УДК 004.652

А. Ю. Берко[1], К. А. Алєксєєва[2]
[2]Національний університет “Львівська політехніка”,
[1]кафедра загальної екології та екоінформаційних систем,
[2]кафедра соціальних комунікаціх та інформаційної діяльності

ОПРАЦЮВАННЯ НЕОДНОРІДНИХ ДАНИХ В ІНФОРМАЦІЙНИХ РЕСУРСАХ web-СИСТЕМ

Описано метод інтегрованого опрацювання неоднорідних інформаційних ресурсів web-систем, який ґрунтується на моделі подання даних як узгодженого поєднання значень, правил їх зображення, правил інтерпретації та структури. Метод передбачає декомпозицію загального процесу на підпроцеси інтеграції значень, синтаксису даних, семантики і структури. Перевагою такого підходу до інтеграційних процесів є можливість їх виконання на рівні метасхем даних, що зменшує кількість звернень до власне даних web-систем, обсяги яких можуть бути значними.

Ключові слова: web-ресурс, значення даних, інтеграція даних, розподілені системи даних, неоднорідні дані.

In the paper the method of integrated processing of heterogeneous information resources web-systems is described. This method is based on the model of data description as a coherent combination of data values, rules of data representation, interpretation rules and data structure. The method involves decomposition of general process into subprocesses of data values integration, data syntax integration, semantics and structure integration. The advantage of this approach is that the integration process can be performed at data metascheme level. It allows to reduce the number of access operation to very large data sets of web-systems.

Key words: web-resource, data value, data integration, distributed data systems, heterogeneous data.

Вступ. Загальна постановка проблеми

Життєвий цикл web-системи складається з послідовності етапів – від розроблення концепції та технологічних засобів її реалізації до припинення використання створеного продукту. На усіх етапах життєвого циклу ключовим елементом web-системи є інформаційний ресурс. Проектування та створення інформаційних ресурсів сучасних web-систем є достатньо складним завданням, яке здебільшого не має однозначного виконання. Одним із чинників скдадності є неоднорідність даних, на основі яких формується такий ресурс. При цьому виникає потреба поєднання в одному середовищі елементів баз даних, баз знань, документів, web-сторінок, текстових даних тощо. Це своєю чергою вимагає організації єдиного функціонального та змістовного простору зберігання, опрацювання та застосування таких неоднорідних складових інформаційного ресурсу.

Вирішення проблем такого характеру можливе двома шляхами: перший – гомогенізація, приведення всіх даних, які утворюють інформаційний ресурс web-системи, до єдиної схеми, другий – інтеграція, забезпечення можливостей спільного опрацювання даних, сформованих та інтерпретованих у різний спосіб. Перший підхід ґрунтується на уніфікації та стандартизації даних і передбачає формування єдиного ресурсу з детермінованою структурою. Для цього потрібно використання додаткових засобів, часових, технологічних та інших витрат, при цьому процес перебудови схем супроводжують ризики зниження якості даних через їх спотворення, виникнення неоднозначностей та суперечностей. Другий підхід ґрунтується на гетерогенній моделі інформаційного ресурсу і дає змогу уникнути додаткових маніпулювань даними, пов’язаних з такими перетвореннями затрат і ризиків, та забезпечує можливість використання даних за принципом "as is" – у форматі їх створення.

Зв'язок висвітленої проблеми із важливими науковими та практичними завданнями

Активний розвиток і запровадження Інтернет-технологій у різноманітних сферах, розширення функціональних можливостей сучасних web-систем, зростання обсягів завдань, які вони виконують, обумовлюють специфіку інформаційних ресурсів таких систем. Однією з характерних особливостей сучасного інформаційного ресурсу є його неоднорідність. Це обумовлюється різноманіттям структури, форматів та змісту вхідних даних, які використовуються web-системою. Найбільш характерною неоднорідність вхідних даних є для систем електронної комерції (насамперед контент-комерції), багатофункціональних Інтернет-порталів, медійних, новинних, науково-технічних ресурсів [1], [3]. Прикладами таких проектів є інтернет-магазини контенту Google Play, Apple AppStore, Amazon.com, Alibaba, новинні портали MSN, Yahoo.com. Актуальність проблематики створення і застосування неоднорідних web-ресурсів складної структури обумовлює активні наукові дослідження, метою яких є розроблення прогресивних методів і технологій для виконання завдань інтеграції неоднорідних даних. Зокрема, а роботах Д. Ланде [5] досліджено основні закономірності формування неоднорідних інформаційних потоків у середовищах web-систем. Загальні, фундаментальні принципи інтеграції інформаційних ресурсів, які досліджено і викладено у роботах М. Р. Когаловського [4] та Л. А. Калініченка [2], продовжують залишатися актуальними, і сьогодні застосовуються у виконанні завдань опрацювання неоднорідних даних, зокрема, web-ресурсів.

