

Оцінка семантичної та часової узгодженості розподілених інформаційних ресурсів

<https://doi.org/10.31713/MCIT.2025.036>

Юрій Геряк

Національний університет «Львівська політехніка»
Львів, Україна
yurii.m.heriak@lpnu.ua

Андрій Берко

Національний університет «Львівська політехніка»
Львів, Україна
andrii.y.berko@lpnu.ua

Анотація — Представлено комплексний підхід до оцінки семантичної та часової узгодженості розподілених інформаційних ресурсів. Описано поняття ключових метрик для вимірювання змістовної та часової відповідності даних між різними вузлами системи, запропоновано формалізовані моделі для автоматичного контролю якості інформації. Проведено аналіз основних труднощів у підтримці консистентності та окреслено перспективи розвитку методів і інструментів моніторингу для підвищення надійності сучасних розподілених систем.

Ключові слова — семантична узгодженість, часова узгодженість, розподілені інформаційні ресурси, контроль якості даних, метрики консистентності, автоматизація, розподілені системи.

I. ВСТУП

Сучасна цифрова трансформація спричинила не тільки зростання обсягів даних, а й появу складних розподілених інформаційних систем, що охоплюють різні вузли, географічні регіони та середовища використання. Якість та цілісність інформації в таких системах є критично важливими для підтримки достовірності даних, коректності аналізу й ефективності процесів прийняття рішень у промисловості, науці, державному управлінні й повсякденному житті. Зокрема, навіть незначні втрати або спотворення даних під час передачі можуть призвести до помилкових висновків, втрат ресурсів або виникнення технологічних ризиків [1].

Забезпечення високого рівня консистентності розподілених інформаційних ресурсів передбачає не лише технічний, а й семантичний та просторовий контроль якості. Семантична узгодженість означає збереження змістової відповідності й уніфікації сенсу даних між незалежними компонентами системи, тоді як просторова консистентність стосується співставлення та узгодженості геопросторової або тимчасової інформації. Сучасні методи підтримки консистентності включають використання алгоритмів порівняння, класифікації та машинного навчання, що дозволяє автоматизувати процеси перевірки відповідності даних і зменшувати

імовірність помилок чи дублювання інформації [2][4].

Удосконалення підходів і методів оцінки семантичної та просторової консистентності розглядається сьогодні одним з найактуальніших напрямів у сфері контролю якості даних. Ефективність рішень у цьому контексті забезпечує підвищення надійності систем, зменшення ризику критичних помилок та сприяє довірі користувачів до сучасних цифрових сервісів і багаторівневих інформаційних інфраструктур [3][4].

II. СЕМАНТИЧНА УЗГОДЖЕНІСТЬ: ПОНЯТТЯ І МЕТРИКИ

Семантична узгодженість у розподілених інформаційних ресурсах – це властивість даних, що забезпечує їхню змістову узгодженість та несуперечливе трактування між різними автономними вузлами системи. На відміну від синтаксичної консистентності, яка стосується структури і формату, семантична акцентує увагу саме на ідентичності змісту, контексті та значеннях інформації. У умовах значної гетерогенності розподілених середовищ виникає ризик втрат або перекручення значень даних через різні підходи до їх опису чи представлення, що особливо критично для інтеграції великих та динамічних даних у складних інформаційних системах [5].

Виявлення несуперечливості сенсу даних між вузлами передбачає застосування спеціалізованих алгоритмів аналізу схожості або відмінностей, які можуть ґрунтуватися як на лінгвістичних, так і на статистичних підходах. Серед найбільш поширених метрик семантичної схожості використовуються Jaccard similarity (порівняння множин елементів), Levenshtein distance (визначення мінімальної кількості операцій редагування між рядками) та WordNet-based similarity (обробка лексичних відносин між словами). Застосування цих підходів дозволяє автоматизувати контроль відповідності контексту та значення інформації, впроваджувати комплексні механізми перевірки та уніфікації змісту даних [5][8].

Кожна з цих метрик має власні переваги та недоліки у розподілених системах. Наприклад, коефіцієнт Жаккара найкраще працює для аналізу схожості структурованих даних (наприклад, тегів, словників або категорій), тоді як Levenshtein distance ефективно визначає рівень схожості між текстовими описами. WordNet-based підхід дозволяє оцінити семантичну близькість із врахуванням ієрархічних мовних зв'язків, що надзвичайно актуально при інтеграції багатомовних або різномовних інформаційних масивів. На практиці доцільно поєднувати кілька алгоритмів для отримання більш комплексної оцінки консистентності [5].

