УДК 004.652

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

МЕТОДИ ІНТЕГРАЦІЇ СИНТАКСИСУ РІЗНОРІДНИХ ДАНИХ У СИСТЕМАХ ЕЛЕКТРОННОГО КОНТЕНТ-БІЗНЕСУ

Розглядаються основні підходи до об’єднання синтаксису різнорідних даних у системах електронного контент-бізнесу на основі трансляції, гомогенізації та уніфікації правил їх зображення.

Basic ways for integration of heterogeneous data syntax in environment of content-business systems based on their translation, homogenization and unifying of data presentation rules are considered in proposed paper.

Вступ

Особливістю сучасних тенденцій розвитку Internet-систем є розширення їх функціональних можливостей щодо надання послуг користувачам, зокрема, такого виду послуг, які отримали назву "інформація за вимогою" (information on demand). Системи електронного контент-бізнесу – це Web – системи поширення інформаційних продуктів на основі Internet-технологій. Сьогодні цей напрям є одним з найбільш перспективних підходів до обслуговування клієнтів як у мережі Internet, так і в інших інформаційних мережах. Особливістю таких систем є те, що продуктом їх застосування, кінцевим результатом діяльності та основних елементом функціонування є інформаційний ресурс (контент). Вимоги до обсягів та якості інформаційних послуг у середовищі Internet невпинно зростають, а реалізація цих вимог все більше ускладнюється внаслідок неможливості простого механічного розширення обсягів інформаційного продукту, що надається користувачеві. Методологія й засоби реінжинірингу інформаційних процесів у мережах аналогічні тим, які використовуються при реінжинірингу бізнес-процесів, однак цей процес у галузі електронного контент-бізнесу має низку особливостей. Одним з напрямів реінжинірингу систем електронного контент-бізнесу є інтеграція. В загальному випадку предметом інтеграції є бізнес-процеси, платформи, прикладні задачі та інформаційні ресурси систем електронного бізнесу.

Як правило, інформаційний ресурс систем контент-бізнесу має гетерогенний характер, що обумовлюється різнобічністю потреб і пропозицій у цій сфері діяльності. Це потребує поєднання в одному середовищі елементів баз даних, документів, web-сторінок, мультимедійних даних тощо. Отже, очевидною стає проблема інтеграції різнорідного контенту в єдиному середовищі зберігання, опрацювання та застосування. Інтеграція інформаційних ресурсів передбачає їх об'єднання, при якому спільне використання є простішим і ефективнішим, ніж локальне застосування кожної складової.

Деякі концептуальні засади вирішення проблем інтеграції даних розглядаються зокрема у [3, 4].

Проблеми інтеграції даних в системах контент-бізнесу

Метою інтеграції даних є отримання єдиної і цілісної картини корпоративної бізнес-інформації 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 у свій час – обчислювальних ресурсів загального користування, корпоративних мереж, розподілених баз даних, сховищ даних тощо. Інтеграція даних є складним, багатогранним та об'ємним процесом, який передбачає, зокрема, реалізацію процедур видобування, перетворення і завантаження (extraction, transformation, loading, скорочено ETL) даних

з різних систем та джерел в єдиний інтегрований набір даних, призначений для опрацювання і аналізу (підготовки звітності), а також узгодження методів та принципів подання, опрацювання та інтерпретації даних. Сховища, вітрини даних, розподілені та інтегровані корпоративні бази даних, інформаційне наповнення Web-систем є найтиповішими прикладами таких наборів, а інструменти ETL – це компоненти процесу інтеграції даних.

Згідно з [4] основними об'єктами у процесах інтеграції даних є глобальна схема інтегрованих даних, вхідна схема джерела даних і деяке відображення, що дає змогу перейти від способу зображення даних у вхідній множині джерел до способу їх глобального зображення в інтегрованому наборі даних. Загальну схему процесів інтеграції даних в гетерогенних системах зображено на рис 1.

