УДК 004.9+504.06

В.Б. Мокін, д-р. техн. наук, проф.,
Ю.М. Коновалюк

Державний вищий навчальний заклад
„Вінницький національний технічний університет“,
м. Вінниця, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОШУКУ ГЕОПРОСТОРОВИХ РІЗНОФОРМАТНИХ ДАНИХ

V.B. Mokin, Dr. Sc. (Tech.), Professor,
Yu.M. Konovaliuk

State Higher Educational Institution
“Vinnytsia National Technical University”, Vinnytsia, Ukraine

INFORMATION TECHNOLOGY OF GEOSPATIAL MULTIFORMAT DATA SEARCH

Вперше розроблено інформаційну технологію пошуку геопросторових різноформатних даних, яка дозволяє формалізувати інформацію у трьох основних форматах даних (картах різних пакетів програм для роботи з ГІС, базах даних та форматах текстових процесорів) на основі єдиного підходу та XML-моделі, а також здійснювати пошук інформації в джерелах цих форматів одночасно, що дасть можливість підвищити швидкість і рівень автоматизації пошуку різноформатної екологічної інформації в електронних джерелах.

Ключові слова: *пошук геопросторової інформації, онтологія, контекстно-вільна граматика*

Актуальність. Останнім часом накопичилось чимало екологічної інформації, тобто даних про об’єкти довкілля та пов’язані з ними процеси і явища – це і дані моніторингу та паспортні відомості про якість та кількість природних ресурсів, і відомості про забруднювачів навколишнього природного середовища, і документи законодавчого та нормативно-методичного характеру, що регламентують діяльність у сфері охорони довкілля та управління і моніторингу природних ресурсів.

Екологічна інформація найбільше поширена у трьох основних видах форматів: карти геоінформаційних систем (просторова інформація про розташування природних об’єктів) [1, 2], бази даних (БД) (атрибутивні дані – кількісні та якісні характеристики про стан природних об’єктів) [1], текстові документи (інформація, яка стосується природних об’єктів – описи їх стану, законодавчі акти тощо). Подання інформації у вигляді графічних зображень чи файлів мультимедіа не розглядаємо, оскільки така інформацію є не структурованою. Для автоматизованої обробки її, як правило, спочатку спеціальними засобами обробляють та структурують у вигляді зазначених вище текстових файлів, баз даних та карт ГІС.

Існуючі засоби пошуку інформації дозволяють здійснювати ефективний пошук у документах тільки окремих типів. Існують також розвинуті системи, які дозволяють здійснювати пошук у документах різних типів, але в таких системах, як правило, відсутня підтримка пошуку в картах геоінформаційних систем (ГІС). Ті системи, які ж все таки дозволяють здійснювати пошук і в базах даних, і в картах ГІС (просторово-орієнтований пошук), (наприклад, система Oracle версії 11), мають суттєве обмеження – уся інформація (текст, бази даних та ГІС) повинна бути спочатку інтегрована в одну систему єдиного специфічного формату. Отже, актуальною є задача створення теоретичних та практичних основ швидкого пошуку екологічної інформації, яка зберігається в різних форматах традиційного типу – текст (doc, htm), електронні таблиці (xls), бази даних (mdb), карти геоінформаційних систем (shp, map, sit тощо). При цьому, модель даних пошуку повинна забезпечувати режим асоціативного пошуку, який би базувався на взаємозв’язках між різними типами об’єктів на карті ГІС, та здійснювати ітераційне уточнення пошукових запитів.

Постановка задачі. У кожному із зазначених вище форматів інформація впорядкована певним чином з такими особливостями:

1. Просторово-орієнтовані об’єкти в ГІС мають топологічні відношення, деякі – з’єднуються в мережі. У той же час, вони ієрархічно впорядковуються у класифікаторі. І все це слід враховувати під час пошуку та формалізації моделі даних ГІС. Самі атрибутивні дані впорядковані за об’єктно-ієрархічною моделлю.
2. Дані по просторових об’єктах (далі будемо називати їх „просторово-орієнтована інформація“) у БД мають реляційні зв’язки і це теж слід враховувати для їх ефективного пошуку.
3. Просторово-орієнтована інформація в текстових файлах подається у вигляді речень, пов’язаних за змістом, в якій згадуються значення атрибутів ГІС (наприклад, назви об’єктів карт) та значення полів БД (наприклад, середня глибина ставка). Традиційно їх формалізують за допомогою БД з онтологіями, що дозволяє швидко знаходити потрібну інформацію з високим значенням релевантності.

