УДК 004.8:004.724 https://doi.org/10.15407/pp2025.02.063

І.П. Сініцин, Ю.В. Рогушина, К.Ю. Юрченко

ІНТЕГРАЦІЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ІЗ ЗАСОБАМИ СЕМАНТИЧНОЇ ОБРОБКИ ЯК ІНСТРУМЕНТ ЦИФРОВІЗАЦІЇ ЗНАНЬ

У роботі розглядається задача автоматизації аналізу, генерації та управління складними природномовними документами на основі інтеграції генеративного штучного інтелекту із семантичними технологіями, зокрема, Semantic MediaWiki. Аналізується, яким чином застосування онтологічних моделей предметних областей та семантичної розмітки дозволяє запобігати таким критичним недолікам великих мовних моделей, як схильність до "галюцинацій" (генерації неправдивих тверджень) та відсутність прозорості у поясненні рішень.

Така інтеграція досліджується на прикладі інструментальної системи “ЛІНЗА”, яка розробляється для автоматизованої інтелектуальної обробки контенту розрізнених документів зі складною слабоформалізованою структурою з метою генерації природномовних звітів за заданими вимогами у різних галузях, таких як публічне адміністрування, юриспруденція, цифровізації знань, сертифікація та стандартизація. Система базується на поєднанні гнучкості та адаптивності великих мовних моделей із формалізованими онтологічними знаннями та підтримкою семантичних запитів щодо пертинентних фактів у середовищі Semantic MediaWiki, або зовнішніх джерел (Retrieval-Augmented Generation). Запропонований підхід дозволить значно знизити ризики помилок, типових для генеративних моделей, та забезпечити фактичну правдивість і прозорість процесу ухвалення рішень. Особлива увага приділяється механізмам прозорості, достовірності та можливості контролю людиною для підвищення довіри до згенерованих даних, що особливо важливо у сферах із підвищеними вимогами до безпеки інформації. Такий підхід також забезпечує більшу довіру до автоматично створених документів.

Багаторівнева архітектура системи характеризує задачі агентів і сервісів, що виконують спеціалізовані функції збору, аналізу, перетворення та перевірки даних, і забезпечує гнучкість, масштабованість та адаптивність системи до зміни вхідних даних і вимог.

Ключові слова: агентні технології, великі мовні моделі, LLM, Semantic MediaWiki, семантичні технології, база знань, формалізовані документи.

I.P. Sinitsyn, Yu.V. Rogushina, K.Yu. Yurchenko

INTEGRATION OF LARGE LANGUAGE MODELS WITH SEMANTIC PROCESSING TOOLS AS AN INSTRUMENT FOR KNOWLEDGE DIGITIZATION

The paper addresses the task of automating the analysis, generation, and management of complex natural language documents based on the integration of generative artificial intelligence with semantic technologies, in particular Semantic MediaWiki. It analyzes how the use of ontological models of subject domains and semantic markup makes it possible to prevent such critical shortcomings of large language models as the tendency to “hallucinations” (generation of false statements) and the lack of transparency in decision explanations.

This integration is explored using the example of the instrumental system “LINZA,” which is being developed for automated intelligent processing of content from heterogeneous documents with complex and weakly formalized structure, with the aim of generating natural language reports according to specified requirements in various domains, such as public administration, jurisprudence, certification, and standardization. The system is based on the combination of the flexibility and adaptability of large language models with formalized ontological knowledge and support for semantic queries about pertinent facts in the Semantic MediaWiki environment or external sources (Retrieval-Augmented Generation). The proposed approach will significantly reduce the risks of typical errors in generative models and ensure factual accuracy and transparency in the decision-making process.

Special attention is paid to mechanisms of transparency, reliability, and the possibility of human control to increase trust in the generated data, which is especially important in areas with high information security requirements, and ensures greater confidence in automatically created documents.

The multi-level architecture of the system defines the tasks of agents and services that perform specialized functions of data collection, analysis, transformation, and verification, and ensures flexibility, scalability, and adaptability of the system to changes in input data and requirements.

Keywords: agent technologies, large language models, LLM, Semantic MediaWiki, semantic technologies, knowledge base, formalized documents.

Вступ

Сучасні інформаційні системи дедалі частіше застосовують технології *генеративного штучного інтелекту* (ГШІ) для автоматизації рутинних завдань, зокрема, створення, підтримки та адаптації документів. У цьому контексті особливої важливості набуває розробка гібридних платформ, що поєднують не лише швидке, а й семантично узгоджене та перевірене генерування документів на основі спеціального підкласу ГШІ – великих мовних моделей (Large Language Models, LLM) з перевіреними структурами збереження знань [1].

Більшість наявних рішень, які зараз використовують штучний інтелект для автоматизованого формування документації, обмежуються використанням шаблонів або спрощених структур для опису потрібних документів [2]. Але розвиток LLM-моделей дозволяє створювати потужніші системи зі значно ширшим функціоналом, здатні генерувати комплексні, формалізовані документи з поясненням логіки ухвалених рішень.

Значна частина чинних систем на базі LLM базуються на графах знань або векторних базах, що забезпечують лише часткову семантичну підтримку [3]. Однак такі підходи мають суттєві обмеження: вони не завжди здатні пояснити результати генерації, а використання LLM без контролю призводить до ризику появи "галюцинацій" — некоректних або вигаданих даних [4].

Попри значні досягнення у галузі обробки *природної мови* (ПМ), сучасні LLM все ще виявляють певні слабкі сторони, особливо в роботі зі спеціалізованими предметними областями (ПрО), що використовують специфічну термінологію та правила побудови документів.

Це пояснюється тим, що робота LLM базується на виявленні статистичних закономірностей у великих обсягах даних, а для таких специфічних ПрО обсяг даних для обробки може бути недостатнім (або ж інформація, релевантна до цієї ПрО, не виокремлена з усього масиву даних, що аналізуються). Використання статистичних моделей дозволяє LLM ефективно генерувати стилістично коректні тексти, проте не завжди забезпечує фактичну правдивість та прозорість процесу ухвалення рішень. Це може призвести до значних перешкод у застосуванні LLM в тих сферах, де потрібен високий ступінь довіри, аудит та можливість експертної інтервенції, зокрема, у сфері підвищеної секретності.

