https://doi.org/10.31649/1997-9266-2025-183-6-127-138

УДК 004.89:004.912

Н. О. Шибаєва[1]
Д. С. Шибаєв[2]
С. І. Гришин[1]
М. Д. Рудніченко[1]
В. В. Вичужанін[1]

ГІБРИДНИЙ ПІДХІД ДО ПОШУКУ ТА ОБРОБКИ СКЛАДНОСТРУКТУРОВАНИХ ДАНИХ ВЕЛИКОГО ОБСЯГУ ДЛЯ ПОБУДОВИ ІНТЕГРОВАНОГО АЛГОРИТМУ АНАЛІЗУ КУЛЬТУРНОЇ СПАДЩИНИ УКРАЇНИ

[1]Національний університет «Одеська політехніка»;
[2]Приватний фаховий навчальний заклад «Одеський коледж комп'ютерних технологій та дизайну «Сервер»

Проблематика збереження та аналізу культурної спадщини України вимагає створення сучасних інтелектуальних інструментів, здатних обробляти складноструктуровані, багатомодальні та різнотипні дані великого обсягу. Традиційні методи пошуку й аналізу інформації здебільшого не враховують багатомовність архівів, наявність рукописних документів, історичних варіацій термінології та необхідність верифікації фактів, що істотно знижує ефективність інтеграції відомостей з різних джерел. Для вирішення цих проблем у роботі запропоновано гібридний підхід на базі розробки інтегрованого алгоритму обробки та аналізу даних, який поєднує парсинг інтернет-ресурсів, методи оптичного та рукописного розпізнавання текстів, технології обробки природної мови, механізми виявлення дублікатів і недостовірних фактів, а також побудову графа знань з подальшим застосуванням алгоритмів кластеризації. Особливістю системи є наявність адаптивного пошукового модуля, що забезпечує автоматичне вилучення, структуризацію та перевірку даних, а також інтерактивна мапа з геоприв'язкою діячів культурної спадщини, реалізована засобами бібліотеки Leaflet і технологій OpenStreetMap. Архітектура системи передбачає багаторівневу обробку інформації — від нормалізації, лематизації та ідентифікації сутностей до семантичного аналізу, асоціативного пошуку та формування прогнозних моделей розвитку культурних процесів. Проведені обчислювальні експерименти підтвердили ефективність запропонованого підходу, що свідчить про придатність її використання у режимі реального часу. Отримані результати демонструють перспективність розробленої інформаційної системи для створення комплексної програмної платформи збору та збереження даних культурної спадщини України. Практичне застосування гібридного підходу охоплює музейну, архівну, освітню та наукову діяльність, забезпечуючи уніфікований доступ до цифрових джерел, підвищення достовірності аналітики й розвиток інфраструктури цифрової гуманітаристики.

Ключові слова: культурна спадщина, обробка даних, аналіз даних, Knowledge Graph, NER, BigData, цифрові архіви.

Вступ

У контексті аналізу сучасної ситуації у сфері аналізу історичних даних, що є частиною цифрової гуманітаристики, спостерігається інтенсивний перехід від традиційних методів дослідження до використання інтелектуальних систем. Цей процес зумовлений експоненційним зростанням обсягів інформації, яка стосується культурної спадщини, її розпорошеністю та різнорідністю. Культурна спадщина України, налічує величезну кількість історичних джерел, архівів, документів, артефактів та інших матеріалів, є унікальним об'єктом для таких досліджень. Проте значна частина цих даних є складноструктурованою, фрагментованою, неповного або суперечливого характеру, що ускладнює її систематизацію та аналіз традиційними методами [1]. В контексті аналізу цієї

проблеми варто зауважити, що сучасні системи аналізу культурної спадщини часто виявляються неефективними через свою орієнтацію на один тип даних (наприклад, тільки текст або тільки зображення). Це, в свою чергу, не дозволяє отримувати цілісне уявлення про об'єкт дослідження та призводить до подальших фундаментальних проблем, які можна вирішити за допомогою гібридного підходу: мультимодальності, суперечливості даних та їхній семантичній неоднорідності [2]. Актуальність дослідження, щодо гібридного підходу до аналізу культурної спадщини України, зумовлена необхідністю подолання фундаментальних викликів цифрової гуманітаристики. Традиційні методи дослідження виявляються неефективними в умовах експоненційного зростання обсягів складноструктурованих даних, які включають текстові, візуальні та інші формати. Це вимагає розробки інтелектуальних систем, здатних не лише агрегувати та оцифровувати інформацію, але й аналізувати її для виявлення прихованих зв’язків та закономірностей. Такий підхід є критично важливим для збереження, систематизації та популяризації унікальної культурної спадщини України в умовах сьогодення [3]. Таким чином розробка та дослідження гібридного підходу до комплексного аналізу культурної спадщини України є актуальним завданням. Такий підхід має поєднувати ефективні методи обробки як структурованих, так і неструктурованих даних, що дозволить забезпечити комплексний підхід до їх систематизації, агрегації для проведення подальшого аналізу. Раціональною є також програмна реалізація інформаційної системи, що дозволить надати зручний функціонал для практичного впровадження запропонованого підходу для практики обробки та аналізу даних культурної спадщини . Зокрема, інтеграція гібридного підходу дозволить вирішити проблему фрагментованості та суперечливості історичних джерел, відкриваючи нові можливості для гуманітарних досліджень.

Метою роботи є розробка алгоритму аналізу культурної спадщини України на базі агрегації методів обробки структурованих та неструктурованих даних.

Завданнями роботи є:

– формалізація проблематики та аналіз наявних підходів до аналізу даних;

– розробка концепції та алгоритмічної складової процесів аналізу та обробки даних;

– реалізація інформаційної системи для апробації алгоритму аналізу даних;

– дослідження особливостей використання пропонованого алгоритму.

Визначення проблеми та аналіз наявних підходів до аналізу складних неоднорідних даних

Аналіз сучасних досліджень у сфері інтегрованого аналізу культурної спадщини свідчить про формування кількох взаємопов’язаних напрямів, які мають безпосереднє відношення до запропонованого гібридного підходу. По-перше, помітна інтенсифікація проєктів, спрямованих на цифрову стабілізацію і оперативний захист культурних матеріалів, що особливо актуалізувалося в умовах воєнних загроз і масової ризикової втрати джерел. Ініціативи міжнародного рівня, зокрема ініціативи Europeana та пов’язані робочі групи, спрямовані на підтримку українського сектора цифрової спадщини, демонструють практичну координацію ресурсів, стандартизацію метаданих та канали для обміну цифровими репліками, що значно посилює потенціал для створення уніфікованих репозиторіїв і підтримує впровадження гібридних аналітичних систем [1].