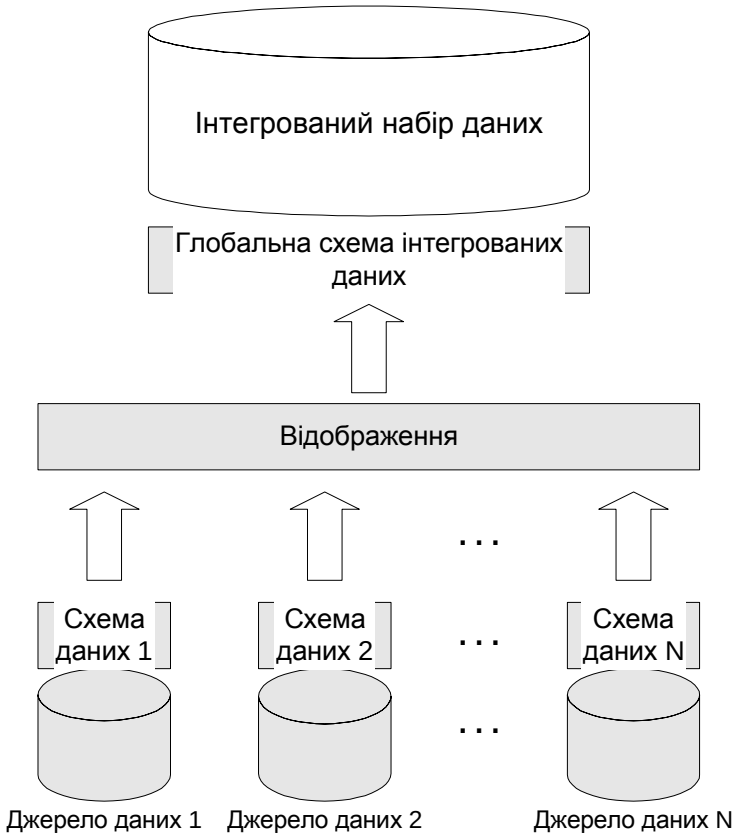

Рис. 1. Загальна модель процесів інтеграції даних

Формально систему інтеграції даних I можна подати як трійку $I=<\Gamma, \Sigma, M>$,

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, задане у вигляді відповідностей $Q_\Sigma \rightarrow Q_\Gamma$ та $Q_\Gamma \rightarrow Q_\Sigma$, у яких – Q_Σ і Q_Γ вирази однакової розмірності, визначені відповідно над вхідною схемою та глобальною схемою [3].

У результаті процес інтеграції даних має забезпечити релевантність виразів, запитів та операцій над даними в інтегрованому наборі до відповідних понять у вхідних джерелах даних. Основні шляхи розв'язання таких проблем на формальному рівні опубліковано, зокрема, у [1, 3].

Порядок та методика інтеграції синтаксису даних

Концептуально дані можна розглядати як деяку формальну мову, яку застосовують для позначення множини понять у середовищі інформаційної системи. Обов'язковими і невід'ємними властивостями даних є синтаксис, семантика та структура. У загальному випадку засоби визначення довільного набору даних D утворюють деяку формальну систему [6] вигляду

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

де V – множина значень, якими зображають множину понять деякої предметної області, G – синтаксис даних, S – структура даних, H – семантика даних.

Процес інтеграції даних – це послідовність дій, що передбачає їх узгодження, перетворення, об'єднання, фільтрування і має на меті утворення кінцевого набору даних D на основі множини початкових наборів:

$$D=I(D_1, D_2, \ldots, D_N),$$

де I – оператор інтеграції даних,

$D_1, D_2, \ldots, D_N$ – множина вхідних початкових наборів даних, які у загальному випадку можуть містити значення, що повторюються, тобто $D_1 \cap D_2 \cap \ldots \cap D_N \neq \varnothing$.

Інтеграція множини різнорідних наборів даних в єдине ціле не є їх простим механічним об'єднанням. Цей процес передбачає низку дій, пов’язаних з їх перетворенням і утворенням нових значень на основі початкових. Враховуючи модель даних, яка ґрунтується на визначенні їх синтаксису, семантики та структури $D=<V, G, S, H>$, формальне визначення процесу інтеграції можна звести до дії над цими компонентами, замінивши опис

$$D=I(D_1, D_2, \ldots, D_N)$$

на деталізований опис всіх складових визначення даних

$$<V, G, S,H> =I(<V_1, G_1, S_1, H_1>,<V_2, G_2, S_2, H_2>, \ldots, <V_N, G_N, S_N, H_N>),$$

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

Тобто проблему інтеграції даних можна декомпонувати на окремі проблеми інтеграції значень даних, інтеграції синтаксису, інтеграції структури та інтеграції семантики. Загальний оператор інтеграції даних I при цьому можна подати як комбінацію

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

де I^V – оператор інтеграції значень, I^G – оператор інтеграції синтаксису, I^S – оператор інтеграції структури даних, I^H – оператор інтеграції семантики. Процес інтеграції при цьому буде декомпоновано на відповідні підпроцеси, які можна описати формальною схемою вигляду

$$<V, G, H, S> =<I^V(V_1, V_2 , \ldots, V_N), I^G(G_1, G_2 , \ldots, G_N), I^S(S_1, S_2 , \ldots, S_N), I^H(H_1, H_2 , \ldots, H_N) >.$$

Взаємне співвідношення цих процесів та їх класифікація за рівнями зображені на рис 2.