З огляду на зазначені обмеження, виникає об'єктивна необхідність у доповненні можливостей LLM інструментами менеджменту знань. Це передбачає створення гібридної архітектури, яка поєднує різні технології обробки інформації для досягнення синергетичного ефекту. LLM можуть ефективно використовуватися для первинного аналізу великих обсягів неструктурованих текстових даних, ідентифікації ключових сутностей та формування початкових версій документів. Системи управління знаннями, такі як SMW, забезпечать формалізоване представлення витягнутих даних, дозволяючи їх структуроване зберігання, легке редагування, валідацію експертами та побудову чітких логічних висновків.

Формулювання задачі

У роботі аналізується доцільність інтеграції семантичних технологій із великими мовними моделями LLM з метою автоматизованого створення документів складної структури на основі гетерогенних даних – природномовних документів, таблиць, баз даних та знань, онтологій та тезаурусів ПрО, мультимедійної інформації тощо. А також вимог щодо подання результатів аналізу та відомостей про користувача. Результуючі документи мають відповідати формалізованим вимогам щодо складу, логіки побудови та оформлення, що визначаються нормативними або внутрішніми регламентами. Прикладами таких задач є побудова спеціального профілю захищеності інформаційної системи, уза-

гальнення досвіду діяльності в певній сфері та його впровадження, генерація персональної траєкторії навчання для здобувача освіти.

Складність кожної конкретної задачі залежить від ступеня формалізованості вхідних даних (особливо – вимог до результату); складності правил, за якими елементи вхідних даних перетворюються на елементи результуючих документів, та від обсягу бази знань ПрО, яка потрібна для побудови цих правил. Але, незалежно від цього, всі подібні задачі потребують однакового набору операцій над вхідними даними (складність задачі впливає лише на час аналізу), і тому для їх розв'язання доцільно створити універсальне інструментальне середовище, що підтримує весь потрібний функціонал. З огляду на те, що набір задач може розширюватися, а інформаційні потреби користувачів – ускладнюватись, доцільно передбачити можливості гнучкого розширення архітектури такої системи, яка дозволить поповнювати її новими модулями без змін вже існуючих.

У найбільш узагальненому вигляді ця задача має наступний вигляд: вхідними даними для аналізу є: 1) набір документів, що містить знання щодо ПрО, – як семантично формалізовані (тезауруси, онтології, бази знань), так і слабо формалізовані (природномовні описи, стандарти, різноманітні сирі дані, проаналізовані приклади тощо); 2) індивідуальні дані користувача, які конкретизують його інформаційні потреби, в тому числі – специфічні вимоги, правила оформлення контенту та термінології; 3) засоби визначення вимог до результуючих документів – формалізовані вимоги, природномовні описи, приклади.

В результаті обробки треба згенерувати документи, структура яких відповідає визначеним вимогам, а контент характеризує вказані користувачем об'єкти із застосуванням знань щодо ПрО. Відповідно до специфіки ПрО, результуючими документами можуть бути спеціалізовані профілі організацій або інформаційних систем; рекомендації, що узагальнюють відомості з сирих даних; аналітичні огляди технічної та наукової інформації, структуровані за певними правилами; оцінки діяльності підрозділів організацій тощо. Для цього необхідно розв'язати низку підзадач: виокремлення структури результуючого документа на основі наявних прикладів та описів, зв'язування на рівні семантики елементів контенту цього документа з відомостями ПрО, співставлення інформації від користувача з терміносистемою ПрО тощо. Найбільш складними елементами є коректне розпізнавання фрагментів документації, що відповідають слабо формалізованим вимогам до результату, та аналіз семантичної коректності отриманого результату.

Важливо розуміти, що використання лише логічного виведення та семантичних запитів є недостатнім для здобуття знань з ПМ-документів, а передача задачі в цілому до LLM (навіть з великою кількістю попередніх налаштувань та значною кількістю ітерацій) виявляється занадто складною для аналізу якості отриманого результату та виокремлення причин, що призвели до некоректних результатів. Тому доцільно інтегрувати в одній інформаційній системі обидві можливості, доповнивши їх гнучкими засобами менеджменту інформації – як на рівні документів, так і на рівні знань.

Але потрібно відзначити, що просте механічне поєднання одного технологічного середовища LLM та Semantic MediaWiki не вирішує поставлене завдання: потрібно чітко визначити етапи обробки інформації, формати збереження даних та моделі обміну між окремими модулям, передбачити засоби керування та зворотного зв'язку.

Така система має забезпечити функціонально повний набір сервісів для перетворення вхідної інформації на результуючий набір документів, що відповідають вимогам користувача. Технологія використання системи повинна визначати порядок застосування сервісів, коректні перетворення інформації та можливість контролю людиною всіх етапів таких перетворень з метою вчасного виявлення та запобігання семантичним неоднозначностям (рис.1).