Наукові праці і практичні кейси свідчать про стійку тенденцію до впровадження семантичних технологій і графових подань знань для моделювання культурних контекстів. Розробки, спрямовані на побудову спеціалізованих онтологій і знаннєвих графів, показали свою ефективність у представленні складних міжоб’єктних зв’язків, уніфікації різнорідних метаданих та забезпеченні семантично обґрунтованих запитів. Приклади таких ініціатив і репрезентативних моделей демонструють, що інтеграція стандартів CIDOC-CRM та суміжних профілів дозволяє не лише здійснювати інтероперабельність між колекціями, а й підтримувати автоматизоване збагачення даних через зв’язування з відкритими ресурсами [2].

Дослідження у сфері мультимодальної обробки даних останніх років свідчать про прагнення до глибшої інтеграції моделей, здатних одночасно опрацьовувати текстову, візуальну та аудіальну інформацію. Роботи, що пропонують трансформерні підходи до багатомодальної ф’юзії, а також експериментальні дослідження з пізнавальними архітектурами для асоціації зображень і тексту, демонструють підходи до спільного простору представлення, що є критично важливим для відновлення контексту фрагментарних джерел і для вирішення проблеми семантичної неоднорідності. У цих працях окреслено практичні стратегії синтезу ознак різних модальностей, включно з поєднанням багаторівневих уявлень і механізмів уваги для збереження модально-специфічної інформації одночасно з виявленням кросмодальних кореляцій [3]

Наявні практичні інструменти для опрацювання історичних рукописів, архівних текстів та метаданих показують, що гібридні системи щільно поєднують автоматичні методи з людською експертизою як у процесі розмітки, так і валідації результатів.

Системи для розпізнавання рукописного тексту та інструменти для напівавтоматичної анотації дозволяють підвищити якість корпусів, водночас залишаючи місце для експертної корекції і контекстуалізації, що є критично важливим в роботі з історичними, мовно і стилістично варіативними джерелами. Практичні ініціативи з архівування українських інтернет-ресурсів історичних записів, виконані в останні роки, демонструють, що оперативні цифрові дії і краудсорсингові практики можуть суттєво посилити повноту і доступність матеріалів для подальшого аналізу [4].

Незважаючи на посилення інтересу і наявність численних експериментальних підтверджень ефективності гібридних підходів, вичерпаності проблем залишаються помітними [5]. До них належать питання узгодження інтероперабельних стандартів у національному масштабі, нестача репрезентативних відкритих мультимодальних датасетів, що відображали б локальні мовні та культурні особливості, а також виклики, пов'язані з апробацією відстежуваних метрик довіри і якісної оцінки згенерованих семантичних зв'язків [6]. До того ж зростаюча складність моделей породжує потребу в прозорих механізмах інтерпретованості й інструментаріях для інтеграції експертних знань у автоматичні ланцюги обробки [7].

Ці проблеми зумовлюють подальші напрями досліджень, які включають розробку адаптивних схем навчання на малих даних, засобів для пояснення рішень моделей та практик для масштабованої валідації результатів у співпраці з музейною та архівною спільнотами [8]. У підсумку, огляд наявних робіт підкреслює, що наукова спільнота розвиває комплементарні технології, які утворюють міцну методологічну базу для побудови запропонованого інтегрованого алгоритму. Наявні ініціативи демонструють практичну доцільність поєднання знаннєвих графів, мультимодальної ф'юзії та людського контролю як ключових компонентів екосистеми для збереження і дослідження культурної спадщини, але масштабована інтеграція цих компонентів у національному контексті вимагає додаткового теоретичного узгодження, стандартизації і емпіричної перевірки. Все це зумовлює актуальність мети роботи та сформованих завдань дослідження.

Розробка концепції та алгоритмічної складової процесів обробки та аналізу даних

Реалізація концепції пропонованого підходу передбачає побудову багаторівневої архітектури інформаційної системи, що інтегрує в себе низку взаємодіючих модулів, де на базовому рівні здійснюється збір і попередня обробка даних з різних джерел, зокрема, текстові документи, рукописи, архівні записи, фотографії, живописні полотна, аудіозаписи та тривимірні моделі архітектурних об'єктів [7]. Важливим аспектом цього рівня є нормалізація і стандартизація даних, що дозволяє забезпечити узгодженість і сумісність різних форматів та структур.

На другому рівні архітектури функціонує модуль аналізу даних, який поєднує методи класичної обробки тексту з алгоритмами машинного навчання, які дозволяють виявляти приховані закономірності, класифікувати об'єкти та прогнозувати зв'язки між елементами культурної спадщини. У межах цього модуля застосовуються алгоритми глибинного навчання для розпізнавання образів і обробки аудіо, що забезпечує автоматичне вилучення інформації з різнорідних медіа-форматів. При цьому інтеграція текстових та візуальних даних здійснюється шляхом побудови кросмодальних репрезентацій [6], що дозволяє алгоритму враховувати взаємозв'язки між різними модальностями і забезпечує комплексний аналіз культурних артефактів.

На наступному рівні архітектури розташовується модуль інтеграції даних, який забезпечує формування єдиного простору знань. Завдяки застосуванню онтологічних моделей та графових структур можлива уніфікація метаданих, забезпечується зв'язність між елементами різних колекцій і створюються умови для формування запитів, які охоплюють кілька типів даних одночасно.

Архітектура передбачає можливість підключення нових джерел даних, додавання нових алгоритмічних модулів та інтеграцію оновлених моделей семантичного аналізу. Це забезпечує довгострокову ефективність системи і дозволяє підтримувати актуальність знань в умовах динамічного розвитку цифрових колекцій та постійного надходження нових історичних матеріалів. Детальна концепція пропонованого підходу показана на рис. 1. Складова логіка розробленого гібридного алгоритму побудована на комбінації композицій математичних перетворень, логіки попередньої обробки інформації та роботі алгоритмів математичного моделювання.

