Academy of Telecommunications, Odesa***
[1]Одеська національна академія зв'язку імені О. С. Попова, м. Одеса
[2]***National University "Odessa Law Academy", Odesa***
[2]Національний університет «Одеська юридична академія», м. Одеса
[3]***Odesa Region Municipal Establishment "Richelieu Lyceum", Odesa***
[3]Комунальний заклад «Рішельєвський ліцей», м. Одеса

Abstract. The creation of archival information retrieval systems is one of the actual directions of the development of the Ukrainian archival system. However, its implementation requires entering a huge amount of information into databases. Moreover, this process is not currently automated and therefore requires significant human resources to enter data manually.

The work **aims** to study the automation of the data preparation process for the geographical block of the archival information retrieval system from sources of various formats: electronic registers, web pages, paper books, handwritten archival documents, etc. A subsystem for data preparation is proposed. It consists of modules for searching for information sources, data extraction, data identification, and entering information into the database. Much of the work in the subsystem is automated and does not require manual data entry. The choice of method of data extraction and pre-processing depends on the source of information. Given the specifics of the task, it can be assumed that the vast majority of sources will be either printed publications or handwritten archival documents. Therefore, the first step to their processing should be scanning and text recognition using common software or neural network. Unstructured text obtained from sources is automatically transformed by syntactical analysis into structured text, which is entered in the table of a certain template. The extracted data must be identified, information about identical administrative units must be combined and entered into a database. The proposed subsystem of data preparation was implemented on the example of the preparation of geographical information for the Mykolayiv region. The significance of the obtained results is that the use of the proposed algorithm will automate the filling of the geographical block with data from other regions for use in regional thematic archival databases and the national archival information retrieval system.

Key words: archival information retrieval system; geographic block; data preparation; information extraction; data preparation automation; syntactic analysis.

Анотація. Створення архівних інформаційно-пошукових систем є одним з актуальних напрямів розвитку української архівної галузі. Проте реалізація його потребує внесення до баз даних величезного обсягу інформації, причому цей процес нині не автоматизований, а тому потребує значних людських ресурсів для введення даних вручну.

Метою роботи є дослідження автоматизації процесу підготовки даних для географічного блока архівної інформаційно-пошукової системи із джерел різного формату: електронних реєстрів, вебсторінок, паперових друкованих видань, рукописних архівних документів тощо. Пропонується підсистема для підготовки даних, яка складається з модулів пошуку джерел інформації, видобування даних, ідентифікації даних і внесення інформації до бази. Значну частину роботи в підсистемі було автоматизовано, тому вона вже не потребує

ручного введення даних. Вибір методу видобування і попереднього опрацювання даних залежить від джерела інформації. З урахуванням специфіки завдання можна припускати, що більшість джерел є або друкованими виданнями, або рукописними архівними документами. Тому першим кроком для їх опрацювання має бути сканування і розпізнавання тексту за допомогою поширених програмних засобів або нейронної мережі. Отриманий із джерел неструктурований текст методом синтаксичного аналізу трансформується у структурований і заноситься в таблиці певного шаблону. Видобуті дані мають бути ідентифіковані, відомості про тотожні адміністративні одиниці – об'єднані та внесені в базу даних. Запропонована підсистема підготовки даних була реалізована на прикладі підготовки географічних відомостей для Миколаївської області.

Практична значимість отриманих результатів полягає в тому, що використання запропонованого алгоритму дозволить автоматизувати заповнення географічного блока даними інших регіонів для використання в регіональних тематичних архівних базах даних і загальнодержавній архівній інформаційно-пошуковій системі.

Ключові слова: архівна інформаційно-пошукова система; географічний блок; підготовка даних; видобування інформації; автоматизація підготовки даних; синтаксичний аналіз тексту.

ПОСТАНОВКА ЗАДАЧІ

Важливим напрямом розвитку архівної галузі України є створення інформаційно-пошукових систем (далі – ІПС), які дозволять швидко виявляти потрібні архівні документи та відомості з них. Актуальність запровадження таких систем зумовлена зростанням зацікавленості науковців та пересічних громадян в історичних, краєзнавчих та генеалогічних дослідженнях і підвищенням суспільного попиту на доступ до першоджерел. Для вдосконалення пошуку доречно передбачити можливість відбору даних за географічною ознакою, створивши у складі архівних ІПС та баз даних спеціальні географічні блоки з урахуванням ієрархічної структури сучасних та історичних адміністративно-територіальних одиниць, підпорядкувань та всіх історичних перейменувань населених пунктів.

На початку 2020 р. Укрдержархів презентував нову ІПС «Архіум». У структуру бази даних (далі – БД), що є основою цієї системи, закладено можливість зберігання основних відомостей про архівні документи (номери та назви фондів, описів, справ, крайні дати документів тощо), а також метаданих, які дозволять пошук не лише за формальними ознаками справ, а й за ключовими словами вмісту документів.

