ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ У СІЛЬСЬКОГОСПОДАРСЬКІЙ ГАЛУЗІ

УДК 004.056.5:004

Васюхін М. І.
доктор технічних наук, професор кафедри комп'ютерних систем і мереж факультету інформаційних технологій НУБІП України

Іванник Ю. Ю.
кандидат технічних наук, асистент кафедри комп'ютерних систем і мереж факультету інформаційних технологій НУБІП України

Сініцин О.В.
аспірант кафедри комп'ютерних систем і мереж факультету інформаційних технологій НУБІП України

МЕТОДИ СИНТЕЗУ ГЕТЕРОГЕННИХ ДАНИХ В ГІС ПРЕЦИЗІЙНОГО ЗЕМЛЕРОБСТВА

***Анотація**. Розглянуто методи синтезу гетерогенних даних: синтез на основі метаданих, контекстуальний синтез гетерогенних даних, синтез гетерогенних даних на основі онтологій. Виявлено найперспективніший, на сьогодні, метод синтезу даних.*

***Ключові слова:** синтез, гетерогенні данні, ГІС, INSPIRE, прецизійне землеробство.*

Вступ

Відповідно до пункту 6 директиви INSPIRE - просторові дані мають зберігатися і підтримуватися на належному рівні і бути доступними.

Просторові дані повинні бути доступні на умовах, які не обмежують їх широкого використання, забезпечують простий пошук і оцінку придатності для конкретної мети, а також мати чітко прописані правила та обмеження використання. Потрібно забезпечити умови для несуперечливого комбінування просторових даних, отриманих з різних джерел, їх вільного розповсюдження між користувачами, а також для розподілу просторових даних, отриманих на одному адміністративному рівні, на всі інші рівні [1].

Процеси синтезу даних мають достатньо широку сферу застосування. Це, зокрема, сховища даних різного типу та спрямування, ГІС, 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 синтезу є створення та застосування єдиних правил і способів зображення гетерогенних даних (далі - ГД) [2].

Синтез ГД забезпечує з одного боку можливість користувачу без спеціальних навичок отримати доступ до актуальних відомостей про геопросторові об'єкти, а з другого дасть поштовх для розробки та інтеграції національної інфраструктури геопросторових даних в інфраструктуру просторової інформації в рамках ЄС (INSPIRE) [3].

Така тенденція призводить до необхідності синтезу ГД, для комплексного вирішення широкої номенклатури задач.

Мета досліджень – аналіз та обгрунтування вибору методів синтезу ГД.

Матеріали та методика досліджень

В даній статті будуть проаналізована одна з основних проблем синтезу даних, а саме створення та застосування єдиних правил і способів зображення ГД.

Для реалізації поставлених задач буде розглянуто наступні методи синтезу ГД [2]:

- синтез на основі метаданих;
- контекстуальний синтез ГД;
- синтез ГД на основі онтологій.

Результати досліджень

Синтез 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. Найпоширенішою структурою метаданих є вимірна схема Захмана, 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;
- спосіб 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.

Множину метаданих Mi деякого набору даних Di можна подати як кортеж Mi = <M0i , Msi , Mpi , Mti , Mgi , Mmi >

