УДК 347.78+04.624

Ю.А. АЛЬОШКІНА, НТУ "ХПІ" (м. Харків),
О.С. ДЕРЕВ'ЯНКО, НТУ "ХПІ" (м. Харків)

КОНСОЛІДАЦІЯ ПАТЕНТНОЇ ІФОРМАЦІЇ З РІЗНИХ ДЖЕРЕЛ

Розглянуті структури даних ряду джерел патентної інформації і засоби доступу до них. Розроблена узагальнена модель даних патентної інформації. Запропонований підхід до приведення початкових структур даних до узагальненої моделі, який забезпечує переніс складності з алгоритмів на дані і базується на технології Extensible Stylesheet Language Transformations (XSLT). Розроблені таблиці стилів XSLT для перетворювання даних з декількох джерел і програмний движок перетворення.

Ключові слова: патентна інформація, модель даних, XSLT.

Постановка проблеми. Невід'ємною складовою частиною процесів надбання та захисту інтелектуальної власності є патентно-кон'юнктурні дослідження (ПКД). Такі дослідження дозволяють визначати патентоспроможність об'єктів промислової власності, що створюються в процесі розробки нової продукції, вирішувати питання про доцільність патентування, визначати умови безперешкодної реалізації промислової продукції на ринку конкретної країни або країн (патентну чистоту), виключати порушення прав третіх осіб, що володіють патентами та виявляти потенційних конкурентів, визначати напрями їх діяльності і вибирати свою ринкову нішу.

Ступінь автоматизації ПКД поки що невисока. Пошук навіть у жорстко структурованих джерелах дає, як правило, великий об'єм результатів, які є нерелевантними, а для пошуку у слабо структурованих джерелах застосовуються пошукові системи загального призначення, у яких об'єм нерелевантних результатів ще більший.

Аналіз літератури. Патентно-кон'юнктурні дослідження [1] є обов'язковою частиною при виконанні науково-дослідних, дослідно-конструкторських, проектно-конструкторських робіт і маркетингу товарів. Джерелами, які розглядаються в ПКД, є патентна інформація (ПІ) і патентно-асоційована інформація (ПАІ).

ПІ отримується з патентних баз даних різних відомств і країн. Як правило, такі бази доступні через Інтернет і мають свої пошукові машини. При істотній відмінності в способах формулювання запитів і зовнішніх форматах результатів, можна відзначити, що результати, отримані з патентних баз даних, легко привести до єдиної загальної структури, оскільки об'єкт (об'єкти), що є результатом пошуку, має один тип – патент (наприклад, [2, 3]). Проте, слід відзначити, що у складі цієї структури є атрибути, які розглядаються в базі даних як атомарні, але фактично такі, що представляють собою повнотекстові документи (повний текст патенту, реферат), які можуть піддаватися додатковій структуризації, виділенню елементів і т.д.

Цього не можна сказати про ПАІ. Вона представляє собою електронні або/та паперові публікації, що містять повнотекстові документи, які не мають явної структури і не обов'язково визначають всі атрибути, які є необхідними для ПІ. Перед особою, що проводить ПКД, стає завдання аналізу змісту такого документа, фактично, його анотування й визначення його релевантності (наприклад, [4, 5]).

Вивчення вказаних джерел, а також інструкцій більш прикладного та конкретного характеру (наприклад, [6, 7]) показує, що проведення ПКД, по-перше, вимагає значних трудовитрат фахівця, по-друге, для часткової автоматизації їх проведення використовується або програмне забезпечення загального призначення, або ж "саморобне" програмне забезпечення, яке не базується на загальноприйнятих, відкритих стандартах.

Втім, сучасний стан інформатики має велику кількість технологій, які вирішують задачі консолідації та інтеграції значної кількості джерел інформації та засобів їх оброблення [8]. Задача автоматизації ПКД складається не стільки у створенні нових методів і технологій, а в адаптації вже існуючих технологій до предметної галузі і інтеграції їх. Зокрема, запропонована у [9] концепція архітектури системи автоматизації ПКД включає у себе ланок оброблення інформації, а саме:

1) ПКД починаються з пошуку у патентних базах даних за емпірично вибраними (підібраними експертами) ключовими словами.