Проте однією із проблем цієї та подібних БД є їх наповнення, яке нині не автоматизоване. Ідеться про ручне введення величезних обсягів даних, які беруться переважно з рукописних архівних джерел або старих машинописних документів. Серед причин, якими пояснюється зволікання у впровадженні архівних ІПС, є і нестача людських кваліфікованих ресурсів для коректного внесення даних у БД. Зважаючи на це, вельми актуальною є автоматизація заповнення географічного блока, яка би дозволила максимально скоротити часові витрати архівістів на введення даних.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Чимало дослідників шукають можливості автоматизації підготовки різного роду даних для подальшого аналізу чи використання у БД. У роботі [1] аналізуються доступні для цього способи і констатується, що, незважаючи на досягнення сучасних технологій роботи з даними, підготовка даних для подальшого аналізу потребує багато ручної роботи та може забирати значну кількість часу та зусиль. Автори статті [2] зазначають, що на підготовку даних спеціалісти витрачають 80% свого часу.

Дослідження [3] присвячено розгляду специфіки підготовки даних, які стосуються наукової активності. Автори [4] пропонують алгоритм, який виділяє з неструктурованих документів назву, автора, реферат, ключові слова, текст та інші елементи, після чого експортує структурований текст у потрібний формат, що відповідає вимогам структурованого пошуку, статистичної класифікації тощо.

У роботі [5] відзначається, що підготовка даних зазвичай вимагає оброблення даних різними методами, залежно від джерел. Для цього, зокрема, застосовуються вирази структурованої мови запитів (SQL) для вилучення й агрегації записів БД, засоби Microsoft Excel для очищення та нормалізації наборів даних, сценарії для виконання складних перетворень (наприклад, у Python).

Значну увагу дослідників привертає проблема автоматизації видобування інформації з вебсайтів у структурованому вигляді. Зокрема, у статті [6] виконано огляд наявних рішень цієї проблеми. У роботі [7] порівнюються різні підходи для отримання структурованих відомостей із вебсайтів. Дослідження [8–9] вивчають різні техніки для збирання та підготовки інтернет-даних для подальшого опрацювання й аналізу. Автор [10] пропонує для видобування даних із вебсайтів використовувати візуальні та структурні особливості елементів вебсторінок, щоб згрупувати їх у семантично схожі кластери.

Синтаксичний аналіз текстових даних є важливою складовою частиною видобування інформації під час підготовки даних, якій присвячено чимало досліджень. Зокрема, у статті [11] надається огляд технік видобування тексту і керування якістю даних у контексті відкритих даних. Дослідження [12] присвячено порівнянню засобів видобування тексту, а у статті [13] запропоновано комплексний підхід до

видобутку даних із текстових новин на основі морфологічного і синтаксичного аналізів. У роботі [14] вивчається автоматизований аналіз тексту, що забезпечує інтеграцію лінгвістичної теорії з конструкціями, які зазвичай використовуються у споживчих дослідженнях.

Ще одним важливим напрямом автоматизації підготовки даних є розпізнавання тексту зі сканованих зображень. Робота [15] присвячена розпізнаванню машинописного тексту, що має пошкодження або шуми, опрацювання якого поширеними програмними засобами видає багато помилок. Увагу дослідників привертає розпізнавання рукописного або змішаного тексту. Так, у роботі [16] вивчаються засоби для автоматизації введення у БД відомостей із рукописних форм: технології сканування даних та машинне навчання для підготовки системи до перетворення паперової копії на структуровані дані. Робота [17] досліджує можливості згорткових нейронних мереж для розпізнавання рукописних символів. Роботи [18–19] пропонують новий метод сегментації рукописного тексту за допомогою конволюційної нейронної мережі (CNN). Автори дослідження [20] використовують мережі короткострокової пам'яті (LSTM) зі згорткою для побудови обмежувальних коробок для кожного символу, після чого передають сегментовані символи в конволюційну нейронну мережу для класифікації, а потім реконструюють кожне слово відповідно до результатів класифікації та сегментації. Автори [21] демонструють результати використання конволюційної нейронної мережі для розпізнавання рукописного тексту з можливістю налаштування на використання для тексту різними мовами.

Одним з етапів підготовки даних до аналізу є очищення і трансформування даних. У статті [22] розглядаються методи очищення і трансформування даних у рамках технології Knowledge Discovery in Databases для прискореного застосування методів інтелектуального аналізу даних.

ВІДОКРЕМЛЕННЯ НЕ ВИРІШЕНИХ РАНІШЕ ЧАСТИН ЗАГАЛЬНОЇ ПРОБЛЕМИ

Аналіз досліджень показав, що підготовка даних, яка охоплює виявлення, вибір, очищення й інтеграцію наявних наборів даних до форми, придатної для подальшого використання, зокрема у БД, вимагає великих часових витрат, і численні дослідники шукають можливості її автоматизації для скорочення часу на її виконання. Попри наявні дослідження, запропоновані в них техніки та підходи до автоматизації підготовки даних або тематично вузько спрямовані, або містять алгоритми розв'язання лише окремих аспектів і можуть використовуватися лише частково для вирішення комплексного завдання автоматизації збирання та підготовки відомостей для географічного блока архівної ІПС.

