The Central Research institute
of the Armed Forces of Ukraine

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

Received: *December 2, 2025* | **Revisions Requested:** - | **Revision Received:** -
Accepted: *January 20, 2026* | **Published Online:** *March 21, 2026*

Licensed (C) by Creative Commons Attribution International License 4.0 (CC BY-NC-SA)

DOI: 10.62524/msj.2025.3.4.25

Сучасні задачі інтеграції даних спеціалізованих інформаційно-комп'ютерних систем

Олександр Додонов
доктор технічних наук, професор
заступник директора з наукової роботи
Інститут проблем реєстрації інформації НАН України
03113, вул. М. Шпака, 2, Київ, Україна
e-mail: dodonovua@gmail.com
ORCID: 0009-0000-7831-6337

Вадим Додонов
кандидат технічних наук
старший науковий співробітник
Інститут проблем реєстрації інформації НАН України
03113, вул. М. Шпака, 2, Київ, Україна
e-mail: v.o.dodonov@nas.gov.ua
ORCID: 0009-0002-3480-2549

Валерій Залужний
Надзвичайний і Повноважний посол України у Великій Британії
Постійний представник України при Міжнародній морській організації (ІМО)
e-mail: uzvk_mou@post.mil.gov.ua
ORCID: 0000-0002-1947-501X

Костянтин Кравчук
науковий співробітник
ІПРІ НАН України
e-mail: k.kravchuk@ipri.kiev.ua
ORCID: 0009-0000-1750-4588

Володимир Юзефович
кандидат технічних наук, доцент
старший науковий співробітник
Навчально-науковий фізико-технічний інститут НТУУ "КПІ ім. Ігоря Сікорського"
03056, Берестейський проспект, 37, Київ, Україна
e-mail: uzefv71@gmail.com
ORCID: 0000-0002-6336-9548

The Central Research institute
of the Armed Forces of Ukraine

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

Ігор Ізварін
кандидат технічних наук, доцент
Інститут комп'ютерних технологій
Відкритий міжнародний університет розвитку людини "Україна"
03115, вул. Львівська, 23, Київ, Україна
e-mail: igor.izvarin@gmail.com
ORCID: 0009-0004-4827-9968

Анотація. Актуальність дослідження зумовлена розвитком автоматизації бізнес-процесів у складних організаційних системах, що призводить до потреби інтеграції множини автономних інформаційно-комп'ютерних систем в єдину систему більш високого порядку. Проблема полягає у функціональній несумісності автономних систем, дублюванні даних та відсутності спільних протоколів обміну, що перешкоджає ефективній взаємодії та знижує якість інформаційного забезпечення. В умовах зростання обсягів інформації задача інтеграції даних стає критичною для підвищення ефективності технологічних процесів та ухвалення обґрунтованих управлінських рішень. Метою дослідження є визначення сучасних шляхів розв'язання задачі інтеграції даних у складних організаційних системах шляхом розроблення технічних, організаційних і семантичних кроків з аналізу, перетворення та об'єднання інформації. Для досягнення поставленої мети використано методи штучного інтелекту для автоматизації виявлення джерел та форматів даних, методи лексичного, синтаксичного та семантичного аналізу для перевірки змістовності інформації. Також застосовано апарат математичного моделювання, зокрема методи кластеризації із застосуванням нечіткої формалізації, дискретну логіку та теорію вимірювань для оброблення нерівноточних даних під час їх об'єднання. Для виправлення помилок використано статистичні методи, аналіз часових рядів та алгоритми машинного навчання. У результаті дослідження сформовано комплексний методичний підхід до процесу інтеграції даних, який структуровано за чотирма етапами: пошук джерел, отримання даних, перевірка та об'єднання. Запропоновано процедури автоматичного визначення систем кодування, ідентифікації об'єктів за мірою близькості значень та механізми відновлення помилкових даних через інтерполяцію, екстраполяцію та реверсний інжиніринг. Це дозволяє мінімізувати втрати корисної інформації та забезпечити усунення дублювання. Практична спрямованість роботи полягає в тому, що запропоновані результати можуть бути використані для створення інтегрованих систем, здатних максимізувати рівень ефективності автоматизації у конкретних прикладних областях. Представлений підхід забезпечує розв'язання питань пошуку потрібних даних, усунення їхньої несумісності та підвищення загальної якості (повноти, точності та достовірності) інтегрованих масивів даних. Отримані висновки створюють підґрунтя для вдосконалення процесів обміну інформацією в реальному часі між сучасними та застарілими системами

Ключові слова: інтеграція даних, інформаційні системи, комп'ютерні системи, оброблення даних, семантична сумісність, аналіз інформації, складні організаційні системи, перетворення даних.

Вступ (Introduction)

Так звана "клаптикова" автоматизація бізнес-процесів у складних організаційних системах, з усіма її перевагами та недоліками, рано чи пізно призводить до необхідності поєднання (інтеграції) множини автономно функціонуючих автоматизованих інформаційно-комп'ютерних систем в автоматизовану систему більш високого порядку, яка здатна максимізувати досяжний рівень ефективності інформатизації та автоматизації у конкретній прикладній області. Складність процесу інтеграції полягає у тому, що автономні системи

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

мають власну (локальну) цільову функцію та, відповідно, функціональне наповнення, використовують власні джерела даних і споживачів інформації. При цьому і функції, і дані можуть багатократно дублюватися. Такі системи переважно не зорієнтовані на обмін даними із іншими автоматизованими системами і, як результат, додатково демонструють несумісність за вхідними та вихідними даними, що ускладнює, а іноді й уможливлює, їхню взаємодію. Отже, однією із проблем інтеграції автономних інформаційно-комп'ютерних систем є інтеграція даних, якими вони оперують, для забезпечення можливості їх використання іншими складовими інтегрованої інформаційно-комп'ютерної системи.