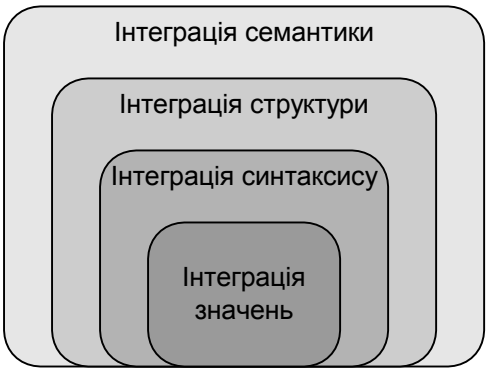

Рис. 2. Рівні інтеграції даних

Згідно з такою схемою кожен наступний рівень інтеграції ґрунтується на результатах попереднього і не може бути реалізований без них. Так, семантична інтеграція даних є можливою лише після інтеграції їх структури, яка, своєю чергою, потребує побудови інтегрованого синтаксису, який визначає способи зображення даних в інтегрованому наборі.

У роботі розглядаються проблеми, пов’язані з інтеграцією синтаксису даних. Такі завдання розв’язують на етапі створення проекту інтегрованого середовища даних, на етапі видобування, перетворення, завантаження даних – ETL, а також у процесах динамічної інтеграції потокових даних чи Internet-ресурсів. У кожному з перелічених напрямів застосування інтегрованого інформаційного ресурсу процеси інтеграції синтаксису даних мають як спільні риси, так і певні особливості і шляхи реалізації. У роботі запропоновано певне узагальнення та формалізація

процесів формування синтаксису інтегрованих даних, що може стати основою для створення інтеграційних технологій різного спрямування і призначення.

Задачі та процеси інтеграції синтаксису даних

Проблема інтеграції синтаксису даних є базовою відносно інтеграції інших складових їх загального опису. Вирішення проблем побудови узагальненої структури та семантики даних є можливим лише на основі єдиної системи їх позначення. Поняття синтаксису даних саме по собі є комплексним і враховує різні аспекти їх зображення у документах, базах даних, сховищах та джерелах даних. Враховуючи це, синтаксис даних можна подати як деяку формальну систему

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

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 впливають зокрема такі фактори, як локалізація середовища опрацювання даних до мови користувачів, характер задач, для вирішення яких застосовують дані, особливості процесів їх збереження, передачі та опрацювання, специфіка змісту 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.

Поняття типу даних може бути визначено як результат класифікації їх 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 та типи даних, що визначає користувач. Така різноманітність типів даних, з одного боку, створює додаткові можливості щодо зображення та опрацювання інформаційного ресурсу, з іншого – значно ускладнює засоби підтримки середовища зберігання даних, процедури їх сумісного застосування, перетворення та об'єднання [1].

Обмеження як елемент синтаксису даних застосовують з метою створення значень, максимально адекватних до понять та величин, які вони зображають. Обмеження синтаксису задають у вигляді кількісних показників, розмірності, форматів, шаблонів, правил формування значень, визначення підмножини допустимих символів тощо. Такі обмеження можна визначити як на рівні систем і технологій підтримки даних, так і на рівні користувачів.

Отже, проблему інтеграції синтаксису даних можна декомпонувати на проблеми інтеграції алфавітів, інтеграції типів та інтеграції обмежень. Співвідношення цих проблем та результатів їх вирішення подано на рис. 3.

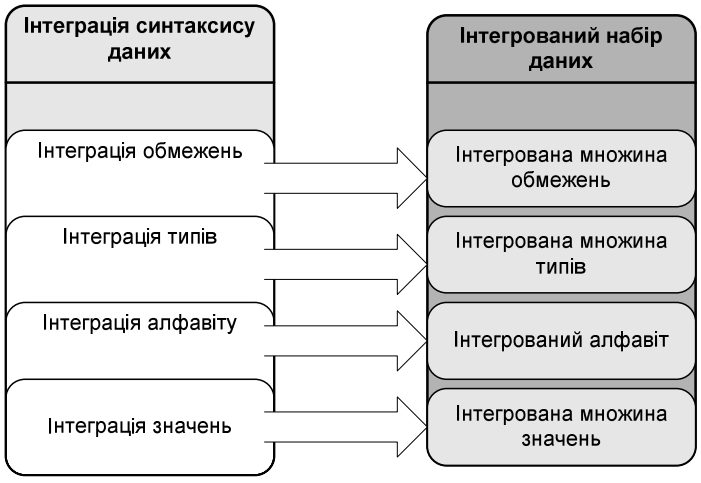

Рис.3. Схема процесу інтеграції синтаксису даних

За такою схемою, синтаксис зображення значень інтегрованого набору G^I даних може бути подано як поєднання 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 $A_1, A_2, \ldots, A_N$;

$T^I = I^T(T_1, T_2, \ldots, T_N)$ – множина типів даних, які застосовують в інтегрованому наборі, отримана як результат інтеграції типів даних, визначених для вхідних даних;

$R^I = I^R(R_1, R_2, \ldots, R_N)$ – множина обмежень інтегрованого набору даних, сформованих інтеграцією обмежень, які застосовано для вхідних даних;

I^A, I^T, I^R – оператори інтеграції відповідно алфавітів, типів даних та обмежень.