МЕТА ДОСЛІДЖЕННЯ

Мета роботи – дослідження засобів автоматизації процесу підготовки даних для географічного блока архівної ІПС із джерел різного формату: електронних реєстрів, вебсторінок, паперових друкованих видань, рукописних архівних документів тощо.

МЕТОДИ, ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Об'єктом дослідження є автоматизація процесу підготовки даних для бази даних архівної ІПС.

Предметом дослідження є методи і засоби автоматизації збору, видобування, ідентифікації та внесення даних до географічного блока ІПС.

ОСНОВНИЙ МАТЕРІАЛ

Процес підготовки даних для географічного блока БД складається із трьох етапів (рис. 1).

Перший із цих етапів виконується людиною. До пошуку, уведення й імпортування даних можуть бути залучені працівники обласних архівів та бібліотек, краєзнавці й історики, які добре знайомі з регіональними адміністративним устроєм та історією, мають досвід пошуку й опрацювання краєзнавчої інформації.

Для імпортування у БД потрібні такі відомості: перелік усіх населених пунктів, розташованих на території України та/або Української Радянської Соціалістичної Республіки (далі – УРСР), їхні історичні назви упродовж періоду існування (або принаймні за останні 250 років), а також інші адміністративні одиниці (райони, повіти, області, губернії); історичні статуси (село, селище, місто, район, повіт,

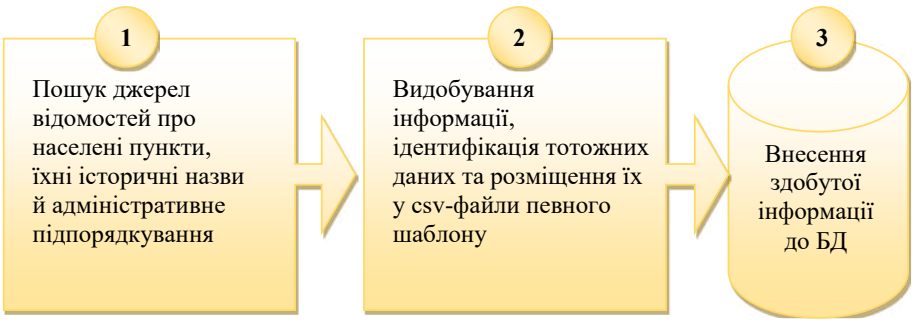

Рис. 1. Етапи підготовки даних

область, губернія тощо); перелік адміністративних підпорядкувань населених пунктів – до волостей та/або повітів (до 1917 р.) і районів (після 1917 р.), волостей – до повітів, повітів/районів – до губерній/областей, губерній/областей – до країн; часові межі існування адміністративних одиниць, назв та підпорядкувань; географічні координати населених пунктів (довгота і широта).

Джерелами сучасних відомостей про населені пункти та їхнє адміністративне підпорядкування можуть бути офіційні реєстри та статистичні дані, наведені на сайті Верховної Ради України. Зокрема, Державний реєстр географічних назв (https://land.gov.ua/info/informatsiia-pro-derzhavnyi-reiestr-heohrafichnykh-nazv/) складається із pdf-файлів, по одному для кожної області й Автономної Республіки Крим. У цих файлах, серед іншого, містяться унормовані назви областей, адміністративних районів, населених пунктів двома мовами (українською й англійською) із зазначенням підпорядкування відповідному району, географічні координати (широта і довгота), причому кожен із них має унікальний ID.

У статистичних даних, що містяться на сайті Верховної Ради України (http://static.rada.gov.ua/zakon/new/NEWSAIT/ADM/zmist.html), для кожної області окремо наведено «Відомості про райони, міста і селища міського типу», звідки можна дізнатися про рік присвоєння населеному пункту статусу міста, селища міського типу тощо. Крім того, там сформований перелік перейменувань населених пунктів із 1986 р. та перелік пунктів, знятих з обліку після 1986 р. Тобто на підставі названих офіційних джерел можна зібрати якщо не вичерпну, то принаймні достатню інформацію щодо сучасного адміністративного устрою країни.

Джерелом географічних даних про УРСР упродовж періоду 1917–1990 рр. можуть бути регіональні або республіканські довідники населених пунктів і архівні документи того часу.