Аналіз останніх досліджень та публікацій

Проблематика формування інформаційних ресурсів web-систем різного призначення на основі неоднорідних даних сьогодні є актуальним об'єктом наукових досліджень та практичних розробок. Особливо істотним є цей аспект у комерційних проектах [1]. Питання спільного опрацювання відмінних за структурою даних розглянуто, зокрема, у монографії Л. Калініченка [2], у роботах Д. Майєра та А. Хелеві (David Maier, Alon Halevy) [7], М. Ленцеріні (Marco Lenzerini) [8] та ін. Інтеграція, як метод формування неоднорідних інформаційних ресурсів web-систем, передбачає їх об'єднання у спосіб, при якому спільне використання є більш простим і ефективним, ніж локальне застосування кожної складової. Метою інтеграції даних є отримання єдиної, цілісної і релевантної інформаційної картини на основі різноманітних за формою та походженням вхідних наборів даних, отриманих з різних джерел. Концепція інтеграції даних є відомою достатньо давно і в різні часи була реалізована у формі вирішень, актуальних свого часу – обчислювальних ресурсів загального користування, корпоративних мереж, розподілених баз даних, сховищ та просторів даних тощо [2], [3]. Вперше формальну модель процесу інтеграції даних запропонував М. Ленцеріні у [8]. У подальшому ця модель отримала загальне визнання, зокрема її використано у [9, 10] та ін. В основу моделі покладено архітектуру, яка ґрунтується на застосуванні глобальної схеми даних та множини локальних джерел даних. У локальних джерелах зосереджено реальні значення даних, у той час як глобальна схема забезпечує узгоджене комплексне віртуальне зображення вмісту цих джерел. Основним завданням такого моделювання є формальний опис зв'язку між джерелами даних та глобальною схемою. Підхід до моделювання процесів інтеграції, запропонований у [8], ґрунтується на концепції двох парадигм інтеграції – ресурсо-центричній *(global-as-view)* і схемо-центричній *(local-as-view)* та описує узагальнений механізм, який є спільним для обох підходів. Формально, таку модель інтеграції даних подано як трійку

$$<\Gamma, \Sigma, M>,$$

де Γ – глобальна схема інтегрованих даних, описана в термінах деякої мови L_Γ над алфавітом A_Γ, Σ – вхідна схема джерела даних описана в термінах мови L_Σ над алфавітом A_Σ, M – відображення, яке задає відповідність виду

$$Q_\Sigma \rightarrow Q_\Gamma \,;\; Q_\Gamma \rightarrow Q_\Sigma \,,$$

у яких Q_Σ і Q_Γ – вирази однакової розмірності, визначені, відповідно, над вхідною та глобальною схемами [8]. У результаті, процес інтеграції даних має забезпечити релевантність виразів, запитів та операцій над даними в інтегрованому наборі до відповідних виразів, запитів та операцій у вхідних джерелах даних (рис. 1).

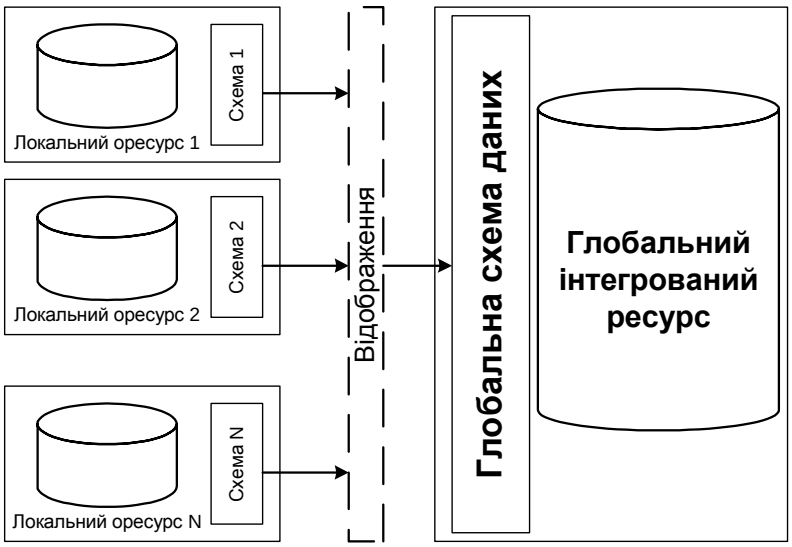

Рис.1. Загальна схема процесу інтеграції даних

Ключовим елементом такої моделі є відображення *M*, яке узгоджує локальні набори даних із глобальною схемою. Залежно від парадигми інтеграції даних, таке відображення має різний характер:

– у ресурсо-центричній інтеграції (global-as-view) призначенням відображення *M* є узгодження складу та змісту глобальної схеми даних із складом та змістом локальних джерел даних;

– у схемо-центричній інтеграції (*local-as-view*) таке відображення має зворотний характер – узгодження способів подання локальних даних із складом та змістом глобальної схеми [7].

