УДК 004.65:504.06

В. Б. МОКІН, Ю. С. БОГОМОЛОВ

Вінницький національний технічний університет, м. Вінниця

НОВІ ПІДХОДИ ДО АВТОМАТИЗОВАНОГО СТВОРЕННЯ ГЕОІНФОРМАЦІЙНОЇ БАЗИ ЕКОЛОГІЧНОЇ ІНФОРМАЦІЇ

Анотація. У статті описуються нові підходи до створення технології автоматизованої ідентифікації структури та наповнення геоінформаційної бази екологічної інформації на основі різноформатних даних. Вирішення даної задачі є дуже актуальним у наш час, оскільки обсяг неформалізованої за єдиними принципами інформації постійно зростає. Авторами пропонується підхід до обробки такої інформації із використанням трьох взаємопов'язаних онтологічних баз даних — бази об'єктів, їх атрибутивних та просторових характеристик, відповідно. Запропонована технологія дозволить підвищити релевантність та обсяг результатів пошуку взаємопов'язаної екологічної інформації щодо об'єктів карт ГІС за рахунок використання трьох типів інформаційних моделей даних.
Ключові слова: автоматизована ідентифікація структури бази даних, екологічна інформація, геоінформаційні системи.

Аннотация. В статье описываются новые подходы к созданию технологии автоматизированной идентификации структуры и наполнения геоинформационной базы экологической информации на основе разноформатных данных. Решение данной задачи является очень актуальным в наше время, поскольку объем неформализованной по единым принципам информации постоянно растет. Авторами предлагается подход к обработке такой информации с использованием трех взаимосвязанных онтологических баз данных — базы объектов, их атрибутивных и пространственных характеристик, соответственно. Предложенная технология позволит повысить релевантность и количество результатов поиска взаимосвязанной экологической информации по объектам карт ГИС за счет использования трех типов информационных моделей данных.

Ключевые слова: автоматизированная идентификация структуры базы данных, экологическая информация, геоинформационные системы.

Abstract. The paper describes a new approach to creating technology of automated identification of the structure and content of a GIS database of environmental information from multiformat data. Solving this problem is very important nowadays, because the amount of informal according to common principles information is growing. The authors proposed an approach to handling such information using three interrelated ontological databases – database of objects, their attribute and spatial characteristics, respectively. The proposed technology will improve the relevance of search results and coherent environmental information on objects of GIS maps through the use of three types of information data models.
Keywords: automated identification of the structure of a database, ecological information, geoinformational systems.

Вступ

Все більша кількість організацій займається екологічними питаннями та дослідженнями у галузі екологічного моніторингу, що призводить до постійного росту обсягу неформалізованої за єдиними принципами інформації, що зберігається у різноформатних джерелах (текстах природною мовою, графіках, таблицях тощо). Саме тому видобування структурованих знань із таких даних, пошук та аналітична обробка цих знань є надзвичайно актуальними задачами сьогодення. Відомо, що для роботи із просторово-орієнтованими даними оптимальним є застосування засобів геоінформаційних систем (ГІС), тому актуальною задачею є ще й прив'язка видобутих знань до об'єктів карт ГІС.

Постановка задачі

В останні роки все більше уваги приділяється екологічним питанням та аспектам усіх сфер нашого життя та дослідженням в галузі екологічного моніторингу. Цими питаннями займається велика кількість державних та громадських інституцій, наукових та освітніх установ тощо, що призводить до появи величезної кількості, переважно неформалізованої за єдиними принципами, інформації, яку все складніше обробляти, порівнювати між собою, та, відповідно, використовувати для прийняття оптимальних управлінських рішень в галузі управління довкіллям та господарством.

Відповідно Орхуської конвенції про доступ до інформації, участь громадськості в процесі прийняття рішень та доступ до правосуддя з питань, що стосуються довкілля, до екологічної інформації відносять будь-яку інформацію в письмовій, аудіовізуальній, електронній чи будь-якій іншій матеріальній формі про: а) стан таких складових навколишнього середовища, як повітря і атмосфера, вода, ґрунт, земля, ландшафт і природні об'єкти, біологічне різноманіття та його компоненти, включаючи генетично змінені організми, та взаємодію між цими складовими; б) фактори, такі як речовини, енергія, шум і випромінювання, а також діяльність або заходи, включаючи адміністративні заходи, угоди в галузі навколишнього середовища, політику, законодавство, плани і програми, що впливають або можуть впливати на складові навколишнього середовища, зазначені вище в підпункті а), і аналіз затрат і результатів та інший економічний аналіз і припущення, використані в процесі прийняття рішень з питань, що стосуються навколишнього середовища; стан здоров'я та безпеки людей, умови життя людей, стан об'єктів культури і споруд тією мірою, якою на них впливає або може вплинути стан складових навколишнього середовища або через ці складові, фактори, діяльність або заходи, зазначені вище в підпункті б).