2) Результати цього пошуку зберігаються у локальному сховищі даних і використовуються як *навчальна вибірка*, на підставі якої засобами інтелектуального оброблення даних будується онтологія предметної галузі.

3) Отримана онтологія є похідним засобом для формулювання запитів на пошук ПАІ і оцінки релевантності результатів цього пошуку.

4) Результати, які отримані на кожному етапі (та на проміжних стадіях кожного етапу) зберігаються у локальному сховищі даних і можуть візуалізуватися та корегуватися фахівцем, який проводить дослідження, і оновлена онтологія може служити похідною для повторення процесу з будь-якої попередньої точки.

Запропонована у [9] архітектура "шлюзів" для перетворення даних між зовнішніми джерелами та локальним сховищем даних і підсистемою їх інтелектуального оброблення. Ці шлюзи фактично є засобом консолідації даних. І першою ланкою у такій консолідації має бути консолідація ПІ. Цінність ПІ полягає у її структурованості й регламентованості (бібліографія, опис, формула, схеми, приклади застосування, втілення), достовірності (заявник несе відповідальність за промислову придатність та достовірність викладених фактів), доступності, неповторності.

Задача консолідації ПІ дещо полегшується тим, що у цій галузі діють стандарти Всесвітньої Організації Інтелектуальної Власності (ВОІВ) [10]. Зокрема, у стандарті ВОІВ ST.9 [11] подані рекомендації, які указують мінімум бібліографічних даних, які повинні друкуватися на титульному листі

патентного документа і публікуватися як частина повідомлень в патентному бюлетені. Ці рекомендації дають перелік приблизно 60 індивідуальних бібліографічних даних, які ідентифікуються за допомогою кодових номерів, так званих "кодів ІНІД" або "номерів ІНІД". (INID – Internationally agreed Numbers for the Identification of (bibliographic) Data). Бібліографічні дані, включені в рекомендації, охоплюють широкий спектр даних – від даних ідентифікації документа, подачі заявки, публікації, даних, пов'язаних з технічною інформацією, до даних, що відносяться до Міжнародних патентних конвенцій. Рекомендації стандартів ВОІВ не вирішують задачі консолідації ПІ, бо вони стосуються тільки змісту даних, а не формату їх подання (тим більше – не формату їх подання в електронній формі) і розглядають тільки титульні дані, а не контент патентного документа, але вони значно спрощують вирішення цієї задачі, фактично представляючи специфікацію результуючих даних для розробника консолідуючих шлюзів.

Ціллю даної статті є розроблення загального підходу до створення шлюзів ПІ для автоматизованої системи підтримки ПКД, вибір засобів для реалізації такого підходу та представлення результатів пілотного проекту реалізації шлюзів для кількох популярних джерел ПІ.

Джерела ПІ. У WWW є багато джерел, що надають достовірні відомості про патенти, опубліковані національними і міжнародними патентними бюро. При здійсненні патентних досліджень можна використовувати різні джерела, але найчастіше вітчизняні фахівці починають роботу з наступних сайтів:

1. USPTO Patent Full-Text and Full-Page Image Databases – патенти США: http://www.uspto.gov/patft/index.html.

2. Esp@cenet – матеріали патентів США, Японії і багатьох інших країн, патентних заявок ЕРО – European Patent Office і ВОІВ: http://ep.espacenet.com.

3. Російська Федеральна служба по інтелектуальній власності, патентам і товарним знакам (Роспатент): http://www.fips.ru/russite.

4. Державне підприємство "Український інститут промислової власності" (Укрпатент): http://www.ukrpatent.org/.