Істотним недоліком такої моделі процесу інтеграції є те, що вона оперує лише схемами даних, не враховуючи участі, власне, даних, які мають певний набір специфічних властивостей, що значною мірою впливають на процеси і процедури інтеграції інформаційних ресурсів web-систем, а також отримані результати.

Виділення проблем

Вхідні дані, на основі яких формується інформаційний ресурс сучасної web-системи, як правило, є неоднорідними за форматами подання, структурою та змістом. Основною проблемою при їх створенні є узгодження способів подання, опрацювання та інтерпретації у єдиному середовищі. Для цього необхідно визначити та обґрунтувати процедури об'єднання та гармонізації синтаксису, семантики та структури таких даних, які забезпечують можливості їх ефективного спільного опрацювання і застосування для виконання ключових завдань web-системи.

Формулювання мети

Метою статті є розроблення узагальненої методики формування цілісних інформаційних ресурсів web-систем, яка ґрунтується на методах розподіленої інтеграції неоднорідних даних, їх синтаксису, структури та семантики. Основними завданнями, які забезпечують досягнення мети, є:

– побудова моделі інтеграції даних, з врахуванням особливостей інформаційних ресурсів web-систем;

– визначення основних принципів та шляхів розподілу процесів інтеграції значень даних, їх синтаксису , структури і семантики;

– розроблення методів інтеграції синтаксису, структур та семантики даних у web-системах та порядку їх застосування.

Формування web-ресурсу шляхом інтеграції неоднорідних даних

Розширена модель інтеграції даних. Подальший розвиток концепції моделювання процесів інтеграції даних можливий шляхом переходу від поняття схеми як об'єкта інтеграції у формальній моделі [2, 3,8] до поняття набору даних. У статті описано модель інтеграції даних, яка узагальнює і

розвиває концепції моделі М. Ленцеріні [8]. Узагальнення ґрунтується на тому, що у процесах інтеграції враховують не лише схему даних, але і, безпосередньо, самі дані, як множину значень, поданих і впорядкованих у визначений спосіб, що мають певний зміст, функції та призначення, і які подають спеціальними засобами. Кожен набір даних являє собою поєднання схеми як деякого формалізованого опису складу і структури даних та множини значень (констант), сформованих відповідно до вимог схеми. У такий спосіб, формальними об'єктами моделі є множина вхідних (локальних) наборів даних, вихідний (глобальний) набір інтегрованих даних і відображення, що встановлює відповідність міх елементами вхідного та вихідного наборів [8]. Таку модель подамо як трійку у вигляді

$$<DS^L, Map(DS^L,DS^I), DS^I>$$

де $DS^L=\{<D_i, \Sigma_i> \mid i=1,\ldots N\}$, – множина *локальних* вхідних наборів даних, Σ_i – схема даних *i*-го вхідного набору, виконана в термінах мови опису вхідних схем L^L, D_i – множина значень (констант) *сформованих* на основі множини символів вхідного алфавіту A^L відповідно до вимог схеми, $DS^I=<D^I,\Sigma^I>$ – глобальний вихідний набір інтегрованих даних, Σ^I – схема глобального набору *інтегрованих* даних виконана в термінах мови опису вихідних схем L^I, D^I – множина значень *вихідного* набору даних, заданих символами вихідного алфавіту A^I, $Map(DS^L,DS^I)$– *відображення* локальних вхідних даних у глобальний вихідний набір інтегрованих даних.

Принциповою відмінністю розширеної моделі інтеграції даних від формальної моделі М. Ленцеріні [8] є поняття глобального набору інтегрованих даних, як результату, отриманого в процесі інтеграції. При цьому цей набір може бути сформовано переміщенням значень вхідних даних у глобальне середовище, і відображенням через віртуальні структури та елементи даних. Загалом, модель такого вигляду більшою мірою відповідає реальним процесам інтеграції, аніж формальна модель [8]. Схему процесу інтеграції згідно з такою моделлю, яка враховує особливості різних методів, показано на рис. 2.

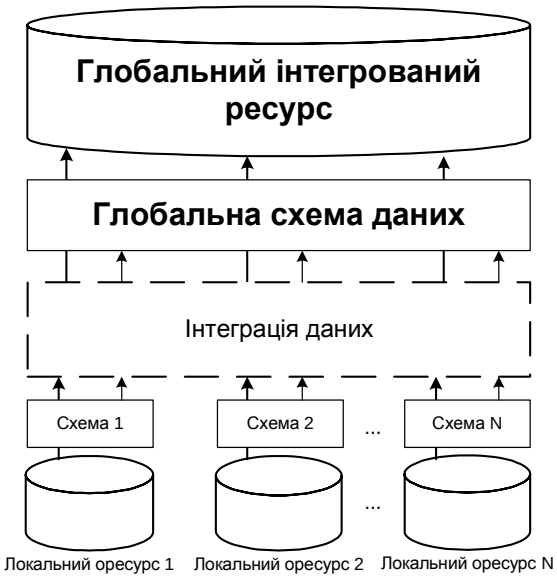

Рис.2. Схема процесу інтеграції даних за розширеною моделлю

Застосування такої моделі, дає можливість побудувати достатньо точний і адекватний формальний опис процесів інтеграції даних.