У широкому сенсі під інтеграцією даних мають на увазі процес, який передбачає:

- визначення джерел потрібних даних із загального переліку джерел, які використовувалися автономними системами, що підлягають інтеграції;
- отримання нових даних від множини визначених джерел, їх трансформування, об'єднання між собою та з уже існуючими в інформаційно-комп'ютерній системі даними.

Така інтеграція даних надає можливість досягти наступних цілей:

забезпечення необхідними даними більш високої якості існуючий, новий, або модифікований технологічний процес (ТП), оскільки для його побудови чи виконання загалом потребуються різні дані від різних джерел;

здійснення аналізу об'єктів, процесів, або явищ на основі більш повної інформації, що покращує якість їх представлення у системі, точність та грануляцію моделей об'єктів, процесів, явищ;

знаходження залежностей між даними, яких раніше неможливо було виявити за відсутності більш повної множини інтегрованих даних;

збір більш повної множини часткових даних, що визначають загальний стан об'єкта (процесу, явища), як, наприклад, збір показників датчиків, що характеризують різні властивості об'єкта зацікавленості, із наступним об'єднанням в один масив даних (консолідація даних);

ухвалення більш ефективних рішень (наприклад, виходячи з набору даних – $НД_1$ може бути ухвалене рішення $Р_1$, а маючи набір даних $НД_2$ (розширений $НД_1$, за рахунок інтеграції даних) може бути ухвалене рішення – $Р_2$, при цьому потенційно рішення $Р_2$ буде більш ефективним, ніж рішення $Р_1$).

Методологічна основа (Methodological Framework)

Дослідження ґрунтується на поєднанні системного, інтелектуального та аналітичного підходів, що забезпечує комплексне розв'язання задачі інтеграції даних у спеціалізованих інформаційно-комп'ютерних системах.

Ключовим інструментом виступають методи штучного інтелекту, які дозволяють автоматизувати процеси виявлення джерел, визначення форматів та інтеграції даних у реальному часі. Застосування ШІ спрощує оброблення великих обсягів як структурованої, так і неструктурованої інформації (довільний текст, мультимедіа), забезпечуючи масштабованість системи та її адаптацію до нових автономних джерел.

Методологія включає багаторівневий аналіз даних, що охоплює лексичний, синтаксичний та семантичний аналіз.

Використання цих методів дозволяє послідовно перевіряти правильність запису кожного окремого елемента, відповідність побудови складних структур заданим форматам та змістовну логічність отриманої інформації з точки зору її відповідності реальності.

Для розв'язання задач ідентифікації та об'єднання об'єктів застосовано апарат математичного моделювання, зокрема методи кластеризації із застосуванням нечіткої формалізації та дискретну логіку. Для оброблення числових значень використано теорію вимірювань (комплексування нерівноточних даних), а для формалізації якісних показників

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

теорію нечітких множин, що у сукупності забезпечує підвищення точності, повноти та достовірності інтегрованих масивів.

У дослідженні використано методи статистичного аналізу та машинного навчання для виявлення аномалій і відновлення помилкових даних.

Застосування алгоритмів регресійного аналізу, методу k-найближчих сусідів (k-NN), а також різних видів інтерполяції та екстраполяції дає можливість прогнозувати відсутні значення та мінімізувати втрати корисної інформації

Поєднання зазначених методів створює багаторівневу методологічну основу, що дає змогу:

ефективно здійснювати пошук джерел через аналіз таксономій та відкритих сховищ;

забезпечувати автоматичне приєднання до систем за різними протоколами обміну (FTP, REST, gRPC);

автоматично визначати та конвертувати системи кодування даних;

усувати дублювання та підвищувати загальну якість інформаційного забезпечення.

Отже, методологічна основа дослідження спирається на інтеграцію сучасних технологій штучного інтелекту з класичними математичними та аналітичними інструментами, що дозволяє сформувати практичний механізм створення інформаційних систем високого порядку у складних організаційних структурах.

Результати (Results)

Сучасний стан розв'язання задачі інтеграції даних

З огляду на складність аналітичної постановки задачі інтеграції даних для її розв'язання нині пропонується застосовувати методи штучного інтелекту (ШІ). Такі підходи можна вважати найбільш перспективними, оскільки, на відміну від традиційних методів, які значною мірою передбачають участь людини (аналітика), технологія на основі штучного інтелекту може спрощувати задачі визначення джерел даних, їхніх форматів, обробляти великі обсяги даних, включаючи структуровані (формалізовані), слабкоструктуровані та неструктуровані (довільний текст та мультимедіа). Крім того, рішення ШІ легко масштабувати, що є однією з найважливіших проблем, коли потрібно отримати інформацію від тисяч, або навіть десятків тисяч автономних джерел даних. ШІ трансформує процес інтеграції даних шляхом автоматизації ключових його складових, таких як виявлення даних, підвищення якості та інтеграція в реальному часі. Завдяки здатності штучного інтелекту адаптуватися до нових джерел даних і вчитися на поточних процесах з'явилась можливість досягти вищої ефективності інтеграції, зменшити кількість помилок й ухвалювати більш обґрунтовані рішення, одночасно досить ефективно інтегруючи як сучасні, так і застарілі системи.

Процес інтеграції даних

Процес інтеграції даних зручно розкласти на декілька етапів, на кожному з яких виконується окрема послідовність дій, що веде до завершення процесу. Пропонується розглядати чотири наступні етапи інтеграції даних: пошук джерел, отримання даних, перевірка даних, об'єднання даних (рис. 1).