Загалом, якість підтримки семантичної консистентності залежить від гнучкості обраної стратегії порівняння, здатності враховувати специфіку домену даних, рівень автоматизації засобів контролю й можливість адаптації метрик під конкретні прикладні завдання. Використання сучасних алгоритмів адаптивного зіставлення надає широкі перспективи для розвитку систем аналізу якості розподілених інформацій, підвищує надійність корпоративних, наукових і державних даних [5] [6].

III. ПРОСТОРОВА І ЧАСОВА УЗГОДЖЕНІСТЬ

Просторова та часова узгодженість є критичними аспектами якості розподілених інформаційних ресурсів, які визначають точність і узгодженість даних щодо їх географічної локалізації та часових характеристик у різних вузлах системи. Геопросторова консистентність означає відповідність координат, геотегів та інших атрибутів просторових об'єктів, що знаходяться в різних фрагментах розподіленої бази, а також ідентичність просторового представлення реальних об'єктів у всіх точках інтеграції інформації. Це актуально для систем моніторингу, картографії, транспортної логістики та інших галузей, де дані можуть надходити з різних джерел і потребують гармонізації та перевірки на співставність [3] [7].

Часова узгодженість стосується узгодженості часових атрибутів даних — наприклад, міток подій, версій документів чи записів транзакцій, особливо коли дані створюються, змінюються або реплікуються в різних часових зонах і різних пристроях. До її показників належать періодичність оновлення, часові затримки між змінами, латентність синхронізації реплік, а також цілісність послідовності подій. Недостатня часова консистентність спричиняє появу застарілих, дубльованих або суперечливих записів, що знижує достовірність результатів обробки інформації та створює ризики для систем реального часу [9].

Основними практичними причинами порушень просторової і часової консистентності є затримки передачі, відмінності в локальних налаштуваннях часу, неоднорідність стандартів зберігання й форматів координат, а також специфіка топології самої розподіленої системи. Наприклад, відхилення в роботі GPS-датчиків, асинхронність оновлень геопросторових шарів або некоректна обробка часових міток можуть призвести до розбіжностей

між частинами бази даних, що знижує ефективність аналітики або автоматизованого управління процесами [9].

Наслідки нехтування просторовою та часовою узгодженістю проявляються у вигляді втрат аналітичної цінності даних, некоректності геопросторової візуалізації, помилок у навігаційних та логістичних системах, а також у підвищенні ймовірності інформаційних колізій при об'єднанні чи актуалізації великих масивів інформації. Відтак регулярний контроль, синхронізація та уніфікація просторових і часових показників є невід'ємною частиною якісного функціонування сучасних розподілених інформаційних ресурсів. Просторова й часова консистентність у розподілених інформаційних системах визначає узгодженість географічних координат, просторових характеристик і часових параметрів даних у різних вузлах системи. Геопросторова узгодженість відповідає за однаковість інтерпретації і відображення просторових об'єктів (наприклад, місцезнаходження, межі територій, маршрути), а також правильне співвіднесення таких даних між різними фрагментами й джерелами інформації. Це набуває особливої ваги для екосистем моніторингу, геоінформаційних платформ, транспортної логістики й цифрової картографії, де навіть незначні невідповідності можуть призводити до критичних помилок [5] [3] [10].

Часова узгодженість означає гармонізацію та точність часових міток, атрибутів оновлення й синхронізації інформації у всіх вузлах системи або середовищах даних. Серед ключових показників — латентність доставки даних, актуальність відомостей, рівень синхронізації між репліками та коректність порядку подій. У зонах реального часу, аналітичних підсистемах і транзакційних службах недостатня часова узгодженість часто спричиняє появу застарілих, недостовірних, дубльованих чи суперечливих записів, що безпосередньо впливає на ефективність управління процесами [5] [10].

До головних джерел розбіжностей у просторових і часових даних відносять локальні відмінності в стандартах: формати й точність координат, часові пояси, асинхронність або затримки у синхронізації, різні режими обробки подій між вузлами, а також збої у роботі сенсорів чи GPS-модулів. Навіть незначне породження відхилень у геопросторових або часових характеристиках може істотно знизити надійність аналітики, ускладнити візуалізацію та порушити автоматизоване управління територіальними чи логістичними системами [3] [10].