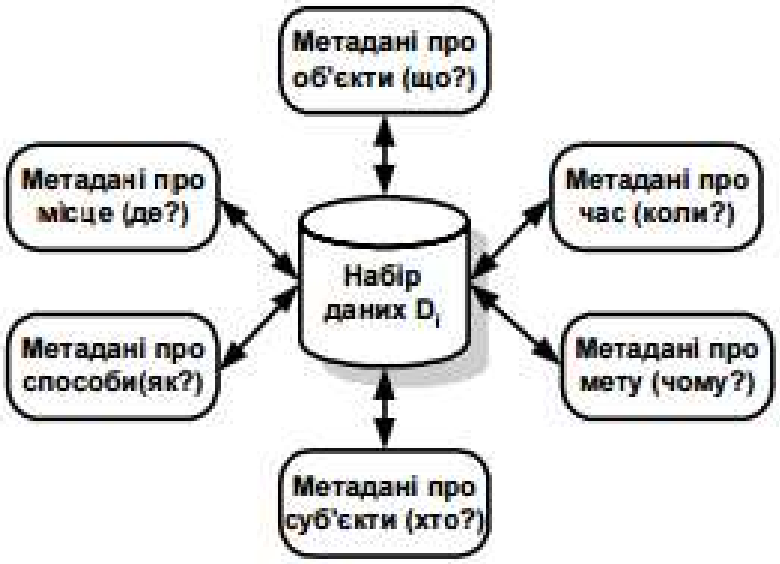

Рис. 1. Схема Захмана організації метаданих.

де: M0i – метадані про об’єкти даних, Msi – метадані про суб’єкти даних, Mpi – метадані про розміщення даних, Mti – метадані про часові показники даних, Mgi – метадані про мету застосування даних, Mmi – 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% . 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 синтезу, та порядок визначення їх збігу є задачею недостатньо формальною, яка потребує участі експерта [6].

Метод контекстуального синтезу ГД. Цей метод, який вперше запропоновано у [4], ґрунтується на змістовому порівнянні інформаційного наповнення наборів даних, що підлягають синтезу. Цей метод дає змогу оцінити можливості синтезу, 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 Ωi та Ωj – множини тезаурусних термінів наборів даних Di та Dj , відповідно, Ωij – множина тезаурусних термінів, які є семантично спільними для двох наборів, |Ωi |, |Ωj |, |Ωij| – 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

$$\mathrm{CSD}(\mathrm{Di}, \mathrm{Dj}) = \frac{|\Omega\Omega ij|}{\mathrm{Min}(|\Omega\Omega i|, |\Omega\Omega j|)} \qquad (1)$$

Вважають [4], що синтез двох наборів даних є можливим, якщо значення функції контекстуальної семантичної віддалі задовольняє умову $\mathrm{CSD(Di,Dj)}^3 \times 0{,}8$. На рис. 2 – зображено загальну схему процесу семантичної

інтеграції двох наборів різнорідних даних із застосуванням тезаурусних термінів та функції семантичної віддалі.

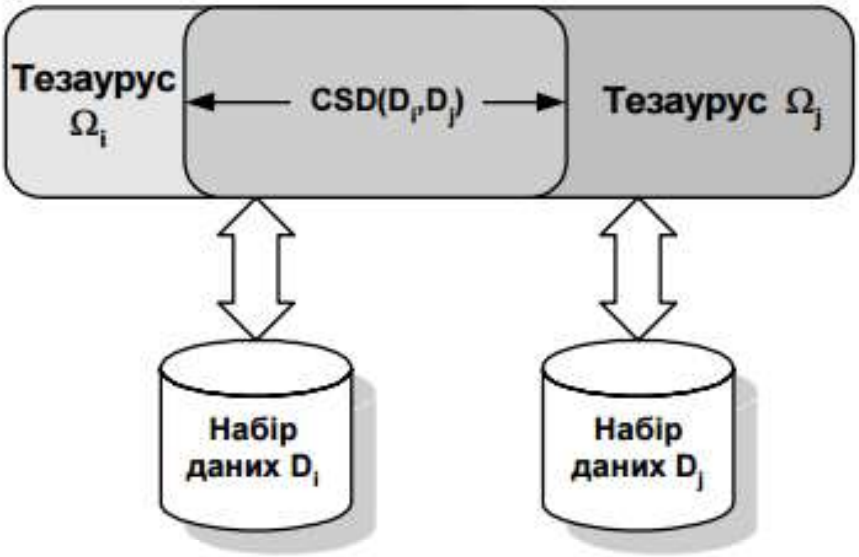

Рис. 2. Загальна схема контекстуального синтезу гетерогенних даних.

