

РОЗШИРЕНА СЕМАНТИЧНА АРХІТЕКТУРА ІНТЕГРАЦІЇ ОНТОЛОГІЙ

Проблема семантичної сумісності (semantic interoperability) залишається критичним бар'єром зі створення глобальних баз знань в аграрному секторі для дослідників. Попередні дослідження системи «Ontology Manager» [1] базувалися на використанні лексичного коефіцієнта Серенсена-Дайса для пошуку ідентичних концептів та аналізу їх атрибутів. Проте практика показала, що лексичні методи не здатні ефективно обробити мультимовні дані та синонімічні ряди, що призводить до втрати багатьох потенційних збігів на етапі первинної фільтрації. Метод Серенсена-Дайса базується на біграмах (спільних парах літер). Якщо терміни записані різними мовами, наприклад, англійською «Sunflower» та французькою «Tournesol», лексичний збіг буде близьким до нуля. Крім того, пряме порівняння «залишків» після лексичного відсіву створює обчислювальний затор складністю $O(N^2)$, що унеможливорює масштабування системи для інтеграції великих сховищ, таких як AgroPortal або FAO.

Для усунення обмежень лексичного аналізу пропонується впровадження рівня векторних ембедінгів (Embeddings). Кожен вузол онтології трансформується у високорозмірний вектор. Оптимальним рішенням було обрано використання мультимовних моделей, наприклад таких як OpenAI text-embedding-3-small. Така модель, завдяки розміру, дозволяє розгортати на локальних серверах, що значно зменшує вартість обробки одного запиту.

Оптимізація обробки даних у Neo4j [2]. Використання вбудованого функціоналу Vector Index дозволяє зберігати обчислені ембедінги безпосередньо у вузлах графа [3,4]. Замість повного перебору, система застосовує алгоритм К-найближчих сусідів (K-Nearest Neighbors). Найінноваційним аспектом нової архітектури є ієрархічний підхід до підтвердження збігів, де кожен наступний «експерт» є точнішим, але дорожчим у використанні. Перший рівень (Автоматичний): Моделі класу «Lightweight LLM» (наприклад, Gemini 2.5 Flash або локальні моделі Gemma 2 [5]). Другий рівень (Контрольний). Якщо «легка» модель повертає статус uncertain (невизначеність), запит передається потужнішій моделі або локальному серверу з підвищеним параметром temperature для глибшого аналізу семантичних нюансів. Третій рівень (Експерт-людина). Дослідник залучається лише для вирішення конфліктів, які не змогла однозначно

розв’язати система LLM. Графове страхування (Similarity Flooding). Для вузлів, які не продемонстрували високої векторної подібності, застосовується топологічний аналіз за допомогою мови Cypher. Порівняльний аналіз архітектурних компонентів представлено в табл.1.

Таблиця 1 – Порівняльний аналіз архітектурних компонентів

Етап процесу	Попередня версія (Sorensen-Dice + LLM)	Оновлена версія (Embeddings + LLM)	Обґрунтування змін
Фільтрація	Символьна (bigrams)	Векторна (OpenAI / Gemma)[4,5]	Підтримка мультимовності та синонімії
Пошук	Повний перебір / Cypher	Vector Index KNN у Neo4j	Прискорення на великих графах
Валідація	Ручна перевірка	Автоматична (Gemini Flash)	Зниження навантаження на експерта
Масштабування	Обмежене $O(N^2)$	Високе (індексований пошук)	Можливість інтеграції глобальних баз

Висновки. Запропонована архітектура системи «Ontology Manager» забезпечує поєднання векторних індексів Neo4j, мультимовних ембеддінгів та каскаду LLM-моделей, що дозволяє автоматизувати інтеграцію українських аграрних даних до міжнародних стандартів, забезпечуючи при цьому повний контроль з боку людини-експерта.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. S. Vladov, A. Bakurova, I. Omelianenko, R. Sirenko, E. Tereschenko. Ontology control system in Neo4j. *Proceedings of the Modern Machine Learning Technologies Workshop (MoMLeT 2025)*. CEUR Workshop Proceedings. Lviv, Ukraine, June 14-15, 2025. Vol. 4004. P. 185–196.
2. Vector Search and Indexing : Neo4j Graph Database Platform Documentation, 2024. URL: <https://neo4j.com/developer/genai-ecosystem/vector-search/> (дата звернення : 16.04.2026).
3. Neosemantics (n10s) : Neo4j Labs Project Documentation. Neo4j Graph Database Platform, 2024. URL: <https://neo4j.com/labs/neosemantics/> (дата звернення : 17.04.2026).

4. Text Embeddings models : OpenAI Documentation, 2024. URL: <https://platform.openai.com/docs/guides/embeddings>. (дата звернення : 16.04.2026).

5. Gemini 2.5 Flash Technical Report : Google DeepMind. 2025. URL: <https://deepmind.google/technologies/gemini/> (дата звернення : 16.04.2026).