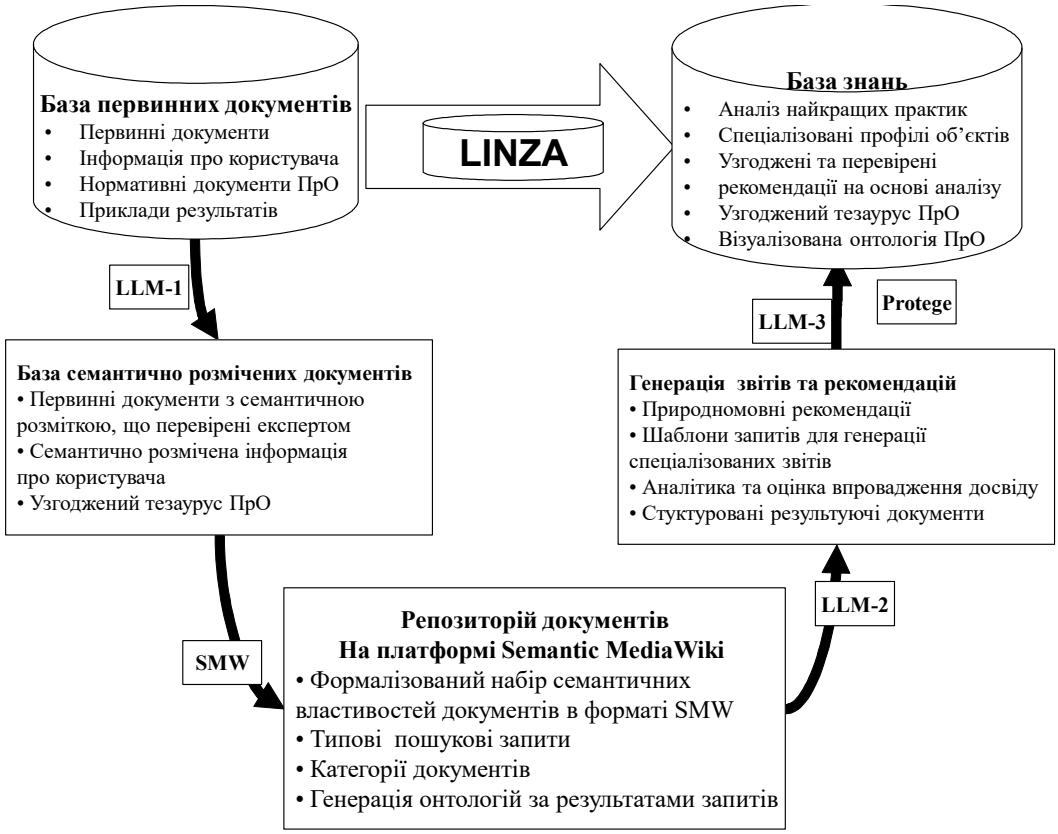

Рис. 1. Узагальнена схема технологічного середовища “Лінза”

На попередньому етапі розробки цього інтегрованого технологічного середовища потрібно чітко визначити його призначення та базовий функціонал, формалізувати основні види перетворень інформації в процесі обробки. Саме це дозволяє визначити як склад цього середовища, так і призначення окремих модулів і засобів взаємодії між ними та користувачами, охарактеризувати необхідні операції у життєвому циклі інформації (Рис.2).

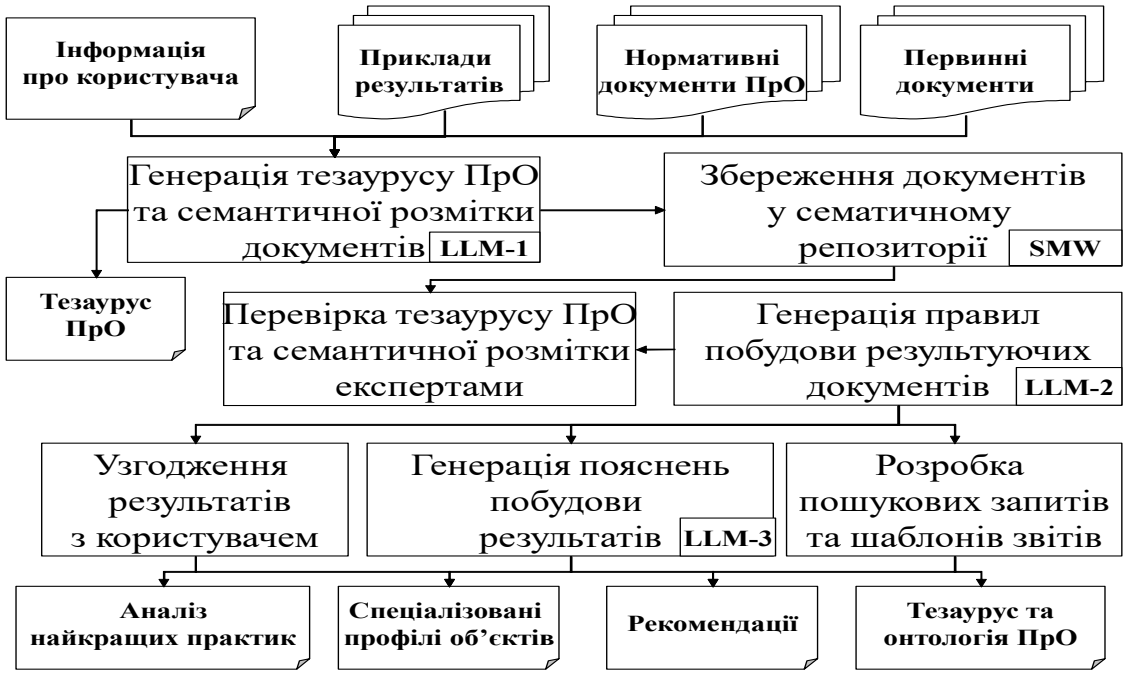

Рис. 2. Життєвий цикл інформації у технологічному середовищі “Лінза”

Для цього потрібно формалізувати вимоги до системи, яка має забезпечити автоматизацію процесу формування таких документів на основі поєднання генеративних можливостей LLM з контрольованістю та прозорістю семантичних платформ, а саме: визначити базові принципи побудови такої системи, її компоненти, механізми інтеграції джерел знань і способи забезпечення відповідності результату очікуваним формальним критеріям.

Теоретичні засади системи аналізу документів "Лінза"

Ключем до успішної реалізації гібридної системи є детермінація та структуризація вхідних і вихідних інформаційних потоків. Вхідні дані в таких задачах здебільшого представлені масивними колекціями неструктурованих чи напівструктурованих документів із ПрО. До них належать нормативні документи та стандарти, зокрема, у сферах з обмеженим або контрольованим доступом до інформації, та емпіричні описи конкретних систем. Ці джерела, як правило, представлені у вільному форматі та містять складні внутрішні структури, термінологію та взаємозв'язки, що вимагають глибокого семантичного аналізу. Обробка цих вхідних даних здійснюється за допомогою інтегрованих LLM та спеціалізованих сервісів, які виконують функції парсингу, ідентифікації ключових сутностей, виявлення реляційних зв'язків та первинної семантичної розмітки.

Вихідними даними системи є формалізовані та структуровані знання, що зберігаються у базі знань на платформі SMW, а також кінцеві згенеровані документи. Зокрема, до вихідних даних належать: семантично збагачені структури знань, де витягнута інформація трансформується у формат, сумісний з SMW (тріади, властивості та класифікації), забезпечуючи високий рівень інтероперабельності та можливість логічного виведення. Зокрема, система продукує верифіковані фрагменти або цілі документи складної структури, правдивість яких гарантується етапом валідації за участю експертів (human-in-the-loop). Також до вихідних даних належать ланцюги логічного виведення та детальні пояснення ухвалених рішень, включаючи ідентифікацію підмножини знань, використаної моделлю, що забезпечує прозорість та можливість аудиту в умовах роботи з конфіденційною інформацією. Таким чином, відбувається процес трансформації неструктурованих вхідних даних у структуровані, верифіковані та пояснювані вихідні знання, що є центральним аспектом функціонування системи, забезпечуючи її високу цінність для автоматизації складних процесів обробки інформації.