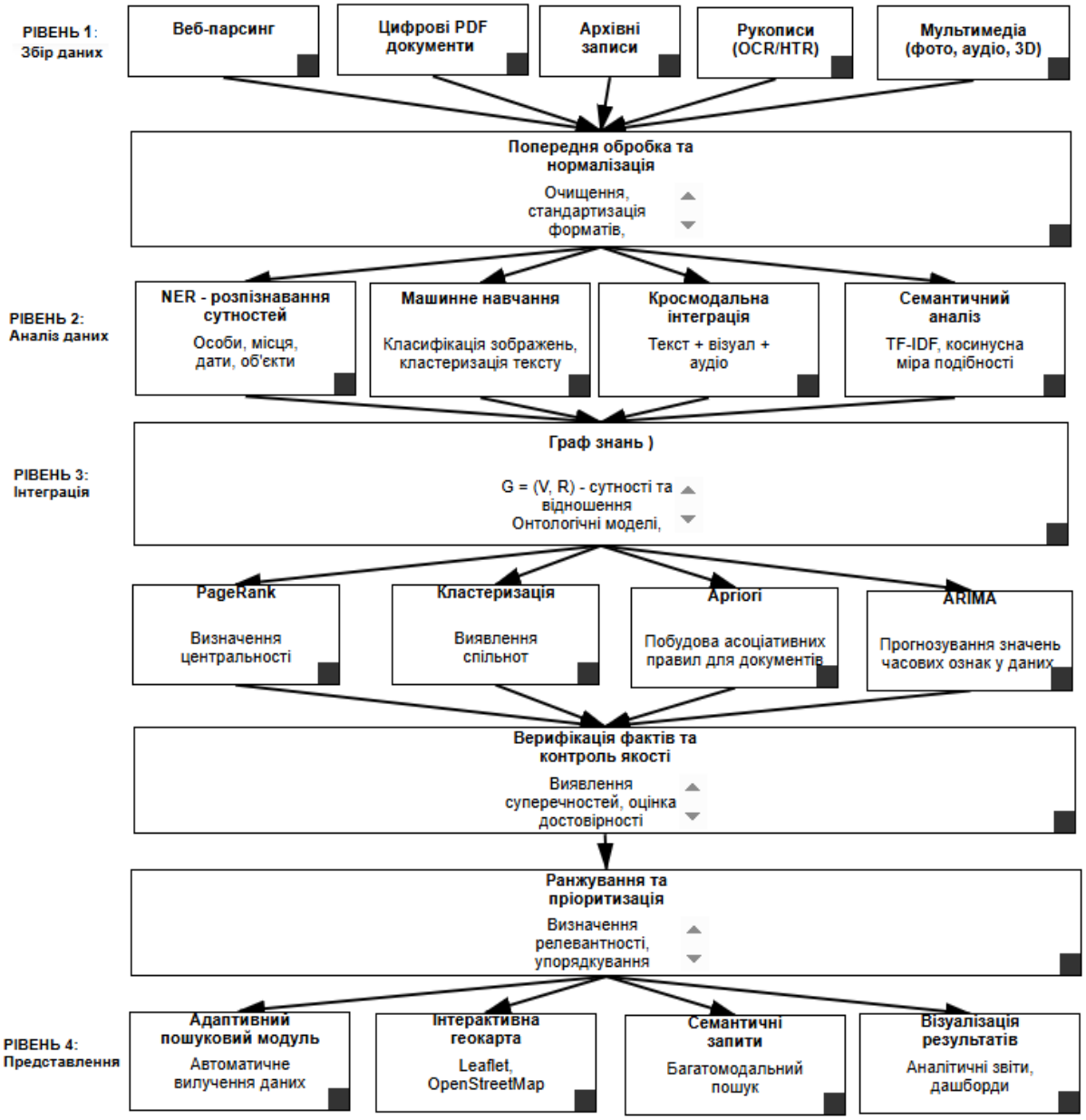

Рис. 1. Схема пропонованої концепції

Розроблений алгоритм ґрунтується на концепції багаторівневої композиції математичних перетворень, що дозволяють здійснювати повний цикл обробки складноструктурованих даних. На першому етапі відбувається формалізація вхідних джерел, що охоплюють цифрові та оцифровані документи, архівні записи і рукописи. За допомогою технологій оптичного та рукописного розпізнавання текстів вхідні дані переводяться у цифровий формат, після чого здійснюється їх очищення від шумів, нормалізація та лематизація. Ці операції створюють підґрунтя для подальшого аналітичного опрацювання і забезпечують уніфіковане представлення текстових даних

$$D = \{d_1, d_2, ..., d_n\}, d_i \in S,$$

де OCR/HTR перетворює документ у текст: $T_i = f_{\left(\frac{OCR}{HTR}\right)}(d_i)$. Очищення тексту: $T_i^{'} = f_{(clean)}(T_i)$, а фаза лематизації має таку логіку: $L_i = f_{(lemma)}(T_i^{'}) = \{l_1, l_2, ..., l_m\}$.

На наступному етапі реалізується розпізнавання сутностей: $E_i = f_{(NER)}(L_i)$, що передбачає формування множини іменованих об’єктів, включно з особами, географічними назвами, хронологічними відмітками та іншими культурно значущими категоріями. Математично це відображається через побудову відображення з простору текстових фрагментів у простір семантично значущих

сутностей, що дозволяє перевести дані на вищий рівень абстракції.

Подальший розвиток алгоритму пов'язаний з формуванням графа знань, який виступає центральною структурою інтеграції даних: $G = (V, R)$. У цьому графі вершинами є сутності, а ребрами — відношення між ними. Використання алгоритмів аналізу графів, таких як PageRank та методи виявлення спільнот, дозволяє визначати центральність, виявляти приховані кластери і структури, які неочевидні на перший погляд,

$$PageRank: PR(v) = \frac{(1-d)}{|V|} + d \cdot \sum \left(\frac{PR(u)}{\deg(u)} \right),\ u \in N(v).$$

Графові перетворення виконують роль математичної основи для формування семантично цілісної системи знань.

Застосування алгоритмів виявлення закономірностей, зокрема Apriori, дозволяє виділяти стійкі асоціативні правила між сутностями, що розширює аналітичний потенціал системи та формує базу для побудови прогнозних моделей: $X \to Y$, if $\frac{\sup port(X \cup Y)}{\sup port(X)} \geq \theta$. Динамічні аспекти аналізу відображаються через використання ARIMA-моделей [5]

$$y_t = c + \phi_1 y_{(t-1)} + ... + \phi_p y_{(t-p)} + \theta_1 \varepsilon_{(t-1)} + ... + \theta_q \varepsilon_{(t-q)} + \varepsilon_t,$$

які забезпечують виявлення часових залежностей, оцінку тенденцій і можливість прогнозування розвитку культурно-історичних процесів. Результати обробки піддаються процедурі ранжування, яка ґрунтується на методах зважування термів TF-IDF — $w_{\{i,j\}} = tf_{i,j} \cdot \log\left(\frac{N}{df_i}\right)$ та оцінці семантичної близькості за допомогою косинусної міри — $\cos(\theta) = \frac{(A * B)}{(|A||| B|)}$. Це дозволяє визначати найрелевантніші елементи у масиві даних, здійснювати пріоритизацію інформації та забезпечувати ефективну навігацію у сформованій системі знань. Завершальним етапом є верифікація фактів, що реалізується у вигляді функції достовірності: $Trust(E) = \sum w_s \cdot l_s(E),\ s \in S$, яка оцінює ступінь узгодженості інформації, виявляє потенційні суперечності та забезпечує контроль якості даних. Верифікація виконує роль механізму інтеграції автоматизованого аналізу з експертною оцінкою, створюючи баланс між алгоритмічною ефективністю та науковою надійністю результатів. Узагальнення описаних перетворень дозволяє подати розроблений алгоритм як композицію математичних операторів, що послідовно реалізують оцифрування, семантичне структурування, графову інтеграцію, асоціативний аналіз, часові прогнози, ранжування та перевірку достовірності: $M: D \to T \to T' \to L \to E \to G \to PR, C, Apriori, ARIMA$. Така композиція забезпечує можливості розподіленої обробки великих обсягів складноструктурованих даних.