Отже, постає задача розробки технології пошуку релевантної інформації в основних форматах ГІС (картах, БД та форматах текстових процесорів) з урахуванням особливостей її структурування та формалізації у цих форматах.

Модель даних основних форматів. У [3] запропоновано модель даних документів основних форматів, яка в загальному випадку має такий вигляд [3]

$$S = \left[\{M\}, \{B\}, \{T\}\right],$$

де *{M}* – множина карт ГІС, *{B}* – множина баз даних, *{T}* – множина текстових документів.

У свою чергу, база даних формалізується двома компонентами – назвою файлу N_B і множиною таблиць $\{T_B\}$, а таблиця має назву N_{TB} та складається з множини полів *{F}* і множини записів *{R}* [3]

$$B = \left[N_B, \{T_B\}\right];$$

$$T_B = \left[N_{TB}, \{F\}, \{R\}\right].$$

Карту можна формалізувати як кортеж, який складається з імені файлу N_M, класифікатора L множини екземплярів об'єктів *{O}* і типу координатної системи T_C [3].

$$M = \left[N_M, L, \{O\}, T_C\right].$$

Формалізація класифікатора включає множину об'єктів *{J}*, які можуть бути відображені на карті, множину шарів *{A}* і множину семантик *{E}* [3], де об'єкт J представляє собою прототип об'єктів, які реально містяться на карті, і формалізується назвою N_J, множиною семантик $\{E_J\}$, шаром A_J, до якого він належить, кодом I_J, типом об'єкта T_J (точковий, лінійний, площинний, тощо), а також позначенням V_J [3]

$$L = \left[\{A\}, \{E\}, \{J\}\right];$$

$$J = \left[I_J, N_J, \{E_J\}, A_J, T_J, V_J\right].$$

У свою чергу, кожен об'єкт, який реально присутній на карті, характеризується ключем (унікальним кодом у межах окремої карти) I_O, типом об'єкта з класифікатора J_O, множиною семантик $\{E_O\}$ і метрикою M_O [3].

$$O = \left[I_O, J_O, \{E_O\}, M_O\right].$$

Текстовий документ, у даному випадку, можна формалізувати у вигляді кортежу, який включає назву файлу N_T і його вміст C_T [3]

$$T = \left[N_T, C_T\right].$$

Вміст, у свою чергу, можна представити у вигляді такої контекстно-вільної граматики [3]

$$C_T \rightarrow text$$
$$text \rightarrow text\ text\_element | text\_element$$
$$text\_element \rightarrow K^W | word | number | charset | delimiter | \zeta | \varepsilon$$
$$word \rightarrow word - word | word\ alpha | alpha$$
$$number \rightarrow digitset | digitset\ decimal\_delimiter\ digitset$$
$$digitset \rightarrow digitset\ digit | digit$$
$$charset \rightarrow charset\ character | character$$
$$character \rightarrow digit | alpha$$
$$digit \rightarrow 0|1|2|3|4|5|6|7|8|9$$
$$alpha \rightarrow а|б|..|я|ь|А|Б|..|Я|Ь$$
$$decimal\_delimiter \rightarrow .|,$$

Формалізація даних основних форматів за єдиним підходом. На основі описаної моделі даних у роботі [4] запропоновано модель бази онтологій, за допомогою якої інформація основних форматів формалізується за єдиним підходом.

Онтологія D може бути формалізована у вигляді множини об'єктів $\{O^D\}$ та зв'язків між ними $\{R^D\}$ [3].

$$D = \left[\{O^D\}, \{R^D\}\right].$$

Кожен об'єкт онтології характеризується назвою N_{OD} та інформацією про характер входження даних про цей тип об'єктів у документи всіх типів [3]:

- для карт ГІС – це код об'єкта в картах IJ і його тип TJ;
- для баз даних – це множина таблиць {TB}, в яких містяться атрибутивні дані про об'єкти цього типу,
- для текстових документів – це множина ключових слів $\{K^W\}$.

