УДК 004.652

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

МЕТОДИ ТА ЗАСОБИ СЕМАНТИЧНОЇ ІНТЕГРАЦІЇ ДАНИХ

Розглянуто способи і можливості створення інтегрованих структур даних для зберігання інформаційного наповнення в гетерогенних системах електронного контент–бізнесу. Запропоновано формальні моделі та методи інтеграції семантики різнорідних даних у системах електронного бізнесу, які ґрунтуються на застосуванні метаданих, контекстного аналізу тезаурусів та онтологій даних.

Some ways and possibilities of integrated data model development for information storage of heterogeneous electronic content–business systems are considered in this paper. Formal models and methods of semantic integration of data based on metadata, context analysis of thesauruses and data ontologies application were proposed.

Вступ

Постійне зростання обсягів інформації та завдань, пов'язаних з її застосуванням в різноманітних галузях, розвиток інформаційних систем великого масштабу викликали потребу інтеграції. Проблема створення 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 подання користувачеві їх об'єднаного подання [1], [3].

Процеси інтеграції даних передбачають розв'язання таких проблем, як інтеграція значень, інтеграція синтаксису, інтеграція структури семантики даних. 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 формування та аналізу критеріїв інтеграції різнорідних наборів даних на основі метаданих, на основі контекстуального аналізу та на основі онтологій.

Аналіз останніх публікацій і досліджень

Основні теоретичні засади семантичної інтеграції сформульовано у [3], [6]. У цій роботі обґрунтовано поняття семантичної інтеграції, а також запропоновано методи контекстуального аналізу та онтології як інструментарію вирішення проблеми побудови інтегрованого змістового простору. 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. Ця схема передбачає формування складу метаданих за певними принципами, які вимагають визначення таких елементів, як метадані про об’єкт даних, суб’єкт даних, часові показники, місцезнаходження та застосування даних, призначення даних та спосіб їх використання.

У [9] описано схему побудови даних, 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. Семантика Дублінського ядра була створена міжнародною міждисциплінарною групою фахівців з комп'ютерних наук, Інтернету, 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.

Згідно з [9] базовий набір елементів метаданих простого Дублінського ядра (Dublin *Core Metadata Element Set; DCMES*) складають 15 одиниць:

1. *Title* – назва;
2. *Creator* – автор;
3. *Subject* – тема;
4. *Description* – опис;
5. *Publisher* – видавник;
6. *Contributor* – учасник створення ресурсу;
7. *Date* – дата;
8. *Type* – тип;
9. *Format* – формат даних;
10. *Identifier* – ідентифікатор;
11. *Source* – джерело даних;
12. *Language* – мова;
13. *Relation* – зв'язки з іншими ресурсами;
14. *Coverage* – область дії ресурсу;
15. *Rights* – авторські права.

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 важливою. 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 складнішими та формалізованішими 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, їх взаємозв'язки і правила (теореми, обмеження), задані у предметній області набору даних [3]. Онтології даних можна описувати різними засобами, і сьогодні відомо достатньо 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 дійсності із застосуванням єдиних, строго визначених зразків (термінів, моделей тощо);

- використання обмеженої кількості базових термінів (сутностей), на основі котрих конструюють всі інші поняття;
- внутрішньої повноти і 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*) може бути застосовано і у процесах семантичної інтеграції даних [6]

Мета та завдання досліджень

Цілі статі. Основними цілями, які поставлено в цій статті, є такі:

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 у Web-системах електронного контент-бізнесу;
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 систем електронного контент-бізнесу. За оцінками експертів [8], лише близько 20 відсотків інформаційного ресурсу в сучасних Інтернет-системах 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.

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. 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. Найвідомішими способами семантичної інтеграції є такі [3]:

- інтеграція на основі метаданих;
- контекстуальна семантична інтеграція;
- інтеграція на основі онтологій.

