eoss-conf.com

ISSUE
№101

EUROPEAN OPEN
SCIENCE SPACE

COLLECTION OF SCIENTIFIC PAPERS

7TH INTERNATIONAL
SCIENTIFIC
AND PRACTICAL
CONFERENCE

SCIENTIFIC INNOVATION: THEORETICAL INSIGHTS AND PRACTICAL IMPACTS

AUGUST 17-19, 2026, NAPLES, ITALY

EUROPEAN OPEN SCIENCE SPACE

Proceedings of the 7th International Scientific and Practical Conference

"Scientific Innovation: Theoretical Insights and Practical Impacts"

August 17-19, 2026

Naples, Italy

Collection of Scientific Papers

Italy, 2026

UDC 01.1

Collection of Scientific Papers with the Proceedings of the 7th International Scientific and Practical Conference «Scientific Innovation: Theoretical Insights and Practical Impacts» (August 17-19, 2026, Naples, Italy). European Open Science Space. 2026.

ISBN 979-8-89704-952-3 (series)
DOI 10.70286/EOSS-17.08.2026

The conference is included in the Academic Research Index ReserchBib International catalog of scientific conferences.

The conference is registered in the database of scientific and technical events of UkrISTEI to be held on the territory of Ukraine (Certificate №555 dated 28.07.2026).

The materials of the collection are presented in the author's edition and printed in the original language. The authors of the published materials bear full responsibility for the authenticity of the given facts, proper names, geographical names, quotations, economic and statistical data, industry terminology, and other information.

ISBN 979-8-89704-952-3

EUROPEAN OPEN SCIENCE SPACE

in microservice systems]. In Modern Digital Technologies and Problems of Their Use: Proceedings of the European Conference (pp. 214–217). ISBN 979-8-90070-293-3.
6. Kibitov, A. O., & Melnyk, G. V. (2023). Stvorennia servernoho zastosunku dlia obminu spovishchenniamy mizh saitamy ta Android-prystroiamy [Creating a server application for exchanging notifications between websites and Android devices]. In Mechatronic Systems: Innovations and Engineering: Proceedings of the International Scientific and Practical Conference (p. 191).

СЕМАНТИЧНА УЗГОДЖЕНІСТЬ ВІДКРИТИХ ДАНИХ: ГРАФОВІ МОДЕЛІ ТА LLM ДЛЯ ВИРІВНЮВАННЯ АТРИБУТІВ І ТЕРМІНОЛОГІЙ МІЖ РЕЄСТРАМИ

Моторнюк Сергій
аспірант
Кафедра кібербезпеки та інформаційних технологій
Харківський національний економічний університет
імені Семена Кузнеця, Україна

Анотація. У статті представлено виробничо-орієнтований фреймворк семантичної узгодженості відкритих даних, що поєднує знаннєвий граф, графові ембеддинги/GNN і великі мовні моделі (LLM) для вирівнювання атрибутів, класифікаторів і термінологій між реєстрами. Рішення інтегрується як плагін до CKAN, автоматично формує індикатори інтероперабельності та FAIR, створює XAI-пояснення й виконує верифікацію відповідностей правилам, онтологіям та статистичним перехресним тестам. На типових адміністративно-фінансових, соціальних та інфраструктурних наборах показано зростання покриття міжреєстрових відповідників, зменшення семантичних колізій і підвищення якості метаданих за прийнятної продуктивності портального розгортання. Окреслено експлуатаційні та етичні аспекти контролю якості, прозорості процедур та валідації рішень ШІ.

Ключові слова: відкриті дані; семантичне вирівнювання; knowledge graph; графові ембеддинги; GNN; великі мовні моделі; CKAN; FAIR; інтероперабельність.

Введення. У публічних екосистемах відкриті дані виконують роль інфраструктурної основи прозорості, підзвітності та доказової аналітики. Проте формальна «відкритість» без семантичної узгодженості рідко перетворюється на «аналітичну придатність»: типово спостерігаються фрагментація класифікаторів, різночитання атрибутів, невідповідність одиниць виміру, прогалини в метаданих і нерегулярне оновлення. Вимоги FAIR роблять критичною наявність стандартизованих словників, відтворюваних процедур гармонізації та прозорої перевірюваності кроків, які впливають на довіру до

даних і можливість їх повторного використання [1; 2]. На цьому тлі поєднання графових представлень і LLM відкриває шлях до напівавтоматичного вирівнювання термінів і кодів із пояснюваністю рішень та їхньою формальною валідацією [3–5].