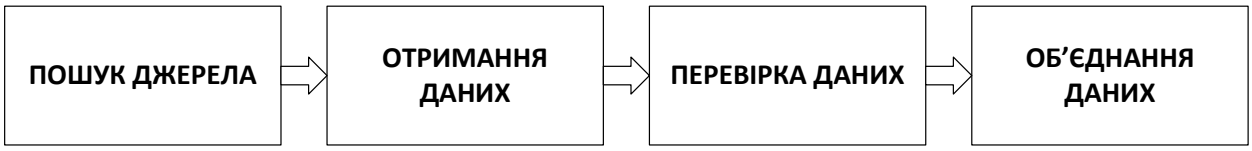


Рисунок 1. Фази процесу інтеграції даних

The Central Research institute
of the Armed Forces of Ukraine

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

Якщо вважати, що перелік даних, потрібних для досягнення вказаних вище цілей, визначено, виникає наступна задача – їх пошук. Оскільки дані оновлюються з плином часу, тобто модифікуються їхні значення та з'являються нові дані, то потрібно знайти не просто статичний масив даних, а визначити їхнє джерело, з якого можна отримувати необхідні дані постійно (із певною періодичністю). Цю задачу можна вважати найскладнішою в усьому процесі інтеграції і в повному обсязі її може бути виконано тільки людиною. Однак є можливості автоматизації деяких складових цього процесу.

Після того, як джерело даних визначено, потрібно зрозуміти, яким чином отримувати дані від нього, і спробувати це зробити один або декілька разів у тестовому режимі. Таким чином буде перевірено технічну можливість отримання даних як у ручному, так і в автоматичному режимах.

Отримані дані, перед об'єднанням з існуючим масивом, потрібно перевірити на відповідність певним вимогам і в подальшому використовувати лише ті дані, які таку перевірку пройшли. Процедура перевірки може бути складною, мати дуже багато різних правил та потребувати певного часу. Але після такої перевірки якість даних, що будуть об'єднані з існуючими, значно підвищиться. Залишиться окреме питання, що робити із тими даними, які після перевірки визначені як помилкові. Це питання буде розглянуто далі.

Останнім етапом інтеграції даних є об'єднання нових отриманих та перевірених даних з існуючим масивом. Знову ж таки, потрібні правила для об'єднання, щоб ця процедура не призвела до руйнування існуючого масиву. Процедура об'єднання також складається з декількох операцій.

Як було з'ясовано, для досягнення цілей потрібно мати конкретні дані, які можна отримати з певних джерел. Тому коли виникає потреба в інтеграції нових або отриманні додаткових даних, перше, що потрібно зробити, це знайти необхідне джерело, з якого можна отримувати дані. Питання пошуку джерела даних багатогранне. Його можна розглядати і з філософського, і з організаційного, і з технічного боку. Тому до його вирішення необхідно підходити комплексно.

На рис. 2 наведено запропонований підхід до процесу пошуку джерела даних.

У найпростішому випадку, коли знаходження даних системі відоме і вони доступні наявними інструментами, задача пошуку втрачає сенс. Якщо ситуація інша, потрібна організація пошуку, і наступним простим варіантом дій є звернення до відкритих джерел даних або до відкритих сховищ даних. Прикладом такого відкритого сховища даних може бути сайт kaggle.com, де можна знайти тисячі різних наборів даних для вільного використання.

Наступним, і теж простим варіантом пошуку можуть бути магазини даних, де потрібні дані продаються для використання. Такі магазини мають каталоги даних, які пропонуються покупцям. Пошук у каталогах можливо організувати автоматично чи автоматизовано, використовуючи ключові слова для опису потрібних даних. Основна відміна магазинів даних від відкритих джерел полягає в тому, що такі магазини будуть пропонувати оновлення даних із зазначеним періодом, що також впливатиме на їхню вартість.

У тому разі, якщо знайти потрібні дані не вдалося за допомогою попередніх способів, пропонується використати для цього аналіз таксономії джерел. Таку таксономію потрібно попередньо побудувати та опублікувати в рамках організації, що потребує нових даних, або у відкритому доступі для загального користування. Аналіз таксономії дозволить зорієнтуватись, де і як можна шукати потрібні дані. Наприклад, якщо потрібне джерело даних про те, скільки кілометрів проїздить авто у місті за один день, таксономія може підказати, що такі дані можна знайти у таксопарках або організаціях, що надають послуги таксі.

Маючи таку початкову інформацію, подальшими кроками можуть бути організаційні дії, а саме – з'ясування адрес та контактів організацій, до яких можна звернутись з офіційним запитом щодо отримання доступу до джерела даних або для з'ясування, чи таке джерело даних взагалі існує.

