УДК 004.652

А. Берко
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ЗАСТОСУВАННЯ ОНТОЛОГІЙ У ПРОЦЕСАХ СЕМАНТИЧНОЇ ІНТЕГРАЦІЇ ДАНИХ

Розглянуто способи і можливості створення семантично інтегрованих структур даних для зберігання контентного ресурсу гетерогенних інформаційних систем. Запропоновано формальні методи інтеграції семантики різнорідних даних, що ґрунтуються на застосуванні інтелектуальних засобів та онтологій даних.

Some ways and possibilities of semantically integrated data structures development for information storage of heterogeneous information systems content are considered in this paper. Formal methods of semantic integration of data based on intelligent tools and data ontology application were proposed.

Вступ

Для сучасного етапу розвитку інформаційних технологій характерним є постійне зростання обсягів інформації та завдань, пов'язаних з її застосуванням у різноманітних галузях. Поширення інформаційних систем великого масштабу, таких як соціальні мережі, корпоративні системи прийняття рішень, сховища даних викликає потребу інтеграції, ресурсів таких систем. Проблематика, пов’язана зі створенням методів та засобів інтеграції, є актуальною як в теоретичному, так і в прикладному аспектах. Сьогодні на ринку інформаційних технологій багато провідних виробників пропонують власні інструментарії вирішення проблем інтеграції на різних рівнях, таких як інтеграція бізнес-процесів *(Business Ptrocess Integration – BPI)*, інтеграція корпоративних застосувань (Entetrprise Application Integration- EAI), інтеграція корпоративних платформ *(Platform Integration – EPI)* інтеграція даних або, як часто її називають, інтеграція корпоративної інформації *(Enterprise Information Integration – EII)*. Інтеграція даних – це завдання об'єднання даних, отриманих з різних джерел з метою їх об'єднаного подання користувачеві [1, 3].

Процеси інтеграції даних передбачають вирішення таких проблем, як інтеграція значень, синтаксису, структури семантики даних [1]. Кожна з цих проблем має свої методи та засоби вирішення. У статті розглянуто формальні методи реалізації одного з напрямів інтеграції даних – семантичної інтеграції. Інтеграція семантики даних є порівняно складним завданням через відсутність достатньо ефективних формальних засобів подання змісту наборів даних. У роботі запропоновано вирішення завдань інтеграції семантики даних на основі застосування онтологій.

Аналіз останніх публікацій і досліджень

Основні теоретичні засади семантичної інтеграції сформульовано у [3, 5, 6] де обґрунтовано поняття семантичної інтеграції, а також запропоновано як інструментарій вирішення проблеми побудови інтегрованого змістового простору даних такі засоби, як контекстуальний аналіз та онтології. У [4] визначено поняття критеріїв семантичної інтеграції даних та поняття семантичної дистанції між наборами даних.

Основні принципи формування та застосування метаданих як засобу визначення семантики даних викладено у [9, 10]. Зокрема, у [10] визначено схему Захмана для організації метаданих, за допомогою яких описують семантику даних у процесах їх інтеграції. Ця схема передбачає формування набору метаданих за певними принципами, які вимагають визначення таких елементів, як відомості про об’єкт даних, суб’єкт даних, часові показники, місцезнаходження та застосування даних, призначення даних та спосіб їх використання.

У [9] описано схему побудови метаданих, що отримала назву "Дублінське ядро" за назвою міста Дублін (штат Огайо, США), де працювала робоча група зі створення відповідного стандарту. *Дублінське ядро* (англ. ***Dublin Core***) — стандарт (формат) метаданих – простий та ефективний набір значень для опису широкого діапазону інформаційних ресурсів. Семантику Дублінського ядра було створено міжнародною міждисциплінарною групою фахівців з комп'ютерних наук, Інтернет-технологій, бібліотечної справи, кодування текстів, музейної справи та інших суміжних галузей.

Стандарт поділяють на два рівні [9]:

- простий (некваліфікований, ***simple***), до складу якого входять 15 елементів;
- компетентний (кваліфікований, ***qualified***), у складі 18 елементів і групи *кваліфікаторів*, що уточнюють семантику елементів з метою підвищення якості пошуку інформаційних ресурсів.

У [9] показано можливості поєднання принципів побудови метаданих на основі Дублінського ядра з такими засобами опису семантики інформаційних ресурсів, як RDF-XML та OWL.