Таким чином, можна записати [3]

$$O^D = \left[N_{OD}, I_J, T_J, \{T_B\}, \{K^W\}\right].$$

У свою чергу, кожен зв'язок R^D між об'єктами O^{DM} і O^{DR} характеризується ключовим словом K^W, яке використовується для ідентифікації зв'язку в текстових документах, і парами полів $[T_B, F]$ таблиць бази даних для ідентифікації зв'язку в БД [3].

$$R^D = \left[O^{DM} \overset{K^W, \{[[T_{BDM}, F_{DM}], [T_{BDR}, F_{DR}]]\}}{\Rightarrow} O^{DR} \right],$$

де $[[T_{BDM}, F_{DM}], [T_{BDR}, F_{DR}]]$ – пара полів (F_{DM} і F_{DR}), по яких зв'язуються таблиці T_{BDM} і T_{BDR}, що відповідають об'єктам O^{DM} і O^{DR}, відповідно.

Таким чином, модель онтології можна записати так [3]

$$D = \left[\begin{array}{l} \left\{ \left[N_{OD}, I_J, T_J, \{T_B\}, \{K^W\} \right] \right\}, \\ \left\{ \left[O^{DM} \overset{K^W, \{[[T_{BDM}, F_{DM}], [T_{BDR}, F_{DR}]]\}}{\Rightarrow} O^{DR} \right] \right\} \end{array} \right].$$

Розроблено XML-модель бази онтологій, яка враховує всі наведені особливості впорядкованої в основних форматах інформації (рис. 1).