Порядок інтеграції алфавітів

Інтеграція алфавіту на етапі проектування інтегрованого середовища даних полягає у створенні деякої множини символів для зображення значень даних з інтегрованого набору – A – інтегрованого алфавіту, такого, що для кожного символу алфавіту A_i, $i=1,2, \ldots, N$, котрий застосовують для зображення значення вхідного набору даних D_i, існує однозначне відображення $T_i : A_i \rightarrow A$, яке кожному символу вхідного алфавіту і-го набору даних $\sigma_i(A_i)$ ставить у відповідність символ інтегрованого алфавіту – $\sigma(A)$. Варіанти співвідношення вхідних алфавітів та інтегрованого алфавіту наведено на рис. 4. Як показано на рисунку, можливими є такі варіанти співвідношення алфавітів вхідних та інтегрованих даних:

- вхідний алфавіт є підмножиною інтегрованого і не має перетинів з іншими вхідними алфавітами (A_1);
- вхідний алфавіт є підмножиною інтегрованого і має непорожній перетин з іншим алфавітом, який є підмножиною інтегрованого (A_5);
- вхідний алфавіт є підмножиною інтегрованого і має непорожній перетин з іншим алфавітом, який має частковий перетин з інтегрованим алфавітом (A_2);
- вхідний алфавіт має непорожній перетин з інтегрованим алфавітом та алфавітом, який є підмножиною інтегрованого ($_{23}$);
- вхідний алфавіт не є підмножиною інтегрованого і має непорожній перетин з іншим вхідним алфавітом, який має частковий перетин з інтегрованим алфавітом (A_4);
- вхідний алфавіт не є підмножиною інтегрованого і не має непорожніх перетинів з іншим вхідним алфавітом (A_6);
- вхідний алфавіт не є підмножиною інтегрованого, але при цьому має непорожній перетин з іншими вхідними алфавітами (A_7, A_8).

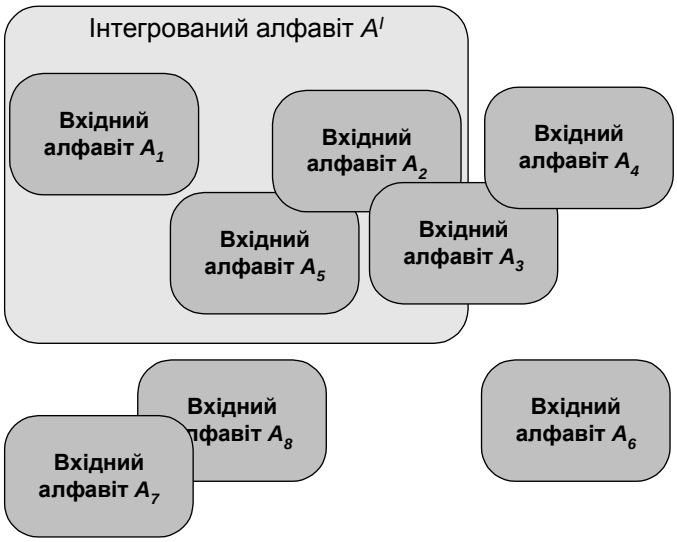

Рис. 4. Діаграма співвідношення вхідних та інтегрованого алфавітів

Процес побудови інтегрованого алфавіту можна подати як послідовність розв'язання взаємопов'язаних задач за такою схемою.

1. Нехай A^0 – деякий початковий набір символів інтегрованого алфавіту.

2. Для кожного зі вхідних алфавітів A_i, $i=1,2, \ldots, N$ перевіряється співвідношення $A_i \subseteq A^0$. Якщо воно виконується, то всі символи алфавіту A_i, що застосовуються для зображення значень набору даних D_i, є припустимими і для зображення відповідних значень в інтегрованому наборі даних D. Тобто можна вважати, що вхідні дані можна вводити до інтегрованого набору без зміни форми їх подання.

3. Однак, у такому випадку може виникати проблема омонімічних символів. Омонімічними називаємо однакові за зображенням символи, що відображають різні значення, наприклад, українська літера "І", латинська літера "I", римська цифра "I" (один), латинські літери, що застосовують для зображення як звуків, так і цифр у шістнадцятковій системі числення тощо. Таке явище може викликати надалі неоднозначну інтерпретацію зображення значень даних та їх змісту. Проблема символів-омонімів має такі варіанти вирішення:

- заборона застосування однакових символів для позначення різних понять; цей спосіб передбачає визначення єдиного застосування для всіх символів, що мають однакове зображення; такий варіант вирішення проблеми омонімічних символів є можливим у випадках, коли їх застосовують для формальних значень, які не мають додаткової (фонетичної, лексичної чи змістової) інтерпретації (наприклад, однотипне застосування латинських та кириличних літер, які збігаються за написанням у реєстраційних номерах автомобілів); у цьому випадку можливими є проблеми при інтерпретації, прочитанні чи фонетизації значень даних;
- використання символів омонімів без обмежень; в такому випадку за кожним зі символів, які мають однакову форму, зберігається власний спосіб застосування; у такому випадку проблема омонімії символів інтегрованого алфавіту не вирішується в процесі інтеграції, а переноситься на рівень застосування даних;
- заміна однакових за формою символів альтернативним зображенням – транслітерація; таке перетворення усуває омонімію символів без втрати можливостей зображення даних.

4. Якщо множина символів вхідного алфавіту не є підмножиною інтегрованого алфавіту – $A_i \not\subset A^0$, то в такому випадку її можна поділити на дві підмножини $A_i^1 = A_i \cap A^0$ та $A_i^2 = A_i \setminus A^0$. До складу першої входять символи, які є елементами інтегрованого алфавіту. Процес інтеграції у цьому випадку описано вище. До складу множини A_i^2 входять символи вхідного алфавіту, які не зустрічаються у поточному інтегрованому алфавіті A^0.

5. У такій ситуації можливим є явище поліморфізму, тобто виникнення символів-синонімів. Синонімами у цьому випадку вважаємо різні за формою зображення, застосовані для позначення однакових понять. Наприклад, великі та малі літери у словах позначають однакові звуки, числові значення можуть бути зображені у різних системах числення, з використанням арабських, латинських цифр чи літер тощо. Поява синонімічних символів у алфавітах може стати причиною неоднозначного сприйняття та інтерпретації значень даних при їх формальному опрацюванні. Щодо проблеми опрацювання синонімічних символів, то можливими є такі варіанти її вирішення:

- заміна поліморфних символів гомоморфними зображеннями, тобто приведення різноманітних за формою позначень до єдиної форми – наприклад, застосування лише великих чи лише малих літер, лише арабських цифр, якими замінюють аналогічні римські тощо;
- паралельне застосування поліморфних зображень символів-синонімів без обмежень; у такому випадку їх інтерпретація залежатиме від змісту та застосування значень даних;
- створення власної інтерпретації та правил застосування символів-синонімів – цей шлях потребує детального аналізу їх властивостей, але дає змогу значно розширити можливості інтегрованого алфавіту щодо зображення даних, наприклад – з великих літер починають власні назви, спеціальними символами позначають оператори чи операції, арабськими цифрами зображають кількісні значення, а римськими – порядкові тощо.

6. Щодо набору символів $A_i^2 = A_i \setminus A^0$ можливими є такі варіанти дій

- заборона використання символів з набору A_i^2 для зображення інтегрованих даних;
- розширення інтегрованого алфавіту за рахунок введення множини до його складу символів A_i^2 з утворенням наступної версії $A^1 = A^0 \cup A_i^2$
- транслітерація – заміна символів, що не є елементами інтегрованого алфавіту символами зі складу алфавіту A^0

7. У результаті запропонованих перетворень має бути сформовано інтегрований алфавіт $A^I = I_A(A_1, A_2, \ldots, A_N)$, що визначає множину символів, призначених для зображення значень даних в інтегрованому наборі.

Порядок інтеграції типів даних

Інтеграція типів даних при побудові інтегрованого набору полягає у формуванні множини типів даних T – такої, що для кожного з типів, застосованих у вхідних наборах даних, існує відображення $F: T_i \rightarrow T$, яке встановлює однозначну відповідність між типами даних $t(T_i)$, що застосовують у вхідному наборі D_i , $i=1,2, \ldots, N$, та $t(T)$ – типами даних, що застосовують в інтегрованому наборі даних D. Взаємне співвідношення різних множин типів даних у процесі інтеграції показано на рис. 5. Як видно з діаграми, наведеної на рисунку, аналогічно до етапу інтеграції алфавітів вхідні набори типів даних можуть мати повний або частковий перетин з інтегрованим чи не мати таких перетинів, а також мати або не мати перетинів один з одним.