Застосування онтологій у процесах семантичної інтеграції даних описано у [4, 5]. Онтологія визначає множину концептів та зв’язків між ними без специфікації предметної області [3]. Загалом, онтології утворюють інфраструктуру, необхідну для визначення семантики даних, яка може бути сприйнятою і опрацьованою програмними засобами. Такий підхід дає змогу коректно вирішувати проблеми, пов'язані зі змістом даних на формальному рівні із застосуванням засобів інформаційних технологій. Відмінність між онтологією, що описує інформаційний ресурс, та метаданими, такими як, наприклад, Дублінське ядро, є, на перший погляд, малопомітною, але істотною. Незважаючи на те, що обидва засоби застосовують для семантичної інтеграції даних, принципова відмінність між ними полягає у ступені участі людини в інтеграційних процесах. Метадані загалом створюють, редагують і інтерпретують люди. Тому суб'єктивні чинники, зокрема обмеження щодо складності їх подання та розуміння є вирішальним фактором. На противагу до метаданих, онтологія є базовою формальною моделлю засобів інтеграції інформаційних ресурсів та реалізації різноманітних додаткових функцій. Отже, застосування онтологій дає змогу оперувати більш складними та формалізованими поняттями, які часто виходять за межі людської компетенції [5].

Онтологію даних розглядають як засіб різнобічної і детальної формалізації знань про дані за допомогою концептуальної схеми. Як правило, до складу такої схеми входить опис структури даних, що містить визначення всіх релевантних класів об'єктів, їхні взаємозв'язки та правила (теореми, обмеження), задані у предметній області набору даних [3]. Онтології даних можна описувати різними засобами, і сьогодні поширені багато мов опису онтологій. Проте, з огляду на те, що в будь-якій онтології визначають терміни і задаються логічні зв'язки між ними, точна семантика опису термінів і зв'язків в різних мовах буде однаковою. Онтологічні системи будують на основі таких принципів [2]:

- формалізації, тобто опису об'єктивних елементів реальності із застосуванням єдиних, суворо визначених зразків (термінів, моделей тощо);
- використання обмеженої кількості базових термінів (сутностей), на основі яких конструюють всі інші поняття;
- внутрішньої повноти;
- логічної несуперечності.

На відміну від звичайного словника, для онтологічної системи характерні внутрішня єдність, логічний взаємозв'язок і несуперечність використовуваних понять.

Одним із напрямів у дослідженні методів і засобів опрацювання даних на основі онтологій, які сьогодні активно розвиваються, є напрям, пов'язаний із застосуванням Web-онтологій [7]. Засоби, створені в цій галузі, такі як XML, RDF (***Resource Definition Framework***) та OWL (***Web Ontology Language***), можна застосувати і у процесах семантичної інтеграції даних [6]

Мета та завдання досліджень

Цілі статі. Основними цілями цієї статті є:

1. Узагальнення та класифікацій підходів до семантичної інтеграції даних на різних рівнях їх подання та сприйняття.

2. Розроблення формальних моделей і методів інтеграції семантики даних, що ґрунтуються на застосуванні інтелектуальних засобів та онтологій.

3. Визначення порядку та місця застосування методів і моделей семантичної інтеграції у процесах проектування структури інтегрованого контенту та управління ним.

4. Розроблення формальних критеріїв семантичної інтеграції різнорідних даних.

Невирішені проблеми. Основні невирішені проблеми виникають через неоднорідність структури, форми та змісту елементів інформаційного ресурсу корпоративних та глобальних інформаційних систем. За оцінками експертів [8], лише близько 20 відсотків ресурсу сучасних інформаційних систем різноманітного спрямування зберігають у базах даних як структуровану інформацію, решту становлять дані, подані у різноманітних форматах без попередньо визначеної чіткої структури – так звані, слабкоструктуровані дані [8]. Різноманітність форм та змісту контенту, своєю чергою, викликає необхідність розроблення методів і технологій, які забезпечують спільне, узгоджене опрацювання і застосування таких неоднорідних даних. Загальний процес інтеграції даних передбачає інтеграцію їх структури, синтаксису та семантики. Проблеми інтеграції на рівні синтаксису та структури значною мірою досліджено, і для їх вирішення існує низка методів та засобів [5]. Натомість проблематика семантичної інтеграції є значно ширшою і менш формалізованою і менш дослідженою. Найістотнішою проблемою інтеграції семантики даних є відсутність єдиного універсального підходу та інтероперабельного інструментарію для її практичної реалізації.

Основні результати досліджень

Методи інтеграції семантики неоднорідних даних. Семантика є невід'ємною властивістю даних, що забезпечує їх змістовність та можливість застосування даних за їх призначенням. Загалом семантику визначає множина відповідностей між формальними позначеннями та реальними поняттями предметної області, що дає змогу однозначно інтерпретувати дані на різних стадіях роботи з ними. В сучасних інформаційних технологіях застосовують багато різноманітних засобів визначення семантики даних. Найпростішими засобами інтерпретації даних є, наприклад, описи стовпчиків таблиць, XML-тегів, розділів документа тощо. У великих системах, до яких належать системи електронного бізнесу, застосовують складніші засоби визначення семантики даних, зокрема, тезауруси (словники даних), метадані та онтології [3].