Джерелами відомостей про населені пункти та їхнє підпорядкування до 1917 р. можуть бути старі друковані довідкові видання, пам'ятні книжки й архівні документи. Зокрема, відомості про Подільську губернію Російської імперії можна зібрати у виданні 1893 р. [23], яке містить, серед іншого, дані про назви повітів, міст, волостей, сіл, їхній тип (статус), підпорядкування. Відомості про адміністративний устрій Херсонської губернії можна знайти на вебсайті «Родове гніздо» (http://rodovoyegnezdo.narod.ru/geografy.htm). Джерелом даних про перейменування може бути багатотомне сучасне видання «Зведений каталог метричних книг».

Загалом, джерела історичних географічних відомостей для кожної області мають встановлюватися індивідуально представниками обласних архівів, залежно від регіональних особливостей. Після підбору джерел для збирання потрібної інформації можна переходити до наступного етапу.

Оскільки обсяг відомостей для введення на другому етапі (рис. 1) заповнення географічного блока даними в рамках країни загалом є досить значним, то саме його реалізація потребує чималих людських ресурсів і гальмує створення архівних ІПС, а тому потребує автоматизації. Для етапів 2 і 3 (рис. 1) підсистема підготовки даних містить модулі видобування інформації, ідентифікації даних та внесення відомостей до БД (рис. 2).

У модулі видобування даних для кожного джерела відомостей створюється власний метод. Інформація з онлайн-реєстрів конвертується до електронних таблиць і опрацьовується засобами Microsoft Excel (видаляються зайві рядки та стовпці, заповнюються порожні комірки для підпорядкування районів та обласного центра). Додаються стовпці, необхідні для відповідності шаблону, який згодом буде імпортований до БД. Окремі таблиці для кожної області можуть бути об'єднані в загальну таблицю.

Наступним кроком є створення аналогічних таблиць, заповнення їх відомостями про перейменування та зняття з обліку населених пунктів зі згаданого сайту Верховної Ради. Після цього матимемо три таблиці, які треба буде звести в одну на етапі ідентифікації даних, об'єднавши рядки, які стосуються тих самих населених пунктів.

Дані з вебсторінок видобувають на цьому етапі шляхом синтаксичного аналізу тексту. Приклад синтаксичного аналізу фрагмента сторінки сайту «Родове гніздо», яка містить відомості про адміністративний склад Херсонського повіту Херсонської губернії, наведено на рис. 3.

