УДК 519.23
DOI https://doi.org/10.24144/2616-7700.2026.49(2).238-244

В. І. Моренець

Київський національний університет ім. Т. Шевченка,
асистент кафедри прикладної статистики,
кандидат фізико-математичних наук, асистент
vmorenets@knu.ua
ORCID: https://orcid.org/0009-0006-0899-4400

PIPELINE-ПІДХІД ДО АВТОМАТИЗОВАНОГО ЗБОРУ ТА ПОПЕРЕДНЬОЇ ОБРОБКИ ВІДКРИТИХ ДАНИХ

У роботі запропоновано pipeline-підхід до автоматизованого збору та попередньої обробки відкритих даних, що інтегрує етапи збору, очищення, трансформації та оцінювання якості в єдиний ітеративний процес. Наукова новизна полягає у трьох складових: (1) формалізована система методологічних принципів (відтворюваність, модульність, розширюваність, вимірюваність), адаптована до гетерогенних відкритих джерел; (2) ітеративний механізм feedback loop, що забезпечує керування процесом обробки на основі кількісної оцінки якості; (3) математична модель оцінювання якості даних $Q(D)$ на основі стандарту ISO/IEC 25012 з теоретично обґрунтованими властивостями обмеженості, монотонності та коректності граничних випадків. Підхід апробовано на реальних даних порталу відкритих даних України (data.gov.ua): з використанням CKAN API зібрано 95 наборів метаданих та проведено їх обробку за основними етапами pipeline. Інтегральний показник якості зріс з $Q(D) = 0.7720$ до $Q(D) = 0.8222$ (приріст $\Delta Q = +6.5\%$). Найбільший внесок у покращення якості забезпечив етап очищення, тоді як показник актуальності залишився низьким ($Q_{\text{tim}} = 0.2886$) через значний середній вік даних (1031 день). Отримані результати демонструють, що запропонований підхід є практично ефективним інструментом для підвищення якості відкритих даних та виявлення системних проблем їх актуальності, що обмежують можливості подальшого аналітичного використання.

Ключові слова: відкриті дані, вебскрейпінг, API, pipeline, якість даних, feedback loop, ISO 25012, CKAN.

1. Вступ. Відкриті дані відіграють важливу роль у задачах аналізу соціально-економічних процесів, підтримці прийняття рішень та розвитку цифрових сервісів [1, 8]. Значна частина таких даних доступна через вебресурси або API, що створює передумови для автоматизації їх збору та обробки. Водночас відкриті дані характеризуються високою гетерогенністю структури, відсутністю єдиних стандартів представлення та нерегулярністю оновлення, що істотно ускладнює їх інтеграцію та подальше аналітичне використання.

Традиційні підходи до обробки даних, зокрема ETL-архітектури, орієнтовані переважно на корпоративні середовища зі стабільною схемою даних та контрольованими джерелами. У контексті відкритих даних такі підходи мають низку обмежень: відсутність ітеративного механізму контролю якості, недостатня адаптивність до гетерогенних і нестабільних джерел, а також відсутність формально обґрунтованої моделі оцінювання якості.

Таким чином, виникає потреба у розробці pipeline-підходу, який поєднує автоматизований збір даних, їх інтеграцію та попередню обробку з формалізованим оцінюванням якості та механізмом зворотного зв'язку.

Метою роботи є розробка pipeline-підходу до автоматизованого збору та попередньої обробки відкритих даних, а також його апробація на реальних даних

порталу відкритих даних України.

Наукова новизна роботи полягає у трьох складових. По-перше, запропоновано ітеративну архітектуру обробки даних з механізмом feedback loop, що забезпечує адаптивне керування pipeline на основі показника якості. По-друге, сформовано систему методологічних принципів — відтворюваність, модульність, розширюваність і вимірюваність, — адаптовану до умов гетерогенних відкритих джерел. По-третє, розроблено математичну модель оцінювання якості даних $Q(D)$ на основі стандарту ISO/IEC 25012 з аналітично доведеними властивостями обмеженості, монотонності та коректності граничних випадків.