The Central Research institute
of the Armed Forces of Ukraine

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

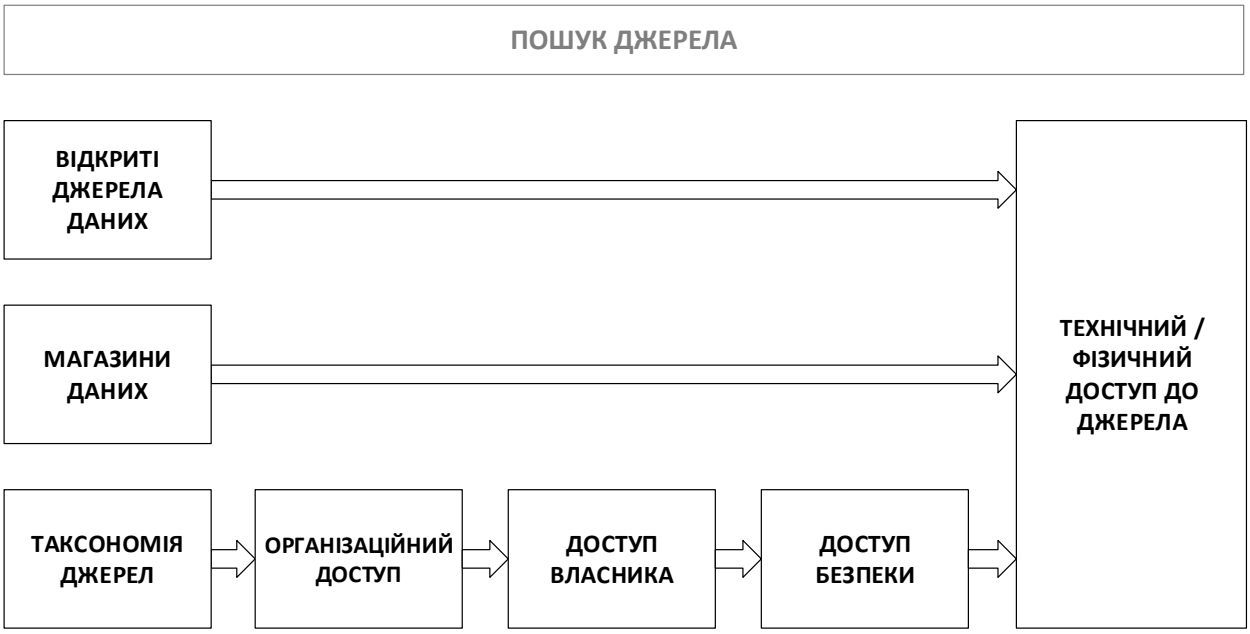


Рисунок 2. Процедура пошуку джерела даних

Слід зауважити, що будь-яке джерело може мати декілька рівнів або ступенів захисту даних, які потрібно пройти, щоб отримати до нього доступ (рис. 3). По-перше, це власник джерела даних або відповідальна особа організації, що володіє або розпоряджається джерелом, – він надає організаційний або правовий доступ до джерела даних та безпосередньо до самих даних. По-друге, це представник організації-власника джерела даних, що відповідає за безпеку даних – він надає дозвіл доступу до джерела та даних з урахуванням усіх технічних та організаційних заходів безпеки. По-третє, це також представник організації-власника джерела даних, що надає фізичний доступ до самого джерела та даних: чи то фізичний доступ до фізичних носіїв, чи то фізичний доступ до електронних носіїв даних.



Рисунок 3. Рівні захисту джерела даних

Тому знання та аналіз організаційної структури установи, до джерела даних якої потрібно отримати доступ, також є необхідною складовою у послідовності кроків пошуку джерел.

Маючи потрібні контактні дані та необхідні дозволи, можна приєднатись до джерел даних для отримання необхідної інформації на регулярній основі.

The Central Research institute
of the Armed Forces of Ukraine

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

Тепер, коли знайдене потрібне джерело, або джерела даних, і коли надано правові, нормативні та організаційні дозволи, та, відповідно, можливий безпосередній доступ до даних, потрібно виконати кроки щодо забезпечення технічної готовності до їх отримання. Послідовність цих кроків зображено на рис. 4.



Рисунок 4. Процедура отримання даних

Отже, для початку потрібно з'ясувати, що собою являє точка доступу до джерела даних із технічної точки зору. Це може бути:

FTP / SFTP сервер для обміну файлами, для доступу до цього сервера потрібні логін та пароль, а також адреса сервера;

фізичний обмін носіями з технічним представником компанії-власника джерела даних;

прямий доступ до бази даних, для чого також потрібно мати адресу СУБД, логін та пароль;

REST API – потрібно мати адресу, логін, пароль та токен доступу;

обмін файлами за допомогою електронної пошти.

Як видно з наведеного переліку, кожен з варіантів доступу до джерела використовує свій протокол доступу та обміну даними. Цей протокол також потрібно з'ясувати, щоб забезпечити автоматичне приєднання до джерела та автоматичне отримання даних. Серед існуючих протоколів можна назвати:

FTP / SFTP (File Transfer Protocol / Secure File Transfer Protocol);

REST (Representational State Transfer);

gRPC (Remote Procedure Call);

SOAP (Simple Object Access Protocol);

EDI (Electronic Data Exchange);

WebSocket.

Користуючись будь-яким з наведених протоколів обміну, можна отримувати дані у тому форматі, який: а) надає власник даних, або б) потрібен отримувачу даних. Для того, щоб це працювало, також необхідно з'ясувати формати обміну даними: чи дані є структурованими, чи неструктурованими, можливо, дані є мультимедійними або надаються у бінарному вигляді. І, зокрема, якщо передача даних відбувається за допомогою обміну файлами, то які формати файлів використовуються, наприклад, Microsoft Word *.docx, *.doc; Microsoft Excel *.xlsx, *.xls, *.csv і т.п.

Ще два питання, відповіді на які потрібно мати перед початком процесу безперервного отримання даних, це, по-перше, питання щодо періодичності обміну / отримання даних – це може бути як доба чи тиждень, чи місяць, так і більш короткий проміжок часу: година, 15 хвилин, хвилина або навіть отримання у реальному часі; та, по-друге, питання щодо об'єму даних, що передаватимуться за один сеанс: можуть бути кілобайти-мегабайти у випадку структурованих даних або десятки-сотні мегабайт у випадку неструктурованих мультимедійних даних.