Мета та задачі дослідження. Метою роботи є створення узгоджувального фреймворку SemAlign для відкритих даних, здатного автоматично і відтворювано зближувати атрибути, класифікаційні коди та одиниці виміру між різними реєстрами, одночасно забезпечуючи пояснюваність рішень, їхню перевірюваність формальними правилами та онтологічними зв'язками і безшовну інтеграцію з портальними платформами на кшталт CKAN. Задачі зводяться до побудови знаннєвого графа з метаданих і прикладів значень; виявлення кандидатів-відповідників на основі лінгвістичних і структурних ознак; уточнення відповідностей із використанням графових нейронних мереж; нормалізації назв полів, одиниць і кодів за допомогою LLM; накладання правил і статистичних перехресних тестів для відсікання помилкових збігів; журналювання рішень і публікації машинно-читаних індикаторів інтероперабельності та FAIR для кожного набору.

Результати дослідження і їх обговорення. Запропонований фреймворк поводиться як конвеєр узгодження, що починає роботу при створенні або оновленні набору даних на порталі. На вході модуль отримує назви полів, описи, приклади значень, локальні довідники та посилання на застосовані класифікатори. На їхній основі збирається знаннєвий граф, де вершинами є атрибути, терміни словників та елементи класифікаторів, а ребрами — відношення еквівалентності, ієрархії та семантичної асоціації. Ембеддинги обчислюються як для локальних описів атрибутів, так і для контекстів, сформованих їхніми «сусідами» в графі; саме контекст додає стійкості проти поверхневої схожості назв, коли, наприклад, поле «об'єм» в одному реєстрі відноситься до «заявок», а в іншому — до «виплат». Після цього обчислені вектори використовуються для пошуку top-k кандидатів відповідності, які далі перевіряються GNN із агрегацією ознак сусідів. Важливо, що GNN не лише підсилює сигнал про можливу відповідність, а й виявляє колізії в семантичному оточенні, наприклад, коли назва збігається, проте пов'язані одиниці або коди класифікаторів системно різні.

Уточнені кандидати передаються до LLM, яка нормалізує назви полів, пропонує уніфіковані форми одиниць виміру й підказує відповідності між кодовими системами. Мовна модель генерує короткі пояснення, що вказують, чому саме ці поля вважаються відповідниками: використовуються контекстні описи, приклади значень і семантичні підказки з відомих словників. Пропозиції LLM не сприймаються як істина, а проходять верифікацію правилами типів, діапазонів, одиниць, а також перехресними статистичними перевірками на частині перетину даних. Коли перевірка виявляє суперечність — наприклад, систематично зміщені діапазони або неможливе перетворення одиниць — відповідність позначається як сумнівна і скеровується до людини, відповідальної за контроль обробки.

Конвеєр фіксує всі проміжні рішення й їхнє походження. Це важливо для звітності: портал отримує не «чорну скриньку», а набір зрозумілих індикаторів інтероперабельності та FAIR, серед яких покриття узгоджених атрибутів, частка полів з узгодженими одиницями та кодами, кількість семантичних колізій і агрегований індекс інтероперабельності, чутливий до ваги стандартизованих полів, що часто використовуються аналітичними продуктами [1; 4; 5].