Практична значущість роботи підтверджується результатами апробації на реальних даних порталу data.gov.ua, де продемонстровано зростання інтегрального показника якості даних та виявлено системні проблеми їх актуальності.

2. Постановка задачі. Нехай задано множину відкритих джерел даних

$$S = \{S_1, S_2, \ldots, S_k\}.$$

Результат збору даних визначається як об'єднання інформації, отриманої з кожного джерела:

$$D_{\text{raw}} = \bigcup_{i=1}^{k} C(S_i), \tag{1}$$

де $C(S_i)$ — оператор збору даних з джерела S_i.

Метою є побудова оператора попередньої обробки P, який забезпечує перетворення сирих даних у якісний набір:

$$D_{\text{clean}} = P(D_{\text{raw}}), \tag{2}$$

тобто реалізується послідовність перетворень

$$S \xrightarrow{C} D_{\text{raw}} \xrightarrow{P} D_{\text{clean}}.$$

Основною вимогою є підвищення якості даних:

$$Q(D_{\text{clean}}) \geq Q(D_{\text{raw}}), \quad \Delta Q = Q(D_{\text{clean}}) - Q(D_{\text{raw}}) \to \max. \tag{3}$$

Для досягнення цієї мети вводиться ітеративна процедура обробки даних з механізмом зворотного зв'язку (feedback loop), що дозволяє адаптивно коригувати параметри обробки залежно від отриманого значення $Q(D)$.

3. Опис pipeline-підходу. Запропонований підхід автоматизованого збору та попередньої обробки відкритих даних базується на інтегрованому pipeline-підході та позитивістській дослідницькій парадигмі. Його метою є формування узгодженого, відтворюваного та придатного для аналізу набору даних D_{clean} з гетерогенних відкритих джерел.

Підхід ґрунтується на чотирьох ключових принципах. По-перше, принцип відтворюваності передбачає, що кожен етап обробки є формалізованим і може бути відтворений незалежним дослідником за фіксованих параметрів. Це забезпечується явним визначенням операторів збору C та обробки P, а також параметрів pipeline.

По-друге, принцип модульності означає, що pipeline представлено як сукупність функціонально незалежних компонентів (збір, інтеграція, очищення, трансформація, оцінювання), кожен з яких може бути змінений без впливу на інші етапи.

По-третє, принцип розширюваності забезпечує можливість додавання нових джерел даних або критеріїв оцінювання якості без зміни загальної архітектури системи, що є критично важливим для відкритих даних з їх динамічною природою.

По-четверте, принцип вимірюваності якості передбачає використання інтегрального показника $Q(D)$ для оцінювання результату кожного етапу обробки. У випадку, якщо $Q(D) < Q_{\min}$, активується механізм зворотного зв'язку, який ініціює повторну обробку даних із скоригованими параметрами.

На відміну від класичних ETL-підходів, запропонований підхід реалізує ітеративну архітектуру. Після кожного циклу обробки обчислюється показник якості $Q(D)$, і у разі невиконання умови $Q(D) \geq Q_{\min}$ параметри обробки адаптивно змінюються, після чого цикл повторюється. Такий підхід перетворює pipeline з лінійного на адаптивний і забезпечує поступове покращення якості даних.

Структурно підхід включає п'ять основних етапів: збір даних (через API або вебскрейпінг), інтеграцію (узгодження форматів і структур), очищення (усунення пропусків і помилок), трансформацію (нормалізація та формування похідних ознак) та оцінювання якості (обчислення $Q(D)$). Послідовність цих етапів утворює єдиний інтегрований процес обробки відкритих даних.

4. Математична модель оцінювання якості даних. Для формалізації процесу оцінювання якості введемо інтегральний показник $Q(D)$, що визначається як зважена сума часткових критеріїв:

$$Q(D) = \sum_{i=1}^{n} w_i Q_i(D), \quad \sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0, \tag{4}$$

де $Q_i(D)$ — часткові показники якості, w_i — відповідні вагові коефіцієнти.