У загальному випадку, визначення довільного набору даних *DS* утворює систему вигляду

$$DS=<D, G, S, H>,$$

де *D* – множина *значень*, якими зображають множину понять деякої предметної області, *G* – формалізоване подання синтаксису даних, *S* – формалізований опис структури даних, *H* – формалізоване подання семантики даних.

У такий спосіб формальне подання набору даних як кортежу виду $<D, \Sigma>$, де D – множина значень , Σ – схема даних, змінюють на кортеж виду $<D, \Theta>$, де $\Theta=<G, S, H>$, формальне подання синтаксису, структури і семантики даних у цьому наборі, яке, в подальшому, будемо називати його метасхемою. Метасхема є узагальнення поняття схеми, утворене шляхом доповнення опису структури та обмежень даних засобами формалізованого опису їх синтаксису та семантики. Введення поняття метасхеми дає змогу побудувати ширший та детальніший, порівняно зі схемою, опис властивостей даних у процесах інтеграції.

Загалом, процес інтеграції даних передбачає низку дій, пов'язаних з їх перетворенням і утворенням нових даних на основі початкових. Його розглядають як послідовність дій, що передбачає узгодження, перетворення, об'єднання та фільтрування даних, і має на меті утворення кінцевого набору даних DS на основі множини початкових наборів, формально це зображає вираз

$$DS=I(DS_1, DS_2, \ldots, DS_N),$$

де, $DS_1, DS_2, \ldots, DS_N$ – множина вхідних початкових наборів даних, N – кількість наборів даних, які беруть участь у процесі інтеграції, I – оператор інтеграції даних – умовне позначення відповідності між множиною локальних вхідних наборів даних та вихідним глобальним набором даних, яке визначає послідовність перетворень, що виконують для отримання результатів інтеграції.

У загальному випадку, такі набори даних можуть містити значення, які повторюються, тобто

$$D_1 \cap D_2 \cap \ldots \cap D_N \neq \varnothing,$$

Враховуючи визначену вище модель даних, яка ґрунтується на специфікації їх синтаксису, семантики та структури

$$DS=<D, \Theta>=<D, G, S, H>,$$

опис інтеграції можна звести до дій над цими компонентами, замінивши опис набору даних, на деталізований опис всіх складових визначення даних

$$<D^I, \Theta^I>=<D^I, G^I, S^I, H^I>=I(<D_i, \Theta_i> \mid i=1,\ldots N)=$$

$$I(<D_1, G_1, S_1, H_1>, <D_2, G_2, S_2, H_2>, \ldots, <D_N, G_N, S_N, H_N>),$$

де $<D_i, G_i, S_i, H_i>$, $i=1,2, \ldots, N$ – деталізоване зображення i-го набору даних.

У такий спосіб задачу інтеграції даних можна розкласти на окремі підзадачі – інтеграції значень даних, інтеграції синтаксису, інтеграції структури та інтеграції семантики. Загальний оператор інтеграції даних I, при цьому подають як комбінацію

$$I=<I^D, I^G, I^S, I^H>,$$

де I^D – оператор інтеграції значень, яким позначено відповідність між множиною значень даних у локальних вхідних наборах та значеннями у вихідному наборі; I^G – оператор інтеграції синтаксису, яким задають відповідність між синтаксисами вхідних даних та синтаксисом вихідного набору; I^S – оператор інтеграції структури даних, що задає відповідність між описом локальних структур вхідних наборів даних та описом глобальної структури вихідного набору даних; I^H – оператор інтеграції семантики, яким позначено відповідність між елементами опису семантики локальних наборів та описом семантики вихідних даних на глобальному рівні. Процес інтеграції при цьому поділяється на відповідні підпроцеси –

$$<D^I, G^I, H^I, S^I>=<I^D(D_1, D_2, \ldots, D_N), I^G(G_1, G_2, \ldots, G_N),$$

$$I^S(S_1, S_2, \ldots, S_N), I^H(H_1, H_2, \ldots, H_N)>.$$

Згідно з таким поданням інтеграції виникає певна послідовність етапів – інтеграція, відповідно, синтаксису, структури і семантики даних.