Коли всю потрібну інформацію зібрано та з'ясовано всі технічні деталі процесу обміну даними, потрібно виконати тестове завантаження, щоб перевірити всі кроки та з'ясувати вірність припущень, перш ніж ставити процес завантаження даних на постійно-періодичне

The Central Research institute
of the Armed Forces of Ukraine
ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

виконання.

Перед тим, як виконати об'єднання даних, отриманих від зовнішнього джерела, з даними "своєї" системи, їх потрібно обов'язково перевірити. Наведемо декілька прикладів даних, які надходять із зовнішніх джерел, щоб пояснити зміст таких перевірок.

Приклад 1. Підготовлений шаблон таблиці у форматі Microsoft Excel було надано для заповнення у регіональні відділення організації. Заповнені шаблони з первинною інформацією надходять до головного офісу на регулярній основі (раз на місяць) та містять сотні рядків даних первинної інформації з регіонів. Один зі стовпчиків таблиці має назву "Наявність автопарку" і в рядки цього стовпчика потрібно вносити відповідну інформацію. Ось приклади змістовно прийнятних відповідей, які можуть бути отримані: "Так/Ні", "так/ні", "1/0", "+/-", "Yes/No" (всі можливі варіанти таких відповідей навести складно. Зрозуміло, що усі ці відповіді потрібно привести до єдиного формату.

Приклад 2. Інший зі стовпчиків тієї самої таблиці має назву "Юридична адреса організації". Приклади відповідей: "Одеса, вул. Остапа Бендера, 8", "вул. Верхня, Івано-Франківськ", "04136, Київ, Мазепи, 2" – знову ж таки, варіантів відповідей безліч. Якщо уявити, що кожного разу з одного регіону надходять файли з даними, які заповнюються з "чистого аркуша" різними виконавцями, – навіть адреса однієї й тієї самої організації щоразу може виглядати по-різному.

Для перевірки даних, що надходять із зовнішнього джерела даних, пропонується виконати послідовність кроків, яку представлено на рис. 5.



Рисунок 5. Процедура перевірки даних

Найнижчий рівень перевірки – це перевірка кодування даних, оскільки вони можуть надходити з різних джерел, підтримуючі системи котрих можуть працювати на різних принципах і тому використовувати різні методи або стандарти кодування. Серед поширених стандартів кодування слід згадати:

ASCII;
UTF-8;
UTF-16;
UTF-32;
UNICODE;
EBCDIC (система кодування, прийнята в комп'ютерах виробництва IBM).

Зважаючи на те, що джерела даних можуть бути знайдені і в автоматичному режимі, потрібно вміти не тільки перетворювати один вид кодування в інший, а й автоматично з'ясовувати, яка система кодування використовується.

Виконавши перетворення отриманих даних у символи, які зрозумілі системі-отримувачу, потрібно виконати лексичний аналіз даних, тобто перевірити правильність запису кожного окремого елемента даних. Якщо, наприклад, очікується ціле число, то воно має складатись тільки із цифр і не містити інших символів: "12345" – вірно, "123a!" – невірно.

Наступним кроком треба визначити вірність побудови складних елементів даних, тобто виконати синтаксичний аналіз, як у прикладі з адресою організації. Скоріш за все, очікувалось, що адреса буде надаватись у вигляді "<індекс>, <місто>, <вулиця>, <будинок>", тому

The Central Research institute
of the Armed Forces of Ukraine
ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

потрібно, щоб отримана інформація відповідала зазначеному формату.

Щоб відповісти на наступне питання – чи має сенс отримана інформація? – потрібно виконати складніший аналіз, а саме – семантичний. Інформація або дані можуть бути надані у правильному вигляді (і лексично, і синтаксично), але не мати жодного сенсу з точки зору людини. Наприклад, якщо вказана серед даних дата народження працівника організації стосується майбутнього.

Після детального аналізу всі дані будуть розділені на два різних набори: 1-й – дані, що пройшли перевірку і їх можна об'єднувати з даними системи, та 2-й – дані з помилками. 2-й набір даних буде розглянуто нижче.

Нарешті отриманий набір даних від зовнішнього джерела, що пройшов повну перевірку, готовий до об'єднання з існуючими даними в системі. Кроки процедури об'єднання (рис. 6) є наступними:

Ідентифікація даних (встановлення факту належності різних даних до одного й того самого об'єкта, процесу, явища);

Вилучення дублікатів даних (якщо дані мають спільне першоджерело);

Покращення якості (точності, достовірності, повноти) даних шляхом спільного оброблення ідентифікованих даних.

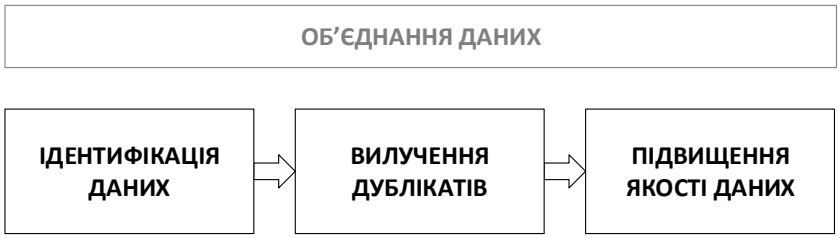


Рисунок 6. Процедура об'єднання даних

Об'єднання даних

Ідентифікація даних

Отримавши набори даних від різних зовнішніх джерел, необхідно здійснити перевірку можливості належності цих наборів до одних і тих самих об'єктів зацікавленості для системи, тобто здійснити їхню ідентифікацію. Це питання дуже просто вирішується, якщо джерела певним чином узгоджують своє функціонування і маркують набори даних унікальними ідентифікаторами об'єктів (явищ, процесів), або у випадку, якщо усі значення елементів даних у наборі повністю збігаються (за умови, що в межах відповідальності системи немає повністю тотожних за ознаками об'єктів). Однак, враховуючи те, що у загальному випадку джерела інформації функціонують незалежно і характеризуються похибками визначення значень елементів даних, без спеціальної процедури ідентифікації кожен набір даних буде сприйматися системою, як дані про різні об'єкти. Тому ідентифікація є першим кроком об'єднання нових даних із існуючими (або такими, що надходять паралельно). Якщо умова відсутності у зоні відповідальності системи тотожних за ознаками об'єктів строго виконується, задачу ідентифікації може бути розв'язано наступним чином.