Інтеграція семантики даних передбачає формування єдиного змістового простору для сприйняття, інтерпретації та застосування даних незалежно від формату їх подання та структури. Основною метою таких процесів є однакова інтерпретація всіх елементів даних, отриманих з різноманітних джерел, та складових деякої об’єднаної інформаційної структури. Однією з основних проблем у цьому процесі є формування критеріїв семантичної інтеграції наборів даних, за допомогою яких можна оцінити можливість чи неможливість об'єднання їх змісту та розв’язати семантичні конфлікти між даними наборів, які підлягають інтеграції.

У [5] визначено зокрема три основні типи семантичних конфліктів:

- *конфлікт неоднозначності* – виникає, коли два поняття виглядають однаковими, але є різними по суті, наприклад, поняття "зобов’язання";
- *конфлікт метрик* – виникає, коли однакові величини вимірюють одиницями різних систем, наприклад, використання різних валют для визначення ціни одного товару;
- *конфлікти імен* – виникає, коли системи іменування понять та об’єктів є принципово відмінним; найчастіше це виявляється у появі синонімів та омонімів [5].

Використовуючи онтології для явного визначення неявних чи прихованих знань про дані та їх зміст, можна вирішити значну частину проблем семантичної неоднорідності даних.

Семантична інтеграція даних на основі онтологій. Найперспективнішим сьогодні підходом до інтеграції семантики даних є *інтеграція на основі онтологій*. Цей метод передбачає використання основних елементів двох попередніх методів – тезаурусу та метаданих, але є значно загальнішим за них та враховує більше аспектів семантики даних. Вперше застосування онтологій як засобу семантичної інтеграції було запропоновано в [4]. Загалом онтологію розглядають як

цілісну формалізовану специфікацію деякої предметної області, яка має на меті забезпечити однакову інтерпретацію знань про цю предметну область на людському та комп'ютерному рівнях. У випадку інтеграції даних об’єктом опису, поданого у вигляді онтології, є певний інформаційний ресурс. Тому доцільно говорити про специфічну категорію онтологій – онтології даних.

Методика семантичної інтеграції даних на основі онтологій передбачає вирішення таких проблем [4]:

- *порядок застосування онтології* – роль та загальна архітектура онтології як засобу опису семантики даних значною мірою впливають на способи її формування та подання;
- *зображення онтології* – залежно від призначення та застосування онтологій, способи їх зображення можуть бути дуже відмінними у різних випадках;
- *застосування рівності понять* – у інтеграційних процесах залежно від специфіки даних, які підлягають інтеграції, рівність може мати різні інтерпретації, зокрема застосовують такі варіанти, як точна рівність, частковий збіг, еквівалентність, подібність тощо;
- *побудова онтологій* – перед початком процесів семантичної інтеграції для кожного локального набору даних необхідно створити власну онтологію, орієнтовану на взаємодію з онтологіями інших наборів даних.

У загальному випадку формальне зображення онтології даних подають як трійку елементів вигляду

$$O = <X, R, F>,$$

де X — скінченна множина понять (класів, концептів) предметної області з їх властивостями (атрибутами); R — скінченна множина відношень (зв’язків, відповідностей) між поняттями; F — скінченна множина функцій інтерпретації (обмежень, аксіом) [3].

Згідно з вимогами стандарту IDEF5 [2], концепти поділяють на класи та значення класів. При цьому класи можуть утворювати ієрархію, тобто значенням класу може бути інший клас (підклас), наприклад, до класу *"документи"* можуть як значення входити підкласи *"текстові документи"*, *"XML-документи"*, *"PDF-документи"* тощо. Зв’язки між концептами поділяють на класифікаційні – між класами і підкласами і структурні, які описують взаємодію класів. Прикладом структурних зв’язків є відповідності між розділами цієї статті, які утворюють цілісний інформаційний ресурс.

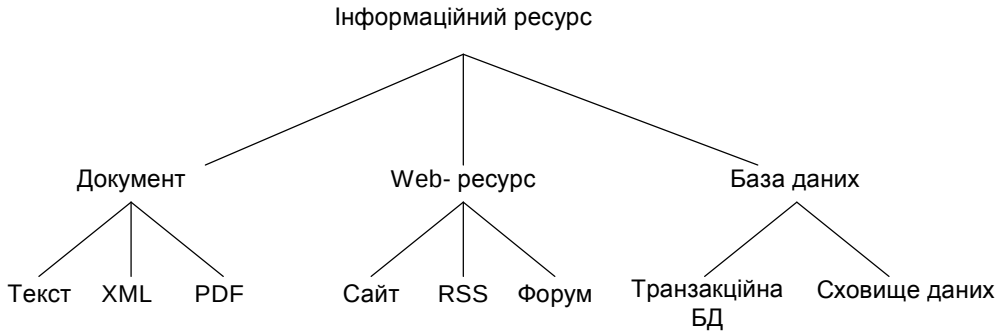

Рис. 1. Приклад визначення класів онтології інформаційного ресурсу

На рис. 1 показано один з варіантів класифікації інформаційних ресурсів з погляду інтеграції. Клас "Інформаційний ресурс" в своєму складі містить підкласи *"Текст"*, *"Web-ресурс"* та *"База даних"*, які, своєю чергою, поділяються на дрібніші підкласи. Кількість рівнів ієрархічної класифікації залежить від конкретних вимог та особливостей процесів інтеграції даних. На рис 2. показано приклад онтології, що описує структуру наукової статті у вигляді концептів, які описують її змістові елементи та зв’язки між ними.