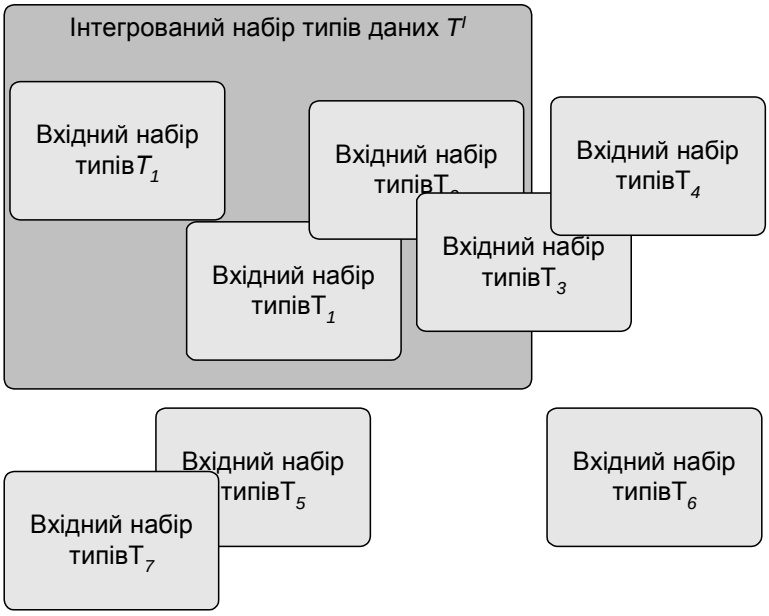

Рис. 5. Діаграма співвідношення вхідних та інтегрованого наборів типів даних

Процес формування інтегрованого набору типів передбачає послідовне вирішення таких проблем.

1. Нехай $T^0 = \{ t_1(T^0), t_2(T^0) , \ldots, t_m(T^0)\}$ – деякий початковий набір типів, що підтримуються в інтегрованому наборі даних.

2. Для кожного набору вхідних даних D_i , $i=1,2, \ldots, N$ необхідно перевірити співвідношення $T_i \subseteq T$, яке визначає узгодження типів вхідного набору з типами інтегрованого набору даних D .

3. Виконання цієї умови ще не означає, що для зображення даних вхідного набору D_i застосовуються типи, дозволені до застосування в інтегрованому наборі, оскільки існує можливість появи типів-омонімів. Типами-омонімами, або омонімічними типами називатимемо однакові позначення типів, різних за способами реалізації. Наприклад, значення типу "дата/час" можуть зображатись як числовими, так і символьними величинами, тип "text" в одних застосуваннях

позначає символьні рядки, в інших – примітки; значення логічного типу зображаються як числові або бітові тощо. Розбіжності такого характеру надалі можуть спричиняти помилки в реалізації, неоднозначну інтерпретацію даних та отримання некоректних результатів.

4. Проблема омонімії типів даних має, зокрема, такі варіанти вирішення

- заміна омонімічних типів даних новими, що за визначенням не збігаються з іншими;
- переформатування значень зі вхідного набору даних до відповідних типів, визначених в інтегрованому наборі даних;

5. Якщо множина типів даних T_i , застосованих у вхідному наборі D_i, виходить за межі множини типів даних інтегрованого набору D, тобто $T_i \not\subset T$, це свідчить про наявність у вхідному наборі даних, належних до таких типів, які не є елементами множини припустимих типів даних інтегрованого набору. Отже, типи даних вхідного набору можна поділити на дві такі категорії:

- підмножина типів $T_i^1 = T_i \cap T^0$, які входять до множини типів інтегрованого набору даних;
- підмножина типів $T_i^2 = T^0 \setminus T_i$, які не входять до множини типів інтегрованого набору даних.

6. У першому випадку процедура узгодження типів описана вище. У випадку, коли дані деякого вхідного набору D_i належать до типів, які не підтримуються у вихідному інтегрованому наборі даних D, можливі такі варіанти подальших перетворень:

- розширення множини типів даних інтегрованого набору за рахунок доповнення її підмножиною T_i^2 типів даних набору D_i , тобто побудова наступної версії множини припустимих типів даних інтегрованого набору $T^1 = T^0 \cup T_i^2$;
- перетворення даних, зображених у форматах типів до відповідних типів з множини, тобто заміна кожного значення даних типу $t(T_i^2) \in T_i^2$ аналогічним значенням, зображеним відповідно до вимог типу $t(T^0) \in T^0$.

7. Особливої уваги заслуговує ситуація поліморфізму або синонімії типів даних в інтегрованому наборі та і вхідних наборах. Синонімічними типами називатимемо різні за формою подання, але однакові за інтерпретацією типи даних (наприклад, типи REAL та FLOAT, BOOLEAN і LOGICAL тощо). У такому випадку можливими є ситуації, в яких дані одного фактичного типу будуть несумісними між собою при виконанні дій над ними, що, своєю чергою, є потенційною причиною помилок та суперечностей в даних. Розв'язання такої проблеми має два можливі шляхи:

- перетворення поліморфних типів до єдиного виду із застосуванням єдиного способу їх визначення за рахунок вилучення таких типів, які фактично збігаються з іншими;
- сумісне застосування всіх можливих варіантів визначення типів даних, що потребує створення і застосування засобів підтримання поліморфізму типів даних, які позначають різними термінами при їх застосуванні.

Перший шлях є простішим в реалізації та підтриманні, зате другий створює додаткові можливості щодо опису та маніпулювання даними в інтегрованому наборі.