ISSN 2071-2227, Науковий вісник НГУ, 2011, № 2

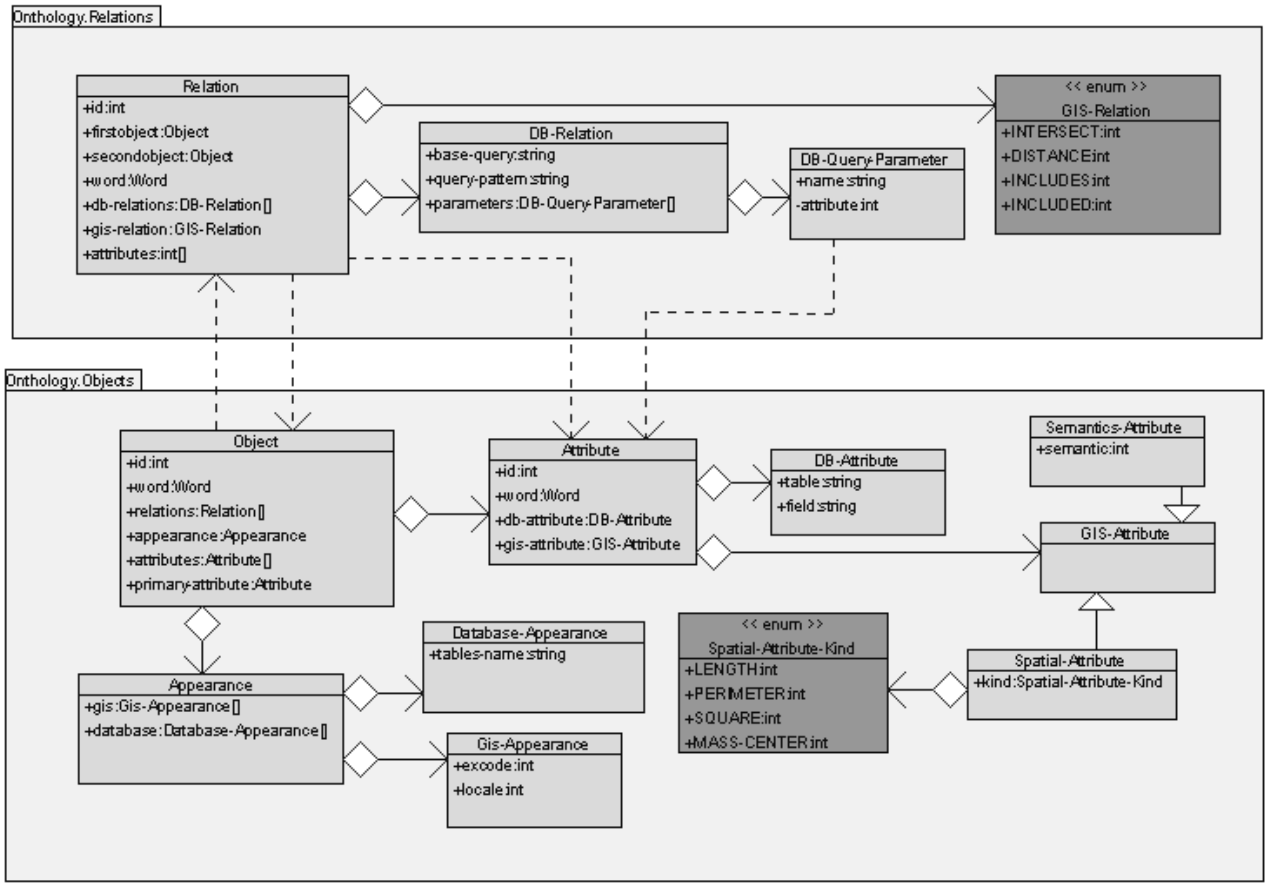

Рис. 1. Інформаційна модель просторово-орієнтованих об'єктів і відношень між ними

Основним модулем є модуль „Onthology.Objects", оскільки він описує об'єкти, ідентифікація яких є відправною точкою при здійсненні різноформатного пошуку. Центральним класом цього модуля є клас „Object". У базовій XML-моделі цей клас виконує 2 основні функції:

- описує, в якому вигляді інформація про об'єкт зустрічається в тому чи іншому типі джерел;
- встановлює зв'язки між об'єктами шляхом посилання на екземпляр класу „Relation".

Перша функція реалізується шляхом створення двох полів – для опису входжень у текстових документах і входжень в інші типи джерел. Для опису входжень у текстових документах достатньо зберігати код слова. Для опису входжень в інші типи джерел передбачено окремий клас „Appearance", який агрегує екземпляри класів, що описують входження в карти ГІС („GIS-Appearance") і в БД („Database-Appearance").

У сукупності з „Onthology.Objects" модуль „Onthology.Relations" складає основу моделі, оскільки описує зв'язки між об'єктами. Як і „Object", клас „Relation" містить поля, які описують належність відношень об'єктам та те, в якому вигляді відношення зустрічаються в документах, але в описі належності відношень об'єктам є одна відмінність. Оскільки відношення встановлюються між двома об'єктами, то посилань на об'єкти в класі – два.

У наведеній моделі особливості об'єктів карти враховуються в класах GIS-Relation, GIS-Appearance, GIS-Attribute та їх нащадках. Так, топологічні зв'язки між об'єктами повністю описуються множиною класів: Object, GIS-Appearance, Relation, GIS-Relation.

У свою чергу, інформація в БД описана класами DB-Relation, DB-Query-Parameter, DB-Attribute, Database-Appearance.

Розробка однопрохідного алгоритму різноформатного пошуку. Для здійснення пошуку пропонується алгоритм однопрохідного різноформатного пошуку (рис. 2).

Опишемо алгоритм однопрохідного пошуку:

1. Виконати розбір вхідного пошукового запиту.
2. Отримати перший об'єкт.
3. Сформувати запит до всіх типів джерел.
4. Якщо немає лексем для обробки, перейти на крок 10.
5. Якщо наступна лексема не позначає атрибут, перейти на крок 9.
6. Запам'ятати поточний атрибут у змінну *A*.
7. Прочитати значення атрибута.
8. Якщо змінна *A* – порожня, додати атрибут за замовчуванням до запитів і перейти на крок 4, ін-

акше – додати атрибут *A* до запитів і перейти на крок 4.

9. Якщо відношення вже оброблено, отримати другий об'єкт і додати його до запитів та перейти на крок 4, інакше – отримати відношення і додати його до запитів та перейти на крок 4.