Наслідки нехтування просторовою та часовою консистентністю проявляються у вигляді зміщення контексту даних, втрати аналітичної цінності, некоректної інтеграції й візуалізації, а також збільшення кількості інформаційних колізій під час оновлення або об'єднання масивів у розподілених сховищах і сервісах. Регулярний аудит, стандартизація форматів і автоматизована синхронізація — ключові інструменти для уникнення цих проблем у сучасних цифрових платформах [5] [11].

IV. Методи й інструменти оцінки

Оцінка якості розподілених інформаційних ресурсів здійснюється за допомогою різних методів — від повністю автоматизованих до ручних, що залежать від специфіки даних та вимог системи. Автоматизовані підходи передбачають інтеграцію спеціалізованих метрик і алгоритмів у процеси збору, моніторингу та порівняння даних, а ручна експертиза надає можливість тонкого налаштування та додаткової перевірки результатів в критичних секторах. В сучасних комплексних системах часто застосовують комбіновану стратегію для підвищення точності оцінки [10].

Для оцінки семантичної узгодженості між даними пропонується використовувати формулу подібності:

$$\text{Sim}(v_i^{\text{source}}, v_i^{\text{target}}) = 1 - D(v_i^{\text{source}}, v_i^{\text{target}}), \quad (1)$$

де $D(v_i^{\text{source}}, v_i^{\text{target}})$ визначено як відстань між відповідними значеннями (наприклад, за методами Jaccard, Levenshtein чи WordNet similarity). Згідно з запропонованим підходом, дані вважаються семантично узгодженими у разі, якщо для більшості пар Sim перевищує поріг 0.8, а середнє значення семантичної консистентності розраховується так:

$$SC = \frac{1}{N} \sum_{i=1}^N \text{Sim}(v_i^{\text{source}}, v_i^{\text{target}}), \quad (2)$$

де N — кількість пар даних для порівняння.

Також семантичну узгодженість можна формально визначити через контекстуальну семантичну відстань (CSD) між множинами термінів, де

$$\text{CSD}(D_i, D_j) = \frac{|\Omega_i \cap \Omega_j|}{\min(|\Omega_i|, |\Omega_j|)} \quad (3)$$

Де, Ω_i, Ω_j — множини семантичних ознак/термінів у відповідних наборах даних. Значення $\text{CSD} \geq 0,8$ свідчить про високу семантичну узгодженість [12].

Для вимірювання просторової узгодженості рекомендується аналізувати співвідношення координат або просторових атрибутів на основі відповідних просторових метрик, що забезпечує виявлення відхилень у розташуванні об'єктів. Щодо часової консистентності, запропоновано використовувати показник затримки реплікації:

$$TC = 1 - \frac{T}{T^{\text{max}}} \quad (4)$$

де T — реальна затримка, а T^{max} — припустиме максимальне значення. Значення, близькі до 1, свідчать про високу часову узгодженість між вузлами розподіленої системи.

Цей підхід доповнюється концепціями часово-просторової узгодженості, які базуються на моделях *timed causal consistency* та *timed linearizability*, що формалізують узгодженість операцій в межах часової затримки Δ .

Формула для *timed causal consistency* має вигляд:

$$\text{TIMEDCAUSALITY}(F, \Delta)$$

$$\equiv \text{CAUSALITY}(F) \wedge \text{TIMEDVISIBILITY}(\Delta) \quad (5)$$

Де:

- F — функція або послідовність операцій системи;
- $\text{CAUSALITY}(F)$ - збереження каузального порядку операцій: якщо операція A впливає на операцію B , всі вузли бачитимуть A перед B , що виконує логічний порядок залежностей;
- $\text{TIMEDVISIBILITY}(\Delta)$ — обмеження в часі, яке гарантує, що будь-який результат операції стане доступний усім вузлам системи не пізніше ніж через інтервал Δ від часу її виконання.