Відповідно до стандарту ISO/IEC 25012 [11] обрано чотири критерії, що є критичними для відкритих даних і можуть бути обчислені без зовнішніх еталонів: повнота, узгодженість, унікальність та актуальність.

Формально визначимо відповідні показники.

$$Q_{\text{comp}}(D) = 1 - \frac{N_{\text{missing}}}{N_{\text{total}}}, \tag{5}$$

$$Q_{\text{cons}}(D) = \frac{N_{\text{consistent}}}{N_{\text{total}}}, \tag{6}$$

$$Q_{\text{uniq}}(D) = 1 - \frac{N_{\text{duplicates}}}{N_{\text{total}}}, \tag{7}$$

де N_{total} — загальна кількість записів, N_{missing} — кількість пропущених значень, $N_{\text{consistent}}$ — кількість записів, що задовольняють правила узгодженості, $N_{\text{duplicates}}$ — кількість дубльованих записів.

Актуальність даних визначається як експоненціально спадна функція часу:

$$Q_{\text{tim}}(D) = \frac{1}{N}\sum_{i=1}^{N} e^{-\lambda(t_{\text{current}}-t_i)}, \quad \lambda > 0, \tag{8}$$

де t_i — момент останнього оновлення запису, λ — параметр, що визначає швидкість втрати актуальності.

Інтегральний показник якості має вигляд:

$$Q(D) = w_1 Q_{\text{comp}} + w_2 Q_{\text{cons}} + w_3 Q_{\text{uniq}} + w_4 Q_{\text{tim}}. \tag{9}$$

Вагові коефіцієнти можуть визначатися різними способами: рівновагово ($w_i = 1/n$), експертно (метод AHP) або адаптивно залежно від характеристик вхідних даних.

Побудована модель має такі властивості. По-перше, обмеженість: $0 \le Q(D) \le \le 1$, оскільки кожен частковий критерій належить інтервалу $[0, 1]$, а $Q(D)$ є їх опуклою комбінацією. По-друге, монотонність: покращення будь-якого з критеріїв не зменшує значення $Q(D)$. По-третє, коректність граничних випадків: для ідеального набору даних $Q(D) = 1$, тоді як при наявності системних дефектів (пропуски, дублікати, застарілість) значення $Q(D)$ зменшується відповідно до їх внеску.

5. Архітектура інтегрованого pipeline та алгоритм обробки даних. Запропонована архітектура pipeline реалізує ітеративний процес обробки відкритих даних з використанням механізму зворотного зв'язку. Основною відмінністю від класичних ETL-підходів є інтеграція оцінювання якості безпосередньо у цикл обробки, що дозволяє адаптивно керувати параметрами процесу.

Pipeline включає п'ять основних етапів: збір даних, інтеграцію, очищення, трансформацію та оцінювання якості. Збір даних здійснюється переважно через API, що забезпечують доступ до структурованих даних у форматах JSON або XML [5], тоді як вебскрейпінг використовується як альтернативний підхід у разі відсутності API [2]. Етап інтеграції передбачає узгодження форматів і структур даних, очищення — усунення пропусків, дублікатів та аномалій [10], а трансформація — приведення даних до узгодженого вигляду відповідно до принципів "tidy data" [9].

Ключовим елементом архітектури є механізм feedback loop. Після виконання повного циклу обробки обчислюється інтегральний показник якості $Q(D)$. Якщо умова $Q(D) \ge Q_{\min}$ не виконується, параметри обробки коригуються, після чого цикл повторюється. Таким чином, забезпечується ітеративне покращення якості даних.

Алгоритм обробки даних можна подати у формалізованому вигляді. Нехай задано множину джерел $S = \{S_1, \ldots, S_k\}$, порогове значення якості $Q_{\min}$ та максимальну кількість ітерацій max_iter. Початково формується набір даних D_{raw} шляхом об'єднання результатів збору з усіх джерел.