Інтеграція синтаксису інформаційних ресурсів web-систем. Проблема інтеграції синтаксису даних (синтаксична інтеграція) є базовою відносно інтеграції інших складових їх загального опису. Вирішення проблем побудови узагальненої структури та семантики даних є можливим лише на основі єдиної узгодженої системи їх позначення. Поняття синтаксису даних саме по собі є комплексним і враховує різні аспекти їх зображення у документах, базах даних,

сховищах та репозиторіях даних тощо [2, 3, 10]. Враховуючи це, синтаксис даних подамо як поєднання трьох складових [2]

$$G=<A, T, R>,$$

де A – алфавіт, T – множина типів даних, R – множина синтаксичних обмежень.

Алфавіт визначає множину символів, які застосовують для зображення значень даних у визначеному середовищі. Як правило, алфавіт складається з літер, цифр, спеціальних та службових символів. Однак, на визначення алфавіту мають вплив, зокрема, такі фактори, як локалізація середовища опрацювання даних, характер задач, для розв'язання яких застосовують дані, особливості процесів їх збереження, передачі та опрацювання, специфіка інтерпретації та застосування різноманітних значень даних. Поряд із традиційними засобами позначення даних, у сучасних системах широко застосовують графічні, звукові, мультимедійні та інші елементи для їх зображення і опрацювання, а також дані складних і комплексних типів, потокові та активні дані, що створює додаткові труднощі у виробленні єдиного узгодженого подання таких даних.

Поняття типу даних визначають як результат класифікації значень за способами зображення та опрацювання. Сьогодні поряд із такими класичними типами, як числові, символьні, логічні, дата-час тощо широко застосовують специфічні типи даних, які відображають особливості їх змісту, опрацювання та застосування. Це, зокрема, такі скалярні типи як "гіперпосилання", "валюта", "об'єкт", "локатор" та інші, комплексні (агрегатні) типи – "масив", "запис", "множина", "XML-документ" тощо, об'єктні типи та типи даних, які визначає користувач. Таке різноманіття типів даних, з одного боку, створює додаткові можливості щодо зображення та опрацювання інформаційного ресурсу, а з іншого, ускладнює засоби підтримки середовища зберігання даних, процедури їх сумісного застосування, перетворення та об'єднання.

Обмеження, як елемент синтаксису даних, застосовують з метою уніфікації форм подання даних та створення значень, адекватних до понять та величин, які вони зображають. Обмеження синтаксису задають у вигляді кількісних показників, розмірностей, форматів, шаблонів, правил формування значень, визначення підмножин допустимих символів тощо. Такі обмеження може бути визначено і на рівні технологій підтримки даних, і на рівні користувача.

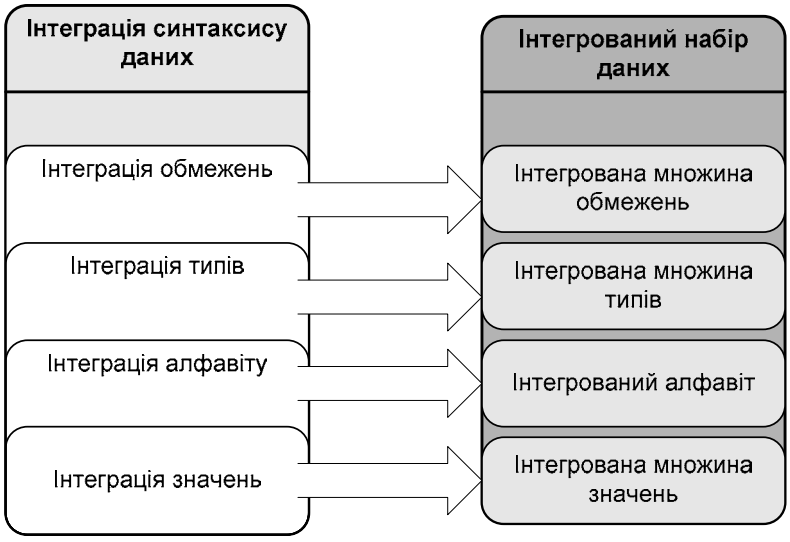

Рис.3. Схема процесу інтеграції синтаксису даних

Отже, процес інтеграції синтаксису розподіляється на послідовність процесів інтеграції алфавітів, інтеграції типів та інтеграції обмежень (рис. 3). Згідно з такою схемою синтаксис зображення значень інтегрованого набору G^I даних утворюється поєднанням трьох складових

$$G^I=<A^I, T^I, R^I>,$$

де $A^I = I^A(A_1, A_2, \ldots, A_N)$ – алфавіт інтегрованого набору даних, утворений шляхом інтеграції алфавітів вхідних наборів даних, $T^I = I^T(T_1, T_2, \ldots, T_N)$ – множина типів даних, які застосовують в

інтегрованому наборі, отримана як результат інтеграції типів вхідних даних, $R^I = I^R(R_1, R_2, \ldots, R_N)$ – множина обмежень інтегрованого набору даних, сформованих інтеграцією обмежень вхідних даних, I^A, I^T, I^R – оператори інтеграції, відповідно, алфавітів, типів даних та обмежень.