Семантична інтеграція на основі метаданих. Цей метод передбачає порівняння складу та змісту метаданих двох наборів з метою визначення можливості їх семантичної інтеграції. Метадані забезпечують формування та застосування опису основних властивостей деякого набору даних (інформаційного ресурсу), зокрема, таких, що визначають його семантичні показники. Найпоши-

ренішою структурою метаданих є вимірна схема Захмана [10], котра передбачає застосування шести категорій метаданих (вимірів), які описують такі властивості інформаційного ресурсу:

- *об'єкти даних* – опис сутностей, які асоціюють зі значеннями з набору даних;
- *суб'єкти даних* – опис осіб, які створюють чи застосовують дані;
- *часові показники* – опис часових моментів чи інтервалів, що характеризують створення, підтримання та застосування даних;
- *розміщення даних* – опис місцезнаходження даних та способів і порядку доступу до ресурсу;
- *призначення даних* – опис функцій та завдань, які застосовують інформаційний ресурс;
- *порядок застосування даних* – правила та обмеження на роботу з інформаційним ресурсом.

Загальну структуру метаданих інформаційного ресурсу побудованих за схемою Захмана подано на рис. 1. Кожен із вимірів метаданих – це деяка множина значень, що характеризує один з аспектів організації, сприйняття та застосування деякого набору даних, зокрема, в процесах семантичної інтеграції з іншими ресурсами.

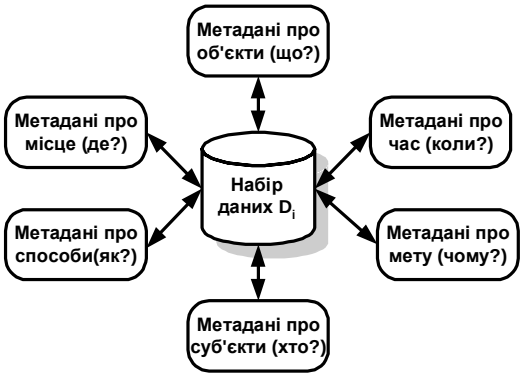

Рис. 1. Схема Захмана організації метаданих

Множину метаданих M_i деякого набору даних D_i можна подати як кортеж

$$M_i = <M^o_i, M^s_i, M^p_i, M^t_i, M^g_i, M^m_i>,$$

де M^o_i – метадані про об'єкти даних, M^s_i – метадані про суб'єкти даних, M^p_i – метадані про розміщення даних, M^t_i – метадані про часові показники даних, M^g_i – метадані про мету застосування даних, M^m_i – метадані про порядок використання даних.

Збіг значень метаданих двох наборів за одним або більше вимірами застосовують як критерій семантичної інтеграції. Міра збігу значень метаданих певної категорії залежить від конкретних завдань, але, як правило, її числовий вираз не повинен бути меншим за 80% [4]. Певним чином цей метод є подібним до методу контекстуальної семантичної інтеграції, але в цьому випадку замість тезаурусу застосовують інший інструментальний засіб – метадані. Визначення категорій метаданих, за якими формують критерії семантичної інтеграції, та порядок визначення їх збігу є задачею недостатньо формальною, яка потребує участі експерта.

Альтернативним до схеми Захмана способом організації метаданих є Дублінське ядро [9]. Основними принципами побудови системи метаданих на основі дублінського ядра є такі:

- простота,
- зрозуміла семантика,
- інтернаціоналізація,
- здатність розширення,
- однозначність,
- спрощення зображення метаданих,
- коректність значень,

- зв'язок зі синтаксисом,
- використання стандартних просторів імен.

Застосування такого переліку властивостей інформаційного ресурсу дає змогу формалізувати та уніфікувати опис даних, що підлягають інтеграції, зокрема їх семантичне наповнення. Метадані, побудовані за принципами Дублінського ядра, мають такі особливості:

- дають змогу побудувати опис інформаційного ресурсу будь-якого виду – від художньої книги до web-сторінок, електронних документів та баз даних;
- забезпечують повний та всебічний опис всіх властивостей даних, зокрема тих, що характеризують їхню семантику;
- можливим є зображення метаданих у форматі XML, що значно спрощує процеси їх формального опрацювання.