Далі виконується ітеративний процес:

$$D^{(t)} \xrightarrow{\text{integrate}} D_1^{(t)} \xrightarrow{\text{clean}} D_2^{(t)} \xrightarrow{\text{transform}} D_3^{(t)}, \tag{10}$$

$$Q^{(t)} = Q(D_3^{(t)}). \tag{11}$$

Якщо $Q^{(t)} < Q_{\min}$, параметри обробки змінюються:

$$\theta^{(t+1)} = f(\theta^{(t)}, Q^{(t)}), \tag{12}$$

після чого ітерація повторюється. Процес завершується при виконанні умови $Q^{(t)} \geq Q_{\min}$ або досягненні обмеження $t = max\_iter$.

Результатом роботи алгоритму є очищений набір даних D_{clean} та відповідне значення показника якості $Q(D)$.

6. Апробація на реальних даних. Для перевірки ефективності запропонованого підходу проведено експеримент на даних порталу відкритих даних України data.gov.ua. Збір здійснювався через публічний CKAN API [12] станом на 10 квітня 2026 року.

Початковий набір даних D_{raw} містив $N = 95$ записів із 15 атрибутами, що відповідають метаданим відкритих наборів даних різних органів державної влади. Дані отримано шляхом виконання запитів типу `GET /api/3/action/package_show`. Для оцінювання якості використано рівновагові ваги $w_i = 0.25$ та параметр $\lambda = 1/365$, що відповідає річному горизонту актуальності.

Для обчислення показника узгодженості Q_{cons} застосовано набір правил, зокрема: непорожність ключових полів, коректність значення стану (`state = active`), наявність ресурсів та узгодженість їх кількості. У початковому наборі виявлено 19 записів із порушеннями, що зумовило значення $Q_{\text{cons}} = 0.8000$.

Результати застосування pipeline наведено у таблиці 1.

Таблиця 1.

Динаміка показників якості даних

Етап	Q_{comp}	Q_{cons}	Q_{uniq}	Q_{tim}	$Q(D)$
D_{raw}	0.9993	0.8000	1.0000	0.2886	0.7720
Інтеграція	1.0000	0.8000	1.0000	0.2886	0.7722
Очищення	1.0000	1.0000	1.0000	0.2886	0.8222
Трансформація	1.0000	1.0000	1.0000	0.2886	0.8222

Загальний приріст якості становив $\Delta Q = 0.0502$, що відповідає відносному зростанню на 6.5%.

Аналіз результатів показує, що найбільший внесок у підвищення якості зробив етап очищення, на якому було усунуто порушення узгодженості, що призвело до зростання Q_{cons} з 0.8000 до 1.0000. Етап інтеграції мав незначний вплив, тоді як трансформація не змінила значення інтегрального показника $Q(D)$, але підвищила аналітичну придатність даних.

Ключовим обмеженням якості виявилася актуальність даних: значення $Q_{\text{tim}} = 0.2886$ залишалося незмінним на всіх етапах. Це пояснюється тим, що значна частина наборів не оновлювалась тривалий час (середній вік записів становить 1031 день). Таким чином, покращення цього показника можливе лише за рахунок повторного збору даних, що підтверджує доцільність використання механізму feedback loop.

7. Висновки. У роботі запропоновано та апробовано pipeline-підхід до автоматизованого збору та попередньої обробки відкритих даних, який відрізняється від класичних ETL-підходів інтеграцією оцінювання якості у процес

обробки та використанням ітеративної архітектури з механізмом зворотного зв'язку.

Ключовим результатом є побудова математичної моделі оцінювання якості даних $Q(D)$ на основі стандарту ISO/IEC 25012, що враховує повноту, узгодженість, унікальність та актуальність даних. Запропонована модель має властивості обмеженості, монотонності та коректності граничних випадків, що забезпечує її теоретичну обґрунтованість і придатність для практичного застосування.

Запропонована архітектура pipeline реалізує ітеративний підхід до обробки даних, у якому показник якості $Q(D)$ виступає керуючим параметром. Це дозволяє адаптивно коригувати параметри обробки та забезпечує поступове підвищення якості даних до досягнення заданого рівня.