Інтеграція структур неоднорідних даних у web-системах. Гетерогенна модель структури інформаційного ресурсу web-систем ґрунтується на інтеграції структурованих (реляційних) даних, що зберігаються у базах даних, та, так званих, нереляційних даних, до яких належать, зокрема, слабкоструктуровані (напіструктуровані) дані, дані без попереднього опису структури, сенсорні і потокові дані, процедурні дані тощо [10]. Загальну модель структури такого єдиного інтегрованого набору неоднорідних даних подамо як

$$C^I = <R, <NR_1, NR_2, \ldots, NR_k>, <J^R, J^N, J^{RN}>>,$$

де C^I – опис загальної структури інтегрованого інформаційного ресурсу, R – опис структури реляційної складової, яку утворюють дані, структуровані у формі таблиць бази даних; NR_1, NR_2, …, NR_k – опис структур нереляційних складових різного типу, J^R – множина зв'язків між реляційними елементами, J^N – множина зв'язків між нереляційними елементами, J^{RN} – множина зв'язків між реляційними та нереляційними елементами.

Основні завдання інтеграції баз даних з даними, структурованими в інші способи, а також напрями і принципи їх реалізації визначено, зокрема, в [3, 10]. Типові підходи до інтеграції структурованих реляційних, слабкоструктурованих і самоструктурованих даних описують такі базові моделі.

1. Інтеграція структурованих даних зі слабкоструктурованими (документальними, текстовими, просторовими, темпоральними, візуальними і мультимедійними) даними. Сучасні системи управління базами даних значною мірою забезпечують виконання таких завдань шляхом використання спеціальних типів даних (темпоральних (часових) типів, символьних та бінарних об’єктів, масивів та колекцій, генерованих типів тощо) та типу даних "XML-документ". Значення таких типів інтегрують у таблиці і доповнюють ними перелік елементарних значень в описах сутностей чи фактів. Структуру таблиць реляційної бази, в яких разом з реляційними зберігають слабкоструктуровані дані, описують як

$$R(A_1, A_2, \ldots, A_k, X_1, X_2, \ldots, X_m,),$$

де A_1, A_2, …, A_k, – стовпці таблиці, які зображають скалярні значення традиційних та спеціальних типів, X_1, X_2, …, X_m – стовпці, які зображають слабкоструктуровані значення.

2. Інтеграція баз даних та процедурних даних. Така модель передбачає інтеграцію власне даних, які зберігаються у базах, та множини об’єктних типів даних разом з методами, які пов’язано з такими даними. У цьому випадку, кожен стовпець таблиці подають парою вигляду (A, M), де A – стовпець таблиці, M – набір методів, пов’язаних з цим стовпцем, які описують певний набір процедур, що активізуються подіями, які впливають на стан даного стовпця. Структуру такої таблиці описує вираз

$$R((A_1, M_1),(A_2, M_2),\ldots,(A_k, M_k)).$$

3. Інтеграція в бази даних тригерів та процедур опрацювання даних. Застосування таких елементів забезпечує реалізацію концепції активної бази даних. Така база даних разом зі значеннями зберігає опис певних правил та дій, які виконують при зміні стану об’єктів бази даних. Таблиці, до складу яких входять традиційні дані та активні елементи, описує модель структури вигляду

$$R(A_1, A_2, \ldots, A_k, T_1, T_2, \ldots, T_m,),$$

де A_1,A_2, …, A_k, – стовпці таблиці, які зображають звичайні типізовані значення, T_1,T_2, …, T_m – набір тригерів, які описують дії, пов’язані зі зміною стану таблиці.

4. Інтеграція статичних даних з потоками та чергами даних. Потокові дані – це множина значень, які не зберігаються на носіях системи, а існують лише в момент їх сприйняття чи застосування. Потік даних формують як результат виконання запитів, пересилання або відбору

даних. Черга – спеціальний тип потоку даних, в якому кожній одиниці надано порядкову ознаку. Структуру потоку даних S у певний момент часу t описує такий вираз:

$$S_t=<S(R_t), S(X_t), S(S_t`), S(W_t)>,$$

де $S(R_t)$ – структура набору значень, отриманих як результат операцій вибору з таблиць реляційної бази даних, $S(X_t)$ – структура набору даних, отриманих як результат вибору з наборів слабкоструктурованих даних, $S(S_t`)$ – структура набору даних, отриманих як результат вибору з інших потоків даних, $S(W_t)$ – структура набору даних, отриманих з web-ресурсів.

Результатом інтеграції статичної та потокової складових є напівдинамічна структура, яка поєднує дані, збережені в базах даних, та дані, які сформовано у вигляді змінного в часі потоку.