Кожна з цих баз даних має власну пошукову систему. Ці пошукові системи різняться можливостями, але усі вони побудовані на східних принципах і прийоми пошукової роботи у них достатньо близькі. Пошукове завдання для них записується у бланку пошуку, який містить у собі поля для вводу значень базових атрибутів патентного документа. Деякі з систем надають кілька варіантів бланків – для звичайного, швидкого, ускладненого пошуків, які різняться кількістю пошукових атрибутів та можливістю застосування нечітких запитів. Також ПІ надається вказаними відомствами на компакт-дисках, де вона записується у різних форматах і супроводжується програмами, які забезпечують доступ до неї і ті ж самі пошукові можливості, що і пошукові движки сайтів. Деякі з таких програм "розуміють" формати кількох відомств, але їх можливості з розпізнавання форматів обмежені.

Узагальнений формат представлення патентних документів. На сьогодні стандартом де-факто представлення патентних документів у різних відомствах є мова XML. Хоча структура тієї інформації, яку представляють відомства може достатньо ефективно описуватися реляційною моделлю даних, відомства обрали мову XML, як ми вважаємо, з міркувань полегшення обміну даними, полегшення представлення даних в Інтернет та, можливо, з міркувань подальшого розширення та ускладнення того складу атрибутів, який використовується у пошукових системах та подається користувачеві. Однак різні патентні відомства використовують різні схеми представлення патентних документів і спроби привести їх до єдиної моделі не мають успіху. Для цілей роботи, яка представлена у даній статті ми вивчили схеми відомств Esp@cenet, Роспатент та Укрпатент. Усі три схеми відповідають ієрархічній моделі даних, але їх дерева, мають різну структуру за кількістю гілок та глибиною вкладення, а також дещо різняться складом атрибутів. За результатами дослідження вказаних схем ми розробили на сформулювали засобами XML Schema узагальнену модель представлення ПІ. Повний опис моделі засобами XML Schema не дозволяє навести обмежений обсяг даної публікації, графічне представлення моделі (обмежене у частині опису варіативності) показано на рис. 1. Іменами типу inidNN на рисунку представлені атрибути, які відповідають стандарту ST.9.