Розроблений алгоритм, у разі його інтеграції до віртуального програмного оточення, дозволяє створити багаторівневий конвеєр обробки даних, що забезпечує інтеграцію складноструктурованих джерел у єдину аналітичну інформаційну систему. Це дозволяє забезпечити автоматизації процесу обробки та аналізу даних шляхом поєднання роботи модулів адаптивного пошуку, багатоструктурного парсингу та верифікації даних, що здійснюється окремим модулем парсингу. У випадку з обробкою даних, що сформована у вигляді HTML-контенту додатково застосовуються модулі вилучення тексту та метаданих з набору тегів, які враховують ієрархічну структуру документа, включаючи дерево об'єктної моделі. Це дає змогу зберігати семантичні маркери, що забезпечують точніше відтворення логіки документа з подальшою класифікацією сутностей.

Для роботи з динамічними веб-ресурсами використовується рендеринг у середовищі headless-браузерів та обробка асинхронних запитів, що дозволяє отримувати актуальну інформацію, приховану за клієнтськими сценаріями. Під час збору даних з API-ресурсів реалізується автоматизована генерація запитів, обробка пагінації та контроль швидкості звернень. Отримані дані перетворюються з форматів JSON або XML у внутрішні структури, які зберігають атрибути сутностей та їхні зв'язки, що створює основу для їх інтеграції у базу знань. Значну увагу приділено роботі з PDF-документами та графічними матеріалами.

Попередня обробка зображень, включно з бінаризацією, шумоподавленням і корекцією перспективи, підвищує якість розпізнавання. Деталізована логіка роботи модуля парсингу показана на рис. 2.

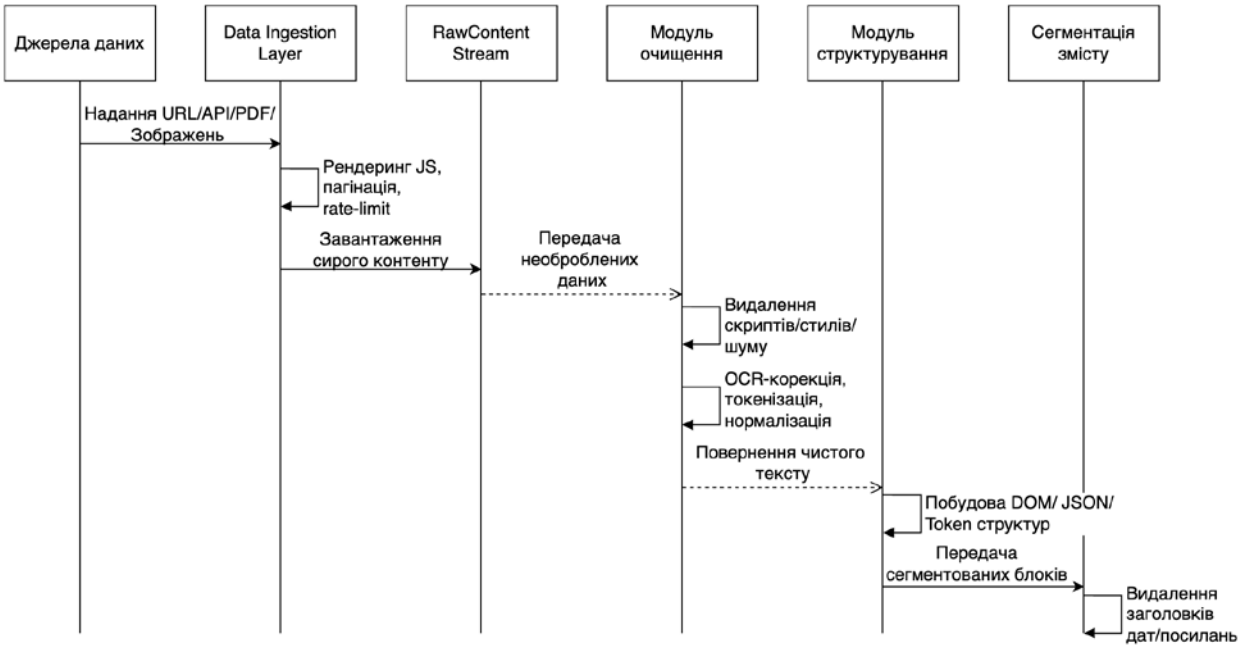

Рис. 2. Схема деталізованого процесу роботи ключових компонентів модуля парсингу даних

Усі проміжні результати використання модуля парсингу зберігаються у структурах, які забезпечують семантичний пошук і подальшу інтеграцію, зокрема асоціативних колекціях багатовимірних словників [4]. Окремо текстові дані кодуються у вигляді векторних представлень, що дає можливість здійснювати пошук на основі змістовної подібності. Для збереження напівструктурованої інформації використовуються універсальні формати, здатні відображати як атрибутивні характеристики сутностей, так і їхні зв'язки. Критично важливим є застосування механізмів кешування, які оптимізують обробку великих потоків інформації та забезпечують високу продуктивність системи.

На базі отриманих результатів функціонування конвеєра виконується пошук дублікатів і перевірка достовірності даних. Виявлення повторів базується на поєднанні методів розмитого порівняння, семантичного аналізу та кластеризації. Це дозволяє не лише вилучати точні дублікати, а й групувати близькі за змістом фрагменти, що знижує рівень інформаційного шуму. Додатково реалізується пріоритизація джерел залежно від їхньої надійності та кількості підтверджень на 3 категорії (низький, середній та високий). Перехресна перевірка фактів забезпечує узгодженість інформації між різними джерелами та дозволяє виявляти хронологічні чи логічні суперечності. Завдяки цьому формується очищена база знань, що складається з унікальних та перевірених записів.

Завершальною фазою є інтеграція структурованих даних до модуля побудови графа знань. Цей процес передбачає створення вузлів, які відображають сутності різних типів, та ребер, що репрезентують відношення між ними. Ієрархічна структура графа дозволяє відтворювати складні взаємозв'язки, включно з організаційними структурами та хронологічними ланцюгами подій. Використання алгоритмів аналізу графів надає можливість визначати ключові сутності, виявляти приховані спільноти та здійснювати пошук шляхів між віддаленими об'єктами.