Наприклад, один з варіантів системи метаданих, що описують інформаційний ресурс, який містить текст цієї статті за принципами Дублінського ядра у форматі XML-документа може бути побудовано так, як це зображено на рис 2.

```
<?xml version="1.0"?>
        <!-- Приклад метаданих побудованих за принципами Дублінського ядра -->
<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/">
        <!-- Назва ресурсу -->
<dc:title> Ця стаття  </dc:title>
        <!-- Автор ресурсу -->
 <dc:creator> Andriy Berko  </dc:creator>
        <!-- Прізвище автора рідною мовою -->
 <dc:creator xml:lang="UA"> Андрій Берко </dc:creator>
        <!-- Тематика за УДК -->
 <dc:subject xsi:type="dcterms:UDC">  УДК 004.652 </dc:subject>
        <!-- Тематика українською мовою -->
<dc:subject xml:lang="UA"> моделі баз даних </dc:subject>
        <!-- Опис змісту ресурсу українською мовою -->
<dc:description xml:lang="UA">
У роботі розглянуто моделі семантичної інтеграції
баз даних та даних довільного формату  на основі метаданих,
контекстного аналізу та онтологій
</dc:description>
        <!-- Дата створення ресурсу-->
<dc:date> 01.10.2008  </dc:date>
        <!-- Тип ресурсу -->
<dc:type> стаття </dc:type>
        <!-- Формат подання -->
<dc:format> MS Word 2003 </dc:format>
        <!-- Ідентифікатор ресурсу -->
<dc:identifier> stattia_10.2008.doc </dc:identifier>
        <!-- Джерело даних -->
<dc:source> http://Andriy.Berko.Lviv.net/stattia_10.2008 </dc:source>
        <!-- Мова подання ресурсу -->
<dc:language> UA </dc:language>
        <!-- Кінець  метаданих-->
</metadata>
```

Рис. 2. Приклад опису інформаційного ресурсу на основі Дублінського ядра у форматі XML

Формально множину метаданих деякого інформаційного ресурсу D, побудованих за принципами Дублінського ядра, можна подати у вигляді поєднання елементів опису, передбачених стандартом :

$$M^{DC}(D) = \{M_1^{DC}(D), M_2^{DC}(D), \ldots, M_{15}^{DC}(D)\},$$

де $M^{DC}(D)$ – множина метаданих, $M_1^{DC}(D)$, $i=1,2,\ldots,15$ значення відповідного елементу базового набору Дублінського ядра.

Критерій семантичної інтеграції даних у такому разі можна сформулювати за допомогою функції збігу елементів опису двох інформаційних ресурсів D_1 та D_2, які визначено важливими для їх поєднання:

$$Map(M_i^{DC}(D_1), M_i^{DC}(D_2)) = true,$$

де $i \hat{I} (1 - 15)$. Залежно від значень $M_i^{DC}(D_1)$ та $M_i^{DC}(D_2)$ функція збігу може бути істинною або хибною. Загальну схему семантичної інтеграції із застосуванням метаданих на основі Дублінського ядра наведено на рис.3.