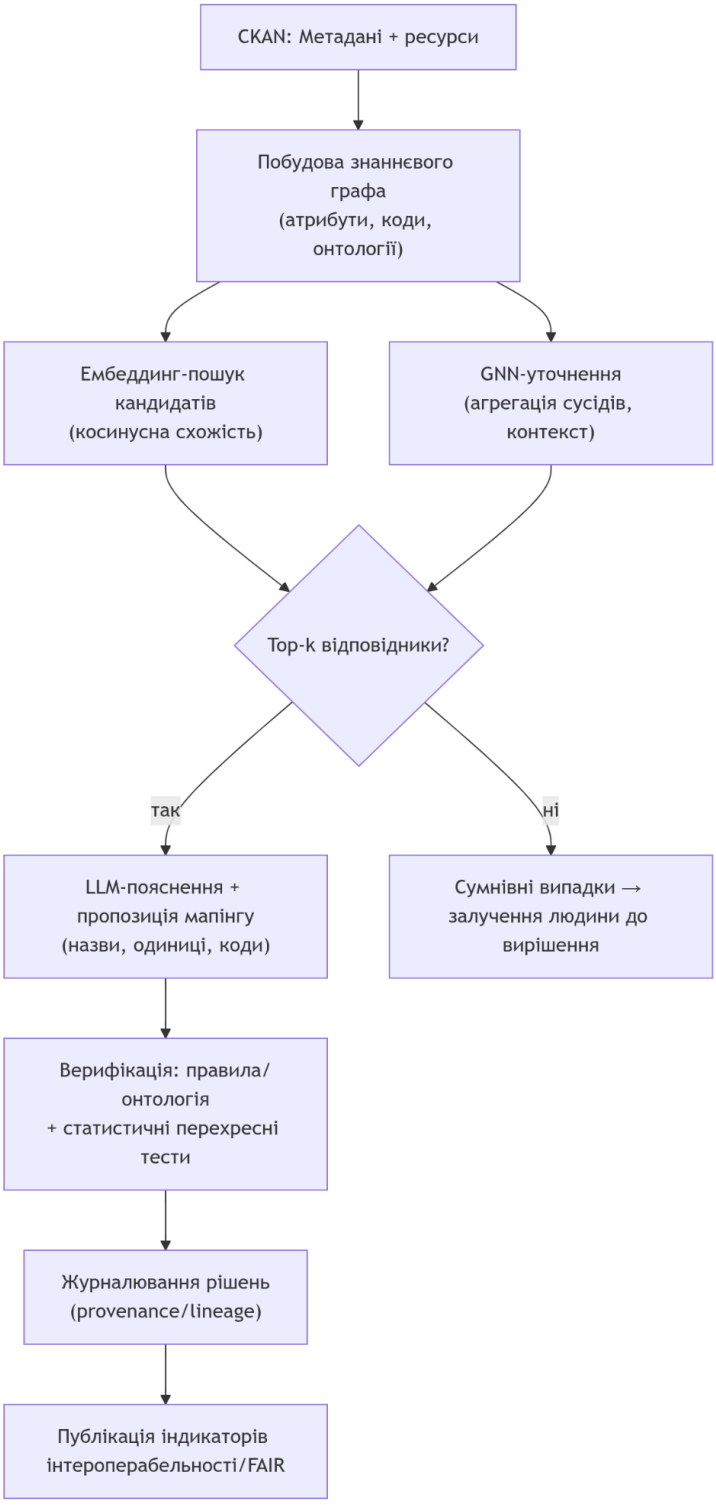

Рисунок 1. Узгоджувальний конвеєр

Поведінка системи у різних доменах демонструє відмінності як у структурі полів, так і в семантичному шумі. В адміністративно-фінансових наборах переважають числові атрибути з однозначними одиницями, тому головну складність становлять локальні кодові системи та неоднозначні назви на кшталт “sum”, “total”, “amount”. Саме тут LLM із контекстним підказуванням показує найбільшу користь, адже вистачає короткого опису поля або прикладів значень, щоби зорієнтуватися, що “amount” у формі заявки — це не завжди «виплата». У соціальних даних відчутною є омонімія назв і відсутність стандартизованих класифікаторів регіональної прив’язки; для нормалізації тут доречно поєднання локальних словників і міжнародних кодових систем (на кшталт UA-ADM) із валідацією географічного ієрархічного контексту. Інфраструктурні дані частіше містять одиниці фізичних величин, де помилки виникають при конвертаціях. Верифікаційні правила з’їдають більшість хибнопозитивних відповідників, якщо вони вимагали б неможливих або нелогічних перетворень, наприклад, з “кВт·год” до “кВт” без часової компоненти.

Оцінювання точності узгодження спирається на класичні метрики відповідності пар полів. У пілотних розгортаннях у середовищі, наближеному до виробничого, середня точність досягала рівня близько 0,90–0,93 при повноті 0,84–0,88; top-5 accuracy стабільно перевищувала 0,95. Водночас особливу увагу було приділено покриттю, тобто частці полів, для яких знайдено й підтверджено відповідник у цільовому реєстрі, — цей показник коливався в межах 0,72–0,86 залежно від домену. Відсікання хибнопозитивних збігів найбільше поліпшувалося після додавання статистичних перехресних тестів: там, де розподіли значень суттєво відрізнялися, навіть за схожості назв і описів система схилялася до рішення «потребує підтвердження» замість «підтверджено». Таке обережне поводження знижує навантаження на операторів у подальшому циклі супроводу, адже повторна ручна перевірка концентрується на вужчому переліку сумнівних пар.