Система розроблена з урахуванням вимог масштабованості та продуктивності, що забезпечує її здатність працювати з великими обсягами даних у режимі реального часу. Механізми шардінгу та реплікації дозволяють підтримувати цілісність графа зі зростанням обсягів інформації, а оптимізація запитів і використання кешування гарантують швидкий доступ до даних навіть за умов високої навантаженості. Додатково впроваджені інструменти моніторингу забезпечують контроль за ефективністю системи та створюють можливість для її подальшого вдосконалення.

Реалізація інформаційної системи для апробації алгоритму аналізу даних

На базі розробленої формалізації концепції пропонованого алгоритму доцільним є його програмна реалізація в рамках окремої інформаційної системи. На базі її використання формується вихідний граф знань в рамках вибраної категорії культурної спадщини. Враховуючи складну логіку, велику кількість даних та динамічність процесів, які можуть впливати на кінцевий результат роботи системи, виконано проєктування компонентної логіки для програмного рішення (рис. 3).

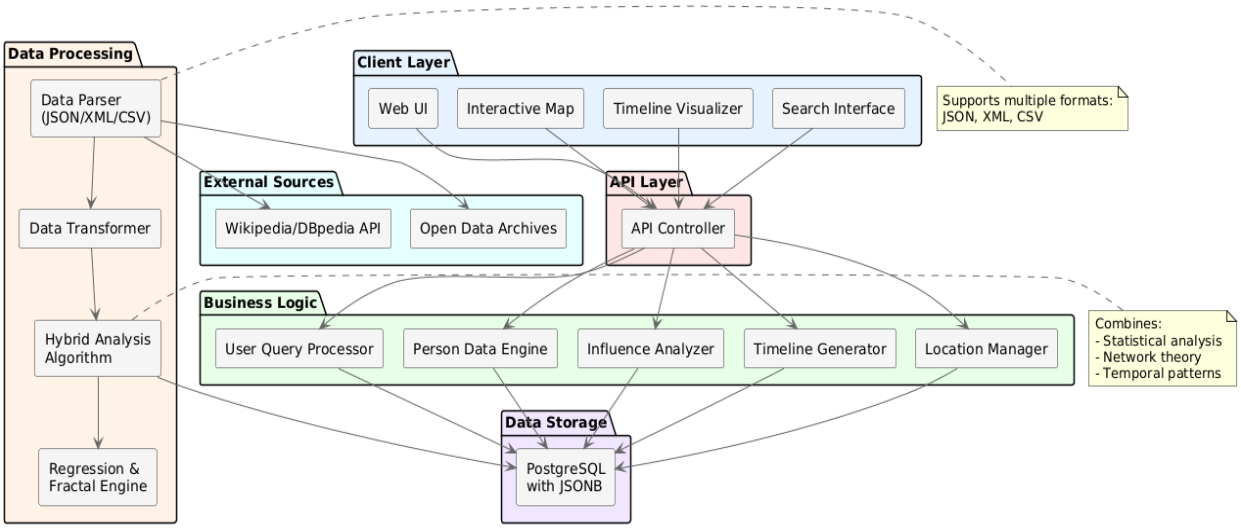

Рис. 3. Діаграма компонентів інформаційної системи

Фінальна реалізація інформаційної системи побудована у вигляді веб-застосунку, що поєднує бекенд аналітичної логіки з інтерактивним фронтендом. Такий підхід дозволяє забезпечити доступність платформи незалежно від місця перебування користувача, створює умови для роботи в реальному часі та підтримує високий рівень інтерактивності під час дослідження даних, рис. 4. Практична цінність інтерактивної мапи проявляється у можливості простежувати динаміку культурно-історичних зв’язків у часі, коли користувач вибирає певний хронологічний інтервал, після чого маркери діячів автоматично перебудовуються відповідно до епохи.

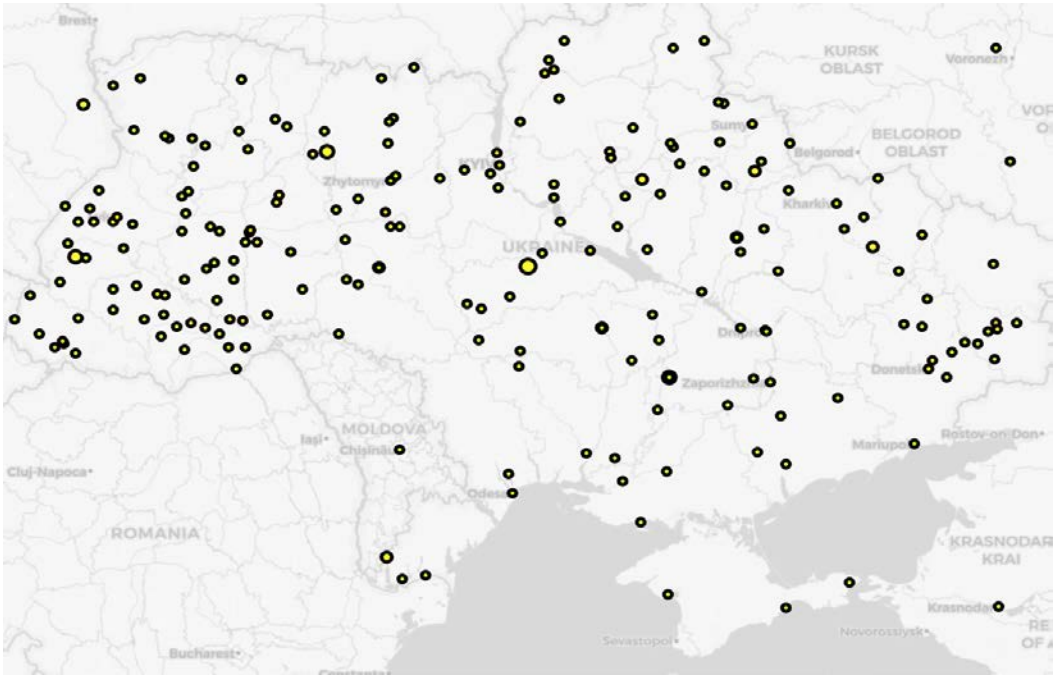

Рис. 4. Інтерфейсна частина розробленого програмного рішення

Наприклад, у разі перемикання шкали 1850–1900 рр. система відображає лише тих персоналій, чия життєдіяльність припадає на вибраний період, а рух повзунка часу демонструє появу та зникнення культурних центрів і зміщення їхніх географічних осередків. Така реалізація забезпечує не лише статичну прив’язку даних до місця, а й дозволяє виявляти історичну еволюцію культурних процесів, що посилює аналітичний потенціал дослідження та підвищує прикладну цінність візуалізаційного модуля. Особливе місце у візуалізації займає модуль GeoMapper, реалізований на основі бібліотеки Leaflet з використанням картографічних даних OpenStreetMap. Це рішення вибране завдяки його універсальності, можливості офлайн-роботи та підтримці інтеграції власних картографічних шарів. Інтерактивна мапа відображає маркери діячів культурної спадщини відповідно до

місць народження або інших геоприв'язок. Користувач може переглядати біографічні атрибути, джерела інформації та здійснювати перемикання між часовими шкалами, що дозволяє простежити еволюцію культурних процесів у динаміці.