10. Здійснити пошук.

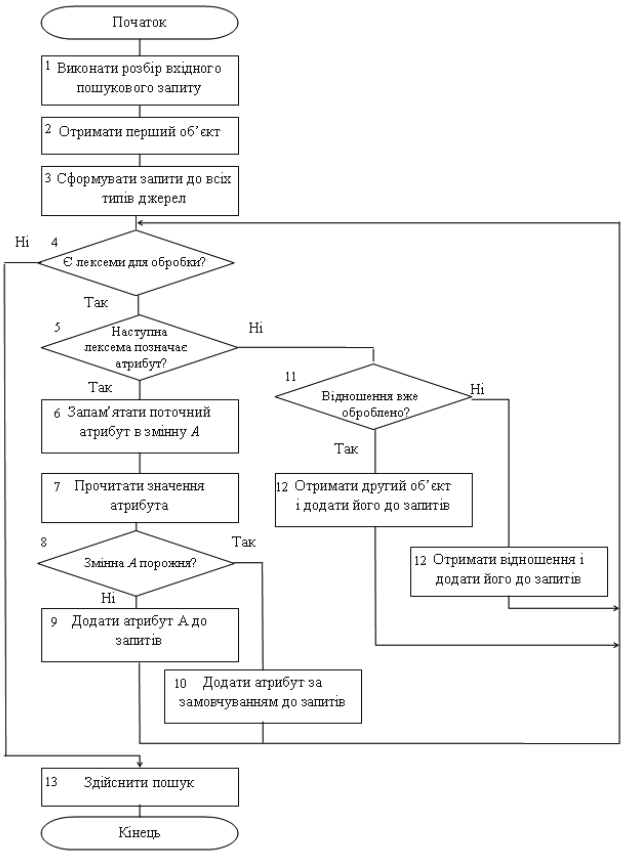

Рис. 2. Схема однопрохідного алгоритму пошуку геопросторової різноформатної інформації

Розробка однопрохідного алгоритму різноформатного пошуку. У [3] запропоновано використовувати контекстно-вільну граматику для розбору вхідних пошукових запитів. З метою розробки багатопрохідного алгоритму пошуку різноформатної інформації необхідно вдосконалити розглянуту граматику та створити породжувальну граматику [5] на її основі.

Виділивши в окреме правило граматики неключові слова, отримаємо

$$Q \rightarrow Q\ query\_element|query\_element;$$
$$query\_element \rightarrow KW|NW|\varepsilon;$$
$$NW \rightarrow charset;$$
$$KW \rightarrow charset;$$
$$charset \rightarrow charset\ character|character;$$
$$character \rightarrow digit|alpha;$$
$$digit \rightarrow 0|1|2|3|4|5|6|7|8|9;$$
$$alpha \rightarrow а|б|..|я|ь|А|Б|..|Я|Ь;$$
$$decimal\_delimiter \rightarrow .|,$$

де NW є будь-якою послідовністю символів, що не є ключовим словом. Таким чином, модель запиту можна записати

$$Q = \left[\left\{K^W\right\}, \left\{N^W\right\}\right].$$

Припустимо, що в результаті пошуку сформовано множину об'єктів {O} і множину пар „атрибут, значення" {[A, V]} для кожного об'єкта Oi. Також позначимо R(Oy, Oz) – функція, що повертає відношення для двох об'єктів з онтологічної БД. Причому, у даному випадку структура об'єкта O матиме такий вигляд

$$O = \left[I_O, K_O, \{[A,V]\}\right],$$

де IO – код об'єкта в онтологічній БД, KO – ключове слово для позначення об'єкта в текстових джерелах (а також пошукових запитах).