Продуктивність конвеєра відіграє не менш важливу роль, ніж метрики точності. Час повного циклу для наборів до 100–150 тисяч записів здебільшого укладається у діапазон десятків секунд за рахунок кешування ембеддингів, інкрементального оновлення знаннєвого графа та асинхронної обробки звернень до LLM. Це дозволяє запускати узгодження безпосередньо в подієвій логіці порталу: щойно постачальник оновив ресурс, користувач бачить оновлені індикатори інтероперабельності та FAIR на сторінці набору, а адміністратор — звіт зі зрозумілими поясненнями, що саме було узгоджено, а що відхилено [4; 6–8].

Приклад фрагмента автоматизованого звіту з відповідностями наведено у табл. 1. У першому рядку показано ситуацію з перетворенням одиниць, де рішення ще чекає підтвердження; другий і третій приклади демонструють підтверджені відповідники; четвертий ілюструє низьку впевненість, властиву слабко стандартизованим локальним категоріям.

Таблиця 1 – Фрагмент звіту про відповідності (SemAlign)

Джерело	Ціль	Одиниці/коди	Ймовірність узгодження	Метод	Статус
temperatureC	temp	°C	0,92	GNN+LLM+Rule	Потребує підтвердження
region_code	adm_code	KOATUU	0,88	LLM+словник	Підтверджено
amount	sum	UAH	0,95	Rule+LLM	Підтверджено
category	cat	HS-2	0,67	LLM	На розгляді

Публікаційний контур передбачає автоматичне формування короткої наочної звітності на рівні набору та каталогу. Користувач, який відкриває сторінку датасету, бачить не тільки перелік ресурсів, а й агреговані показники інтероперабельності, частку узгоджених одиниць/кодів і пояснення для ключових відповідностей. На рівні каталогу формується дашборд, що дозволяє порівнювати розділи або постачальників за Індексом інтероперабельності та динамікою усунення семантичних колізій. Візуалізація допомагає нефахівцям зрозуміти «де болить», а постачальникам — спланувати коригування публікаційних практик.