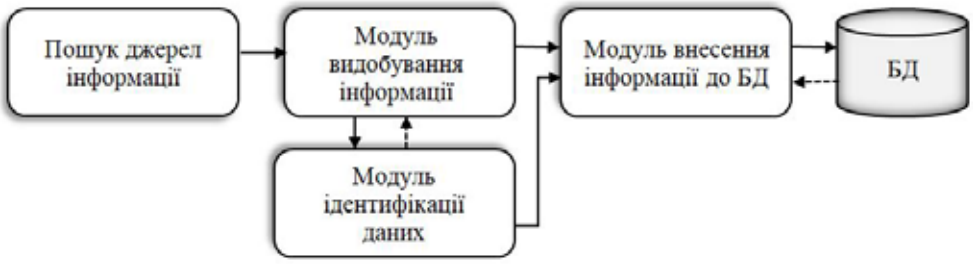

Рис. 2. Схема підсистеми підготовки даних для архівної ІПС

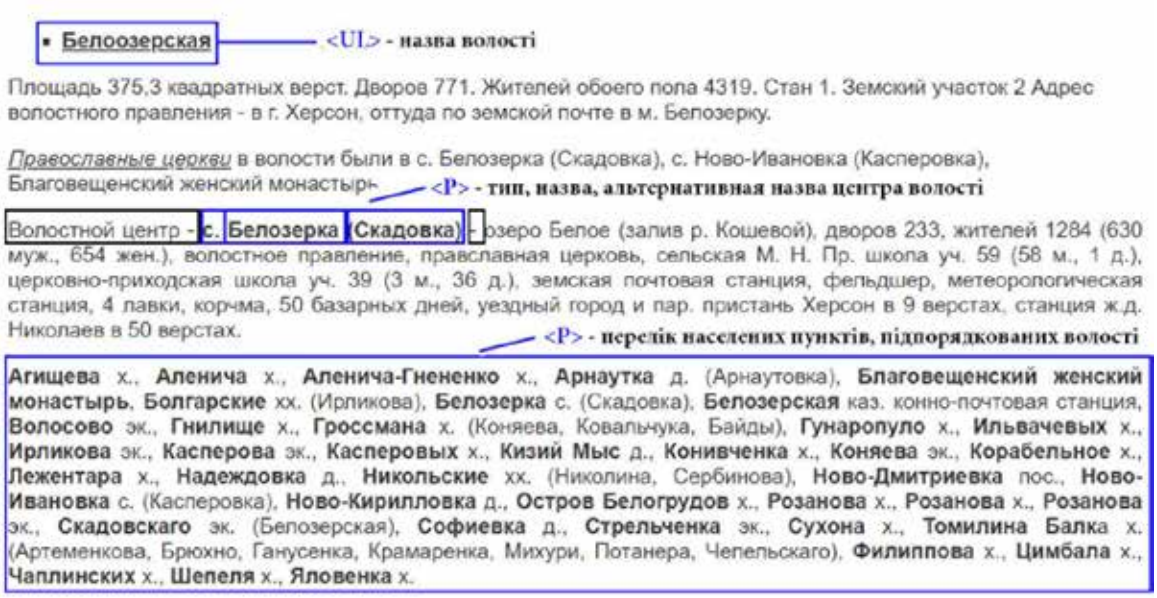

Рис. 3. Приклад синтаксичного аналізу тексту вебсторінки

Скановані копії друкованих видань, які містять чіткий машинописний текст, попередньо проходять розпізнавання засобами спеціального програмного забезпечення, зокрема програмою Abbyy FineReader, результатом якого є текстовий документ із неструктурованими даними. Наступним кроком є вичищення розпізнаного тексту від шуму – зайвих пробілів та сторонніх символів, розривів сторінок тощо.

Далі інформація видобувається завдяки проведенню синтаксичного аналізу тексту. Зокрема, перелік населених пунктів Подільської губернії містить текст, поданий у такому форматі: на початку кожного абзацу вказана назва населеного пункту, а в дужках подана альтернативна назва, яка супроводжується словами «тожъ» або «иначе», після коми вказано тип населеного пункту, далі вказується в родовому відмінку, до якого повіту і волості належить населений пункт (усе відокремлюється комами із пробілами).

Найскладніше видобування даних із рукописних джерел, якими є багато архівних документів. Для успішного видобування таких даних необхідна заздалегідь створена та натренована нейронна мережа для конвертації рукописного тексту (з урахуванням регіональних мовних особливостей) до текстового файлу, що далі аналізується аналогічно до попереднього.

Суттєво, що відомості про підпорядкування, які потраплять до таблиці за такого підходу до заповнення даних, мають дискретний характер і зазвичай не містять визначених часових меж. Тому доцільно для кожного джерела задати як такі межі принаймні рік, до якого належить джерело, а далі розширити межі за можливості під час ідентифікації.

Залежно від джерел відомостей, імовірно, до таблиць потраплять не всі історичні та народні назви населених пунктів. За можливості варто додати їх до таблиці вручну, переклавши трьома мовами. Аналогічно треба зробити з підпорядкуваннями, які не потрапили до таблиці через недосконалість вибраних джерел.

Не обов'язково встановлювати абсолютно всі варіанти історичних підпорядкувань, здебільшого достатньо вказати одне підпорядкування до 1917 р., одне чи декілька – часів УРСР, а також сучасне. Однак неприпустимо для спрощення підпорядковувати населені пункти безпосередньо губерніям і областям, якщо вони насправді підпорядковувалися повітам або районам. Варто також звернути увагу на підпорядкування дрібних селищ волостям (зазвичай ця інформація міститься в дореволюційних списках населених пунктів).

Після виконання модуля видобування даних матимемо певну кількість таблиць, у яких міститься окремо інформація про адміністративний устрій до 1917 р., часів УРСР та сучасний. Якщо відомості про сучасний стан із реєстру можна одразу занести у БД, то історичні дані мають обов'язково пройти попередню ідентифікацію.

Модуль ідентифікації даних призначений для встановлення тотожності населених пунктів, відомості про які здобуті з різних джерел у різні моменти часу. Головною метою цього модуля є встановлення тотожним населеним пунктам однакових значень ID. У разі наявності джерел різними мовами (наприклад, українська та російська або українська та польська) бажано заздалегідь перекласти всі назви цими мовами і зберегти їх в одній таблиці. Спочатку варто провести автоматичну (програмну) ідентифікацію населених пунктів, для яких зібрано відомості про перейменування. Ще одну частину даних можна ідентифікувати програмними засобами, зіставляючи назви

і підпорядкування, за умови, що заздалегідь встановлене відношення між губерніями й областями, проте це можливо лише для унікальних і малопоширених назв. Решту даних доведеться ідентифікувати за безпосередньої участі людини (зазвичай спеціаліста-краєзнавця або архівіста).

Залежно від підбору джерел даних, імовірно, у таблиці, до якої буде зведено всі відомості, все ще залишиться чимало незаповнених комірок. Відомості, які критично необхідні, але досі не потрапили до таблиці, доцільно ввести вручну (якщо вони відомі фахівцям).