Процеси семантичної інтеграції даних передбачають створення для кожного вхідного набору даних D_i онтології $O(D_i)$, яка формує однозначний опис семантики як всього інформаційного ресурсу, так і окремих його елементів

$$O(D_i) = <X(D_i), R(D_i), F(D_i)>,$$

де $X(D_i)$ – множина концептів, які описують одиниці даних, їх зміст, властивості та належність до певного класу чи категорії; $R(D_i)$ – множина зв'язків і відношень між одиницями даних, що визначають порядок їх взаємодії та взаємного застосування; $F(D_i)$ – множина семантичних обмежень та функцій інтерпретації даних, які пов'язують їх з реальними поняттями та об'єктами предметної області, а також регламентують порядок визначення таких відповідностей.

Така онтологія описує семантичний зв'язок визначених і специфікованих елементів даних з поняттями предметної області, утворюючи цілісну структуру "дані–зміст". Оскільки об'єктом опису онтології у випадку семантичної інтеграції є дані, то її можна класифікувати як прикладну онтологію, реалізовану у формі метаданих спеціального виду. У такий спосіб проблему семантичної інтеграції даних можна звести до проблеми виявлення відповідностей та суперечностей між їх онтологіями.

Проблематика формування та застосування онтологій як засобу опису семантики даних у процесах інтеграції гетерогенних інформаційних ресурсів має декілька підходів [5], для кожного з яких характерними є власні моделі, способи та засоби вирішення. Це зокрема такі.

Інтеграція даних на основі єдиної онтології. У цьому випадку для явної специфікації семантики різних наборів даних формують єдину глобальну онтологію зі спільними узгодженими розподіленими ресурсами [5]. Єдину онтологію може бути сформовано двома методами.

(1) Шляхом розподілу – при цьому утворюється глобальний опис концептів, відношень та функцій інтерпретації з розподіленими словниками, який застосовують для специфікації семантики кожного з наборів даних, які підлягають інтеграції. При цьому глобальну онтологію O^G інтегрованого набору даних DI можна подати як єдину узагальнену систему:

$$O^G(DI)=O(D_1, D_2, \ldots, D_N)=< X(D_1, D_2, \ldots, D_N), R(D_1, D_2, \ldots, D_N), \ldots, F(D_1, D_2, \ldots, D_N)>,$$

де $O^G(DI)$ – глобальна онтологія інтегрованого ресурсу, яка одночасно є спільною онтологією для всіх вхідних локальних інформаційних ресурсів; $X(D_1, D_2, \ldots, D_N)$ – спільна множина концептів вхідних ресурсів $D_1, D_2, \ldots, D_N$; $R(D_1, D_2, \ldots, D_N)$ – спільна множина відношень вхідних ресурсів $D_1, D_2, \ldots, D_N$; $F(D_1, D_2, \ldots, D_N)$ – спільна множина правил інтерпретації вхідних ресурсів $D_1, D_2, \ldots, D_N$.

(2) Шляхом інтеграції – такий спосіб передбачає формування та поповнення глобальної онтології як результатів узгодженого об'єднання словникових ресурсів локальних онтологій, сформованих для наборів даних, які підлягають інтеграції. У цьому випадку порядок формування глобальної онтології O^G інтегрованого набору даних DI можна описати виразом:

$$O^G(DI)=I^O(O_1(D_1), O_2(D_2), \ldots, O_N(D_N))=< I^X(X_1,X_N,\ldots,X_N), I^R(R_1,R_N,\ldots,R_N), I^R(F_1,F_N,\ldots,F_N)>$$

де $O^G(DI)$ – глобальна онтологія інтегрованого ресурсу; I^O – оператор інтеграції онтологій; $O_1(D_1), O_2(D_2), \ldots, O_N(D_N)$ – локальні онтології вхідних інформаційних ресурсів $D_1, D_2, \ldots, D_N$,; I^X – оператор інтеграції концептів; $X_1,X_N,\ldots,X_N$ – локальні множини концептів вхідних інформаційних ресурсів $D_1, D_2, \ldots, D_N$; I^R – оператор інтеграції відношень; $R_1,R_N,\ldots,R_N$ – локальні множини відношень вхідних інформаційних ресурсів $D_1, D_2, \ldots, D_N$; I^F – оператор інтеграції функцій інтерпретації; $F_1,F_N,\ldots,F_N$ – локальні множини функцій інтерпретації вхідних інформаційних ресурсів $D_1, D_2, \ldots, D_N$.