Тоді таку граматику можна записати у вигляді

$$Q \rightarrow object\ relation\ object$$
$$object \rightarrow KO|KO\ attribute$$
$$attribute \rightarrow A\ V$$
$$relation \rightarrow R(O1, O2)$$
$$KO \rightarrow key\_word$$
$$A \rightarrow key\_word$$
$$R \rightarrow key\_word$$
$$V \rightarrow charset$$
$$key\_word \rightarrow alphaset$$
$$alphaset \rightarrow alphaset\ alpha|alpha$$
$$charset \rightarrow charset\ character|character,$$
$$character \rightarrow digit|alpha,$$
$$digit \rightarrow 0|1|2|3|4|5|6|7|8|9,$$
$$alpha \rightarrow а|б|..|я|ь|А|Б|..|Я|Ь.$$

Використовуючи таку граматику, можна генерувати пошукові запити для пошуку об'єктів, відношень між ними, з указанням атрибутивної інформації про об'єкти. На базі даної контекстно-вільної граматики можна розробляти розширені граматики для пошуку за складнішими запитами.

Розробка багатопрохідного алгоритму різноформатного пошуку. Алгоритм багатопрохідного пошуку має вигляд (рис. 3):

1. Отримати початковий пошуковий запит.
2. Здійснити різноформатний пошук.
3. Обрати результат пошуку.
4. Якщо обраний результат відсутній у БД, перейти на крок 6.
5. Сформувати список варіантів пошуку запиту для БД і перейти на крок 7.
6. Якщо вибраний результат є в ГІС, сформувати список варіантів доповнення запиту для ГІС, інакше – сформувати список варіантів доповнення запиту для текстових документів.
7. Обрати один із варіантів пошуку для уточнення запиту.
8. Сформувати запит.
9. Якщо потрібно виконати повторний запит, перейти на крок 2.

ISSN 2071-2227, Науковий вісник НГУ, 2011, № 2

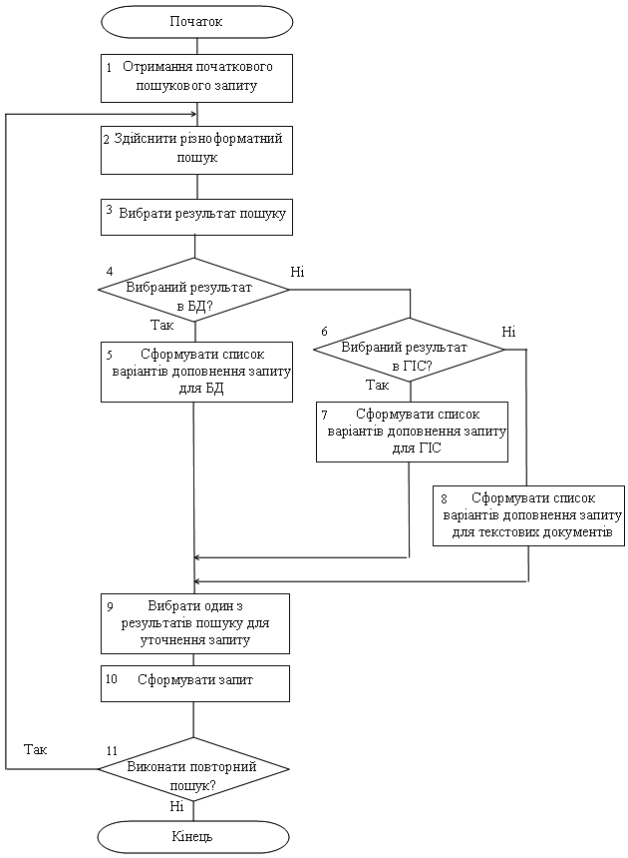

Рис. 3. Схема багатопрохідного алгоритму пошуку геопросторової різноформатної інформації

Приклад пошуку різноформатної інформації. Для демонстрації ітеративного пошуку можна навести такий приклад.

Для першого пошуку використано пошуковий запит „місто Липовець“ (рис. 4). Після здійснення пошуку отримано результати, які розміщено в списку за релевантністю. Для продовження пошуку зі списку можливих запитів для уточнення обрано пошуковий запит „річка протікає місто Липовець“. Отримаємо наступні результати (рис. 5).

Розробка програмного модуля. Для апробації технології створено типовий програмний модуль різноформатного пошуку екологічної інформації.