8. Результатом виконання описаних вище кроків є узагальнений та узгоджений перелік типів даних, які застосовують при визначенні одиниць інтегрованого набору $T^I = I_T(T_1, T_2, \ldots, T_N)$.

Порядок інтеграції синтаксичних обмежень даних

Інтеграція обмежень, що застосовують для формування значень даних деяких вхідних наборів, своєю чергою, передбачає утворення такої множини обмежень $R=(r_1(R), r_1(R) , \ldots , r_1(R))$, що для кожного обмеження $r(R_i) \in R_i$, яке застосоване у деякому вхідному наборі даних D_i , $i=1,2, \ldots, N$, існує однозначна відповідність $r(R_i) \rightarrow r_j(R)$. Однак, на відміну від алфавіту та типів даних, обмеження не є вільними елементами, їх формулюють і застосовують лише до конкретних типів, категорій чи значень даних. Тому кожне обмеження, яке застосовують для деякого набору даних D_i, може бути визначено як умова виду $r(R_i, t(T_i), D^j_i)$, котру визначають такі фактори, як належність до множини обмежень R_i певного набору даних – D_i, прив'язка до певного типу даних – $t(T_i)$, а також область застосування – деяка підмножина набору даних $D^j_i \in D_i$. Отже, проблема інтеграції множини обмежень вхідних наборів даних у єдину множину обмежень інтегрованого набору може бути вирішена лише після вирішення задач інтеграції алфавіту та типів даних. Співвідношення наборів обмежень вхідних даних та інтегрованого набору даних подано на рис. 6. Як показано на рисунку, вхідні набори обмежень можуть мати часткові перетини між собою, бути підмножинами

одне одного, бути повністю незалежними, повністю чи частково входити до складу інтегрованого набору, який утворений шляхом їх інтеграції чи не мати перетину і не входити до його складу.

Загальна послідовність дій з інтеграції наборів синтаксичних обмежень, метою якої є створення несуперечного, узгодженого та повного набору обмежень, застосованих до значень з інтегрованого набору даних, може бути подана у вигляді етапів, що виконують згідно з викладеною схемою.

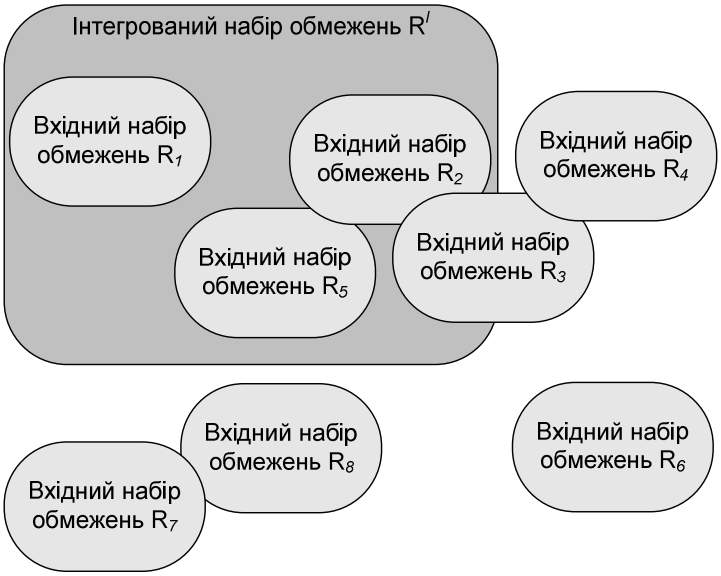

Рис. 5. Діаграма співвідношення вхідних та інтегрованого наборів обмежень

1. Нехай $R^0=(r_1(R^0), r_1(R^0) , \dots , r_1(R^0))$ – початкова множина обмежень деякого інтегрованого набору даних D.

2. Для кожної із множин обмежень R_i вхідних наборів даних D_i , $i=1,2, \dots, N$ виконуємо перевірку на відповідність $R_i \in R^0$

3. Виконання такої відповідності означає, що кожне з обмежень вхідного набору даних існує в інтегрованому наборі. Але для остаточного визначення можливості застосування обмежень до інтегрованих даних необхідно додатково виконати такі перевірки:

- наявність серед типів даних інтегрованого набору типів, щодо яких застосовано обмеження, тобто для кожного з обмежень $r^j(R_i) \in R_i$ перевірити співвідношення $t^j(T_i) \in T$, де $t^j(T_i)$ – тип даних, до якого застосовано обмеження, T – множина типів інтегрованого набору даних;
- наявність серед множини значень інтегрованого набору даних значень, щодо яких застосовано обмеження, тобто для кожного з обмежень $r(R_i) \in R_i$ треба перевірити співвідношення $D^j_i \in D$, де D^j_i – підмножина значень набору даних D_i, до якого застосовано обмеження, T – множина типів інтегрованого набору даних;
- виконання таких вимог дає змогу зробити висновок про можливість застосування обмеження до даних інтегрованого набору, а відсутність узгодження типів і/або значень даних, відповідно, унеможливлює перенесення даного обмеження зі вхідного до інтегрованого набору даних

4. Якщо серед множини обмежень R_i деякого вхідного набору D_i є такі, що не узгоджуються з обмеженнями інтегрованого набору даних D, тобто $R_i \not\subset R^0$, і множина обмежень може бути поділена на підмножини $R_i^1 = R_i \cap R^0$ і $R_i^2 = R_i \setminus R^0$.