Формула для *timed linearizability* має вигляд:

$$\text{TIMEDLINEARIZABILITY}(F, \Delta)$$

$$\equiv \text{SINGLEORDER} \wedge \text{TIMEDVISIBILITY}(\Delta)$$

$$\wedge \text{RVAL}(F) \quad (6)$$

де: SINGLEORDER вказує, що операції сприймаються у системі в єдиному послідовному порядку, логічно узгодженому для всіх користувачів і процесів (лінійна зупинка чи *linearizability*); $\text{TIMEDVISIBILITY}(\Delta)$ - те саме часово-затримкове обмеження видимості оновлень як і вище; $\text{RVAL}(F)$ - вимога, що значення, які повертаються операціями читання, відповідають останньому лінійно узгодженому оновленню.

Ці моделі спільно гарантують:

- збереження порядку операцій відповідно до їх логічних або причинно-наслідкових зв'язків,
- забезпечення того, що всі зміни у системі будуть помічені іншими вузлами не пізніше ніж через Δ часу,
- підтримку узгодженого та свіжого стану даних для користувачів у розподіленій системі.

Іншими словами, застосування цих моделей дозволяє врахувати не лише логіку взаємозв'язків операцій, а й реальні часові характеристики системи, що критично для забезпечення високої якості та надійності розподіленого інформаційного процесу [13].

Щоб визначити агрегований індекс якості для узгодженості даних, у роботі пропонується формула:

$$C_{\text{agg}} = w_1 C_{\text{col}} + w_2 C_{\text{schema}} + w_3 C_{\text{por}} \quad (7)$$

де C_{col} , C_{schema} , C_{por} — це часткові показники для повноти колонок, схеми й популяції, а $w_1 + w_2 + w_3 = 1$; ваги добираються з урахуванням специфіки прикладного сценарію.

Практична реалізація запропонованих методів передбачає використання сучасних інструментів збору й обробки метрик: Prometheus, Zabbix, Power BI, Tableau, а також автоматизованих REST API для

інтеграції з хмарними й потоковими платформами. Це дозволяє оперативно отримувати оцінки якості даних у великих розподілених системах, забезпечуючи гнучке реагування на потенційні загрози консистентності та підвищення надійності інформаційних процесів [10].

V. СКЛАДНОСТІ ТА ПЕРСПЕКТИВИ ВПРОВАДЖЕННЯ

Оцінка узгодженості розподілених інформаційних ресурсів стикається з низкою суттєвих складнощів, що зумовлені специфікою розподілених систем. По-перше, відсутність централізованого контролю створює виклики у синхронізації даних між численними незалежними вузлами, де затримки, мережеві відмови та асинхронні оновлення можуть призводити до несуперечливих станів баз даних. По-друге, неоднорідність структур і форматів даних, різні політики доступу і обробки інформації ускладнюють стандартизацію показників якості, що ускладнює побудову універсальних моделей та алгоритмів оцінки [10].

Іншою проблемою є відсутність єдиних норм та стандартів для визначення і вимірювання якості даних у розподілених середовищах, через що організації часто залежать від власного досвіду та встановлених практик. Це породжує ризики появи непослідовних або навіть суперечливих оцінок якості, що може призвести до некоректного використання інформації й зниження довіри до системи. Крім того, значна частина таких вимірювань носить багатовимірний, суб'єктивний характер і потребує балансу між точністю автоматизованих процесів та експертним аналізом [10].

Перспективи розвитку методів оцінки якості розподілених даних пов'язані з удосконаленням алгоритмів машинного навчання, штучного інтелекту і семантичної аналітики, які здатні покращити розуміння контексту та сенсу інформації, усувати помилки та забезпечувати адаптивність систем оцінки до різних доменів і сценаріїв. Майбутні напрями досліджень орієнтовані на посилення інтеграції різних аспектів якості, в тому числі точності, повноти, узгодженості та актуальності, у комплексні моделі з урахуванням специфіки розподілених систем і їх масштабованості [2].

Важливими також є розробка стандартизованих фреймворків та протоколів для контролю якості, які зможуть забезпечити уніфікований підхід до оцінки й прогнозування стану даних. Впровадження механізмів відновлення консистентності та самокорекції даних, а також автоматизованих систем моніторингу допоможуть мінімізувати ризики деградації інформаційних ресурсів у реальному часі. Ці тенденції відкривають широку перспективу для наукових і практичних розробок у сфері систем контролю якості розподілених інформаційних ресурсів [10].