Особливістю семантичної інтеграції даних на основі єдиної онтології є спільне використання її ресурсів для опису семантики кожного вхідного набору даних.

Перевагою такого способу інтеграції даних є однотипність визначення та інтерпретації концептів в усіх вхідних наборах даних, відсутність неоднозначностей формулювання понять та суперечностей імен і метрик. Це, своєю чергою, спрощує процеси формування єдиного семантичного простору для інтегрованого набору даних, зменшує обсяги самих онтологій та додаткових метаданих.

Проблемою такого підходу є саме формування єдиної глобальної онтології, оскільки загалом не завжди можна забезпечити єдність концептуалізації усіх вхідних даних

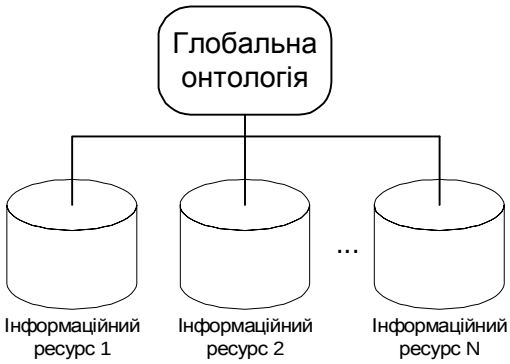

Рис. 2. Схема семантичної інтеграції даних на основі єдиної онтології

Інтеграція даних на основі множини онтологій. Цей підхід застосовують, коли побудова єдиної глобальної онтології для опису семантики всіх вхідних наборів є неможливою чи складною. У такому випадку кожен вхідний набір даних для семантичної інтеграції описують власною онтологією, яка не пов'язана з іншими і оперує власними нерозподіленими словниковими ресурсами. Процес семантичної інтеграції у цьому випадку ґрунтується на узгодженні, взаємодії та обміні ресурсами локальних онтологій (рис. 3).

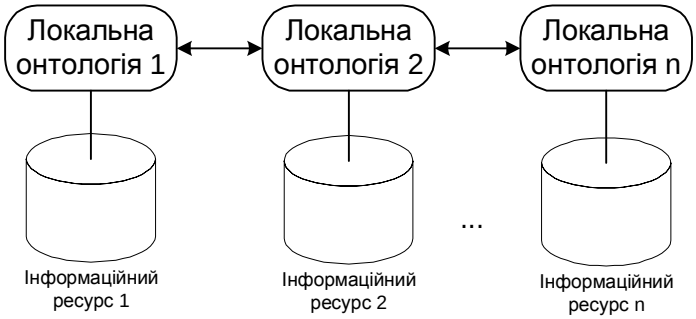

Рис. 3. Схема семантичної інтеграції даних на основі множини онтологій

Семантична інтеграція даних при цьому вимагає застосування методів та засобів побудови і опрацювання онтологій, що забезпечують їх спільне застосування у формуванні єдиного семантичного простору інтегрованих даних.

Основною перевагою інтеграції на основі множини онтологій є відносна автономність засобів опису семантики кожного вхідного набору та відсутність потреби переходу від специфічних способів концептуалізації даних до уніфікованих. Це, в свою чергу дозволяє зменшити кількість перетворень на етапі формування єдиного семантичного простору інтегрованого набору даних.

Однак, застосування множини онтологій породжує іншу категорію проблем – проблеми узгодженого застосування локальних онтологій, а саме узгодження:

- визначень концептів у локальних онтологіях;
- імен;
- метрик;
- обмежень;
- інтерпретацій.

Лише за умови вирішення цих проблем можна застосовувати множини локальних онтологій як засіб визначення семантики інтегрованого набору даних. Узгоджену множину локальних онтологій у цьому випадку можна розглядати як деяку віртуальну федеративну онтологію, яка є об'єднанням онтологій локальних вхідних ресурсів $O_1(D_1)$, $O_2(D_2)$, … , $O_N(D_N)$,

$$O^{GV}=O_1(D_1)\grave{E}\ O_2(D_2)\grave{E}\ \ldots\ \grave{E}\ O_N(D_N).$$

При цьому функції такої віртуальної глобальної не поширюються одночасно на всі вхідні локальні інформаційні ресурси, що підлягають інтеграції.

Гібридний підхід до інтеграції даних на основі онтологій. Такий спосіб семантичної інтеграції поєднує особливості двох попередньо описаних методів (рис. 4).