Верифікація даних інтегрована безпосередньо у бекенд-запити, що забезпечує фільтрацію інформації ще на етапі обробки. Модуль перевірки дублікатів застосовує метод TF-IDF у поєднанні з косинусною мірою для виявлення семантично подібних записів. Завдяки цьому користувач працює вже з очищеними даними, де дублікати або суперечливі відомості усуваються автоматично.

Бекендова частина реалізована на Python з використанням FastAPI як основного веб-фреймворка. Така архітектура дозволяє масштабувати систему під зростаючі обсяги даних і водночас підтримувати асинхронну обробку запитів. Для роботи з великими масивами передбачено використання черг завдань, що забезпечує стабільність та рівномірний розподіл ресурсів.

Дослідження особливостей використання пропонованого алгоритму

Для експериментальних досліджень отримано доступ до платформ Europeana, Wikimedia Ecosystem та OpenStreetMap, з яких зібрано біля 70 % даних (включають документи, літописи, рукописи, фотографії, картини, карти та числові характеристики, а також метаінформацію), інші 30 % даних сформовано на базі автоматизованого парсингу з близько 350 вітчизняних веб-ресурсів. У перелічених платформах здійснено відбір тематично релевантних інформаційних носіїв за ключовими параметрами належності до української культурної спадщини з подальшим вилученням цифрових репрезентацій через API доступ. Після отримання первинних даних виконувалася їхня очистка та впорядкування, що забезпечувало зменшення шумових компонентів і приведення до єдиної текстово-форматної структури. Далі відбулася інтеграція матеріалу у спільний інформаційний простір за рахунок зіставлення сутностей, узгодження ідентифікаторів і відтворення логічних зв'язків між одиницями знань. В результаті сформовано масив структурованих записів, що репрезентує узгоджений датасет, сумісний із графом знань, NER-системою та модулями прогнозної аналітики.

Таблиця 1

Опис структуру зібраних даних на базі використання розробленого модуля парсингу

Тип даних	Джерело	Формати збереження	Обсяг, GB	Частка у загальному наборі, %
Текстові документи (літописи, оцифровані рукописи, архівні описи)	Wikisource, Europeana	TXT, JSON, PDF-OCR	4,15	29,53
Візуальні матеріали (фотографії артефактів, картини, картографічні зображення)	Wikimedia Commons, Europeana IIIF	JPEG, PNG, TIFF	3,8	27,04
Геопросторові та числові дані про об'єкти спадщини	OpenStreetMap Overpass-WD Linked	GeoJSON, CSV, PARQUET	2,5	14,23
Текстові та мультимедійні дані	Веб-ресурси у вільному доступі	JSON, TXT, JPEG	4,1	29,20

Запропонований алгоритм поєднує етапи OCR/HTR-розпізнавання, лематизації, нормалізації та ідентифікації сутностей (NER), що забезпечує уніфікацію та структуризацію даних. Для оцінки ефективності реалізованої інформаційної системи проведено обчислювальні експерименти, спрямовані на дослідження її роботи у реальних умовах функціонування веб-застосунку. Основна увага приділялася перевірці коректності алгоритмічних модулів, масштабованості конвеєра обробки даних та інтеграції результатів у середовище.

На першому етапі досліджувалася робота адаптивного парсингу, який забезпечує автоматизоване отримання даних із різних джерел. Підтверджено, що система здатна коректно розпізнавати як статичні HTML-сторінки, так і динамічний контент, що генерується асинхронними запитами. Середній час вилучення даних із джерел середньої складності становив 1,8 секунди, тоді як для динамічних ресурсів з рендерингом через headless-браузери цей показник зростав до 3,4 секунди. Обробка запитів до API у випадках з великою кількістю сторінок показала стабільність роботи: середня швидкість становила 120 запитів за хвилину без перевищення обмежень сервера.

Другим об'єктом дослідження став модуль перевірки дублікатів і достовірності фактів. Використання TF-IDF у поєднанні з косинусною мірою дозволило виявляти семантично близькі записи з середньою точністю 92 %, при цьому рівень хибнопозитивних результатів не перевищував 7 %. Для корпусу обсягом у 100 тисяч документів середній час повної перевірки становив близько 14

хвилин, що є прийнятним показником для інтеграції у конвеєр з підтримкою черг завдань.

Подальші експерименти зосереджено на роботі візуалізаційного модуля GeoMapper. Інтерактивна мапа, реалізована на Leaflet з використанням OpenStreetMap, показала здатність у реальному часі відображати понад 5 тисяч маркерів без відчутних затримок у роботі інтерфейсу. Під час навантажувального тестування на рівні 500 одночасних користувацьких сесій середній час відгуку інтерфейсу залишався у межах 0,9 секунди, що підтверджує ефективність застосованих механізмів кешування та оптимізації запитів. Оцінка масштабованості бекендової логіки, реалізованої на Python з використанням FastAPI, показала, що система здатна обробляти понад 2 тисячі запитів на хвилину за середнього часу відповіді 320 мілісекунд. Використання черг завдань забезпечило рівномірний розподіл навантаження, і навіть у пікові моменти час відповіді не перевищував 1 секунди. Фрагмент побудованого графу вихідного графу знань на базі використання розробленого алгоритму засобами реалізованої інформаційної системи показано на рис. 5.