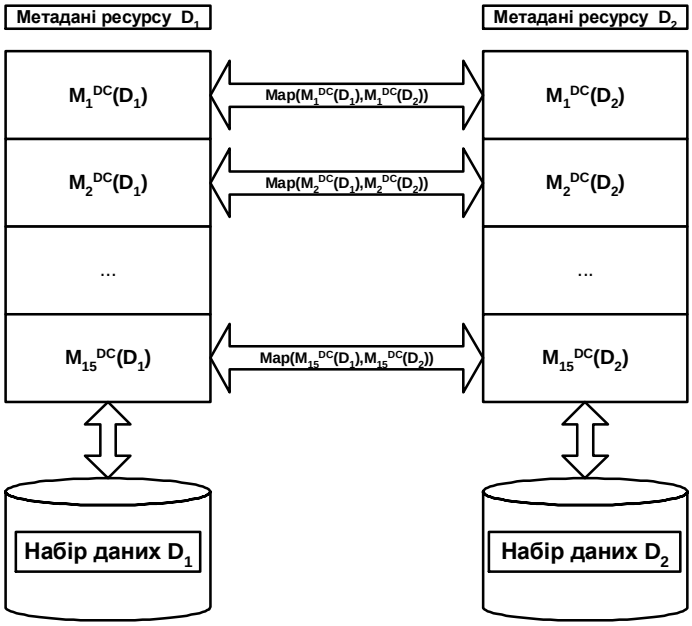

Рис. 3. Схема семантичної інтеграції даних на основі Дублінського ядра

Перелік істотних елементів метаданих залежить від конкретного випадку та завдань їх семантичної інтеграції. Вважають, що два набори даних, для яких функція збігу за істотними елементами метаданих набуває значення "істина", є придатними для семантичної інтеграції. В загальному випадку це може бути як один, так і більше елементів Дублінського ядра. Важливою проблемою такого підходу є те, що поняття збігу значень у загальному випадку означає їхні рівності, тому чітке визначення функції збігу є завданням, яке не завжди може бути виконане без участі людини-експерта. Для цього застосовують спеціальні таблиці відповідності значень метаданих, які дають змогу робити висновки про відповідність змісту окремих елементів Дублінського ядра. Наприклад, тематика набору даних "Internet" за змістом відповіднає тематиці "World Wide Web". Потреба участі експерта у формуванні критеріїв інтеграції інформаційних ресурсів значно зменшує універсальність та масштаби застосування метаданих у процесах семантичної інтеграції.

Метод контекстуальної семантичної інтеграції. Цей метод, який вперше запропоновано у [4], ґрунтується на змістовому порівнянні інформаційного наповнення наборів даних, що підлягають інтеграції. Цей метод дає змогу оцінити можливості інтегрування як структурованих (реляційних) даних, так і слабкоструктурованих – поданих у довільних форматах. Істотним

аспектом визначення семантики слабкоструктурованих даних є їх контекст [4]. Контекст може мати різноманітні форми, такі як текст і гіперпосилання на Web-сторінці, ім'я каталогу, в якому зберігають дані, супутні анотації і коментарі до даних, зв'язки з фізично або логічно близькими елементами даних, перелік ключових слів і понять тощо. У таких застосуваннях контекст допомагає інтерпретувати зміст даних. Слабкоструктуровані дані часто є менш точними, ніж у традиційних базах даних. Той фактор, що їх отримано з неструктурованого тексту, робить такі дані семантично різнорідними або чутливими до умов, при яких вони були зафіксовані.

Оскільки в більшості випадків семантичний аналіз повного вмісту інформаційних ресурсів є складним, а часто і неможливим, то в інтеграційних процесах його замінюють контекстуальним аналізом тезаурусних термінів наборів даних. Тезаурусні терміни – це перелік ключових понять, пов'язаних з даними та внесеними до спеціального переліку – тезаурусу [4]. Їх застосовують для опису семантичної відповідності між лексичними одиницями цього набору даних та конкретними значеннями з предметної області. Основою формування тезаурусу можуть бути, наприклад, перелік описів стовпчиків таблиці бази даних, XML-теги в документі, назви розділів та пунктів текстового документа, гіперпосилання на Web-сторінці тощо.

Критерієм семантичної інтегрованості двох наборів даних у цьому випадку є функція контекстуальної семантичної віддалі між ними (CSD-функція) [4]. Вираховують значення CSD-функції у такий спосіб. Нехай та W_i та W_j – множини тезаурусних термінів наборів даних D_i та D_j, відповідно, W_{ij} – множина тезаурусних термінів, які є семантично спільними для двох наборів, $|W_i|$, $|W_j|$, $|W_{ij}|$ – потужності відповідних множин. Тоді значення функції контекстуальної семантичної віддалі між наборами даних вираховують як частку спільних значень у меншій за об'ємом з множин тезаурусних термінів двох наборів даних

$$CSD(D_i, D_j) = \frac{|W_{ij}|}{Min(|W_i|, |W_j|)}$$

Вважають [4], що семантична інтеграція двох наборів даних є можливою, якщо значення функції контекстуальної семантичної віддалі задовольняє умову

$$CSD(D_i, D_j) \geq 0{,}8 .$$

На рис 4. зображено загальну схему процесу семантичної інтеграції двох наборів різнорідних даних із застосуванням тезаурусних термінів та функції семантичної віддалі.