Як уже зазначалося, значення даних від різних джерел щодо одного й того самого об'єкта не будуть збігатися через похибки їх визначення (можна зазначити, що ймовірність повного збігу визначених значень елементів даних на деякій їх множині прямує до нуля при збільшені розміру такої множини). Однак, якщо похибки визначення значень відомі (чи можуть бути оцінені), ідентифікацію наборів даних можна здійснити шляхом задавання певної міри близькості таких наборів за значеннями множини елементів, що їх складають. Якщо значення міри близькості за значенням кожного елемента даних менше похибок їх визначення – може

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

бути ухвалене рішення, що такі набори даних стосуються одного й того самого об'єкта. Задачу ідентифікації об'єктів за аналізом міри близькості між значеннями елементів даних у наборі може бути розв'язано методами кластеризації із застосуванням нечіткої формалізації елементів даних якісного змісту. Нечітка формалізація таких даних (окрім тих, що визначаються номінальними шкалами) дозволить визначати відстань між їхніми значеннями. Для номінальних шкал може бути застосовано дискретну логіку: якщо значення даних різних наборів збігаються – відстань між ними мінімальна. Якщо ні – максимальна. Задача ідентифікації достатньо якісно розв'язується для наборів даних, що стосуються фізичних об'єктів, якщо вони містять дані про місце їх розташування (координати). Фактично координати об'єктів є головними і "надійними" ідентифікуючими ознаками множин даних.

Зазначимо, що якщо серед об'єктів зацікавленості системи можуть бути різні об'єкти із повністю тотожними значеннями ознак (наведена вище умова не виконується) – однозначного розв'язання задачі ідентифікації не існує.

Вилучення дублікатів даних

Беручи до уваги викладені вище умови, можна констатувати, що наявність двох наборів даних у системі із повністю тотожними значеннями їхніх елементів є однозначною ознакою того, що дані сформовані одним і тим самим першоджерелом, навіть якщо отримані системою від декількох різних джерел. Такі дані хибно вважати доповнюючими (такими, що зменшують невизначеність щодо дійсних характеристик об'єкта) і їх має бути вилучено з подальшого аналізу.

Покращення якості даних

Набори даних, що пройшли ідентифікацію як такі, що належать до одного й того самого об'єкта зацікавленості, підлягають спільному обробленню з метою підвищення якості даних. Першим етапом такого оброблення є просте доповнення даних про об'єкт до повного списку елементів. Тобто, якщо у двох наборах даних є унікальні елементи – їх має бути додано до загальної множини елементів об'єднаної множини. Тобто – кількість елементів даних про об'єкт зросте до кількості унікальних елементів у кожному з наборів даних. Таким чином, очевидно збільшиться повнота даних про об'єкт. Якщо дані містять значення одного й того самого елемента – до них має бути застосовано процедуру комплексування даних – отримання узагальненого значення елемента. Для об'єднання числових значень, що фактично є результатом вимірювань, може бути застосовано методи оброблення, у загальному випадку, нерівноточних даних із відповідної теорії вимірювань, що у середньостатистичному сенсі збільшить точність об'єднаних значень даних. Для даних якісного характеру у числовій, лінгвістичній чи іншій формі може бути застосовано формалізацію у вигляді нечітких множин із подальшим використанням різних методів їх (множин) об'єднання. У результаті такого оброблення може бути підвищено як точність, так і достовірність об'єднаних значень даних якісного характеру.

Отже, у результаті об'єднання даних, що надходять від множини джерел, може бути підвищено якість об'єднаних даних (повнота, точність та достовірність) та вилучено дублюючі дані. При цьому мінімізується можливість того, що дані про один і той самий об'єкт будуть існувати у системі як такі, що стосуються різних об'єктів.

Помилки під час інтеграції

Як зазначалося вище, під час виконання процедур перевірки та об'єднання даних можуть бути виявлені дані із помилками, що не підлягають автоматичній інтеграції з даними системи-отримувача. Зміст таких відфільтрованих даних доцільно додатково проаналізувати та спробувати виправити помилки.

Для виправлення помилок можна виконати наступні дії:

- Виявити аномальні дані (наприклад, методами кластеризації, статистичними методами чи методами аналізу часових рядів);

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

- Заповнити відсутні дані середнім значенням, отриманим на основі аналізу всіх інших або сусідніх записів набору даних (найпростіша інтерполяція);
- Використати більш складну інтерполяцію (поліноми, сплайни) і розрахувати потрібні значення на основі сусідніх точок, чи екстраполяцію – отримати наступні значення даних за попередніми;
- Застосувати машинне навчання (наприклад, k-Nearest Neighbors (k найближчих сусідів), або регресійний аналіз) для прогнозування відсутніх значень;
- Використати реверсний інжиніринг – спробу відновити дані за допомогою інших джерел або метаданих.

Зазначені дії дозволять, принаймні частково, відновити дані, що містять помилки, і надалі застосувати до них автоматизовану або навіть автоматичну процедуру їх інтеграції з іншими існуючими у системі даними.