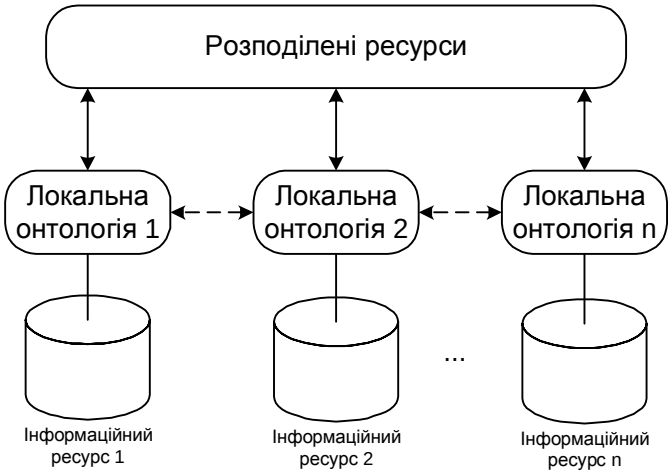

Рис. 4. Схема гібридної семантичної інтеграції даних

За аналогією з єдиною онтологією, у цьому випадку створюють спільний, узгоджений розподілений ресурс, але використовують цей ресурс для специфікації семантики вхідних наборів даних через їх власні локальні онтології.

Аналогічно до попереднього випадку семантику інформаційного ресурсу, що підлягає інтеграції, описує окрема онтологія. Але для сумісності локальних онтологій створюють глобальні розподілені словникові ресурси, в яких зосереджено базові терміни і поняття, спільні для предметної області інтегрованих даних. Гібридний підхід поєднує як переваги, так і недоліки обох попередніх підходів.

Критерії семантичної інтеграції. Важливим елементом процесу семантичної інтеграції є визначення можливості спільного застосування тих чи інших одиниць та елементів вхідних наборів даних у вихідному інтегрованому наборі. Призначенням критеріїв є визначення можливості та порядку узгодження складових онтологій вхідних наборів даних. Для цього пропонується застосувати систему норм, виконання яких є необхідною передумовою об'єднання вхідних даних в єдиний набір.

Критерії семантичної інтеграції даних для усіх трьох підходів можна сформулювати як послідовність вимог щодо попарного узгодження елементів онтологій даних: два набори даних D_i та D_j вважають придатними до семантичної інтеграції , якщо для двох онтологій O_i та O_j , які відповідають цим наборам даних, виконуються правила:

- у множинах концептів $X(D_i)$ та $X(D_j)$

(1) немає однакових понять, описаних різним способом;

(2) немає понять різного змісту, описаних однаковим способом;

- у множинах зв’язків $R(D_i)$ та $R(D_j)$

(1) відсутні зв’язки протилежного напряму та змісту між однаковими концептами;

(2) відсутні однотипні зв'язки, що не можуть бути реалізованими одночасно;

- у множинах функцій інтерпретації F_i та F_j

(1) немає функцій, одночасна реалізація яких призведе до неоднозначності інтерпретацій;

(2) з однотипними концептами різних онтологій не пов'язано обмежень, які не можуть бути виконані одночасно.

Перевірити зазначену низку критеріїв семантичної інтеграції даних можна як на формальному, так і на експертному рівні, при цьому результат має бути однаковим. Виконання всієї множини вимог дає змогу зробити висновок про можливість інтеграції двох наборів даних на рівні

їх змісту з отриманням семантично коректного результату. Ключова властивість онтологій створювати однозначне сприйняття змісту даних як на людському рівні, так і на рівні інформаційних технологій забезпечує основну перевагу методу семантичної інтеграції на основі онтологій:

(1) можливості її технічної реалізації за допомогою спеціалізованих програмних засобів;

(2) формування та аналіз критеріїв семантичної інтеграції на формальному рівні;

(3) отримання семантично коректного результату без безпосередньої участі людини експерта.

Застосування онтологій для інтеграції невизначеностей в даних. Для опису невизначеностей як структурної та семантичної одиниці даних пропонується застосувати спеціальний концепт – клас *"Невизначеність"*, елементами якого є різноманітні варіанти невизначеностей, наприклад

$$Undefined=\{U_1\text{»"неможливо"}, U_2\text{»"невідомо"},$$
$$U_3\text{»"не існує"}, U_4\text{»"не визначено"}, \dots\}.$$

На рис. 5 подано приклад структури класу онтології, який описує семантику та природу невизначеностей у складі даних.