Одним із важливих аспектів екологічної інформації є її просторова прив'язка, оскільки більшість екологічних даних стосується об'єктів, які є реальними фізичними сутностями (річки, водосховища, джерела скидів, викидів, відходів, місця видалення відходів, заповідники тощо) із певними географічни-

ми координатами. Такі об’єкти ще називаються просторовими [1]. А їх характеристики, як правило, поділяють на атрибутивні (різні характеристики, стан тощо) та просторові (географічні координати та інша інформація, яка позиціонує об’єкт на карті) [1]. Світовий досвід довів, що для зберігання та обробки інформації про просторові об’єкти оптимальним є використання геоінформаційних систем (ГІС) та технологій. У той же час, більшість інформації про екологічні об’єкти формується та зберігається у різних форматах: текстова, бази даних, електронні таблиці тощо [2, 3]. Отже, актуальними є методи, технології та засоби поєднання електронних карт ГІС та різноформатної інформації в єдиній системі, яка дозволятиме зручно та швидко здійснювати пошук та обробляти цю інформацію [2, 3]. Однак, якщо подібні підходи вже існують (наприклад, з використанням онтологічних моделей та баз даних [2-6]), то методи автоматизованого або, краще, автоматичного формування та ідентифікації таких геоінформаційних баз даних ще не розроблені.

Таким чином, постає задача розробки підходів до формалізації за єдиними принципами та розробки інформаційної технології автоматизованої ідентифікації структури та наповнення геоінформаційної бази екологічної інформації на основі різноформатних даних, у в т.ч. електронних карт ГІС.

Розв’язання задачі

Існує багато підходів для формалізації різноформатних даних з метою їх подальшої обробки [4, 5]. Найбільш поширеним є застосування онтологічних баз даних [2-4, 6]. Також часто застосовують універсальні пошукові методи та засоби, які створюють індекси для кожного типу джерел даних та здійснюють пошук по створених індексах одночасно [7].

Ідея технології, яка пропонується, полягає, по-перше, в адаптації відомих методів та технологій формалізації різноформатної екологічної інформації шляхом їх автоматизованої прив’язки до просторових об’єктів, з урахуванням відношень між ними. По друге, у формуванні та ідентифікації комплексу із трьох взаємопов’язаних інформаційних моделей:

- модель Х просторових різнотипних об’єктів з певним кодуванням;
- модель АХ атрибутивних характеристик (знань) цих об’єктів та відношень між ними у вигляді певних аналітичних залежностей та ін.;
- модель РХ просторових характеристик (знань) цих об’єктів та топологічних відношень між ними.

Такий поділ інформаційних моделей на окремі є подібним до поділу таблиць на окремі таблиці, відповідно до правил нормалізації, із приведенням їх до третьої нормальної форми [8].

У разі використання онтологічних баз даних для формалізації інформації та формування індексних масивів пропонується створювати окремо три бази даних:

База об’єктів та їх простих характеристик (назва, тип, шар карти, унікальні ідентифікатори та інше кодування, умовне позначення на карті тощо).

База просторових характеристик об’єктів (координати, глибина, довжина, геометричні розміри, площа тощо), яка ідентифікується по картах ГІС, та топологічних відношень між ними (об’єкт А знаходиться на об’єкті Б, об’єкт А впадає в об’єкт Б, від об’єкта А до об’єкта Б 22 км тощо).

База атрибутивних характеристик (рівень викидів в атмосферу, кількість опадів тощо) та аналітичних відношень між ними (характеристика А залежить від характеристики В як залежність А = F(B) тощо), яка ідентифікується по різноформатних джерелах даних.

Для прикладу розглянемо такі географічні об’єкти як «річка Південний Буг», «річка Соб» та підприємства-джерела стічних вод, які надходять у ці річки. У кожній із трьох баз може зберігатись така інформація:

У базі об’єктів: а) об’єкти із назвами «Південний Буг», «Соб», тип — «річка», їх закодовані унікальні ідентифікатори та умовне позначення у вигляді синьої тонкої лінії тощо; б) підприємства-джерела стічних вод, які здійснюють скиди у води вищезгаданих річок (їх назви, коди та умовні позначення).