Не менш важливим є пошук ефективних технологічних рішень, зокрема, застосування агентів та сервісів для цілісного та безперервного перетворення інформації між різними компонентами системи та визначення надійних джерел отримання знань щодо ПрО. У контексті поточної роботи предметна область охоплює документи і стандарти, приклади конкретних систем та їхній опис. Критично важливим аспектом є розуміння специфіки задачі, зокрема питання щодо допустимості розміщення опису предметної області у відкритому доступі, оскільки передача чутливих відомостей до зовнішніх LLM є неприпустимою з огляду на інформаційну безпеку та конфіденційність. Це додатково доводить доцільність локального розгортання та інтеграції компонентів і застосування концепції "human-in-the-loop" для контролю та валідації.

Метою роботи є аналіз доцільності інтеграції семантичних технологій з LLM для автоматизованого створення комплексних природномовних документів, що включають текстові, графічні та табличні дані, відповідають формальним вимогам до структури та враховують специфічні запити замовника.

Аналіз конкретних прикладних задач та шляхів їх розв'язання спрямований на те, щоб дослідити вимоги до такої автоматизованої системи. Запропонований підхід має дозволити розв'язати проблему довіри та достовірності результатів, притаманні сучасному генеративному ШІ.

Дослідивши сучасний стан розробок у сфері інтелектуальної обробки інформації, ми виявили доцільність застосовувати у "Лінзі"**:**

- Агентні технології та сервіс-орієнтоване програмування [5] – для персоніфікованої динамічної обробки даних у веб-середовищі: інтелектуальні агенти дозволяють відображати цілі та задачі різних суб'єктів системи, щоб знаходити та активувати найбільш прийнятні сервіси для перетворення та аналізу інформації;
- Семантичні технології та онтологічний аналіз [6] – для семантичної інтерпретації контенту вхідних даних системи з використанням зовнішніх джерел знань (онтологій, тезаурусів, таксономій), їх структурування формалізованими мовами для інтероперабельності та автоматичної обробки;
- Великі мовні моделі – як засіб глибокого семантичного аналізу природномовних документів для співставлення їхнього контенту з онтологічними моделями ПрО класифікації та структурування (наприклад, LLM дозволяють автоматизовано генерувати семантичну розмітку сирих природномовних даних тегами, що відповідають поняттям та відношенням з обраної онтології).
- Семантичні вікі [7] – як платформи для зберігання та надання користувачам доступу до семантично структурованого контенту.

Сучасний стан досліджень у сфері інтеграції LLM із семантичними технологіями

У більшості сучасних інтелектуальних систем, що базуються на LLM, користувач позбавлений доступу до процесів структурування знань, джерел первинного контенту та логіки формування рекомендацій. Алгоритми обробки залишаються непрозорими, що обмежує довіру до системи, ускладнює валідацію результатів та викликає труднощі у експертній - перевірці.[4]

У цьому контексті особливий інтерес становить застосування вікі-технологій, зокрема, таких, як Semantic MediaWiki та WikiData, у поєднанні з LLM. Саме вікі-підхід дозволяє зробити структуру знань відкритою, зрозумілою та доступною для редагування, що, своєю чергою, пояснює причини формування системних висновків і підвищує прозорість ухвалення рішень.

У науковій літературі представлено низку проєктів, у яких здійснюється інтеграція LLM із WikiData або Wikipedia [8, 9]. Їхній аналіз виділяє ряд закономірностей, характерних для мультиагентних систем із залученням LLM.

Більшість описаних рішень реалізують мультиагентну архітектуру, де кожен агент виконує окрему спеціалізовану функцію — від семантичного аналізу . до генерації тексту [3] та перевірки правдивості даних [4] Такий підхід дозволяє формалізувати робочий процес і значно знижує ризики помилок, типових для генеративних моделей. У всіх системах застосовується *Retrieval-Augmented Generation* (RAG) [10] — техніка, що дозволяє LLM поєднувати генеративні можливості з релевантною фактологічною інформацією із семантичних структур або зовнішніх джерел. У цих підходах активно використовують техніки покрокового уточнення запитів (workflow orchestration, prompt chaining), що наближає модель до логічного міркування й багатоетапної побудови контенту та забезпечує структуровану генерацію документів на основі формалізованих знань, зниженні частоти «галюцинацій» моделей, високій гнучкості в розподілі обов'язків між агентами і можливості адаптації під різні доменні задачі. Зокрема, використання семантичних шаблонів дозволяє LLM автоматично виводити інформаційні вимоги до доку-

мента, тоді як у CLAIR граф знань забезпечує багатоетапне логічне виведення фактів для підготовки технічної документації. DocAgent використовує багатогранну систему оцінювання згенерованого результату, що дозволяє об'єктивно оцінювати повноту, корисність та фактичність документації [9].

Водночас, існують і певні обмеження. Системи залишаються чутливими до обмежень контексту моделей — навіть при використанні довгих вікон (16K+ токенів) генерація в межах великих кодових баз або графів знань може втрачати релевантність. Деякі системи, зокрема, DocAgent, орієнтовані переважно на статичний аналіз, що обмежує обробку динамічних аспектів програмних систем. Ефективність генерації значною мірою залежить від якості семантичного опису — неповні або суперечливі шаблони можуть призводити до помилок або втрати важливої інформації. Також слід зазначити, що інтеграція з платформами на кшталт SMW потребує додаткових зусиль у побудові запитів та формалізації знань у придатному для LLM вигляді.

Крім того, у сучасних дослідженнях активно вивчаються підходи до поєднання великих мовних моделей із семантичними базами знань, зокрема, у форматі Wikidata. Одним з таких прикладів є LLM Store [11], що виступає як проміжне середовище між LLM та структурованими даними, дозволяючи трансформувати природномовні запити у твердження у форматі RDF-триад [12]. Архітектурно система реалізована як плагін до KIF (Knowledge Integration Framework), що забезпечує трансляцію відповідей моделі у формат, придатний до розміщення у Wikidata. Зокрема, її ефективність була продемонстрована в задачах LM-KBC (Language Model-based Knowledge Base Construction), де LLM Store показав високі результати точності генерації. Найвищих показників F1-score (до 91%) вдалося human-in-the-loop досягти шляхом додавання контексту до запитів, що підтверджує ключову роль модуля генерації контексту. Попри це, автори підкреслюють слабке місце системи — невисоку точність для "вузьких" специфічних відносин та помилки в ідентифікації сутностей.