ЛІТЕРАТУРА

- [1] V. S. Balatska, Enhancing the Effectiveness of Personal Data Protection in the Context of Digital Informatization of State Registers in Ukraine, PhD dissertation in specialty 125 "Cybersecurity," Lviv Polytechnic National University, Lviv, 2025. [Online]. Available: <https://lpnu.ua/sites/default/files/2025/radaphd/32586/disertacyiabalatskavaleriiiasergiivna.pdf>
- [2] Ye. R. Kovylin, Model of Answers Generation in Search Engines Based on an Unstructured Knowledge Base, PhD dissertation in specialty 01.05.02 – Mathematical Modeling and Computational Methods, Oles Honchar Dnipro National University and National Metallurgical Academy of Ukraine, Dnipro, 2020. [Online]. Available: https://nmetau.edu.ua/file/diss_kovilin.pdf
- [3] Ye. A. Borovskova, "Determining Optimal Caching Strategies for Improving the Performance of Server Applications on NestJS," Scientific Notes of V.I. Vernadsky Taurida National University. Series: Technical Sciences, vol. 35 (74), no. 6, 2024, pp. 18–26. DOI: 10.32782/2664-6468/2024.6-2.04.
- [4] O. Ye. Arkusha, "Comparison of Hashing Methods in Supporting Consistency of Distributed Databases," in Advanced Systems and Computing, vol. 40, 2022. DOI: 10.20535/1560-8956.40.2022.261646.
- [5] M.K. Aguilera, D.B. Terry, "The Many Faces of Consistency," IEEE Database Engineering Bulletin, vol. 39, 2016, pp. 3–13. [Online]. Available: <http://sites.computer.org/debull/A16mar/p3.pdf>
- [6] P. V. Buslov, Models and Methods for the Differentiation of Consolidated Information Data for Decision Support Systems, PhD dissertation in specialty 05.13.06 – Information Technologies, Kharkiv National University of Radio Electronics, Cherkasy State Technological University, 2021. [Online]. Available: https://er.chdtu.edu.ua/bitstream/ChSTU/2067/12/dis_%D0%91%D1%83%D1%81%D0%BB%D0%BE%D0%B2%20%D0%9F.%D0%92..pdf
- [7] V. Duchenchuk and V. Boublik, Consistency Models for Distributed Virtual Reality Systems. NaUKMA Research Papers. Computer Science. 3, 2020, pp. 7-11. DOI: 10.18523/2617-3808.2020.3.7-11.
- [8] A. S. Tanenbaum, and M. V. Steen, Distributed systems: Principles and paradigms. Createspace Independent Publishing Platform, 2016.
- [9] G. Zholtkevych, Mathematical Modeling of Data Replication Processes in Distributed Datastores [Electronic resource]: qualification scientific work for the degree of Candidate of Science in specialty 01.05.02 – Mathematical modeling and computational methods (Technical Sciences) / Galina Grigoriyivna Zholtkevych; supervisor Kirilo Markovych Rukkas. – Kharkiv: V. N. Karazin Kharkiv National University; A. M. Pidgorny Institute for Mechanical Engineering Problems of NAS of Ukraine, 2021. – 116 p. [Online]. Available: <https://ipmach.kharkov.ua/wp-content/uploads/2021/03/%D0%94%D1%96%D1%81%D0%B5%D1%80-Copy.pdf>
- [10] Y. Heriak and A. Berko, "The System of Data Quality Assessment Criteria in Distributed Information Systems," Scientific Journal of Lviv Polytechnic, vol. 16, 2024. DOI: 10.23939/sisn2024.16.191.
- [11] A. M. Hlybovets, T. A. Chernova, and M. M. Hlybovets, "Development of a Methodology for Transaction Implementation in Distributed Systems with Microservice Architecture," in Problemy Programuvannya, no. 1, 2024, pp. 64–72. DOI: 10.15407/pp2024.01.64
- [12] A. Berko, "Methods and Means of Semantic Data Integration," Scientific Journal, Lviv Polytechnic National University, 2009. [Online]. Available: https://vlp.com.ua/files/29_10.pdf
- [13] P. Viotti and M. Vukolić, "Consistency in Non-Transactional Distributed Storage Systems," arXiv preprint arXiv:1512.00168, 2015. [Online]. Available: <https://arxiv.org/pdf/1512.00168>