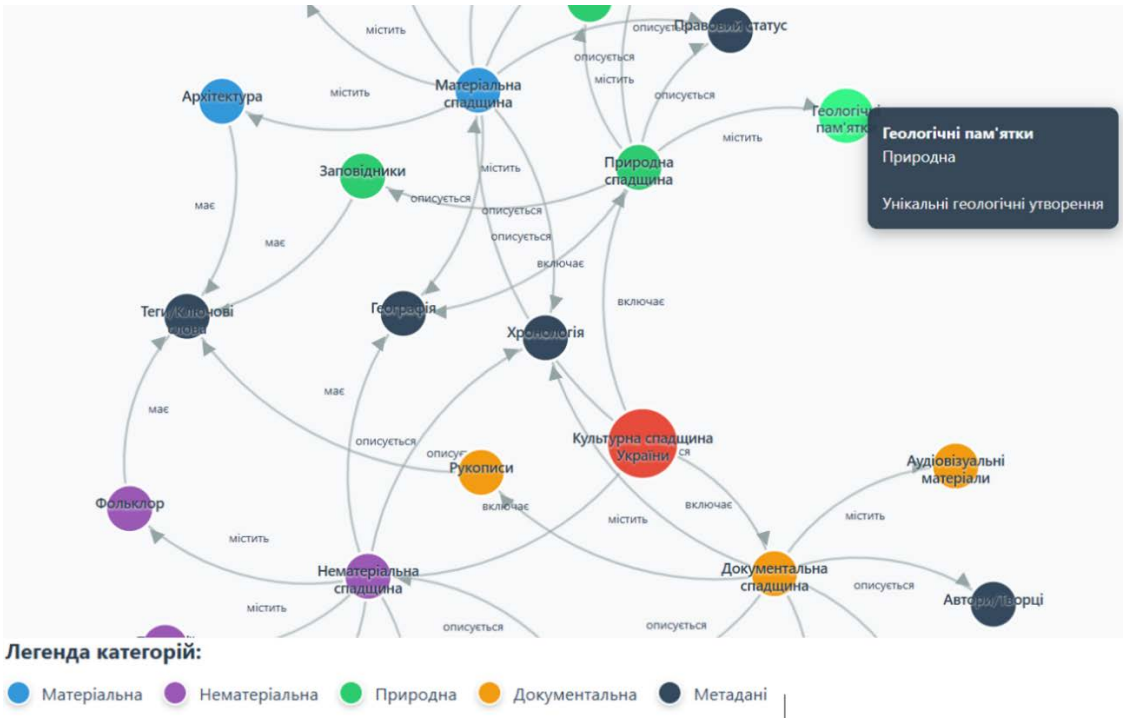

Рис. 5. Фрагмент побудованого графу вихідного графу знань

Додатково проведено експерименти з інтеграцією користувацьких сценаріїв. Встановлено, що формування пошукового запиту, його проходження через модулі адаптивного пошуку, перевірки фактів, інтеграції у граф знань та відображення результату на мапі становить в середньому 2,7 секунди для корпусу невеликого розміру і до 5,2 секунди для великих колекцій. Це підтверджує, що система є придатною для інтерактивної взаємодії, коли користувач очікує результатів у реальному часі. Для наочності результати обчислювальних експериментів подано у вигляді табл. 2, що узагальнює ключові показники ефективності основних модулів системи.

Таблиця 2

Результати експерименту роботи системи

Модуль / Етап	Час обробки, с	Пропускна здатність, запитів/хв	Точність / Помилки, %
Адаптивний парсинг (статичні сторінки)	1,8	—	—
Адаптивний парсинг (динамічні сторінки)	3,4	—	—
API запити (великі колекції)	—	120	—
Перевірка дублікатів та фактів (100 000 документів)	14 хв	—	92 / хибно-позитив ≤ 7
Візуалізація (5000 маркерів, 500 сесій)	0,9	—	—
Бекенд FastAPI	0,32	2000	—
Повний сценарій (малий корпус)	2,7	—	—
Повний сценарій (великий корпус)	5,2	—	—

Ця таблиця включає середній час обробки запитів та сценаріїв, пропускну здатність модулів в роботі з великими обсягами даних, а також показники точності перевірки дублікатів і достовірності фактів. Як видно з поданих даних, адаптивний парсинг забезпечує ефективне вилучення як статичного, так і динамічного контенту із зовнішніх джерел з часом обробки від 1,8 до 3,4 секунд, що підтверджує швидкодію механізму під час роботи з різними типами ресурсів. Модуль перевірки фактів і дублікатів демонструє високий рівень точності (92 %) за відносно низького рівня хибнопозитивних результатів (≤ 7 %), що свідчить про надійність алгоритмів семантичного аналізу на великому корпусі документів. Результати тестування візуалізаційного модуля GeoMapper підтверджують здатність системи відображати значні обсяги даних у режимі реального часу без суттєвих затримок інтерфейсу. Аналіз бекендової логіки показав високу пропускну здатність (до 2000 запитів на хвилину) зі збереженням стабільного часу відповіді. Таким чином, проведені обчислювальні експерименти підтвердили працездатність і ефективність розробленої архітектури. Вони показали, що система здатна виконувати інтегровані завдання збору, верифікації та візуалізації даних з високим рівнем точності, залишаючись стійкою до високих навантажень і масштабованою в умовах роботи з великими масивами культурологічної інформації.

Порівняння з відомими аналогами, зокрема Europeana, Archives Portal Europe та OpenHeritage, показує, що більшість існуючих систем забезпечують доступ до цифрових фондів, однак не реалізують інтегрованого мультимодального аналізу з використанням графів знань і механізмів часової аналітики. Запропонований гібридний підхід з реалізованим алгоритмом вирізняється можливістю автоматичного вилучення даних із неструктурованих джерел, динамічною побудовою семантичних зв’язків та підтримкою прогнозування культурно-історичних тенденцій.

Таким чином, наукова новизна роботи полягає у створенні інтегрованого гібридного підходу, що поєднує мультимодальне вилучення, семантичну обробку та верифікацію складноструктурованих даних культурної спадщини з подальшим формуванням онтологічного графа знань по пріоритетних категоріях. Здійснено поєднання адаптивного багаторівневого парсингу, OCR/HTR-розпізнавання, NER-аналізу, графових алгоритмів у межах єдиної обчислювальної архітектури. Запропонований підхід забезпечує підвищену точність обробки та інтеграції даних з різнорідних джерел, що дозволяє формувати узгоджені аналітичні представлення культурно-історичних даних та відкриває можливість їх прогнозного аналізу.

Висновки

Запропонований алгоритм в рамках гібридного підходу забезпечує повний цикл обробки даних — від їх попередньої підготовки до глибокого аналізу та виявлення прихованих закономірностей. Цей підхід дозволяє ефективно долати ключові проблеми, такі як лакунарність, фрагментованість та суперечливість історичних джерел, які є типовими для цієї галузі.

На практичному рівні реалізовано інформаційну систему у вигляді веб-застосування з модулями для обробки текстів, зображень, аудіоджерел та інтеграції даних, що забезпечує узгоджене представлення культурної спадщини. Запропоноване рішення поєднує автоматизовані етапи обробки та аналізу даних, підвищуючи достовірність та повноту інформації. Це відкриває можливості створення цифрових архівів, інтерактивних баз знань і платформ для міждисциплінарних досліджень, сприяючи збереженню та популяризації культурного надбання. Практична значущість полягає у застосуванні в музейній справі, архівознавстві, освіті та формуванні інфраструктури цифрової гуманітаристики, що інтегрує українську культурну спадщину у міжнародний науковий і освітній простір.