В іншій системі — Scholarly Wikidata, LLM використовується для напівавтоматизованого вилучення метаданих наукових конференцій із веб-сайтів та текстів матеріалів. Результатом стало суттєве збагачення бази Wikidata: додано тисячі сутностей та нових властивостей, що охоплюють організаційні ролі, прийняті статті, доповіді, тощо. Перевагою цього підходу є висока ефективність витягування структурованої інформації, однак автори вказують на схильність LLM до помилок під час роботи з датами, назвами треків або складними залежностями між сутностями. Для таких випадків необхідне втручання людини (human-in-the-loop), що підвищує загальні витрати на підтримку системи [8].

Дослідження Каверинського, Літвіна та Палагіна [13] пропонують інноваційний підхід до керованої генерації природної мови. Основна ідея роботи полягає в концепції "зворотного синтезу", яка, на відміну від традиційного аналізу природної мови, зосереджена на продукуванні текстових висловлювань з наявних формалізованих онтологічних представлень. Ці онтологічні представлення, які є центральною ланкою методології, автоматично будуються на основі аналізу речень науково-технічних текстів за допомогою раніше розроблених програмних засобів. Вони інкорпорують сутності, ідентифіковані в тексті, та типізовані семантичні зв'язки між ними, формуючи таким чином структуровану мережу знань. Для взаємодії з великою мовною моделлю, зокрема ChatGPT, автори розробили спеціально структуровані інструкції-підказки (prompts), які направляють модель для генерації тексту, що точно відповідає заданій семантичній структурі. Проведена серія експериментів із синтезу природномовних висловлювань підтвердила ефективність запропонованого підходу. Ця методологія є розв'язанням проблем, пов'язаних із неконтрольованою генерацією та низькою прозорістю LLM, забезпечуючи значно вищий рівень правдивості та прозорості згенерованих висловлювань через

їхню опору на верифіковані онтологічні структури та керовані підказки.

У контексті інтеграції великих мовних моделей із семантичними технологіями, значний інтерес становить досвід застосування семантичних вікі-систем для управління знаннями. Зокрема, [14] описує використання MediaWiki та Semantic MediaWiki для створення онтологічно-орієнтованих репозиторіїв знань, що підтверджує доцільність застосування таких платформ для формалізації предметних областей, забезпечення однозначної інтерпретації термінології та підтримки семантичного пошуку у складних, динамічних інформаційних масивах. Запропонована архітектура, що інтегрує різні джерела знань та сервіси, слугує вагомим підґрунтям для розробки гібридних систем, які прагнуть поєднати генеративні можливості LLM із верифікованими семантичними структурами.

Отже, вище зазначені системи демонструють спільну тенденцію: використання великих мовних моделей як джерела знань, які потім зберігаються у формалізованій структурі типу Wiki. Проте ці моделі мають спільне обмеження — дані, які були згенеровані великими мовними моделями, стають частиною бази знань без механізмів прозорої верифікації. Крім того, наразі, запит у пошуковій системі не дає жодного релевантного посилання, що свідчить про новизну запропонованого напряму (“LLM and SMW”). Тож можна стверджувати, що поєднання нейронних моделей із SMW утворює синергетичну модель, де LLM виступає як інтерпретатор природної мови та генератор контенту, а SMW забезпечує структуроване середовище для верифікації, представлення та розширення знань. Такий підхід дозволяє забезпечити покращення для традиційних систем, підвищуючи точність і надійність згенерованих даних завдяки формалізованим семантичним шаблонам і механізмам перевірки.

Таким чином, системи на основі мовних моделей і семантичних структур демонструють значний потенціал для автоматизованої побудови та структурування даних у різних галузях — від публічного адміністрування до програмної інженерії. Подальший розвиток доцільно спрямувати на вдосконалення механізмів інтеграції з джерелами знань, покращення інтерпретованості результатів і розширення підтримки динамічних сценаріїв використання.

Система «ЛІНЗА»: архітектура, призначення та інтелектуальні механізми

Інформаційна система лінгвістичної обробки нормативних документів «ЛІНЗА» призначена для автоматизації створення та обробки складних природномовних документів, що інтегрують розрізнені дані та відповідають жорстким формальним вимогам.

Система є розробкою Інституту програмних систем НАН України та має значний потенціал застосування в тих сферах, де критично важлива точність, структура та правдивість інформації.

“ЛІНЗУ” доцільно застосовувати в тих галузях, де необхідно забезпечити семантичне структрування великих обсягів інформації з їх подальшим аналізом та перетворенням на документи зі складною, наперед визначеною структурою, а також генерації пояснюваної інформації щодо створених документів та забезпеченні високої правдивості даних. Водночас значна частина правил перетворення та принципів сруктурування подається неявно, тобто ці знання потрібно здобувати з відповідних документів, пов’язаних між собою. Сферами застосування системи можуть бути: державне управління, де вона може значно спростити підготовку нормативно-правових актів, звітів, протоколів та іншої офіційної документації; юриспруденція – для аналізу величезних обсягів юридичних документів, ефективного вилучення ключової інформації та автоматичного формування декларацій; сертифікація та стандартизація – як іструмент автоматизованої підготовки звітної документації на вимогу сертифікаційниих органів.

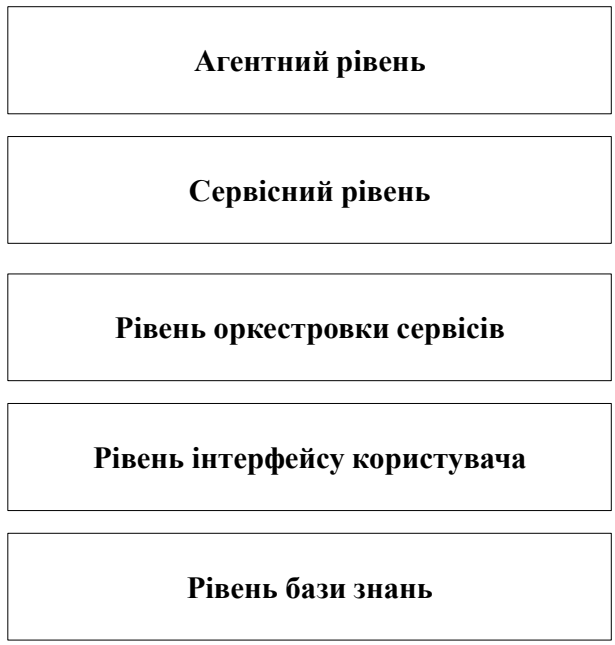

Рис. 3. Багаторівнева архітектура технологічного середовища “Лінза”