Обговорення (Discussions)

Результати проведеного дослідження підтверджують, що перехід від “клаптикової” автоматизації до інтегрованих систем високого порядку є критичним етапом еволюції складних організаційних структур. Запропонований комплексний підхід, що охоплює етапи від пошуку джерел до об’єднання даних, корелює із сучасними світовими тенденціями. Зокрема, використання методів штучного інтелекту (ШІ) для автоматизації інтеграції узгоджується з підходами, описаними у стратегіях моніторингу глобальних процесів, де ШІ виступає ключовим інструментом оброблення великих масивів неструктурованих даних.

Порівняно з існуючими рішеннями, такими як користувачо-центровані платформи Big Data або системи промислової інтеграції на базі туманних обчислень, запропонована методика акцентує увагу на багаторівневій перевірці даних (лексичній, синтаксичній та семантичній). Важливим аспектом, що виділяє це дослідження, є фокус на автоматичному визначенні систем кодування та відновленні помилкових даних за допомогою алгоритмів машинного навчання (k-NN, регресійний аналіз) та реверсного інжинірингу. Це дозволяє мінімізувати втрати інформації, що є критичним для систем біології та фармакології, де точність даних має вирішальне значення.

Однак у процесі дослідження було виявлено певні дискусійні моменти. По-перше, попри високий рівень автоматизації, етап пошуку джерел даних залишається найбільш складною задачею, яка в повному обсязі може бути виконана лише людиною. Це створює певний “вузький прохід” для повністю автономних систем. По-друге, розв’язання задачі ідентифікації об’єктів за допомогою міри близькості та нечіткої логіки є ефективним лише за умови відсутності в системі абсолютно тотожних за ознаками об’єктів; у протилежному випадку однозначного математичного розв’язку не існує.

Критичне міркування щодо впровадження методів ШІ полягає в їхній масштабованості. Хоча ШІ легко масштабується для роботи з тисячами джерел, технічна реалізація точок доступу (REST API, gRPC, FTP) вимагає постійної підтримки актуальності протоколів обміну, що може бути ускладнено у застарілих системах.

На основі отриманих результатів можна надати такі рекомендації:

організаціям слід попередньо розробляти та публікувати таксономії джерел даних для полегшення автоматизованого пошуку;

застосовувати процедури комплексування нерівноточних даних для підвищення достовірності інтегрованих масивів.

Перспективним напрямом подальших досліджень є розроблення алгоритмів автоматичного подолання семантичних розбіжностей у реальному часі без участі аналітика, а також інтеграція запропонованого підходу з технологіями блокчейн для забезпечення незмінності та простежуваності джерел даних. Представлені результати доводять наукову

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

коректність обраних позицій та створюють надійне підґрунтя для побудови інтегрованих інформаційних просторів у спеціалізованих доменах.

Висновок (Conclusion)

У статті розглянуто задачі інтеграції даних спеціалізованих інформаційно-комп'ютерних систем у складних організаційних структурах, що має критичне значення для подолання "клаптикової" автоматизації та створення систем управління вищого порядку.

Аналіз показав, що функціональна несумісність автономних компонентів, дублювання функцій, відсутність єдиних протоколів обміну вимагають розроблення комплексних методичних рекомендацій, які охоплюють технічні, організаційні, семантичні аспекти процесу інтеграції.

Запропоновано системний підхід, який передбачає чотири послідовні етапи: пошук джерел, отримання даних, їх перевірку та об'єднання.

Ключовою особливістю підходу є поєднання методів штучного інтелекту для автоматизації виявлення джерел і масштабного оброблення гетерогенної інформації з багаторівневим аналізом (лексичним, синтаксичним та семантичним) для забезпечення змістовної коректності даних.

Важливою складовою методики є застосування апарату математичного моделювання, зокрема методів кластеризації з нечіткою формалізацією та теорії вимірювань для комплексування нерівноточних даних.

Це дозволяє ефективно здійснювати ідентифікацію об'єктів та вилучати дублікати, що виникають при отриманні інформації з незалежних джерел.

Окрему увагу приділено механізмам виправлення помилок за допомогою алгоритмів машинного навчання, регресійного аналізу та реверсного інжинірингу, що дозволяє мінімізувати втрати корисної інформації.

Отримані результати підтверджують, що впровадження запропонованих кроків забезпечує підвищення загальної якості інтегрованих даних, їхньої повноти, точності та достовірності.

Зазначене створює надійне підґрунтя для ухвалення більш ефективних управлінських рішень та оптимізації технологічних процесів у конкретних прикладних областях.

Перспективними напрямами подальших досліджень є розроблення алгоритмів автоматичного подолання семантичних розбіжностей у реальному часі, розширення переліку підтримуваних протоколів автоматичного обміну та інтеграція запропонованого підходу з технологіями інтелектуального моніторингу для роботи із надвеликими масивами даних.

Конфлікт інтересів (Conflict of Interest)

Одним із авторів статті є член редакційної колегії. Для уникнення конфлікту інтересів редакційний процес здійснював Олег Семененко.