Модуль внесення даних до географічного блока БД передбачає імпорт csv-файлів і рознесення відомостей у відповідні таблиці. Реалізується таке або безпосередньо у БД за допомогою SQL (збережених процедур), або ззовні бази засобами php, Node.JS тощо. У коді має бути обов'язково закладено перетворення текстових назв районів, областей та країн на відповідні ID. Має забезпечуватись цілісність даних і зв'язність із контентом, який уже міститься у БД. Для кожної пари населених пунктів, якщо задані їхні географічні координати і це передбачено структурою БД, треба обчислити відстань за формулою [24]:

$$\Delta\sigma = \arctan \frac{\sqrt{\left(\cos\varphi_2 \sin(\Delta\lambda)\right)^2 + \left(\cos\varphi_1 \sin\varphi_2 - \sin\varphi_1 \cos\varphi_2 \cos(\Delta\lambda)\right)^2}}{\sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(\Delta\lambda)}, \quad (1)$$

де λ_1, φ_1 і λ_2, φ_2 – географічні широта і довгота в радіанах двох точок 1 і 2;

$\Delta\lambda$ і $\Delta\varphi$ – їхні абсолютні різниці;

$\Delta\sigma$ – центральний кут між ними.

Кожна з пар населених пунктів (A, B) має згадуватися лише один раз: (A, B) або (B, A).

Після виконання всіх зазначених етапів база даних буде заповнена країнами, областями, губерніями, округами, районами і повітами, історичними до 1917 р., сучасними пунктами України, їхніми назвами в минулому та (можливо, частково) історичними варіантами їхнього адміністративного підпорядкування з дискретними часовими межами.

Після заповнення географічного блока має відбутися процес географічного маркування записів архівної бази. Цей аналіз даних також можна довірити програмі: система буде шукати в назвах фондів і справ повнотекстові збіги географічних назв із даними, що містяться у створеному географічному блоці, та пропонувати варіанти для прив'язування справ, зокрема й залежно від назви фонду та часових меж документів. Проте остаточний вибір локалі для маркування все одно має залежати від людини.

ВИСНОВКИ

У роботі досліджено можливість автоматизації процесу підготовки даних для заповнення географічного блока архівної інформаційно-пошукової системи, що використовує відомості з державних реєстрів, друкованих довідкових видань, архівних справ та інтернет-ресурсів і мінімізує витрати часу архівістів на введення даних про адміністративні одиниці та їхнє підпорядкування. Запропоновано підсистему для підготовки даних, яка складається з модулів пошуку джерел інформації, видобування даних, ідентифікації даних, внесення інформації до БД. У підсистемі вирішено автоматичне виконання досить значного обсягу роботи, без ручного введення даних. Проте роль архівістів і краєзнавців у цьому процесі є також важливою, оскільки частина роботи (пошук джерел та коригування помилок) виконується вручну.

У рамках дослідження було автоматизовано підготовку набору географічних даних, апробовано її на прикладі Миколаївської області України. Запропонований алгоритм дозволить автоматизувати заповнення географічного блока даними інших регіонів для використання в регіональних тематичних архівних базах даних і в загальнодержавній архівній інформаційно-пошуковій системі.

REFERENCES

[1] Abdallah, Z.S., Du, L., Webb, G.I. (2017). Data Preparation. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, MA. DOI: 10.1007/978-1-4899-7687-1_62.

[2] Paton, N.W. (2019). Automating Data Preparation: Can We? Should We? Must We? *Proceedings of the 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*. 2019. Retrieved from: http://ceur-ws.org/Vol-2324/Paper00-InvTalk2-NPaton.pdf.

[3] Zagorulko, Y.A., Akhmadeeva, I.A., & Sery, A.S. (2015). Avtomatizaciya sbora informacii o nauchnoj deyatelnosti dlya tematicheskih intellektualnyh nauchnyh internet-resursov [An Automatization of Collection of Information about Scientific Activity for Thematic Intelligent Scientific Internet Resources]. *Analitika i upravlenie dannymi v oblastyah s intensivnym ispolzovaniem dannyh: XVII Mezhdunarodnaya konferenciya DAMDID/RCDL'2015. Obninsk: IATE NIYaU MIFI*, 2015, pp. 105–111. [in Russian].

[4] Chen J., Chen H. (2013). A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning. *Journal of Software,* vol. 8, № 1, pp. 55–62, 2013. DOI: 10.1007/978-3-642-01891-6_5.

[5] Narayanan, S., Jaiswal, A., Chiang, Y., Geng, Y., Knoblock, C.A., & Szekely, P. (2014). Integration and Automation of Data Preparation and Data Mining. *IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 1076–1085. DOI: 10.1109/ICDMW.2014.44.

[6] Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques. *Know.-Based Syst*. 70, C (November 2014), 301–323. DOI: 10.1016/j.knosys.2014.07.007.

[7] Bin Mohd Azir, M.A., Ahmad, K.B. (2017). Wrapper approaches for web data extraction : A review. *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, Langkawi, 2017, pp. 1–6. DOI: 10.1109/ICEEI.2017.8312458.

[8] Parvez, M.S., Tasneem, K.S.A., Rajendra, S.S., & Bodke, K.R. (2018). Analysis Of Different Web Data Extraction Techniques. *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, Mumbai, 2018, pp. 1–7. DOI: 10.1109/ICSCET.2018.8537333.