Можливість аналізу та реалізації досвіду для таких інтелектуальних систем, як "ЛІНЗА", має ключове знання. Це є основою для вдосконалення функціоналу системи: регулярний аналіз результатів її роботи, зворотний зв'язок від користувачів, експертів, що дозволяє виявляти недоліки та оптимізувати процеси. Такий підхід сприяє ефективному накопиченню знань, оскільки система дозволяє структурувати та зберігати отриманий досвід, перетворюючи його на структуровані онтології та інші формати, доступні для подальшого використання та аналізу. Крім того, у динамічних сферах, таких як захист інформації, де постійно виникають нові виклики та вимоги, постійні адаптації до змін є життєво необхідними.

“Лінза” є складною системою, що поєднує різні технології та методи обробки інформації. Тому для її розробки застосовується багаторівнева архітектура, яка дозволяє перетворити задачу розробки системи на набір простіших підзадач (Рис.3).

Взаємодія складових системи "ЛІНЗА" координується через сервіс оркестровки [15] (Orchestration Service), який координує інтелектуальну обробку документації за участі агентів і сервісів, що виконують спеціалізовані функції збору, аналізу, перетворення та перевірки даних. Такий підхід забезпечує гнучкість, масштабованість та адаптивність системи до зміни вхідних даних і вимог.

На рівні взаємодії з користувачем система реалізує веб-інтерфейс (UI - user interface), який уможливлює завантаження документів, формулювання запитів природною мовою та перегляд аналітичних результатів і згенерованих - декларацій.

Сервісний рівень виконує наступні функції: завантаження та попередню обробку документів, семантичне анотування, онтологічне моделювання, генерацію відповідей за допомогою LLM, та інтеграцію знань за підходом Retrieval-Augmented Generation (RAG). Для збереження, оновлення та валідації знань використовується Semantic MediaWiki, що дає можливість семантичного доступу через SPARQL-запити.

Такий архітектурний підхід дозволяє ефективно інтегрувати експертні знання з мовними моделями, що суттєво покращує якість аналізу, обґрунтування рішень та пояснення складних нормативних понять.

Функціональні можливості “Лінзи”

Функціональна архітектура системи "ЛІНЗА" інтегрує низку спеціалізованих модулів, кожен з яких виконує критично важливі операції в процесі інтелектуальної обробки інформації, забезпечуючи її перетворення від неструктурованих вхідних даних до формалізованих та пояснюваних знань.

Обробка вхідних документів: Відповідає за ініціальний етап життєвого циклу інформації в системі. Він здійснює інжест неструктурованих документів, представлених у типових офісних форматах (наприклад, PDF, DOCX). Основною функцією є не лише ідентифікація формату, а і їхня трансформація у стандартизований текстовий формат, придатний для подальшого автоматизованого аналізу. Додатково цей модуль інтегрує засоби оптичного розпізнавання символів (OCR) для конвертації графічних представлень тексту, а також алгоритми попереднього очищення даних, спрямовані на усунення аберацій та нормалізацію текстового контенту для забезпечення високої якості вхідної інформації для подальших етапів обробки.

Семантичне анотування (Semantic Annotation): Центральний елемент інтелектуальної обробки, що реалізує парадигму семантичного збагачення даних. Шляхом інтеграції передових LLM та семантичних технологій, цей модуль здійснює не лише ідентифікацію номенклатурних сутностей у тексті, а й виявляє складні когнітивні та функціональні зв'язки між ними. Результатом є побудова деталізованої семантичної моделі документа, що включає контекстуалізовані відносини між об'єктами знань. Ця структура зберігається у структурованому сховищі знань, реалізованому на базі Semantic MediaWiki, формуючи онтологічно збагачений репозиторій для подальшого аналізу та запитів.

Перетворення природномовних запитів (Natural Language Query Transformation): Виконує функцію інтерфейсу між природномовним запитом користувача та формалізованою базою знань. Його призначення полягає у трансляції неформальних запитів, висловлених природною мовою, у структуровані запити мовою SPARQL або ASK – мовою для запитів до SMW. Цей процес вимагає глибокого семантичного розуміння запиту користувача та його відображення на онтологічну структуру сховища, забезпечуючи високу релевантність та точність отриманих результатів.

Генерації декларацій та аналітичних звітів (Declaration and Analytical Report Generation): Даний процес синтезує отримані знання у формі, зручній для кінцевого користувача. Він використовує не лише безпосередньо витягнуті факти, а й складні механізми логічного виведення (reasoning) для формування обґрунтованих та когерентних відповідей. Застосування підходу Retrieval-Augmented Generation (RAG) дозволяє інтегрувати можливості генерації тексту LLM з доступом до верифікованих знань з внутрішньої бази даних. Це забезпечує продукування не просто відповідей, а повноцінних аналітичних звітів та декларацій, які є прозорими, посилаються на джерела та відповідають встановленим стандартам звітності.

Управління знаннями (Knowledge Management): Є фундаментом для забезпечення життєвого циклу знань у системі. Він надає функціональність для створення, персистенції, модифікації та версифікації знань, представлених у вигляді онтологій та статей Semantic MediaWiki. Онтології забезпечують формалізоване представлення предметної області, тоді як статті Semantic MediaWiki слугують зручним інтерфейсом для взаємодії зі знаннями. Система версифікації критично важлива для відстеження еволюції знань у динамічних доменах, гарантуючи їхню актуальність та цілісність.

Експертна валідація (Expert Validation Module): Ця функція впроваджує парадигму Human-in-the-Loop, забезпечуючи високий рівень достовірності результатів. На відміну від автоматизованої обробки, система надає інтерфейс для ручної верифікації та коригування витягнутих сутностей та згенерованих декларацій.

Можливість гнучкої корекції та аналізу експертами дозволяє мінімізувати потенційні помилки, підвищити точність інформації та забезпечити відповідність високим галузевим стандартам, особливо у сферах, де помилки неприпустимі.