Використані джерела:

1. Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006) (pp. 9–16)*. VLDB Endowment.
2. Lenzerini, M. (2002). Data integration: A theoretical perspective. *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002)* (pp. 233–246). https://doi.org/10.1145/543613.543644
3. Fillinger, S., de la Garza, L., Peltzer, A., Kohlbacher, O., & Nahnsen, S. (2019). *Challenges of big data integration in the life sciences*. Analytical and Bioanalytical Chemistry. https://doi.org/10.1007/s00216-019-02038-7
4. Jovanovic, P., Nadal, S., Romero, O., Abelló, A., & Bilalli, B. (2020). *Quarry: A user-centered*

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

big data integration platform. Information Systems Frontiers. https://doi.org/10.1007/s10796-020-10031-9
5. Le Fèvre, C., Poty, L., & Noël, G. (2017). Big data, generalities and integration in radiotherapy. *Cancer/Radiothérapie, 21* (6–7), 503–508. https://doi.org/10.1016/j.canrad.2017.05.008
6. Wang, J., Zheng, P., Lv, Y., Bao, J., & Zhang, J. (2019). Fog-IBDIS: Industrial big data integration and sharing with fog computing for manufacturing systems. *Engineering, 5* (4), 685–698. https://doi.org/10.1016/j.eng.2018.12.013
7. Hämäläinen, E., & Inkinen, T. (2019). Industrial applications of big data in disruptive innovations supporting environmental reporting. *Journal of Industrial Information Integration, 15*, 100–110. https://doi.org/10.1016/j.jii.2019.04.001
8. Ma'ayan, A., Rouillard, A. D., Clark, N. R., Wang, Z., Duan, Q., & Kou, Y. (2014). Lean big data integration in systems biology and systems pharmacology. *Trends in Pharmacological Sciences, 35* (9), 450–460. https://doi.org/10.1016/j.tips.2014.06.004
9. Meaningful big data integration for a global COVID-19 strategy. (2020). *IEEE Computational Intelligence Magazine, 15* (4), 16–24. https://doi.org/10.1109/MCI.2020.3020978
10. Nash, S. S. (2019). *The role of big data and machine learning in the integration and implementation of historical, current, and continuously gathered Earth data*. Proceedings of the Geological Society of China.

Modern Challenges of Data Integration in Specialized Information and Computer Systems

Oleksandr Dodonov
Doctor of Technical Sciences, Professor
Deputy Director for Research
Institute for Information Recording of the National Academy of Sciences of Ukraine
03113, M. Shpaka St., 2, Kyiv, Ukraine
e-mail: dodonovua@gmail.com
ORCID: 0009-0000-7831-6337

Vadym Dodonov
PhD in Technical Sciences
Senior Research Fellow
Institute for Information Recording of the National Academy of Sciences of Ukraine
03113, M. Shpaka St., 2, Kyiv, Ukraine
e-mail: v.o.dodonov@nas.gov.ua
ORCID: 0009-0002-3480-2549

Valerii Zaluzhnyi
Ambassador Extraordinary and Plenipotentiary of Ukraine to the United Kingdom
Permanent Representative of Ukraine to the International Maritime Organization (IMO)
e-mail: uzvk_mou@post.mil.gov.ua
ORCID: 0000-0002-1947-501X

Kostyantyn Kravchuk
Senior Research Fellow
Institute for Information Recording of the National Academy of Sciences of Ukraine
e-mail: mailto:k.kravchuk@ipri.kiev.ua
ORCID: 0009-0000-1750-4588

The Central Research institute
of the Armed Forces of Ukraine

ISSN (print): 2786-9318
ISSN (online): 2786-9326
Military Science. Vol. 3 No. 4 (2025)

Volodymyr Yuzefovych
PhD in Technical Sciences, Associate Professor
Senior Research Fellow
Educational and Scientific Institute of Physics and Technology
Igor Sikorsky Kyiv Polytechnic Institute
03056, Beresteiskyi Avenue, 37, Kyiv, Ukraine
e-mail: uzefv71@gmail.com
ORCID: 0000-0002-6336-9548

Ihor Izvarin
PhD in Technical Sciences
Associate Professor Institute of Computer Technologies
Open International University of Human Development "Ukraine"
03115, Lvivska St., 23, Kyiv, Ukraine
e-mail: igor.izvarin@gmail.com
ORCID: 0009-0004-4827-9968

Abstract. The relevance of the study is determined by the development of business process automation in complex organisational systems, which leads to the need to integrate multiple autonomous information and computer systems into a unified higher-order system. The problem lies in the functional incompatibility of autonomous systems, data duplication, and the absence of common exchange protocols, which hinder effective interaction and reduce the quality of information support. In the context of increasing volumes of information, the task of data integration becomes critical for improving the efficiency of technological processes and making well-founded managerial decisions. The aim of the study is to identify modern approaches to solving the problem of data integration in complex organisational systems through the development of technical, organisational, and semantic steps for analysing, transforming, and consolidating information. To achieve this aim, artificial intelligence methods were employed to automate the identification of data sources and formats, as well as lexical, syntactic, and semantic analysis methods to verify the meaningfulness of information. In addition, mathematical modelling tools were applied, including clustering methods using fuzzy formalisation, discrete logic, and measurement theory for processing imprecise data during integration. Statistical methods, time series analysis, and machine learning algorithms were used to correct errors. As a result of the study, a comprehensive methodological approach to the data integration process was developed, structured into four stages: source identification, data acquisition, validation, and integration. Procedures for the automatic detection of encoding systems, identification of objects based on value similarity measures, and mechanisms for restoring erroneous data through interpolation, extrapolation, and reverse engineering were proposed. This makes it possible to minimise the loss of useful information and eliminate duplication. The practical significance of the work lies in the fact that the proposed results can be used to create integrated systems capable of maximising the efficiency of automation in specific application domains. The presented approach ensures the resolution of issues related to data retrieval, elimination of incompatibility, and improvement of the overall quality (completeness, accuracy, and reliability) of integrated datasets. The findings provide a foundation for improving real-time information exchange processes between modern and legacy systems.

Keywords: data integration, information systems, computer systems, data processing, semantic interoperability, information analysis, complex organizational systems, data transformation.