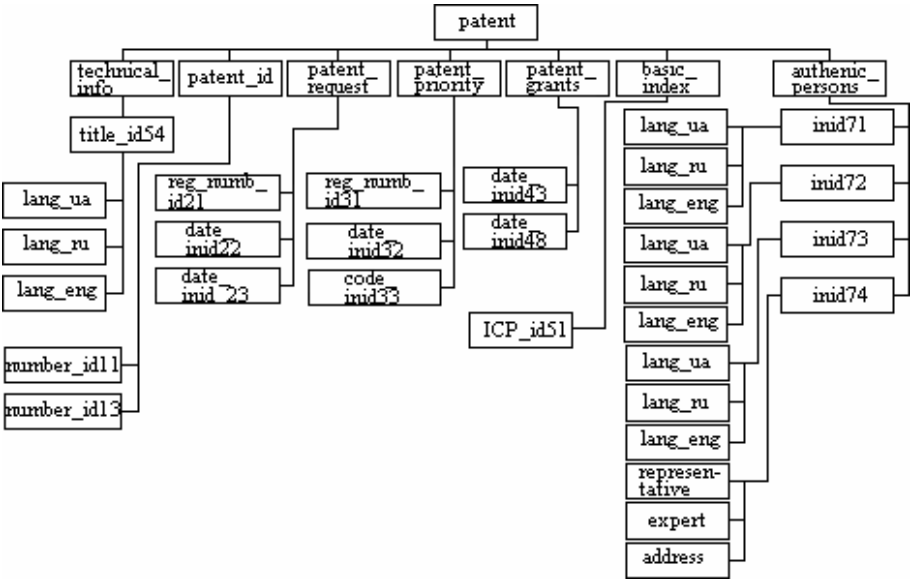

Рис. 1. Узагальнена модель даних

Модель приведення даних до єдиної схеми. Як було відзначено, названі вище патентні відомства є відправними точками для фахівців у патентному пошуку, але перелік джерел ПІ ними не вичерпується і досить часто глибокий пошук потребує звернення до інших патентних відомств (канадського, японського і т.д.). Інші джерела мають свої схеми представлення даних і свої пошукові засоби. Для програміста не складе великих зусиль написання програмного модуля-шлюзу для нового джерела ПІ, але написання великої кількості таких шлюзів та керування ними може значно ускладнити систему. З цих міркувань ми вважаємо, що процес приведення похідних патентних даних до єдиної моделі має бути керованим даними – тобто "сценарій" перетворення даних має не бути жорстко закладеним у програмному модулі, а має описуватися якимись простими засобами і описи різних сценаріїв може зберігатися у тій самій локальній базі даних автоматизованої системи, у якій зберігаються самі дані. При такому підході програмна частина складає єдиний програмний процесор, який "розуміє" мову опису сценаріїв і для якого сценарій перетворення є похідними даними разом з тими даними, які підлягають перетворенню. Такий підхід, звичайно, потребує фахівця для створення опису сценарію, але управління великою кількістю сценаріїв, кожен з яких є просто окремим запитом у базі даних, значно спрощується і може бути легко автоматизовано. Схема моделі приведення даних показана на рис. 2.

Рис. 2. Приведення даних

Вибір засобу опису сценарію. Маючи на увазі те, що похідні дані подаються мовою XML, і те, що у [9] запропоновано мати локальне сховище у гібридній (реляційній + XML) базі даних, засоби перетворення слід шукати у технологіях XML. І такі засоби серед технологій XML є – це мова трансформацій XSLT [12]. Документ, який записаний мовою XSLT і який містить правила трансформацій має назву таблиця стилів (stylesheet). Таблиця стилів XSLT складається з ряду шаблонних правил, кожне з яких має форму "якщо така-то умова зустрічається на вході, то генерувати такий-то вихід". Сценарій, виражений через XSLT, описує правила перетворення початкового дерева документа в кінцеве дерево. Перетворення будується шляхом зіставлення зразків і шаблонів. Зразок порівнюється з елементами початкового дерева, а шаблон використовується для створення частин кінцевого дерева. Кінцеве дерево відокремлене від початкового дерева. Структура кінцевого дерева може повністю відрізнятися від структури початкового дерева. В ході

побудови кінцевого дерева елементи початкового дерева можуть піддаватися фільтрації і переупорядкуванню, також може додаватися нова структура. Порядок правил неістотний, і є алгоритм вирішення конфліктів, який застосовується, якщо декілька правил можна застосовувати для одного входу. Вхід не обробляється послідовно, рядок за рядком. Натомість похідний XML-документ розглядається як структура дерева, і кожне шаблонне правило застосовується до вузла дерева. Шаблонне правило може само вирішувати, який вузол повинен оброблятися наступним, так що вхід не обов'язково сканується в оригінальному порядку документа.

Сама таблиця стилів також є документом XML, отже може зберігатися і оброблюватися так само, як і інші дані у форматі XML.

Нами були розроблені таблиці стилів для трьох форм похідних документів, відповідних до схем відомств Esp@cenet, Роспатент та Укрпатент. Результуючою є запропонована нами узагальнена схема патентних даних. Таблиця стилів містить 3 – 7 сторінок нещільного тексту. Це значно менш тих обмежень на розмір даних типу XML, які накладає гібридна СУБД IBM DB2 v.9 [13].

Реалізація програмного процесора. Для реалізації програмного процесора, який би втілював модель приведення даних, ми обрали мову програмування Java. Такий вибір обґрунтовується відкритістю технологій Java, її незалежністю від платформи і, що найважливіше, наявністю у платформі Java ряду пакетів javax.xml.transformation, які забезпечують базові засоби оброблення перетворювань XSLT [13]. Діаграма класів прикладення, яке є процесором приведення даних, показана на рис. 3