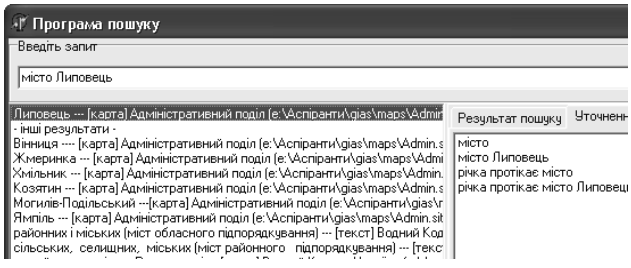

Рис. 4. Результат різноформатного пошуку

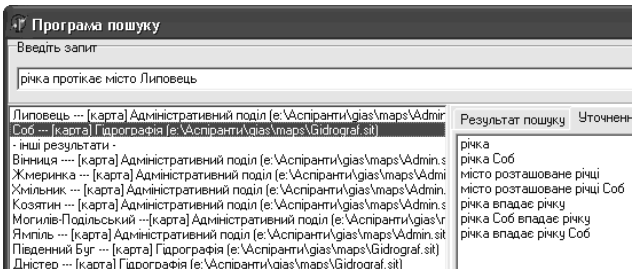

Рис. 5. Результат уточнення різноформатного пошуку

Як видно, вдалося визначити, що через місто Липовець дійсно протікає річка Соб (рис. 6). Звичайно, такий висновок простіше було би отримати просто із ГІС, але даний приклад мав за мету продемонструвати приклад роботи ітераційного методу різноформатного пошуку, а не здійснити повноцінний складний пошук інформації.

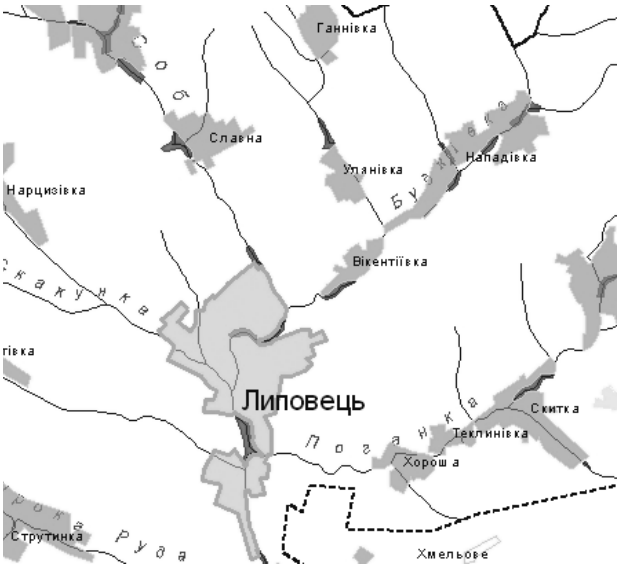

Рис. 6. Річка Соб, яка протікає через місто Липовець

Розроблений типовий програмний модуль різноформатного пошуку може функціонувати як самостійна програмна одиниця, а також в якості інструменту в складі будь-якої програми-оболонки для геоінформаційних систем. Запропонуємо два підходи до інтеграції цього модуля в ГІС.

Перший підхід передбачає створення автономного програмного модуля пошуку та включення його до складу інсталяційного пакету

Другий підхід до інтеграції програмного модуля пошуку передбачає включення його до складу програми-оболонки ГІС в якості одного з інструментів. Така інтеграція дає можливість використовувати засоби програми-оболонки для доступу до БД та карт ГІС, а також відображення результатів пошуку по картах ГІС.

Створені програмні засоби інтегровані в такі інформаційні автоматизовані системи екологічного

моніторингу, розроблені у Вінницькому національному технічному університеті за участі авторів і впроваджені в установах та організаціях України (на створені чи вдосконалені комп'ютерні програми отримано свідоцтва про реєстрацію авторського права №№ 26733, 28115, 28117, 28118, 28120, 28122 у Держдепартаменті інтелектуальної власності Міносвіти і науки України):

– геоінформаційні системи моніторингу стану та підтримки прийняття рішень для інтегрованого управління водними ресурсами басейнів річок Південний Буг, Дністер, Тиса, Сіверський Донець, Кальміус, Прип'ять та водних ресурсів Вінницької та Львівської областей на замовлення підрозділів ООН, ЮНЕСКО, ОБСЄ, Мінприроди та Держводгоспу України (2007–2010 рр.);