[9] Kohan, A., Yamamoto, M., & Artho, C. (2016). Automated Dataset Construction from Web Resources with Tool Kayur. *2016 Fourth International Symposium on Computing and Networking (CANDAR)*, Hiroshima, 2016, pp. 98–104. DOI: 10.1109/CANDAR.2016.0029.

[10] Grigalis, T. (2013). Towards web-scale structured web data extraction. In *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM'13)*. Association for Computing Machinery, New York, NY, USA, pp. 753–758. DOI: 10.1145/2433396.2433491.

[11] Azeroual, O., Saake, G., Abuosba, M., & Schöpfel, J. (2018). Text data mining and data quality management for research information systems in the context of open data and open science. *ArXiv, abs/1812.04298*.

[12] Kaur, A., Chopra, D. (2016). Comparison of text mining tools. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 186–192. DOI: 10.1109/icrito.2016.7784950.

[13] Cherenkov, I.A., Orekhov, S.V. (2012). Dobycha dannyh iz tekstovyh novostej na primere rynka polimerov [News Data Mining Based On Example Of Polymer Market]. *Sistemi obrobki informaciyi*, 2012, № 9 (107), pp. 224–227. [in Russian].

[14] Humphreys, A., Wang, R. (2018). Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, Volume 44, Issue 6, April 2018, pp. 1274–1306. DOI: 10.1093/jcr/ucx104.

[15] Rashid, S.F., Shafait, F., & Breuel, T.M. (2012). Scanning Neural Network for Text Line Recognition. *In Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems (DAS'12)*. IEEE Computer Society, USA, pp. 105–109. DOI: 10.1109/DAS.2012.77.

[16] Rizvi, M.A., Shrivastava, M., & Sahu, M. (2012). Artificial Neural Network Based Character Recognition Using Backpropagat. *Bioinformatics,* 2012. DOI: 10.24297/ijct.v3i1c.2777.

[17] Kovalchuk, A.M., Marchuk, G.V., & Marchuk, D.K. (2019). Zastosuvannya zgortkovoyi nejronnoyi merezhi dlya rozpiznavannya rukopisnih simvoliv [Application of a convolutional neural network for the recognition of handwritten characters]. *Vcheni zapiski TNU imeni V.I. Vernadskogo.* Branch of science: technical sciences. Volume 30 (69). № 4, Part 1, pp. 68–73. DOI: 10.32838/2663-5941/2019.4-1/13. [in Ukrainian].

[18] Jo, J., Koo, H.I., Soh, J.W., & Cho, N.I. (2019). Handwritten Text Segmentation via End-to-End Learning of Convolutional Neural Network. arXiv:1906.05229v1 [cs.CV] 12 Jun 2019.

[19] Wang, T., Wu, D.J., Coates,A., & Ng, A.Y. (2012). End-to-end text recognition with convolutional neural networks. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Tsukuba, 2012, pp. 3304–3308.

[20] Balci, B., Saadati, D., Shiferaw, D. (2017). Handwritten text recognition using deep learning. *CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University Project, 2017*.

[21] Such, F., Peri, D., Brockler, F., Hutkowski, P., & Ptucha, R. (2018). Fully Convolutional Networks for Handwriting Recognition. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Volume 1, pp. 86-91. DOI: 10.1109/ICFHR-2018.2018.00024.

[22] Shichkina, Y.A., Degtyarev, A.B., & Koblov, A.A. (2017). Tehnologiya ochistki i transformirovaniya dannyh v ramkah Knowledge Discovery in Databases (KDD) dlya uskorennogo primeneniya metodov Data Mining [Technology of cleaning and transforming data using the Knowledge Discovery in Databases (KDD) technology for fast application of Data Mining methods]. *CEUR Workshop Proceedings*, 1787, 428–434. [in Russian].

[23] Huldman, V.K. (1893). Naselennye mesta Podolskoj gubernii: (alfavitnyj perechen naselennyh punktov gubernii, s ukazaniem nekotoryh spravochnyh o nih svedenij). Kamenec-Podolskij : Tip. Podolskogo gub. pravleniya, 1893. [2], II, IV, 636 p. [in Russian].

[24] Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*, 23 (176), pp. 88–93. DOI: 10.1179/sre.1975.23.176.88.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Abdallah Z.S., Du L., Webb G.I. Data Preparation. *Encyclopedia of Machine Learning and Data Mining* / C. Sammut, G.I. Webb (eds.). Springer, Boston, MA, 2017. DOI: 10.1007/978-1-4899-7687-1_62.

[2] Paton N.W. Automating Data Preparation: Can We? Should We? Must We? *Proceedings of the 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*. 2019. URL: http://ceur-ws.org/Vol-2324/Paper00-InvTalk2-NPaton.pdf.

[3] Загорулько Ю.А., Ахмадеева И.Р., Серый А.С. Автоматизация сбора информации о научной деятельности для тематических интеллектуальных научных интернет-ресурсов. *Аналитика и управление данными в областях с интенсивным использованием данных :* XVII Международная конференция DAMDID/RCDL'2015. Обнинск : ИАТЭ НИЯУ МИФИ, 2015. С. 105–111.