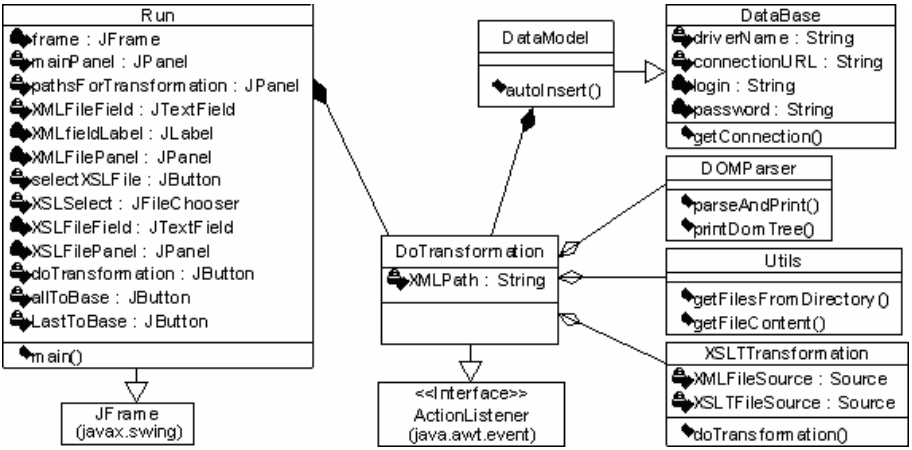

Рис. 3. Діаграма класів

У наступній таблиці наведені описи основних класів процесора.

Клас Run	
Коментар	Головний клас, який містить метод main().
Клас XSLTTransformation	
Коментар	Клас, який здійснює трансформацію файлів за шаблоном.
Атрибути	XMLFileSource : Source – шлях до файлу. XSLTFileSource : Source – шлях до шаблону.
Операції	doTransformation() – здійснення трансформації файлів.
Клас DoTransformation	
Коментар	Клас, який оброблює подію трансформації файлів.
Атрибути	XMLPath : String – шлях до файлу.
Клас Utils	
Коментар	Допоміжний клас.
Операції	getFilesFromDirectory() – отримання файлів з каталогу. getFileContent() – отримання вмісту файлу.
Клас DOMParser	
Коментар	Клас, який формує дерево вузлів після аналізу файлу.
Операції	parseAndPrint() – сформувати й показати. printDomTree() – показати дерево вузлів.
Клас DataModel	
Параметр	Значення.
Коментар	Клас, який здійснює запис файлів до бази даних.
Операції	autoInsert() – вставити дані в базу.
Клас DataBase	
Параметр	Значення.
Коментар	Клас, який являє собою базу даних.
Атрибути	driverName : String – назва драйвера. connectionURL : String – шлях для з'єднання з базою даних. login : String – ім'я користувача. password : String – пароль користувача.
Операції	getConnection() – отримання з'єднання з базою даних.

У пілотному проекті процесор виконаний як окреме прикладення з графічним інтерфейсом, для застосування у складі системи підтримки ПКД він може бути оформлений згідно з вимогами тієї платформи, на якій будується система (наприклад, як Web-сервіс або модуль UIMA).

Висновки. На основі аналізу джерел ПІ запропонована узагальнена модель даних ПІ та підхід до програмного приведення даних до узагальненої моделі. Можливість втілення запропонованої моделі та підходу доказана реалізацією програмного виробу, який може бути застосований як окрема програма консолідації ПІ або як модуль системи підтримки ПКД.

Список літератури: 1. *Косенко С.* Патентна інформація [Електронний ресурс] / Юридичний журнал. – К.: Європейський університет, 2006. – Вип. 8. – Режим доступу: http://www.justinian.com.ua/article.php?id=2354. **2.** *Рагойша А.А.* Патентные базы данных в Интернете [Електронний