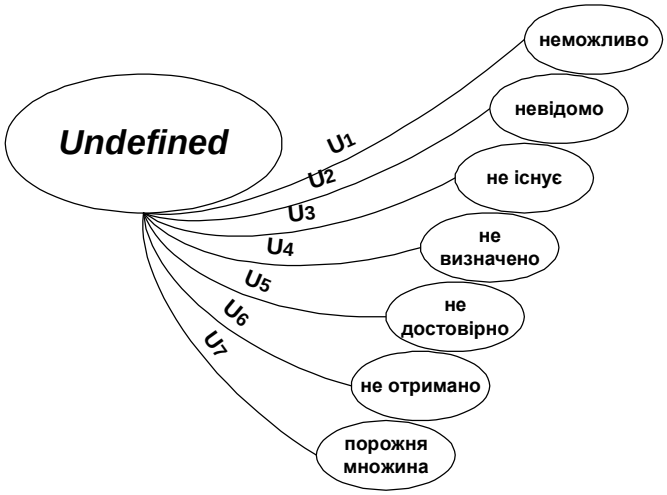

Рис. 5. Структура класу "Невизначеність"

Така класифікація невизначеностей необхідна для уникнення суперечностей, які виникають

- при побудові глобальної онтології для вхідних наборів даних за умови їх неповноти чи неточності;
- при організації обміну і взаємодії локальних онтологій вхідних наборів даних з невизначеностями;
- при побудові розподілених словників загального користування у поєднанні з локальними онтологіями;
- при визначенні способів подальшого опрацювання невизначеностей;
- при виконанні запитів до інтегрованих даних, в складі яких присутні невизначеності.

Окремо в складі онтології передбачають множину функцій інтерпретації невизначеностей, які встановлюють відповідність між конкретними невизначеностями в даних та елементами відповідного класу в складі онтології

$$F_i():UN \to Undefined,\ i=1, 2, 3, \dots$$

де $UN=\{null, "", </>, „, \dots\}$ – множина можливих позначень невизначеностей; *Undefined* – відповідний клас онтології, який описує їх природу та характер.

Визначення такого класу для типізованих даних [1] означає створення відповідного типу, який є підтипом усіх можливих типів. У складі такого типу визначають множину відповідних констант, кожна з яких позначає належність даних до тієї чи іншої категорії невизначеностей з можливістю їх подальшої інтерпретації.

Залежно від стратегії семантичної інтеграції даних [5] – на основі спільної онтології, на основі локальних онтологій чи на основі гібридного підходу клас *"Невизначеність"* може бути застосований по-різному. У випадку інтеграції на основі спільної єдиної онтології клас є спільним для опису невизначеностей в усіх вхідних наборах даних, які підлягають інтеграції. Такий підхід забезпечує однорідність інтегрованих даних та однозначність інтерпретації невизначеностей для усіх вхідних наборів. Однак побудова спільної глобальної онтології не завжди є можливою [5].

У випадку застосування локальних онтологій для семантичної інтеграції даних, у кожній з них описують клас *"Невизначеність"*. При цьому можливими є варіанти:

- клас типу *"Невизначеність"* є однаковим в усіх локальних онтологіях,
- клас типу *"Невизначеність"* є специфічним в кожній локальній онтології;
- класи типу *"Невизначеність"* є спільними в частині локальних онтологій.

У першому випадку спільний клас *"Невизначеність"* забезпечує єдину інтерпретацію та єдиний порядок опрацювання невизначеностей як у всіх вхідних наборах даних, так і в інтегрованому наборі. Такий спосіб значною мірою збігається з випадком семантичної інтеграції на основі єдиної онтології.

У другому випадку для спільного опрацювання даних, описаних за допомогою специфічних локальних онтологій, виникає потреба формування механізму їх взаємодії, зокрема, в частині інтерпретації та опрацювання невизначеностей. Це може бути реалізовано шляхом визначення відповідності між підкласами класів типу *"Невизначеність"* та їх атрибутами за принципом

$$O_i.UT_i.UST^k_i \ \text{®} \ O_j.UT_j.UST^k_j ,$$

де $O_i.UT_i.UST^k_i$ – посилання на підклас UST^k_i класу *"Невизначеність"* UT_i онтології O_i; для $i,j=1,2,\ldots,N,\ i \neq j$, де N – кількість локальних онтологій вхідних наборів даних.

Така відповідність дає змогу узгодити інтерпретацію невизначеностей у різних вхідних наборах даних та визначити єдину стратегію їх опрацювання в інтегрованому наборі.

У випадку збігу та відмінностей у визначенні класу *"Невизначеність"* у різних локальних онтологіях, для узгодження роботи з ними і інтегрованих даних застосовують комбінацію двох попередніх підходів.

Гібридний підхід до семантичної інтеграції неповних чи неточних даних, який поєднує принципи інтеграції на основі глобальної та локальних онтологій, дає змогу застосувати такі варіанти опрацювання невизначеностей як специфічного класу:

- формування класу *"Невизначеність"* як спільного розподіленого ресурсу,
- формування класу *"Невизначеність"* як власного специфічного ресурсу локальних онтологій,
- формування класу *"Невизначеність"* як у складі локальних так і розподілених ресурсів.

Перший випадок використання такого опису невизначеностей повністю збігається за принципами зі стратегією семантичної інтеграції на основі єдиної глобальної онтології.