[4] Chen J., Chen H. A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning.. *Journal of Software.* 2013. Vol. 8. № 1, P. 55–62. DOI: 10.1007/978-3-642-01891-6_5.

[5] Integration and Automation of Data Preparation and Data Mining / S. Narayanan et al. *IEEE International Conference on Data Mining Workshop*, Shenzhen. 2014. P. 1076–1085. DOI: 10.1109/ICDMW.2014.44.

[6] Web data extraction, applications and techniques / F. Ferrara et al. *Know.-Based Syst*. 2014. № 70. P. 301–323. DOI: 10.1016/j.knosys.2014.07.007.

[7] Bin Mohd Azir M.A., Ahmad K.B. Wrapper approaches for web data extraction : A review. *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*. Langkawi, 2017. P. 1–6. DOI: 10.1109/ICEEI.2017.8312458.

[8] Tasneem Analysis Of Different Web Data Extraction Techniques / M.S. Parvez et al. *International Conference on Smart City and Emerging Technology (ICSCET)*. Mumbai, 2018. P. 1–7. DOI: 10.1109/ICSCET.2018.8537333.

[9] Kohan A., Yamamoto M., Artho C. Automated Dataset Construction from Web Resources with Tool Kayur. *2016 Fourth International Symposium on Computing and Networking (CANDAR)*. Hiroshima, 2016. P. 98–104. DOI: 10.1109/CANDAR.2016.0029.

[10] Grigalis T. Towards web-scale structured web data extraction. *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM'13)*. Association for Computing Machinery. New York, 2013. P. 753–758. DOI: 10.1145/2433396.2433491.

[11] Text data mining and data quality management for research information systems in the context of open data and open science / O. Azeroual et al. *ArXiv, abs/1812.04298*. 2018.

[12] Kaur A., Chopra D. Comparison of text mining tools. *5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2016. P. 186–192. DOI: 10.1109/icrito.2016.7784950.

[13] Черенков И.А., Орехов С.В. Добыча данных из текстовых новостей на примере рынка полимеров. *Системи обробки інформації*. 2012. № 9 (107). С. 224–227.

[14] Humphreys A., Wang R. Automated Text Analysis for Consumer Research. *Journal of Consumer Research*. 2018. Vol. 44, Iss. 6. April 2018. P. 1274–1306. DOI: 10.1093/jcr/ucx104.

[15] Rashid S.F., Shafait F., Breuel T.M. Scanning Neural Network for Text Line Recognition. *In Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems (DAS'12)*. IEEE Computer Society, USA. 2012. P. 105–109. DOI: 10.1109/DAS.2012.77.

[16] Rizvi M.A., Shrivastava M., Sahu M. Artificial Neural Network Based Character Recognition Using Backpropagat. *Bioinformatics.* 2012. DOI: 10.24297/ijct.v3i1c.2777.

[17] Ковальчук А., Марчук Г., Марчук Д. Застосування згорткової нейронної мережі для розпізнавання рукописних символів. *Вчені записки Трійського національного університету імені В.І. Вернадського. Серія «Технічні науки»*. 2019. Т. 30 (69). Ч. 1. № 4. С. 68–73. DOI: 10.32838/2663-5941/2019.4-1/13.

[18] (2019). Handwritten Text Segmentation via End-to-End Learning of Convolutional Neural Network / J. Jo et al. arXiv:1906.05229v1 [cs.CV]. 12 Jun 2019.

[19] End-to-end text recognition with convolutional neural networks / T. Wang et al. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. Tsukuba, 2012. P. 3304–3308.

[20] Balci B., Saadati D., Shiferaw D. Handwritten text recognition using deep learning. *CS231n : Convolutional Neural Networks for Visual Recognition, Stanford University Project.* 2017.

[21] Fully Convolutional Networks for Handwriting Recognition / F. Such et al. *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018. Vol. 1. P. 86–91. DOI: 10.1109/ICFHR-2018.2018.00024.

[22] Шичкина Ю.А., Дегтярев А.Б., Коблов А.А. Технология очистки и трансформирования данных в рамках Knowledge Discovery in Databases (KDD) для ускоренного применения методов Data Mining. *CEUR Workshop Proceedings*. 2017. С. 428–434.

[23] Гульдман В. К. Населенные места Подольской губернии: (алфавитный перечень населенных пунктов губернии, с указанием некоторых справочных о них сведений). Каменец-Подольский : Тип. Подольского губ. правления, 1893. [2], II, IV, 636 с.

[24] Vincenty T. Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review*. 1975. № 23 (176). P. 88–93. DOI: 10.1179/sre.1975.23.176.88.

Дата надходження статті до редакції: 12.06.2020
Дата затвердження статті до друку: 10.07.2020