Експорт результатів (Results Expor): Завершальний етап функціонального ланцюга, що забезпечує інтеграцію системи із зовнішніми середовищами та бізнес-процесами. Даний модуль відповідає за форматування та експорт вихідної звітної документації у стандартизованих, загальноприйнятих форматах (наприклад, PDF, DOCX). Це дозволяє безперешкодно інтегрувати результати роботи "ЛІНЗИ" в існуючий документообіг, що забезпечує можливість передачі сформованих документів до відповідних сертифікаційних або регуляторних органів.

Основні модулі системи

Сервіс-орієнтована архітектура системи "ЛІНЗА" забезпечує її модульність та гнучкість. Вона складається з чотирьох основних функціональних груп сервісів: модулі взаємодії з користувачем, сервіси обробки документів та управління LLM, сервіси управління знаннями, сервіси валідації та експорту, процеси взаємодії яких узагальнено представлено на рис. 4.

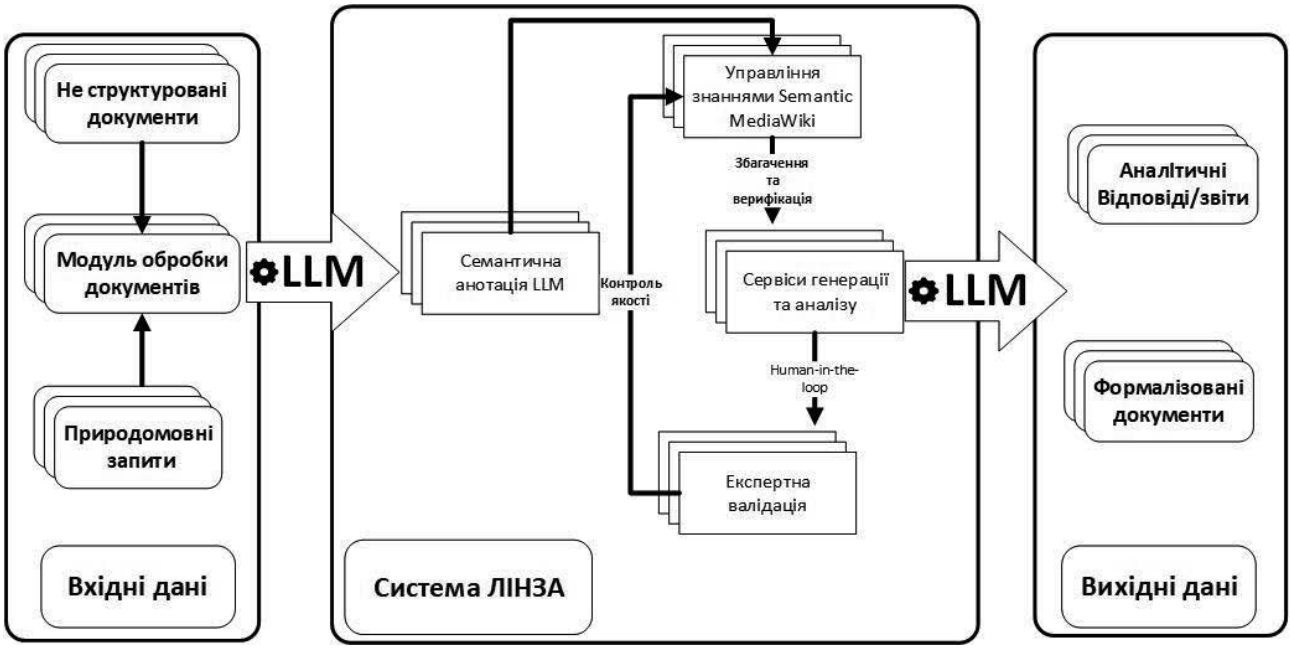

Рис. 4. Узагальнена схема функціонування системи "ЛІНЗА"

Модулі взаємодії з користувачем

- **UI Service** — це графічний інтерфейс користувача, через який відбувається завантаження документів, подання запитів, перегляд результатів та декларацій.
- **API Gateway Service** — єдина точка входу до системи, що виконує функції маршрутизації запитів, автентифікації та авторизації.

Сервіси обробки документів та управління LLM

- **Document Parsing Service** — відповідає за трансформацію вхідних файлів у структурований текст із попереднім очищенням і OCR.
- **LLM Orchestration Service** — головний координаційний модуль, що розподіляє завдання між агентами та LLM, виконує reasoning і генерацію за ReAct-підходом.
- **NLQ-to-SPARQL** — трансформує природномовні запити у формальні запити до бази знань.

- **Response Generation** — виконує генерацію відповідей на основі результатів SPARQL-запитів.

Сервіси управління знаннями

- **Knowledge Base Service (Semantic MediaWiki Core)** — реалізує сховище структурованих знань та підтримує SPARQL-запити.
- **Persistence Service** — забезпечує довготривале зберігання вхідних документів, онтологій, логів та інших службових даних.
- **Vector Database Service** — векторне сховище embedding-представлень знань для семантичного пошуку в задачах типу RAG.

Сервіси валідації та експорту

- **Validation Service** — надає інтерфейс для експертної перевірки витягнутих сутностей та декларацій, дозволяє ручну корекцію та затвердження.
- **Export Service** — відповідає за експорт результатів у відповідні стандартизовані формати (PDF, DOCX), з можливістю подальшої передачі сертифікаційним органам.

Вхідні та вихідні дані системи

Вхідні дані:

- **Неструктуровані текстові документи**: Завантажуються користувачами у форматах PDF, DOCX.
- **Природномовні запити**: Формулюються користувачами для отримання пояснень чи витягів з бази знань.
- **Редагування експертів**: Валідовані уточнення, зауваження, вручну додані знання.

Вихідні дані:

- **Аналітичні відповіді**: Сформовані системою результати запитів, представлені у зрозумілій для користувача формі.
- **Семантичні сутності**: Витягнуті та семантично пов'язані поняття, збережені у базі знань.
- **Декларації та звіти**: Автоматично згенеровані документи у форматах PDF, DOCX, придатні для подальшого використання.
- **Embedding-представлення**: Векторні форми знань для швидкого пошуку релевантної інформації.

Висновки

У роботі досліджено потенціал синергетичної інтеграції великих мовних моделей та семантичних технологій для автоматизованого створення й обробки складних природномовних документів. Розглядається концепція побудови гібридної інформаційної системи лінгвістичної обробки документів «ЛІНЗА», яка поєднує великі мовні моделі (LLM), семантичні технології та мультиагентні архітектурні рішення. Основна мета дослідження полягає в розробці автоматизованої платформи для обробки складних природномовних документів, що потребують високого ступеня точності, структурного узгодження та пояснювання.