Порівняно з методикою синтезу на основі метаданих, метод контекстуального аналізу дає змогу перевірити критерії синтезу 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 синтезу, 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.

Синтез на основі онтологій. Цей метод передбачає використання основних елементів двох попередніх методів – тезаурусу та метаданих, але є значно загальнішим за них та враховує значно більше аспектів семантики даних. Одним із прикладів застосування онтологій як засобу семантичної інтеграції було запропоновано в [4]. Розглянемо онтологію 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 синтезу даних об'єктом опису поданого у вигляді онтології є певний інформаційний ресурс. У загальному випадку формальним зображенням онтології є трійка

$$< O = X, R, F >, \qquad (2)$$

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) [2, 5]. Згідно до вимог стандарту IDEF5 [6], 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 ієрархії, які утворюють цілісний інформаційний ресурс. На рис. 3 показано один з варіантів класифікації інформаційних ресурсів з погляду інтеграції.

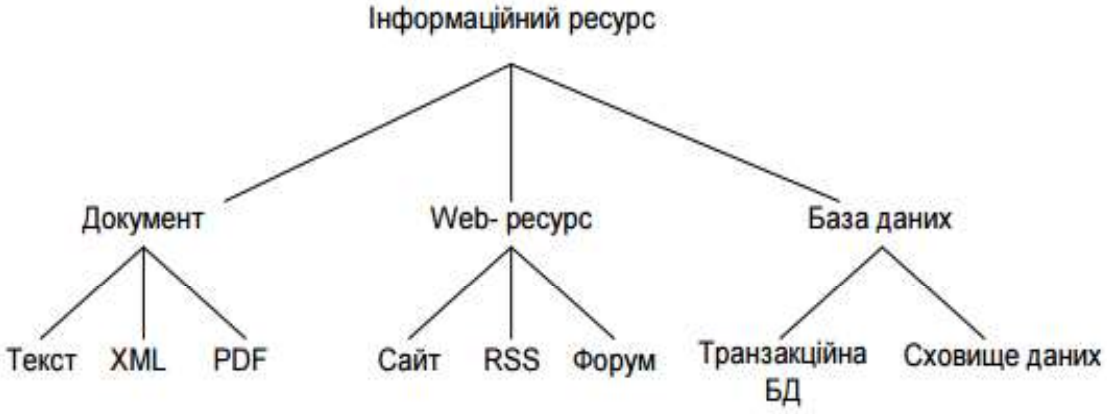

Рис. 3. Приклад визначення класів онтології інформаційного ресурсу

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. Процеси синтезу даних передбачають створення для кожного вхідного набору даних Di власної онтології O(Di), яка формує однозначний опис семантики як всього інформаційного ресурсу, так і окремих його елементів

$$O(Di) = < X(Di), R(Di), F(Di)>, \quad (3)$$

де: X(Di) – множина концептів, які описують одиниці даних, їх зміст, властивості та належність до певного класу чи категорії; R(Di) – множина зв’язків і відношень між одиницями даних, що визначають порядок їх взаємодії та взаємного застосування; F(Di) – 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 ГД, 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 синтезу у цьому випадку можна сформулювати як послідовність вимог щодо елементів двох онтологій даних: два набори даних Di та Dj вважають придатними до синтезу, якщо для двох онтологій Oi та Oj , які відповідають цим наборам даних, виконуються правила:

у множинах концептів X(Di) та X(Dj):

- немає однакових понять, описаних по-різному;
- немає понять різного змісту, описаних однаково;

у множинах зв’язків R(Di) та R(Dj):

- відсутні зв’язки протилежного напряму та змісту між однаковими концептами;
- відсутні однотипні зв'язки, що не можуть бути реалізованими одночасно;

у множинах функцій інтерпретації Fi та Fj:

- немає функцій, одночасна реалізація яких призведе до неоднозначності інтерпретацій;
- 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 машинному рівні забезпечує основну перевагу методу синтезу на основі онтологій:

- можливості її технічної реалізації за допомогою спеціалізованих програмних засобів;
- формування та аналіз критеріїв семантичної інтеграції на формальному рівні;
- отримання семантично коректного результату без безпосередньої участі людини-експерта [2, 5].