У базі просторових характеристик: інформація про довжину обох річок, середню глибину, про те, яка з річок є притокою якої річки (або яка з річок впадає у яку), координати підприємств-джерел стічних вод та в яку річку вони здійснюють скид вод тощо.

У базі атрибутивних характеристик: інформація про рівень забруднення річок по роках чи місяцях, обсяг скидів по кожному джерелу стічних вод, певні аналітичні залежності, наприклад, залежність витрат вод у р. Південний Буг від витрат води (мало- чи повноводності) її притоки р. Соб у місці, де одна впадає в іншу, тощо.

Таким, чином, модель геоінформаційної бази екологічної інформації І, яка може бути ідентифікована та наповнена на основі різноформатних даних, матиме вигляд:

$$I = D\{X, A_X, P_X\},$$

де D — деяка функція співвіднесення знань щодо атрибутивних A_X та просторових P_X характеристик із об'єктом X на карті ГІС.

Узагальнений алгоритм та етапи застосування інформаційної технології, яка пропонується, подано на рис. 1.

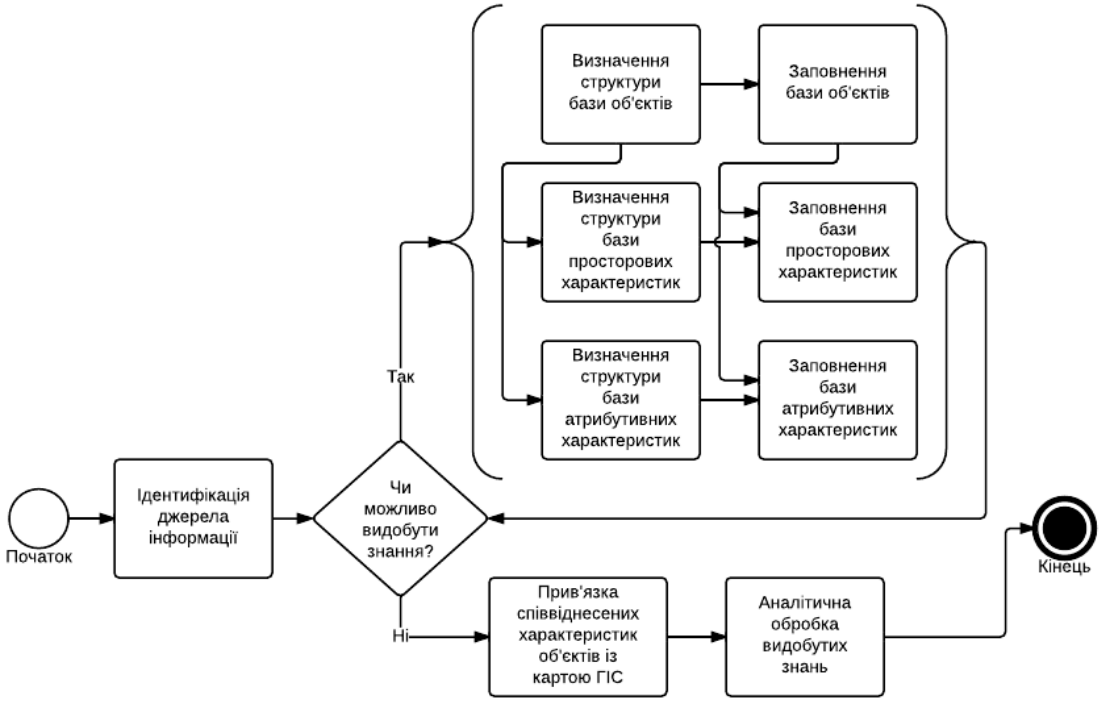

Рисунок 1 — Узагальнений алгоритм застосування інформаційної технології, яка пропонується

Основними етапами алгоритму, зображеному на рис. 1, є наступні:

Ідентифікація джерела інформації. На даному етапі виконується ідентифікація параметрів моделі, що використовується для видобування знань із джерела інформації.

Циклічна ідентифікація та наповнення бази об'єктів, бази просторових та бази атрибутивних характеристик. На даному етапі виконується циклічна перевірка «чи можливо видобути із джерела інформації нові (ще невідомі) знання?», та, за умови позитивної відповіді, виконується уточнення параметрів інформаційних моделей баз об'єктів, їх просторових та атрибутивних характеристик, а також наповнення баз новими видобутими знаннями. Наприклад, на першій ітерації виявляється об'єкт, на другій — ідентифікуються його параметри та топологічні співвідношення з іншими об'єктами, на третій — зв'язок його параметрів з параметрами інших об'єктів тощо.