Запропонована система має багаторівневу архітектуру, яка містить модулі попередньої обробки документів, семантичного анотування, генерації документів і аналітичних звітів, управління знаннями та експертної валідації. Особливу увагу приділено використанню Retrieval-Augmented Generation як механізму інтеграції генеративних можливостей LLM із формалізованими базами знань на основі Semantic MediaWiki.

На етапі архітектурного проєктування авторами запропоновано багаторівневу модель інформаційної системи «ЛІНЗА», що поєднує агентні компоненти з використанням LLM для глибокого семантичного аналізу, набір сервісів для динамічної обробки даних і платформу Semantic MediaWiki для структурованого зберігання та верифікації контенту.

Запропонована архітектура демонструє високий потенціал ефективного застосування в критичних доменах, де особливо важливими є достовірність, прозорість і безпека інформаційної обробки. Залучення формалізованих семантичних шаблонів, механізмів перевірки та підходу

human-in-the-loop забезпечує підвищення точності, надійності та контрольованості результатів.

Застосування системи «ЛІНЗА» охоплює сфери публічного управління, юриспруденції, сертифікації, стандартизації та спрямоване на забезпечення прозорої, контрольованої і формалізованої обробки документів, що відкриває перспективи масштабованої автоматизації складних інформаційних процесів з явним використанням знань предметної - області.

Подальший розвиток системи доцільно спрямувати на вдосконалення механізмів інтеграції з джерелами знань, покращення інтерпретованості результатів і розширення підтримки динамічних сценаріїв використання, а також ширшу інтеграцію зі стандартами проєкту Semantic Web [16], що стосуються подання сервісів, програмних агентів та онтологій.

Література

1. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Mian, A. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*. 2023. URL: https://dl.acm.org/doi/pdf/10.1145/3744746.

2. Liang, X., Zhou, B., Jiang, L., Meng, G., Xiu, Y. Collaborative pursuit-evasion game of multi-UAVs based on Apollonius circle in the environment with obstacle. *Connection Science*. 2023. Vol. 35, Iss. 1. P. 1–24. DOI: https://doi.org/10.1080/09540091.2023.2168253.

3. Musumeci, E., Brienza, M., Suriani, V., Nardi, D., Bloisi, D. D. LLM-based multi-agent generation of semi-structured documents from semantic templates in the public administration domain. In: *Proceedings of the International Conference on Human-Computer Interaction (HCII 2024)*. Cham: Springer, 2024. P. 98–117.

4. Eichhorn, T. CLAIR: Generating on-demand low-code application documentation through knowledge graph and LLM-based multi-agent system integration. Master's thesis. University of Twente, 2025.

5. Плескач В. Л., Рогушина Ю. В. Агентні технології: Монографія. Київ : Київ. нац. торг.–екон. ун–т, 2005.

6. Рогушина Ю. В., Гладун А. Я., Осадчий В. В., Прийма С. М. Онтологічний аналіз у Web: Монографія. Мелітополь : МДПУ ім. Богдана Хмельницького, 2015. 407 с. URL: http://www.dut.edu.ua/uploads/l_2148_67675988.pdf. ISBN 978-617-7346-27-1.

7. Vrandečić, D., Krötzsch, M. Semantic MediaWiki. In: *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. P. 171–179.

8. Chen, J., Lu, X., Du, Y., Rejtig, M., Bagley, R., Horn, M., Wilensky, U. Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024. P. 1–18.

9. Yang, D., Simoulin, A., Qian, X., Liu, X., Cao, Y., Teng, Z., Yang, G. DocAgent: A multi-agent system for automated code documentation generation. *arXiv preprint arXiv:2504.08725*. 2025.

10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 9459–9474.

11. Machado, M., Rodrigues, J. M., Lima, G., Fiorini, S. R., da Silva, V. T. LLM Store: Leveraging large language models as sources of Wikidata-structured knowledge. In: *Proceedings of the International Semantic Web Conference (ISWC 2024)*. Cham: Springer, 2024 (to appear).

12. Mihindukulasooriya, N., Tiwari, S., Dobriy, D., Nielsen, F. Å., Chhetri, T. R., Polleres, A. Scholarly Wikidata: Population and exploration of conference data in Wikidata using LLMs. In: *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024)*. Cham: Springer, 2024. P. 243–259.

13. Каверинський В. В., Літвін А. А., Палагін О. В. Зворотний синтез природномовних висловлювань на основі їх онтологічного представлення з використанням великої мовної моделі. *Проблеми програмування*. 2024. № 2-3. С. 359. DOI: https://doi.org/10.15407/pp2024.02-03.359.

14. Рогушина Ю. В., Гладун А. Я., Аніщенко О. В., Прийма С. М. Семантичні технології як інструмент інформаційного забезпечення професіоналізації андрагогів. *Проблеми програмування*. 2024. № 2-3. С. 441. DOI: https://doi.org/10.15407/pp2024.02-03.441.

15. Rotsos, C., King, D., Farshad, A., Bird, J., Fawcett, L., Georgalas, N., Hutchison, D. Network service orchestration standardization: A technology survey. *Computer Standards & Interfaces*. 2017. Vol. 54. P. 203–215.

16. Patel, A., Jain, S. Present and future of semantic web technologies: a research statement. *International Journal of Computers and Applications*. 2021. Vol. 43, Iss. 5. P. 413–422.

Одержано: 19.07.2025
Внутрішня рецензія отримана:26.07.2025
Зовнішня рецензія отримана: 28.07.2025

Про авторів:

Сініцин Ігор Петрович,
доктор технічних наук, професор,
член-кореспондент НАН України,
https://orcid.org/0000-0002-4120-0784,
ips@nas.gov.ua

Рогушина Юлія Віталіївна,
кандидат фіз.-мат.наук, с.н.с., доцент,
http://orcid.org/0000-0001-7958-2557,
ladamandraka2010@gmail.com

Юрченко Костянтин Юрійович,
аспірант, м.н.с.
https://orcid.org/0000-0003-3150-0027,
urchikak8@gmail.com

Місце роботи авторів:

Інститут програмних систем
Національної академії наук України,
03187, м. Київ-187,
просп. Академіка Глушкова, 40,
корпус 5.