ресурс] / Режим доступу: http://www.abc.chemistry.bsu.by/patent/default.htm. **3.** *Кузнєцов Ю.М.* Основи патентознавства та авторського права. – К.: ТОВ "ЗМОК", 2001. – 206 с. **4.** *Аветисов А.Р. и др.* Современный патентный поиск: использование традиционных источников и возможностей сети Интернет [Електронний ресурс] / Белорусский медицинский журнал. – 2004. – Вип. 3. – Режим доступу: http://bsmu.by/bmm/03.2004/41.html. **5.** *Брагарник О.* Використання Internet як засобу патентного пошуку [Електронний ресурс] / Державний департамент інтелектуальної власності. – Режим доступу: http://www.sdip.gov.ua/ukr/help/statti/ dopuintei/bragarnyk. **6.** *Клейман А.М.* Проведение патентных исследований в Интернете и оформление отчёта о поиске. Методические указания [Електронний ресурс] / Каф. ЮНЕСКО по авторскому праву и другим отраслям права интеллектуальной собственности. – Режим доступу: http://mgtu-sistema.ru/triz/patent_metodika.pdf. **7.** Інструкції по використанню інформаційно-пошукової системи [Електронний ресурс] / Державне підприємство Український інститут промислової власності (Укрпатент). – Режим доступу: http://www.ukrpatent.org/old/instructions.html. **8.** *Деревянко А.С., Солощук М.Н.* Технологии и средства консолидации информации. – Харьков: НТУ "ХПИ", 2007. – 432 с. **9.** *Дерев'янко О.С., Сомхієва О.С.* Застосування засобів інтелектуального оброблення даних у патентно-кон'юнктурних дослідженнях / Матеріали XVI Міжнар. науково-практичної конф. "Інформаційні технології: наука, техніка, технологія, освіта, здоров'я", 2008. – С. 313. **10.** Стандарти Всесвітньої організації інтелектуальної власності (ВОІВ) [Електронний ресурс] / Державний департамент інтелектуальної власності. – http://www.sdip.gov. ua/ukr/laws/418. **11.** "Руководство по информации и документации в области промышленной собственности". Стандарты – ST.9 [Електронний ресурс] / Державне підприємство Український інститут промислової власності (Укрпатент). – Режим доступу: http://www.ukrpatent.org/atachs/ sst9.doc. **12.** XSL Transformations (XSLT) Version 1.0. W3C Recommendation. [Електронний ресурс] / World Wide Web Consortium. Режим доступу: http://www.w3.org/TR/xslt. **13.** *Саракко С.М.* Что нового в DB2 Viper? [Електронний ресурс] / Режим доступу: http://khpi-iip.mipk.kharkiv.edu/ library/extent/dbms/viper/vip3.html **14.** *Bruke E.M.* Java and XSLT. – O'Raily, 2001. – 495 с.

УДК 347.78+04.624

Консолидация патентной информации из различных источников / **Алешкина Ю.А., Деревянко А.С.** // Вестник НТУ "ХПИ". Тематический выпуск: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2008. – №. 49. – С. 3 – 10.

Рассмотрены структуры данных ряда источников патентной информации и средства доступа к ним. Разработана обобщенная модель данных патентной информации. Предложен подход к приведению исходных структур данных к обобщенной модели, который обеспечивает перенос сложности с алгоритмов на данные и базируется на технологии Extensible Stylesheet Language Transformations (XSLT). Разработаны таблицы стилей XSLT для преобразования данных из нескольких источников и программный движок преобразования. Ил.:3. Библиогр.: 14 назв.

Ключевые слова: патентная информация, модель данных, XSLT.

UDK 347.78+04.624

Consolidation of a patent information from different sources / **Alioeshkina Y.A., Derevyanko A.S.** // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkiv: NTU "KhPI". – 2008. – №. 49. – P. 3 – 10.

Data structures for several sources of a patent information are considered. The generalized data model for patent information is proposed. An approach for source data transformation to the generalized model which transfers the complexity from algorithms to data and based on the Extensible Stylesheet Language Transformations (XSLT) technology is proposed. XSLT stylesheets for data transformation from several sources and transformation program engine are designed. Figs: 3. Refs: 14 titles.

Key words: patent information, data model, XSLT.

Поступила до редакції 10.10.2008