Прив'язка співвіднесених характеристик до об'єктів карт ГІС. На даному етапі виконується співвіднесення та прив'язка отриманих характеристик із безпосередніми об'єктами карт ГІС, що дозволить використання інструментальних засобів ГІС для обробки цих знань.

Аналітична обробка видобутих знань. На даному етапі відбувається аналітична обробка знань, видобутих на етапі 2, засобами математичної статистики, інструментами ГІС тощо.

Важливо відмітити, що для ідентифікації запропонованих баз даних і знань слід використовувати тільки джерела із достовірною інформацією. Оскільки, потрапляння в систему недостовірних відомостей може суттєво вплинути на достовірність висновків системи в цілому.

Найкраще використовувати дані державної звітності та систематизовані за рік аналітичні доповіді, якими, наприклад, є Національна доповідь про стан навколишнього природного середовища. В Україні як правило, вони розміщується на сайтах Міністерства екології та природних ресурсів України (Національна доповідь) та його територіальних органів (регіональні доповіді).

Запропоновані підходи та технологія дозволить підвищити релевантність та обсяг результатів пошуку взаємопов'язаної екологічної інформації щодо об'єктів карт ГІС за рахунок використання трьох типів інформаційних моделей даних.

Висновки

В роботі запропоновано нову інформаційну модель та нові підходи до створення інформаційної технології автоматизованої ідентифікації структури та наповнення геоінформаційної бази екологічної інформації на основі різноформатних даних. Охарактеризовано основні аспекти її застосування, з урахуванням географічної прив’язки об’єктів екологічного профілю до карт геоінформаційних систем. Наведено узагальнений алгоритм та етапи застосування технології, а також відзначено окремі аспекти її практичного застосування.

Список літератури

1. Комп’ютеризовані регіональні системи державного моніторингу поверхневих вод: моделі алгоритми програми : [монографія] / Мокін В. Б., Боцула М. П., Горячев Г. В. та ін.; під ред. В. Б. Мокіна. — Вінниця: УНІВЕРСУМ–Вінниця, 2005. — 315 с. — ISBN 966-641-132-6.
2. Мокін В. Б. Новий метод пошуку різноформатної екологічної інформації на основі онтологічної бази даних та її XML-представлення / В. Б. Мокін, Ю. М. Коновалюк // Вісник Вінницького політехнічного інституту. — 2009. — № 2. — С. 66—69.
3. Мокін В. Б. Розробка моделей вхідних даних для ітеративного методу пошуку різноформатної екологічної інформації / В. Б. Мокін, Ю. М. Коновалюк // Вісник Вінницького політехнічного інституту. — 2011. — № 5. — С. 44—47.
4. Симаков К. В. Модели и методы извлечения знаний из текстов на естественном языке: Автореферат дис. канд. техн. наук / К. В. Симаков; Московский государственный технический университет имени Н. Э. Баумана – М., 2008. — 16 с. — рус.
5. Kao, A., and Poteet, S. Natural Language Processing and Text Mining / A. Kao, S. Poteet — Springer, 2006. — 277 с.
6. Вороной А.С. Використання онтологій для підвищення якості пошуку інформації для поповнення баз знань інтелектуальних систем / А.С. Вороной // Матеріали Міжнародної науково-технічної конференції «Комп'ютерні науки і інженерія 2009» (Львів, 14 – 16 травня 2009 р.). – Львів, 2009. – С. 364-366.
7. Hatcher, E., Gospodnetic, O. Lucene in Action. / E. Hatcher, O. Gospodnetic. — Manning Publications, 2004. — 456 с. — ISBN 978-1932394283.
8. An Introduction to Database Systems. Русскоязычное издание: К. Дж. Дейт. Введение в системы баз данных. — 8-е изд. — М.: Вильямс, 2006. — 1328 с. — ISBN 0-321-19784-4

Відомості про авторів

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри комп’ютерного еколого-економічного моніторингу та інженерної графіки, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, тел. 59-82-91, e-mail: vbmokin@gmail.com.

Богомолов Юрій Сергійович — аспірант кафедри комп’ютерного еколого-економічного моніторингу та інженерної графіки, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, тел. (0432) 43-77-22, e-mail: yuriy.bogomolov@gmail.com.