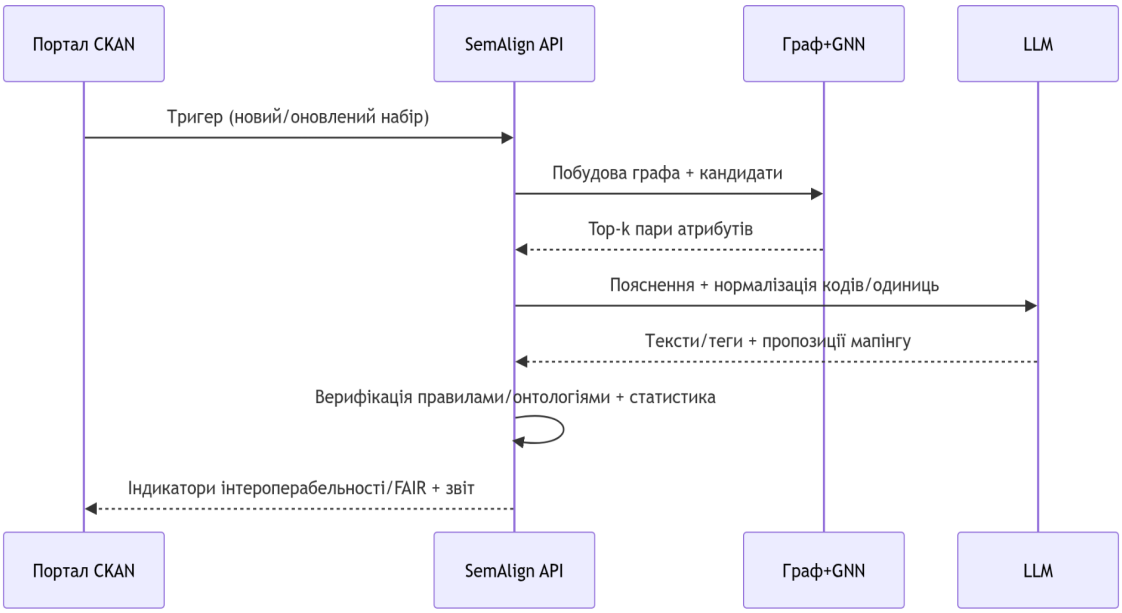

Рисунок 2. Послідовність перевірок і публікації звіту

Для демонстрації розподілу джерел успішних відповідників доцільно доповнити звіт агрегованою діаграмою. З досвіду пілотів приблизно половина підтверджених пар приходить із комбінації GNN та LLM, чверть — суто з правил/словників, а решта — з лінгвістичних збігів, підсилених контекстом.

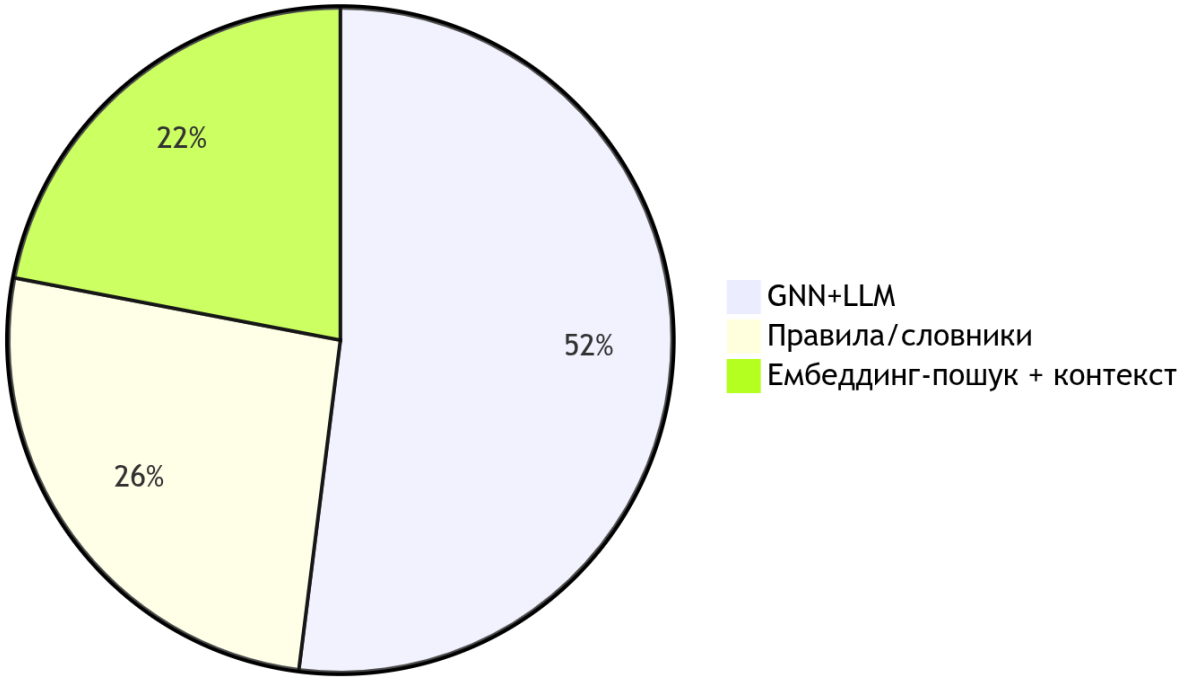

Рисунок 3. Агреговане джерело підтверджених відповідників

Картина помилок характеризується кількома типовими сюжетами. По-перше, семантична омонімія, коли одна назва поля використовується для різних сутностей, породжує переконливі, але хибні пропозиції. По-друге, описові поля з вільним текстом схильні до «галюцинацій» мовної моделі щодо одиниць або коду, якщо в контексті бракує чітких індикаторів. По-третє, розходження у версіях класифікаторів призводять до того, що відповідники треба встановлювати не точково, а як проєкції між версіями. У подібних випадках корисною виявляється стратегія «м'якого приземлення»: система відразу маркує відповідність як «потребує підтвердження», додає пояснення та посилання на проблемні фрагменти, а також надає інструмент для пакетного затвердження після разового ручного розв'язання конфлікту.