Висновки

У статті проведено порівняльний аналіз та обгрунтування вибору найкращого з методів синтезу ГД в ГІС прецизійного землеробства.

Встановлено, що синтез ГД даних передбачає формування єдиного змістового простору для сприйняття, інтерпретації та застосування цих даних незалежно від формату їх подання та структури. Однією з основних проблем у цьому процесі є формування критеріїв синтезу ГД даних, за допомогою яких також можна оцінити можливість чи неможливість об'єднання їх змісту.

Досліди показали, що найперспективнішим, на сьогодні, методом синтезу даних є синтез ГД 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.

Список літератури

1. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). – 180c.
2. Lenzerini M. Data Integration: A Theoretical Perspective. / Marco Lenzerini // Proc. of the ACM Symp. on Principles of Database. Systems (PODS), 2002. – pp. 233 – 246.
3. Васюхін М.І. Аналіз концепції побудови геопорталів як основної складової ГІС прецизійного землеробства / М.І. Васюхін, О.В. Сініцин, Ю.Ю. Іваник // Науковий вісник НУБіП України. Серія «Техніка та енергетика АПК» / редкол.: С.М. Ніколаєнко (відп. ред) та ін. – 2016. – Вип. 240. – С. 227 – 235.
4. Tierney B. Contextual Semantic Integration for Ontologies / Brendan Tierney, Mike Jackson // www.macs.hw.ac.uk/ BNCOD21/DC/Tierney.pdf , 2005
5. Палагин А.В. Онтологические методы и средства обработки предметных знаний: монография / А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. – Луганск: изд-во ВНУ им. В. Даля, 2012. – 324 с.
6. Берко А. Методи та засоби семантичної інтеграції даних / А. Берко // Вісник Національного університету "Львівська політехніка". – 2009. – № 638 : Комп'ютерні науки та інформаційні технології. – С. 190-199.

Васюхин М. И.
доктор технических наук, профессор кафедры компьютерных систем и сетей факультета информационных технологий НУБИП Украины;

Иванник Ю. Ю.
кандидат технических наук, ассистент кафедры компьютерных систем и сетей факультета информационных технологий НУБИП Украины;

Синицын А. В.

аспирант кафедры компьютерных систем и сетей факультета информационных технологий НУБИП Украины.

МЕТОДЫ СИНТЕЗА ГЕТЕРОГЕННЫХ ДАННЫХ В ГИС ПРЕЦИЗИОННОГО ЗЕМЛЕДЕЛИЯ

Аннотация. *Рассмотрены методы синтеза гетерогенных данных: синтез на основе метаданных, контекстуально синтез гетерогенных данных, синтез гетерогенных данных на основе онтологий. Выявлено перспективное, на сегодня, метод синтеза данных.*

Ключевые слова: *синтез, гетерогенные данные, ГИС, INSPIRE, прецизионное земледелие.*

M. Vasyuhin
doctor of technical sciences, professor of the department of computer systems and networks Faculty of information technology NUBYP Ukraine;

Y. Ivanyk
candidate of technical sciences, assistant of the department of computer systems and networks Faculty of information technology NUBYP Ukraine;

A. Sinitsyn
graduate student the department of computer systems and networks Faculty of information technology NUBYP Ukraine.

METHODS OF SYNTHESIS HETEROGENEOUS DATA GIS PRECISION AGRICULTURE

Abstract. *The methods of synthesis of heterogeneous data: synthesis based on metadata, contextual heterogeneous data synthesis, synthesis of heterogeneous data based on ontologies. Found the most promising, to date, the method of synthesis of data.*

Keywords*: synthesis, heterogeneous data, GIS, INSPIRE, precision agriculture.*