Результати апробації на 95 реальних наборах даних порталу data.gov.ua показали зростання інтегрального показника якості з 0.7720 до 0.8222, що відповідає приросту $\Delta Q = 0.0502$ або 6.5%. Найбільший внесок у підвищення якості зробив етап очищення, який забезпечив повне усунення порушень узгодженості даних.

Водночас встановлено, що ключовим обмеженням якості відкритих даних є їх низька актуальність ($Q_{\text{tim}} = 0.2886$), зумовлена значним часом відсутності оновлень (середній вік записів — 1031 день). Це свідчить про системний характер проблеми та підкреслює необхідність періодичного повторного збору даних, що природно реалізується в межах запропонованого механізму feedback loop.

Отримані результати підтверджують ефективність запропонованого підходу для роботи з гетерогенними відкритими даними та його придатність для застосування у задачах соціального аналізу, економічних досліджень і побудови аналітичних систем.

Подальші дослідження можуть бути спрямовані на розширення набору критеріїв якості відповідно до стандарту ISO/IEC 25012, автоматизацію визначення вагових коефіцієнтів та параметра λ, а також інтеграцію методів машинного навчання для підвищення ефективності процесів очищення та обробки даних.

Конфлікт інтересів

Автор заявляє, що не мають конфлікту інтересів щодо даного дослідження, включаючи фінансовий, особистий, авторський або будь-який інший, який міг би вплинути на дослідження, а також на результати, представлені в даній статті.

Фінансування

Дослідження було проведено без фінансової підтримки.

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автор підтверджує, що при створенні даної роботи він не використовував

технології штучного інтелекту.

Список використаної літератури

1. Dong, X. L., Halevy, A. (2014). Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. *Proceedings of the 20th ACM SIGKDD.* https://doi.org/10.1145/2623330.2623623
2. Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web.* O'Reilly Media.
3. McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* O'Reilly Media.
4. Krotov, V., Johnson, L., Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems,* 47. https://doi.org/10.17705/1CAIS.04724
5. Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures.* Doctoral dissertation, University of California, Irvine. https://roy.gbiv.com/pubs/dissertation/fielding_dissertation.pdf
6. Bray, T. (2017). The JavaScript Object Notation (JSON) Data Interchange Format. *RFC 8259.* https://doi.org/10.17487/RFC8259
7. World Wide Web Consortium (W3C). Extensible Markup Language (XML) 1.0. https://www.w3.org/TR/xml/
8. Open Knowledge Foundation (2015). *Open Data Handbook.* https://opendatahandbook.org
9. Wickham, H. (2014). Tidy Data. *Journal of Statistical Software,* 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10
10. Rahm, E., Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin,* 23(4), 3–13.
11. ISO/IEC 25012:2008. Data quality model. International Organization for Standardization.
12. CKAN Association. CKAN API Documentation. https://docs.ckan.org/en/latest/api/

Morenets V. I. Pipeline-based approach to automated collection and preprocessing of open data.

This paper presents a pipeline-based approach to automated collection and preprocessing of open data. Three original components distinguish it from classical ETL approaches: (1) a feedback loop mechanism that makes the pipeline iterative; (2) a system of methodological principles for heterogeneous open data sources; (3) a data quality model $Q(D)$ based on ISO/IEC 25012 with analytically justified properties.

The proposed approach was validated on 95 real datasets from the data.gov.ua portal using the CKAN API (April 2026). The results demonstrate an improvement of the integral quality indicator from 0.7720 to 0.8222, corresponding to $\Delta Q = 6.5\%$. The dominant limitation of data quality was timeliness ($Q_{\text{tim}} = 0.2886$, mean record age 1031 days), with 51.6% of datasets not updated for more than two years. This result confirms the practical importance of the proposed feedback loop mechanism for iterative data quality improvement.

Keywords: open data, API, CKAN, pipeline, data quality, feedback loop, ISO/IEC 25012, ETL.

Отримано: 28.03.2026 Прийнято: 15.04.2026 Опубліковано: 30.04.2026