5. Інтеграція сенсорних даних та даних сенсорних мереж. Сенсорна мережа – це мережа розподілених пристроїв, кожен з яких є джерелом даних певного характеру. Така форма подання даних дає змогу реалізувати можливості та переваги мережних обчислень. До складу даних сенсорних мереж можуть входити і структуровані об’єкти, і слабкоструктуровані набори даних, чи набори даних без попередньо визначеної структури. Нереляційні дані сенсорної мережі можуть мати і статичний, і динамічний характер, тобто подаватися і у формі наборів, і потоків даних. В такому випадку сенсорну мережу поділяють на дві частини – статичну і динамічну, між якими визначено зв’язки та порядок взаємодії. Загальну структуру даних сенсорних мереж описує вираз

$$SN^I=< R, SN, SN_t, J_R, J_{SN}, J_{SNt}, J_{RSNt}>,$$

де R – структура реляційної складової сенсорних даних, SN – структура статичної складової сенсорної мережі, SN_t – структура динамічної складової сенсорної мережі в момент часу t, J_R – схема зв’язків між елементами реляційної складової, J_{SN} – схема статичних зв’язків між елементами сенсорної мережі, J_{SNt} – схема динамічних зв’язків між елементами сенсорної мережі в момент часу t, J_{RSNt} – схема зв’язків між елементами реляційної, статичної і динамічної складових сенсорної мережі в момент часу t.

6. Сховище чи простір даних, як засіб інтегрованого подання і опрацювання даних, передбачає утворення деякої глобальної структури, на яку відображають структури локальних вхідних ресурсів. Залежно від способу інтеграції, така схема може бути статичною (у разі застосування процедур "видобування-перетворення-завантаження") або створюватися динамічно (для сховищ на основі On-Line інтеграції). Глобальну структуру сховища чи простору C^G подають як об'єднання відображень n локальних структур на глобальну, виконаних за попередньо визначеною процедурою

$$C^G=Str_1(C^G)\cup Str_2(C^G)\cup \ldots \cup Str_n(C^G),$$

де C^G – глобальна структура сховища даних, $Str_i\ (C^G)$ – відображення локальної структури i-го вхідного ресурсу на глобальну структуру (i=1,2, …, n). Структура, утворена у такий спосіб, є результатом інтеграції множини неоднорідних вхідних ресурсів, і забезпечує можливість їх спільного застосування, управління та доступу до них.

Загальну гетерогенну модель структури неоднорідного інформаційного ресурсу, із врахуванням базових інтеграційних моделей, поданих у п.п. 1–6, може бути трансформовано до вигляду

$$C^I=<<R(A_1, A_2, \ldots, A_k, X_1, X_2, \ldots, X_m,), R((A_1, M_1),(A_2, M_2),\ldots,(A_k, M_k)), R(A_1, A_2, \ldots, A_k, T_1, T_2, \ldots, T_m,)$$
$$>,< S_t , SN^I , C^G >, < J^R, J^N, J^{RN} >>,$$

де реляційну складову утворюють таблиці, інтегровані зі слабкоструктурованими даними, методами та тригерами, *нереляційну* складову – потокові, сенсорні дані та дані, зосереджені у сховищах і просторах. Застосування гетерогенної моделі структурування даних дає позитивний ефект у процесах формування інформаційних ресурсів web-систем. Це досягається, насамперед, через скорочення затрат на додаткові перетворення структур даних на етапі побудови інформаційного ресурсу та в процесах його життєвого циклу, підвищення надійності і адекватності даних.

Інтеграція семантики даних у web-ресурсах. Найперспективнішим сьогодні підходом до інтеграції семантики даних є інтеграція на основі онтологій [9]. Застосування онтологій як ефективного засобу семантичної інтеграції обґрунтовано в [3], [9]. 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 – онтологію даних. Узагальненим формальним зображенням онтології є трійка вигляду [9]

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

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) [9].

Процедури семантичної інтеграції даних у процесах проектування інформаційних ресурсів web-систем передбачають побудову для кожного вхідного набору даних D_i (i=1,2, …, n) власної онтології $O(D_i)$, яка утворює повний, узгоджений та однозначний опис семантики інформаційного ресурсу та 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.

Онтологія такого характеру 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 і подавати її у вигляді системи метаданих спеціального виду. У такий спосіб, задачу семантичної інтеграції даних можна звести до побудови деякої глобальної онтології інтегрованого інформаційного ресурсу, шляхом поєднання множини вхідних локальних онтологій, виявлення при цьому відповідностей між ними, а також усунення змістових суперечностей і конфліктів між онтологіями. Для цього процес семантичної інтеграції даних будують на основі виконання низки попередньо визначених умов, які мають на меті забезпечити коректне, з погляду змісту, поєднання локальних вхідних даних в глобальному вихідному ресурсі.