Подальший розвиток пропонованого підходу передбачає його масштабування та поглиблення, основним напрямом може бути інтеграція додаткових мультимодальних джерел даних, таких як аудіозаписи народної творчості або 3D-моделі архітектурних пам’яток, що розширить спектр досліджуваної культурної спадщини. Це потребує доповнення, вдосконалення та оптимізації алгоритмів для роботи з великими обсягами даних, а також використання розподілених обчислювальних ресурсів для підвищення ефективності. З наукової точки зору, подальші дослідження можуть бути спрямовані на вдосконалення механізмів верифікації фактів та вирішення проблеми семантичної неоднорідності. Це включатиме розробку алгоритмів, що використовують технології блокчейну для створення довірених записів про походження артефактів, а також застосування передових методів обробки природної мови для уточнення значень історичних термінів у динамічному контексті.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] S. Barzaghi, A. Moretti, I. Heibi, and S. Peroni, “CHAD-KG: A knowledge graph for representing cultural heritage objects and digitisation paradata,” *arXiv preprint*, 2025. [Electronic resource]. Available: https://arxiv.org/abs/2505.13276. Accessed: 09-Oct-2025.

[2] M. T. Biagetti, “An ontological model for the integration of cultural heritage information: CIDOC-CRM,” *Italian Journal of Library*, 2016. [Electronic resource]. Available: https://www.cidoc-crm.org/Resources/an-ontological-model-for-the-integration-of-cultural-heritage-information-cidoc-crm. Accessed: 09-Oct-2025.

[3] H. El-Hajj and M. Valleriani, “Representing and validating cultural heritage knowledge graphs in CIDOC-CRM ontology,” *Future Internet*, vol. 13, no. 11, p. 277, 2021, https://doi.org/10.3390/fi13110277 .

[4] M. Puren, and P. Vernus, “Towards a domain ontology for the analysis of ancient fabrics: The SILKNOW Project and the case of European silk heritage,” *arXiv preprint*, 2021. [Electronic resource]. Available: https://arxiv.org/abs/2112.15341. Accessed: 09-Oct-202.

[5] P. Fafalios, A. Kritsotaki, and M. Doerr, “The SeaLiT Ontology — an Extension of CIDOC-CRM for the Modelling and Integration of Maritime History Information,” *arXiv preprint*, 2023. [Electronic resource]. Available: https://arxiv.org/abs/2301.04493. Accessed: 09-Oct-2025.

[6] Z. Wang, and H. Song, “A fusion model for artwork identification based on convolutional neural networks and transformers,” *arXiv preprint*, 2025. [Electronic resource]. Available: https://arxiv.org/abs/2502.18083. Accessed: 09-Oct-2025.

[7] T. Fan, H. Wang, and S. Deng, “Intangible cultural heritage image classification with multimodal attention and hierarchical fusion,” *Expert Systems with Applications*, vol. 231, 2023, https://doi.org/10.1016/j.eswa.2023.120555 .

[8] H. El-Hajj, and M. Valleriani, “CIDOC2VEC: Extracting information from atomized CIDOC-CRM humanities knowledge graphs,” *Information*, vol. 12, no. 12, p. 503, 2021, https://doi.org/10.3390/info12120503 .

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Стаття надійшла до редакції 30.10.2025

Шибаєва Наталя Олегівна — канд. техн. наук, доцент, доцент кафедри інформаційних технологій, e-mail: n.o.shybaieva@op.edu.ua ;

Шибаєв Денис Сергійович — викладач приватного фахового навчального закладу «Одеський коледж комп’ютерних технологій та дизайну «Сервер», e-mail: denscreamer@gmail.com ;

Гришин Сергій Іванович — канд. техн. наук, доцент, доцент кафедри інформаційних технологій, e-mail: grishin@op.edu.ua ;

Рудніченко Микола Дмитрович — канд. техн. наук, доцент, доцент кафедри інформаційних технологій, e-mail: nickolay.rud@gmail.com ;

Вичужанін Володимир Вікторович — д-р техн. наук, професор, завідувач кафедри інформаційних технологій, e-mail: v.v.vychuzhanin@op.edu.ua .

Національний університет «Одеська політехніка», Одеса

N. O. Shibaeva[1]
D. S. Shibaev[2]
S. I. Grishin[1]
M. D. Rudnichenko[1]
V. V. Vychuzhanin[1]

Hybrid Approach to Searching and Processing of Complex Structured Big Data for Building an Integrated Algorithm for Ukraine’s Cultural Heritage Analyzing

[1]National University “Odesa Polytechnic”;

[2]Private professional educational institution “Odessa College of Computer Technologies and Design “Server”

The issue of preserving and analyzing Ukraine’s cultural heritage requires the development of the advanced intelligent tools capable of processing complex, multimodal, and heterogeneous data. Traditional methods of information retrieval and analysis often fail to account for the multilingual nature of archives, the presence of handwritten and poorly digitized documents, historical variations in terminology, and the necessity of fact verification, which significantly reduces the effectiveness

of data integration from diverse sources. To address these challenges, this study proposes a hybrid approach that combines multilevel web parsing, optical and handwritten text recognition (OCR/HTR), natural language processing (NLP) techniques, mechanisms for detecting duplicates and unreliable facts, and the construction of a knowledge graph employing clustering algorithms, PageRank, Apriori, and ARIMA. A distinctive feature of the proposed system is an adaptive search module enabling automated extraction, structuring, and verification of data, as well as an interactive map with geospatial visualization of cultural heritage figures, implemented using the Leaflet library and OpenStreetMap technologies. The architecture of the system supports multilayer data processing — from normalization, lemmatization, and named entity recognition to semantic analysis, associative search, and predictive modeling of cultural and historical dynamics. Computational experiments confirmed the efficiency and scalability of the approach, demonstrating stable system performance in real-time conditions. The obtained results highlight the potential of the developed model as the foundation for a unified national information and retrieval system for Ukraine's cultural heritage. The practical value of this hybrid framework extends to museum studies, archival science, education, and digital humanities research, ensuring standardized access to cultural data, enhancing analytical reliability, and fostering the integration of Ukrainian heritage into the global digital ecosystem. Further development of the system may involve the incorporation of multimodal data sources such as 3D models, audio archives, and blockchain-based provenance verification to strengthen data authenticity and long-term digital preservation.

Keywords: cultural heritage, data processing, data analysis, Knowledge Graph, NER, BigData, digital archives.

Shibaeva Natalia O. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of Information Technologies, email: n.o.shybaieva@op.edu.ua ;

Shibaev Denys S. — Lecturer of the Private Professional Educational Institution "Odessa College of Computer Technologies and Design "Server", e-mail: denscreamer@gmail.com ;

Grishin Serhiy I. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of Information Technologies, e-mail: grishin@op.edu.ua ;

Rudnichenko Mykola D. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of Information Technologies, e-mail: nickolay.rud@gmail.com ;

Vychuzhanin Volodymyr V. — Dr. Sc. (Eng.), Professor, Head of the Chair of Information Technologies, e-mail: v.v.vychuzhanin@op.edu.ua