– системи впроваджено у понад 20-ти державних установах України – держуправліннях охорони навколишнього природного середовища у Вінницькій та Донецькій областях, відповідних басейнових управліннях водними ресурсами та облводгоспах країни та ін.;

– Єдина автоматизована система Державної екологічної інспекції та спеціальних підрозділів Мінприроди України з отриманням результатів вимірювань стану забруднення довкілля, викидів, скидів і відходів, їх накопичення, оброблення та аналізування (АСУ „ЕкоІнспектор") (2007–2010 рр.) впроваджена в усіх 53-х обласних та регіональних підрозділах Державної екологічної інспекції Мінприроди України, у т.ч. м. Київ, Севастополь та Автономна Республіка Крим, може бути адаптовано до підприємств та до інших відомств;

– інші системи, у т.ч. ГІС державного моніторингу довкілля Вінницької області, впроваджено в основних суб'єктах системи моніторингу області (2010 р.);

– ГІС моніторингу та обліку корисних копалин або мінеральних ресурсів (2009 р.) – створено для Вінницької та Донецької областей, системи впроваджені у відділі промислової та інвестиційної політики Головного управління економіки Вінницької обласної держадміністрації, у ДРГП „Донецькгеологія" та у Держуправлінні охорони навколишнього природного середовища у Донецькій області, відповідно.

Висновки.Таким чином, уперше розроблено інформаційну технологію пошуку інформації про просторово-орієнтовані об'єкти карт геоінформаційних систем у базах даних, текстових файлів на основі єдиної XML-моделі та онтологічної бази даних цих об'єктів, що дозволяє підвищити швидкість та рівень автоматизації пошуку різноформатної екологічної інформації в електронних джерелах.

Список літератури

1. Бусыгин Б.С. Инструментарий геоинформационных систем (справочное пособие). / Бусыгин Б.С., Гаркуша И.Н. – Киев, ИРГ „ВБ". – 2000. – 172 с.
2. Постанова Кабінету Міністрів України від 21 листопада 2007 р. №1021-р "Про схвалення Концепції проекту Закону України „Про національну інфраструктуру геопросторових даних" // Офіційний вісник України від 03.12.2007. – 2007 р., № 89, стор. 81, стаття 3280.
3. Мокін В.Б. Розробка моделей вхідних даних для ітеративного методу пошуку різноформатної екологічної інформації / Мокін В.Б., Коновалюк Ю.М. // [Вісник ВПІ]. – (Прийнята до друку)
4. Мокін В.Б. Новий метод пошуку різноформатної екологічної інформації на основі онтологічної бази даних та її XML-представлення / Мокін В.Б., Коновалюк Ю.М. // [Вісник ВПІ]. – Вінниця : „УНІВЕРСУМ-Вінниця", – 2009. – № 2. – С. 66–69.
5. Grune D. Parsing techniques: a practical guide. Springer / Grune D. and Jacobs C.J., New York, 2008.
6. Hopcroft J.E. Introduction to automata theory, languages, and computation, (2nd edition), Addison-Wesley, Reading, MA / Hopcroft J.E., Ullman J.D., 1979.

Впервые разработана информационная технология поиска геопространственных разноформатных данных, позволяющая формализовать информацию в трех основных форматах данных (картах разных пакетов программ для работы с ГИС, базах данных и форматах текстовых процессоров) на основе единого подхода и XML-модели, а также осуществлять поиск информации в источниках этих форматов одновременно, что даст возможность повысить скорость и уровень автоматизации поиска разноформатной экологической информации в электронных источниках.

Ключевые слова: *поиск геопространственной информации, онтология, контекстно-свободная грамматика*

For the first time information technology of geospatial multiformat data search has been developed. It allows formalizing of information in three main data formats (maps of different GIS application packages, databases and formats of some text processors), using common approach and XML-model and exercise simultaneous information search in data sources of above-mentioned format. It will allow increasing speed and improve automation level of search of multiformat information on ecology in electronic sources.

Keywords: *geospatial data search, ontology, context-free grammar*

Рекомендовано до публікації докт. техн. наук В.Г. Петруком. Дата надходження рукопису 25.02.11