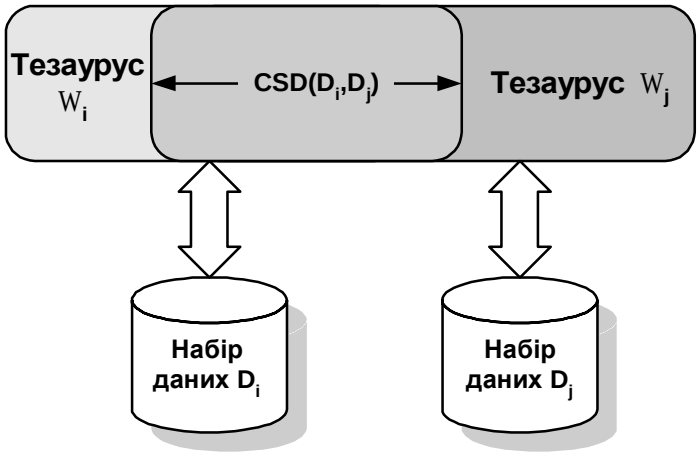

Рис. 4. Загальна схема контекстуальної семантичної інтеграції

Порівняно з методикою семантичної інтеграції на основі метаданих, метод контекстуального аналізу дає змогу перевірити критерії семантичної інтеграції на формальному рівні і не потребує безпосередньої участі експерта. Найслабшим місцем цього методу є формування тезаурусу для набору даних, який містить певний інформаційний ресурс. Через неоднорідність форматів та структур даних, що підлягають інтеграції, створення наборів ключових термінів може бути

достатньо трудомістким і малоефективним. Окрім того, формування тезаурусу значною мірою залежить від суб'єктивного людського фактора, що значно впливає на універсальність методу та його незалежність від конкретних умов.

Семантична інтеграція даних на основі онтологій. Найперспективнішим, на 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. У загальному випадку формальним зображенням онтології є трійка

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

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 до вимог 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.

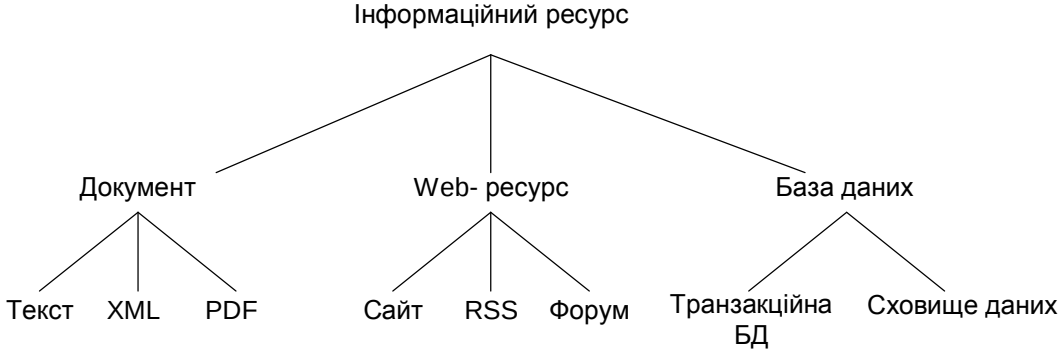

Рис. 5. Приклад визначення класів онтології інформаційного ресурсу

На рис. 5 показано один з варіантів класифікації інформаційних ресурсів з погляду інтеграції. Клас "Інформаційний ресурс" містить підкласи *"Текст"*, *"Web-ресурс"* та *"База даних"*, які, своєю чергою, поділяють на дрібніші підкласи. Кількість рівнів ієрархічної класифікації залежить від конкретних вимог та особливостей процесів інтеграції даних. На рис 6. 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 власної 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.

Рис. 6. Приклад визначення онтології, що описує склад інформаційного ресурсу

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 вигляду. Тобто, проблему семантичної інтеграції даних можна звести до проблеми виявлення відповідностей та суперечностей між їх онтологіями.

Критерії семантичної інтеграції у цьому випадку можна сформулювати як послідовність вимог щодо елементів двох 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 по-різному;

(2) немає понять різного змісту, описаних однаково;

- 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.

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.

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.- [Електронний ресурс].- www.let.uu.nl/~Paola.Monachesi/personal/papers/wache.pdf, 2001. 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: Монографияія / 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:// tea-code.com/concept/eor/dc.html, 2004. 10. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, развитие / Э. Спирли.– М. : Издательский дом "Вильямс", 2001. – 400 с.