5. Множина обмежень R_i^1 є узгодженою з множиною R^0 і процес її інтеграції описано вище, натомість множина обмежень R_i^2 має такі варіанти інтеграції:

- обмеження з множини R_i^2 стосуються значень і/або типів даних, що не входять до складу інтегрованого набору;
- обмеження з множини R_i^2 стосуються значень і типів даних, що входять до складу інтегрованого набору.

6. У першому випадку кожне з обмежень, що не має об'єкта застосування, може бути без втрат вилучене з множини обмежень інтегрованого набору. У другому треба застосувати процедуру узгодження додаткових обмежень R_i^2 та множини обмежень R^0 інтегрованого набору даних. Узгодити ці множини можна за рахунок:

- вилучення обмежень R_i^2 з подальшого застосування у множині обмежень інтегрованого набору даних *D*;
- трансформація обмежень, що входять до множини R_i^2 шляхом заміни їх на еквівалентні за змістом та застосуванням зі складу множини R^0 за принципом – кожному з обмежень $r(R_i^2) \in R_i^2$ відповідає обмеження $r_1(R^0) \in R^0$, яке є допустимим щодо типів та значень інтегрованого набору даних;
- розширення множини обмежень R^0 інтегрованого набору даних за рахунок доповнення її складовими множини обмежень R_i^2 з утворенням нової версії $R^1 = R^0 \cup R_i^2$

7. Результатом послідовності дій інтеграції множини синтаксичних обмежень вхідних наборів даних є формування такого набору правил подання даних, які можна застосувати для визначення властивостей значень інтегрованого набору даних.

Шляхом виконання послідовності дій з інтеграції алфавітів, типів даних та синтаксичних обмежень за описаною вище схемою буде утворено повний та несуперечливий набір елементів інтегрованого синтаксису даних $G=<A, T, R>$, який застосувати як спосіб і засіб зображення інтегрованих даних, отриманих внаслідок реалізації процесів видобування, перетворення та завантаження (ETL) даних у сховищах, а також при їх динамічній інтеграції в оперативних Web-системах.

Висновки

Процеси інтеграції даних мають достатньо широку сферу застосування. Це, зокрема, сховища даних різного типу та спрямування, 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 синтаксису. В основу роботи покладено формалізацію даних як об'єкта, властивостями якого є синтаксис, структура та семантика. Проаналізовано зокрема особливості створення інтегрованих способів зображення даних на етапі їх переміщення від джерел утворення до інтегрованого кінцевого набору. Синтаксис даних у статті подається як деяка формальна система, що складається з окремих самостійних складових – алфавіту, набору типів даних, множини обмежень. Запропоновані способи поєднання гетерогенних за походженням даних, що ґрунтуються на гомогенізації, узгодженні або розширенні їх синтаксису. Такий підхід значно розширює можливості поєднання в єдиному середовищі значень різної природи та форми, зокрема числових, текстових, графічних, звукових чи мультимедійних даних.

Запропоновані вирішення можуть слугувати базисом для створення алгоритмів та методів організації процесів видобування, перетворення та завантаження (ETL) даних у різноманітних сховищах, вітринах даних чи розподілених базах даних.

1. Berko A. Consolidated data models for electronic business systems. Proceedings of IXth Internationale Conference CADSM 2007. – Lviv, 2007. – pp.341–342. 2. Berko A. Models and methods of data integration for content-business systems. Proceedings of Internationale Conference ACSN 2007. – Lviv, 2007.– -p. 112–113. 3.Lenzerini M. Data Integration: A Theoretical Perspective. In: Proc. of the ACM Symp. on Principles of Database. Systems (PODS), 2002, pp. 233–246. 4. C. White. Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise (Report Excerpt). http://www.tdwi.org/Publications/ WhatWorks/display.aspx?id=7979, 2007. 5. Берко А.Ю., Висоцька В.А. Моделі та методи проектування інформаційних систем електронної комерції // Автоматизовані системи управління та прилади автоматики: Всеукраїнський міжвідомчий науково-технічний збірник. – Харків, 2007. – № 138. – С.55–66. 6. Смальян Р. Теория формальных систем. – М.: Наука, 1981.