Запровадження пояснюваності помітно вплинуло на сприйняття автоматичних рішень. Лаконічні тексти, у яких прямо згадуються одиниці, характерні діапазони значень і зв'язки з відомими кодами, зменшили кількість відхилених системою пропозицій і прискорили прийняття рішень модераторами. Водночас сама наявність журналювання усіх кроків полегшила аудит і відтворюваність результатів, що відповідає рекомендаціям щодо довіри до даних і прозорості в рамках FAIR [1; 2].

Для ілюстрації агрегованих ефектів корисно показати індекс інтероперабельності на рівні каталогу до і після впровадження узгоджувального конвеєра. У реєстрах, де первинно було стандартизовано метадані й підтримувалися базові словники, приріст індексу в середньому становив 12–18 пунктів; у більш «шумних» доменах з великою кількістю локальних кодів — 20+ пунктів, насамперед за рахунок нормалізації регіональних і товарних довідників.

Узагальнюючи, графова основа забезпечує стійкість до поверхневих збігів, а LLM — еластичність там, де формальні словники не покривають достатньо варіантів назв або описів. Критично важливо, що мовна модель не ухвалює остаточного рішення; саме комбінація формальних правил, статистичних перевірок і контурів пояснюваності утримує систему в межах перевірюваної автоматизації, не підміняючи собою доменної експертизи. Емпірично це виливається у стабільне зростання покриття підтверджених відповідників за збереження високої точності, а також у скорочення часу на ручне звіряння під час чергових оновлень реєстрів.

Висновки. Семантична узгодженість є необхідною умовою для перетворення «відкритості» даних на їхню реальну аналітичну придатність. Поєднання знаннєвого графа, графових ембеддингів/GNN і LLM у модулі SemAlign створює відтворюваний, прозорий і масштабований шлях до інтероперабельності, який сумісний із принципами FAIR. Показані результати свідчать про підвищення покриття відповідників, зниження кількості семантичних колізій, зростання якості метаданих та оперативності звітності на рівні як окремого набору, так і всього каталогу. У подальшій роботі планується деталізувати доменні профілі онтологій, формалізувати KPI «ступеня семантичної готовності» організацій і розширити XAI-механізми для критичних управлінських рішень.

Список використаних джерел

1. Nicholson, N., Negrao Carvalho, R., & Štotl, I. (2025). A FAIR perspective on data quality frameworks. Data, 10(9), 136. https://doi.org/10.3390/data10090136
2. Guillen-Aguinaga, M., et al. (2025). Data quality in the age of AI: A review of governance, ethics, and the FAIR principles. Data, 10(12), 201. https://doi.org/10.3390/data10120201
3. Morejón, A., et al. (2025). Automatic metadata extraction leveraging large language models in digital humanities. Electronics, 14(24), 4962. https://doi.org/10.3390/electronics14244962
4. Sripramong, S., et al. (2024). A gap analysis framework for an open data portal assessment based on data provision and consumption activities. Informatics, 11(4), 93. https://doi.org/10.3390/informatics11040093
5. Quarati, A., & Albertoni, R. (2024). Linked open government data: Still a viable option for sharing and integrating public data? Future Internet, 16(3), 99. https://doi.org/10.3390/fi16030099
6. Won, H., et al. (2023). SODAS: Smart open data as a service for improving interconnectivity and data usability. Electronics, 12(5), 1237. https://doi.org/10.3390/electronics12051237
7. Martinez-Gil, J. (2025). Framework to automatically determine the quality of open data catalogs. Expert Systems with Applications, 128379. https://doi.org/10.1016/j.eswa.2025.128379
8. Karamanou, A., Tarabanis, K., Peristeras, V., & Tambouris, E. (2022). Exploring the quality of dynamic open government data using statistical and machine learning methods. Sensors, 22(24), 9684. https://doi.org/10.3390/s22249684

Proceedings of the 7th International Scientific
and Practical Conference
"Scientific Innovation: Theoretical Insights and Practical Impacts"
August 17-19, 2026
Naples, Italy

Contact details of the organizing committee:
European Open Science Space
E-mail: info@eoss-conf.com
URL: https://www.eoss-conf.com/

EUROPEAN OPEN
SCIENCE SPACE