Другий випадок реалізують за принципом опрацювання невизначеностей при застосування стратегії локальних онтологій із визначенням специфічних класів типу *"Невизначеність"* для кожної з них.

Третій випадок інтеграції даних з невизначеностями передбачає можливості як локального, так і глобального їх опису. При цьому способи опрацювання класу *"Невизначеність"* потребують узгодження як на рівні локальних онтологій, як це описано вище, так і на рівні взаємодії локальних і розподілених ресурсів.

У такий спосіб для кожного невизначеного елемента у вхідних наборах даних може бути задано часткову інтерпретацію, яка не приводить до усунення невизначеності, але дає можливість охарактеризувати її природу, походження, зміст, вплив на інші значення, а також визначає способи і засоби опрацювання відповідних даних в інтегрованому наборі.

Згідно з таким підходом, в інтегрованому наборі даних від позначень типу Null, порожніх чи відсутніх елементів пропонується перейти до їх умовного позначення відповідно до належності до відповідного підкласу класу *"Невизначеність"*. Тобто кожне значення буде подано у вигляді

відповідної константи (псевдоконстанти) як величини, що належить до спеціального типу, який є підтипом усіх можливих типів даних.

Отже, часткова інтерпретація невизначеностей за допомогою визначення спеціальних класів в онтологіях даних переводить проблему їх опрацювання на принципово інший рівень – від опрацювання невідомих, неповних, недостовірних понять до опрацювання явно визначених концептів.

Висновки

Процеси інтеграції даних мають достатньо широку сферу застосування. Це, зокрема, сховища даних різного типу та прямування, корпоративні ERP та CRM системи, інформаційні Web-системи, системи електронного бізнесу тощо. Інформаційні ресурси таких систем передбачають одночасне застосування значної кількості різноманітних за формою, структурою, змістом, способами подання і застосування даних. Однією з основних проблем інтеграції є створення та застосування єдиних правил і способів зображення таких різнорідних даних. Така проблема може бути вирішена за рахунок формування інтегрованого синтаксису даних, утвореного на основі синтаксичних методів і засобів вхідних даних.

У запропонованій роботі розглянуто низку питань, пов’язаних з одним із принципових аспектів інтеграції даних – інтеграції їх змісту. В основу запропонованого вирішення покладено формальне подання даних як системи, семантику елементів якої описують за допомогою спеціальних засобів, придатних для програмного сприйняття та опрацювання. Проаналізовано зокрема особливості створення інтегрованих інформаційних ресурсів із застосуванням метаданих, контекстуального аналізу та онтологій.

У такий спосіб проблему інтеграції семантики різнорідних наборів даних можна звести до узгодження їх онтологій і вирішити за допомогою засобів і технологій формального комп'ютерного опрацювання знань.

Запропоновані вирішення можуть слугувати базисом для створення алгоритмів та методів організації процесів видобування, перетворення, завантаження даних та інших технологій інтеграції у сховищах, вітринах даних чи інтегрованих або розподілених базах даних.

1. Berko A. Consolidated data models for electronic business systems. / Andriy Berko // Proceedings of IXth Internationale Conference CADSM 2007. – Lviv, 2007. pp. 341 – 342. 2. IDEF5 Ontology Description Capture method. – [Електронний ресурс].- http://www.idef.com/IDEF5.html, 2006. 3. Lenzerini M. Data Integration: A Theoretical Perspective. / Marco Lenzerini // Proc. of the ACM Symp. on Principles of Database. Systems (PODS), 2002. – pp. 233 – 246. 4. Tierney B. Contextual Semantic Integration for Ontologies / Brendan Tierney, Mike Jackson // www.macs.hw.ac.uk/ BNCOD21/DC/Tierney.pdf , 2005. 5. Wache H. Ontology-Based Integration of Information – A Survey of Existing Approaches / H.Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner.- Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle, USA, August 4-5, 2001, pp.108-118. 6. White C. Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise (Report Excerpt)./ C. White // http://www.tdwi.org/ Publications/ WhatWorks/ display.aspx?id=7979, 2007. 7. Ландэ Д.В. Основы интеграции информационных потоков: Монография / Дмитрий Ландэ. – Киев. : Инжиниринг, 2006. – 240 с. 8. Литовский К.Ю. Слабоструктурированные данные: некоторые методы их представления и обработки запросов / К.Ю. Литовский, Г.С. Томусяк Г.С. // Московская Секция ACM SIGMOD.– http:// synthesis.ipi.ac.ru/ sigmod/seminar/s20000224, 2000.– с.1 – 2. 9. Манцивода А.В. Система метаописаний Dublin Core.- [Електронний ресурс].- http://teacode.com/concept/eor/dc.html, 2004. 10 Спирли Э. Корпоративные хранилища данных. Планирование, разработка, развитие / Эрик Спирли.– М. : Издательский дом "Вильямс", 2001. – 400 с.