Умови можливості семантичної інтеграції вхідних локальних наборів даних формулюють як послідовність вимог щодо узгодженого спільного застосування елементів різних онтологій вхідних даних. Вхідні набори даних D_i та D_j вважають семантично інтегрованими, придатними для сумісного використання у формуванні глобального інтегрованого ресурсу, якщо для двох онтологій $O(D_i)$ та $O(D_j)$, 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, різного змісту та напряму між однаковими концептами,
  (2) відсутні однотипні зв'язки, що не можуть бути реалізованими одночасно;
- у множинах функцій інтерпретації $F(D_i)$ та $F(D_i)$
  (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.

Перевірку зазначених умов семантичної інтеграції даних у єдиному webресурсі може бути реалізовано і на формальному, і на експертному рівні, при цьому результат має бути однаковим. Виконання всієї множини вимог дає підстави для висновку про можливість поєднання двох наборів даних на рівні їх змісту з отриманням семантично коректного результату. Ключова властивість онтологій – створювати однозначне подання змісту даних і на людському рівні, і на рівні

інформаційних технологій – становить істотну перевагу семантичної інтеграції на основі онтологій перед іншими підходами, а саме:

(1) побудова змістовно повного та узгодженого опису семантики вхідних локальних наборів неоднорідних даних та вихідного інтегрованого інформаційного web-ресурсу;

(2) можливість застосування уніфікованих засобів та процедур незалежних від предметної області і застосування проектованої web-системи;

(3) можливості реалізації процедур формування інтегрованого web-ресурсу за допомогою відповідних програмних засобів;

(4) визначення та аналіз умов семантичної інтеграції web-ресурсів на формальному рівні;

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

Як наслідок, в результаті процесу семантичної інтеграції формується семантично узгоджений web-ресурс, який об'єднує змістовне наповнення набору неоднорідних вхідних даних.

Висновки і перспективи подальших наукових розвідок

Застосування методу інтеграції неоднорідних інформаційних ресурсів web-систем, який ґрунтується на розподілі процесів інтеграції синтаксису, структур та семантики даних, створює додаткові можливості та переваги у процесах побудови та опрацювання таких ресурсів. Це досягається, зокрема, через скорочення витрат на надлишкові перетворення даних і на етапі формування інформаційного ресурсу, і в процесах його життєвого циклу; зменшення ризиків втрати та спотворення даних у процесах формування нових структур; повноту, адекватність і достовірність відтворення структурних, семантичних та синтаксичних особливостей вхідних даних в інформаційних ресурсах web-систем. Особливістю описаного в роботі підходу є можливість розв'язання значної частини задач створення та маніпулювання інформаційним ресурсом на рівні метаданих, які містять опис власне даних та їх ключових властивостей. Скорочення кількості звернень до значень даних, обсяги яких часто є доволі істотними здебільшого значно спрощує та прискорює процеси побудови неоднорідних інформаційних ресурсів web-систем.

1. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л: НУЛП, 2009. – 612 с. 2. Калиниченко Л.А. Методы и средства интеграции неоднородных баз данных. / Леонид Калиниченко. – М.: Наука, 1983.– 424 с. 3. Кеберле Н. Г. Огляд сучасних систем інтеграції неоднорідних баз даних і знань [Електронний ресурс] / Н. Г. Кеберле.– Режим доступу: http://shcherbak.net/z-yedinix-pozicij-proanalizovano-suchasnij-stan-sistem-integracii-neodnoridnix-baz-danix-i-znan/ 4. Когаловский М. Р. Методы интеграции данных в информационных системах. Институт проблем рынка РАН // М. Р. Когаловский. – М., 2010. – 74 с. 5. Ландэ Д. Основы интеграции информационных потоков: монография / Д. Ландэ. – К.: Інжиніринг, 2006. – 240 с. 6. Bergamaschi S. Semantic integration of semistructured and structured data sources / S. Bergamaschi, S. Castano, M. Vincini // ACM SIGMOD Record. – March 1999. – Volume 28. – Issue 1. – P. 54–59. 7. Halevy A. From Databases to Dataspaces: A New Abstraction for Information Management / Michael Franklin, Alon Halevy, David Maier // SIGMOD Record. – Dec. 2005. – Vol. 34. – No. 4. – P. 27–33. 8. Lenzerini M. Data Integration: A Theoretical Perspective / Marco Lenzerini // Proc. of the ACM Symp. on Principles of Database. Systems (PODS), 2002. – P. 233 – 246. 9. Ontology-Based Integration of Information – A Survey of Existing Approaches / H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Hubner // Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle, USA, August 4-5, 2001. – P. 108–118. 10. The Lowell Database Research Self-Assessment Meeting: Lowell, Massachusetts, 4–6 May 2003 // Communications of the ACM (CACM). – 2005. – Vol. 48. – No. 5. – P. 111–118.